Análisis estadístico y modelos econométricos de elección discreta con correlación espacial en transporte y economía urbana: aplicación a los modelos de predicción de la localización residencial

---

Autor: Jose-Benito Perez-Lopez

Directores: Alfonso Orro Arcay
Margarita Novales Ordax

Tesis doctoral/2022
Programa de Doctorado en Ingeniería Civil

UNIVERSIDADE DA CORUÑA

A Eva, Paloma y Miguel

Agradecimientos

Quiero dar las gracias a Marga y Alfonso por dirigir científica y personalmente esta tesis. Agradezco su conocimiento, su experiencia y su compromiso. Y agradezco, sobre todo, su humanidad, confianza y apoyo.

Me gustaría agradecer a mis compañeros y compañeras de investigación sus aportaciones, enseñanzas y compañerismo. En esta tesis, en los proyectos de investigación, en los artículos y en las ponencias. Desde la Universidade da Coruña, en mi grupo de investigación en Ferrocarriles y Transportes, en CartoLab, en OMA y en la Facultad de Economía y Empresa. Desde la Universidad de Cantabria, en el grupo de investigación en Sistemas de Transporte y en el grupo de investigación en Movilidad Sostenible e Ingeniería Ferroviaria. Desde la Universidad Politécnica de Cataluña, en el Centro de Innovación del Transporte. Desde la Universidad de Sevilla, en el grupo de investigación en Ingeniería de los Transportes.

Mi agradecimiento al Gobierno de España, por el apoyo financiero mediante los proyectos de investigación del plan nacional de I+D+i IMPROVEBUS (RTI2018-097924-B-I00 MCIU/AEI/FEDER, UE), TRANSPACE (TRA2012-37659) y SIMETRIA (FOM/3864/2008-63).

Por supuesto, quiero agradecer a toda mi familia, amigos/as y compañeros/as (durante mis estudios, los actuales en la Universidade da Coruña y los de mi carrera profesional anterior en las distintas áreas de Hewlett Packard y en The Boston Consulting Group). Por su cariño, por su apoyo y por su confianza. Y por existir.

Resumen

En esta tesis se propone un nuevo modelo de elección discreta entre alternativas de naturaleza geográfica, el modelo spatially correlated nested logit (SCNL). En este tipo de elecciones se prevé un elevado número de alternativas y la presencia de correlación, al menos espacial, entre ellas. La literatura actual en modelos de elección discreta considera dos enfoques para este tipo de elecciones, que son analizados en esta tesis. El modelo propuesto combina ambos enfoques sin añadir parámetros desconocidos, presenta una estructura matemática cerrada y es compatible con especificaciones mixtas que permitan variaciones en los gustos de quienes deciden. El modelo SCNL se formula y analiza en esta tesis, donde también se aplica a un caso real, para modelizar la elección de localización residencial en contexto urbano, que es un problema clave en planificación del transporte y economía urbana. Los resultados empíricos con el nuevo modelo mejoran significativamente los obtenidos con los modelos actuales. Además, se propone una métrica espacial para zonificaciones del espacio con diferentes tamaños y formas irregulares, muy habituales en áreas administrativas de ciudades europeas. Los modelos especificados con la métrica espacial propuesta obtuvieron resultados empíricos significativamente mejores que con el resto de métricas espaciales consideradas.

Abstract

A new model of discrete choice between alternatives of a geographical nature is proposed in this thesis, the spatially correlated nested logit (SCNL) model. This type of choices are generally characterized by the existence of a high number of alternatives and the presence of correlation, at least spatial, between them. The current literature on discrete choice models considers two approaches for this type of choice, which are analyzed in this thesis. The proposed model combines both approaches without adding unknown parameters, it presents a closed mathematical structure and it is compatible with mixed specifications that allow variations in the tastes of those who decide. The SCNL model is formulated and analyzed in this thesis, where it is also applied to a real case, to model the choice of residential location in an urban context, which is a key problem in transport planning and urban economics. The empirical results with the new model significantly improve those obtained with the current models. In addition, a spatial metric for spatial zoning with different sizes and irregular shapes is proposed, very common in administrative areas of European cities. The models specified with the proposed spatial metric obtained empirical results that were significantly better than those obtained with the rest of the spatial metrics considered.

Resumo

Nesta tese proponse un novo modelo de elección discreta entre alternativas de natureza xeográfica, o modelo spatially correlated nested logit (SCNL). Neste tipo de eleccións espérase un elevado número de alternativas e a presenza de correlación, polo menos espacial, entre elas. A literatura actual sobre modelos de elección discreta considera dous enfoques para este tipo de eleccións, que se analizan nesta tese. O modelo proposto combina ambos enfoques sen engadir parámetros descoñecidos, presenta unha estrutura matemática pechada e é compatible con especificacións mixtas que permiten variacións nos gustos de quen decide. O modelo SCNL formúlase e analízase nesta tese, onde tamén se aplica a un caso real, para modelizar a elección da localización residencial nun contexto urbano, que é un problema clave na planificación do transporte e na economía urbana. Os resultados empíricos co novo modelo melloran significativamente os obtidos cos modelos actuais. Ademais, proponse unha métrica espacial para zonificacións do espazo con diferentes tamaños e formas irregulares, moi común en áreas administrativas das cidades europeas. Os modelos especificados coa métrica espacial proposta obtiveron resultados empíricos significativamente mellores que co resto de métricas espaciais consideradas.

Resumen extenso

La predicción de demanda de usos del suelo y transporte es un área de investigación de gran importancia en la planificación del transporte de la ingeniería civil, en economía urbana y en otros ámbitos científicos y profesionales. En la actualidad, el marco conceptual más ampliamente utilizado para abordarlo es el de los modelos matemáticos de interacción entre el sistema de usos del suelo y el sistema de transporte (LUTI). El enfoque econométrico desagregado de los modelos de elección discreta es una herramienta fundamental en modelización LUTI. Algunos de los elementos del contexto LUTI requieren modelizar la elección entre alternativas que son de naturaleza geográfica, como es el caso de los modelos de elección de la localización residencial. Estos modelos de elección espacial presentan problemáticas específicas, principalmente la presencia de dependencia espacial entre alternativas y el elevado número de las mismas.

La principal aportación de esta tesis es la propuesta de un nuevo modelo de elección discreta multinomial, específico para el contexto de la elección entre localizaciones espaciales. El modelo propuesto, denominado spatially correlated nested logit (SCNL), combina los dos enfoques actuales para abordar la modelización de elección espacial, sin necesidad de añadir parámetros desconocidos. Ambos enfoques son analizados en esta tesis. El primer enfoque no es específico de elección espacial. Este enfoque se basa en agrupaciones de las alternativas, llamados nidos, que son diseñadas por el analista para reflejar la correlación entre las alternativas. El segundo enfoque es específico de elección espacial. Este enfoque se basa en métricas espaciales, que deben recoger la correlación espacial entre las alternativas. En esta tesis se propone también una métrica espacial, denominada common-border, basada en la proporción de frontera común entre alternativas. Esta métrica se propone para contextos de aplicación urbanos, basados en áreas administrativas de diferentes tamaños y formas irregulares, muy habituales en las ciudades europeas.

El modelo SCNL que se propone en esta tesis forma parte de la familia de modelos logit denominada generalized extreme value (GEV), y tiene, por tanto, una estructura matemática cerrada. Además, es compatible con una especificación mixta con coeficientes aleatorios, que le permite incorporar variaciones en los gustos entre los individuos decisores. En esta tesis se formula y analiza el modelo SCNL propuesto, y se aplica a un caso real de modelización de la elección de la localización residencial en planificación urbana, en la ciudad de Santander (España). También se aplican, en el mismo contexto empírico, el resto de modelos de elección discreta presentes en la literatura y que son compatibles con elección espacial, que se analizan en esta tesis. La capacidad explicativa y predictiva de todos los modelos se compara empíricamente bajo las mismas condiciones.

El modelo SCNL obtiene, en el contexto empírico de aplicación, resultados de capacidad explicativa y predictiva significativamente mejores que el resto de modelos analizados. Por tanto, el modelo SCNL propuesto en esta tesis demuestra empíricamente, en la aplicación desarrollada, que es capaz de combinar los dos enfoques de correlación entre alternativas compatibles con elección espacial. Además, esta integración permite

incorporar eficientemente la correlación (tanto espacial como no espacial) existente entre alternativas de elección espacial, sin necesidad de añadir parámetros desconocidos adicionales.

Los modelos especificados con la métrica espacial common-border propuesta mejoran los resultados empíricos de capacidad explicativa y predictiva obtenidos con los especificados con el resto de métricas analizadas. Las especificaciones mixtas de los modelos GEV analizados mejoran la capacidad explicativa y predictiva de su núcleo GEV, lo que indica la presencia de variaciones en los gustos de los individuos decisores. La especificación mixta con coeficientes aleatorios del modelo SCNL mejora significativamente los resultados de capacidad explicativa y predictiva de su núcleo GEV, y del resto de núcleos GEV y especificaciones mixtas aplicadas a lo largo de la tesis.

Extended abstract

Land-uses and transportation demand forecasting is a research area of great importance in transportation planning of civil engineering, in urban economics and in other scientific and professional fields. Currently, the most widely used conceptual framework to address it is that of the mathematical models of interaction between the land-uses system and the transportation system (LUTI). The disaggregated econometric approach of discrete choice models is a fundamental tool in LUTI modeling. Some of the elements of the LUTI context require modeling the choice between alternatives that are geographical in nature, as is the case with residential location choice models. These spatial choice models present specific problems, mainly the presence of spatial dependency between alternatives and the high number of them.

The main contribution of this thesis is the proposal of a new multinomial discrete choice model, specific for the context of choice between spatial locations. The proposed model, called spatially correlated nested logit (SCNL), combines the two current approaches to address spatial choice modeling, without the need to add unknown parameters. Both approaches are analyzed in this thesis. The first approach is not specific to spatial choice. This approach is based on groupings of the alternatives, called nests, which are designed by the analyst to reflect the correlation between the alternatives. The second approach is spatial choice specific. This approach is based on spatial metrics, which must collect the spatial correlation between the alternatives. This thesis also proposes a spatial metric, called common-border, based on the proportion of common border between alternatives. This metric is proposed for urban application contexts, based on administrative areas of different sizes and irregular shapes, very common in European cities.

The SCNL model proposed in this thesis is part of the family of logit models called generalized extreme value (GEV), and therefore has a closed mathematical structure. In addition, it is compatible with a mixed specification with random coefficients, which allows it to incorporate variations in tastes among decision makers. In this thesis, the proposed SCNL model is formulated and analyzed, and applied to a real case of modeling the choice of residential location in urban planning, in the city of Santander (Spain). The rest of the discrete choice models present in the literature and that are compatible with spatial choice, which are analyzed in this thesis, are also applied in the same empirical context. The explanatory and predictive capacity of all the models is empirically compared under the same conditions.

The SCNL model obtains, in the empirical context of application, significantly better explanatory and predictive capacity results than the rest of the analyzed models. Therefore, the SCNL model proposed in this thesis empirically demonstrates, in the developed application, that it is capable of combining the two correlation approaches between alternatives compatible with spatial choice. In addition, this integration allows efficient incorporation of the correlation (both spatial and non-spatial) between spatial choice alternatives, without the need to add additional unknown parameters.

The models specified with the proposed common-border spatial metric improve the empirical results of explanatory and predictive capacity obtained with those specified with the rest of the analyzed metrics. The mixed specifications of the analyzed GEV models improve the explanatory and predictive capacity of their GEV kernel, which indicates the presence of variations in the tastes of individual decision makers. The mixed specification with random coefficients of the SCNL model significantly improves the results of the explanatory and predictive capacity of its GEV kernel, and of the rest of the GEV kernels and mixed specifications applied throughout the thesis.

Resumo extenso

O uso do solo e a predición da demanda de transporte é unha área de investigación de gran importancia na planificación do transporte de enxeñaría civil, na economía urbana e noutros campos científicos e profesionais. Actualmente, o marco conceptual máis empregado para abordalo é o dos modelos matemáticos de interacción entre o sistema de uso do solo e o sistema de transporte (LUTI). O enfoque econométrico desagregado dos modelos de elección discreta é unha ferramenta fundamental no modelado LUTI. Algúns dos elementos do contexto LUTI requiren modelar a elección entre alternativas de natureza xeográfica, como é o caso dos modelos de elección de localización residencial. Estes modelos de elección espacial presentan problemas específicos, principalmente a presenza de dependencia espacial entre alternativas e o elevado número delas.

A principal contribución desta tese é a proposta dun novo modelo de elección discreta multinomial, específico para o contexto de elección entre localizacións espaciais. O modelo proposto, chamado spatially correlated nested logit (SCNL), combina os dous enfoques actuais para abordar o modelado de elección espacial, sen necesidade de engadir parámetros descoñecidos. Ambos enfoques son analizados nesta tese. O primeiro enfoque non é específico para a elección espacial. Este enfoque baséase en agrupacións das alternativas, denominadas niños, que son deseñadas polo analista para reflectir a correlación entre as alternativas. O segundo enfoque é específico da elección espacial. Este enfoque baséase en métricas espaciais, que deben recoller a correlación espacial entre as alternativas. Esta tese tamén propón unha métrica espacial, denominada common-border, baseada na proporción de límite común entre alternativas. Esta métrica proponse para contextos de aplicación urbana, baseada en áreas administrativas de diferentes tamaños e formas irregulares, moi comúns nas cidades europeas.

O modelo SCNL proposto nesta tese forma parte da familia de modelos logit denominados generalized extreme value (GEV), polo que ten unha estrutura matemática pechada. Ademais, é compatible cunha especificación mixta con coeficientes aleatorios, o que lle permite incorporar variacións de gustos entre os que toman decisións individuais. Nesta tese formúlase e analízase o modelo SCNL proposto, e aplícase a un caso real de modelización da elección da localización residencial no planeamento urbanístico, na cidade de Santander (España). O resto de modelos de elección discreta presentes na literatura e compatibles coa elección espacial, que se analizan nesta tese, tamén se aplican no mesmo contexto empírico. A capacidade explicativa e preditiva de todos os modelos compárase empíricamente nas mesmas condicións.

O modelo SCNL obtén, no contexto empírico de aplicación, resultados de capacidade explicativa e predictiva significativamente mellores que o resto dos modelos analizados. Polo tanto, o modelo SCNL proposto nesta tese demostra empíricamente, na aplicación desenvolvida, que é capaz de combinar os dous enfoques de correlación entre alternativas compatibles coa elección espacial. Ademais, esta integración permite a incorporación eficiente da correlación (tanto espacial como non espacial) entre

alternativas de elección espacial, sen necesidade de engadir parámetros adicionais descoñecidos.

Os modelos especificados coa métrica espacial common-border proposta melloran os resultados empíricos de capacidade explicativa e preditiva obtidos cos especificados co resto das métricas analizadas. As especificacións mixtas dos modelos GEV analizados melloran a capacidade explicativa e preditiva do seu núcleo GEV, o que indica a presenza de variacións nos gustos dos tomadores de decisións individuais. A especificación mixta con coeficientes aleatorios do modelo SCNL mellora significativamente os resultados da capacidade explicativa e preditiva do seu núcleo GEV, e do resto dos núcleos GEV e especificacións mixtas aplicadas ao longo da tese.

Índice general

Índice de figuras

Índice de tablas

Lista de abreviaturas

AIC: Akaike information criterion.

ALRI: adjusted likelihood ratio index.

BSCL: common-border spatially correlated logit.FG: fitting geometric.

GDSCL: gravitational-distance spatially correlated logit.

GEV: generalized extreme value.

GIS: geographic information system.

GNL: generalized nested logit.

GM: geometric mean.

GoF: goodness-of-fit.

GSCL: generalized spatially correlated logit.

L: likelihood function.

LL: natural logathim of the likelihood function.

LRI: likelihood ratio index.

LRT: likelihood ratio test.

LUTI: land-uses and transportation interaction mathematical model.

M-NL: mixed nested logit.

MBSCL: mixed common-border spatially correlated logit.

MBSCNL: mixed common-border spatially correlated nested logit.

MGDSCL: mixed gravitational-distance spatially correlated logit.

MGEV: mixed generalized extreme value.

MMNL: mixed multinomial logit.

MNL: multinomial logit.

MRNL: mixed restricted nested logit.

MSCL: mixed spatially correlated logit.

MSCNL: mixed spatially correlated nested logit.

NL: hierarchical or nested logit.

PCL: paired combinatorial logit.

PG: predicting geometric.

PGNL: paired generalized nested logit.

RI: relative influence.

RNL: restricted nested logit.

RUM: random utility maximization.

SCL: spatially correlated logit.

SCL-b: SCL-based.

SC: standardized coefficient.

SCNL: spatially correlated nested logit.

SD: standard deviation.

SE: standard error.

SL: simulated likelihood function.

SLL: natural logarithm of the simulated likelihood function.

1. Introducción

1.1. Motivación

El crecimiento de la población mundial, y especialmente de la población que reside en las áreas urbanas y metropolitanas, constituye un reto para diferentes campos científicos y profesionales. El informe de perspectivas de urbanización mundial de las Naciones Unidas (WUP, 2018) recoge que la población mundial ha crecido hasta los 7.875 millones de personas. Su ritmo de crecimiento es actualmente de 1% anual, incluso superior en las ciudades, como muestra la figura 1.1. Este proceso de urbanizacióin se inició en Europa en el siglo XIX como consecuencia de la revolución industrial, y se ha ido acelerando en las últimas décadas. En el año 1950 la población urbana era el 30% de la población total. Actualmente la población urbana es el 57% del total, que asciende al 79% en el caso de los países de las regiones más desarrollados (81% en España), como se puede ver en la figura 1.2.

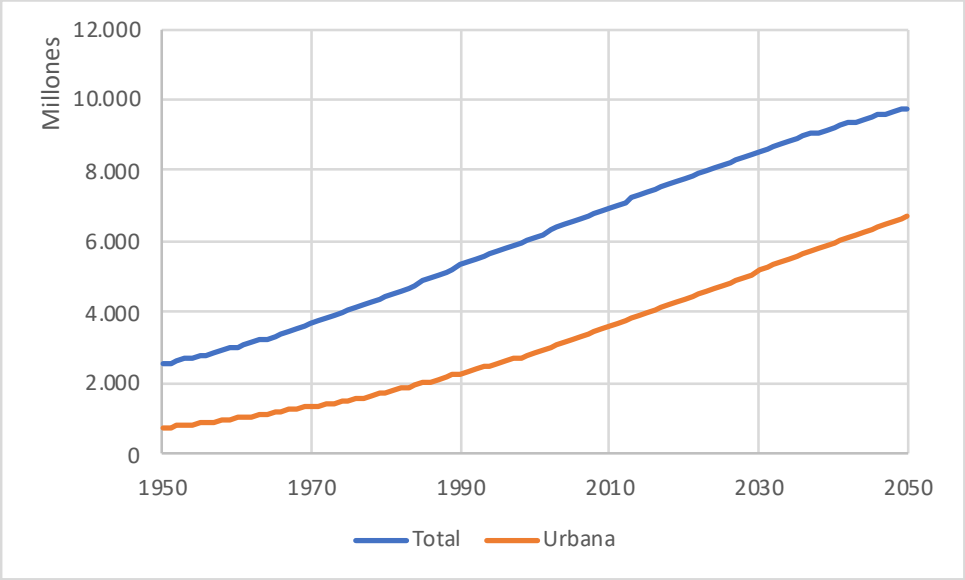

Figura 1.1. Gráfico de secuencia de la serie anual 1950-2050 de la población mundial. Elaboración propia a partir del WUP (2018) de Naciones Unidas.

Uno de los retos de la ingeniería civil y la economía urbana ante este crecimiento es la gestión de la movilidad y de los usos de suelo. La planificación urbana eficiente necesita elaborar modelos de predicción de la demanda urbana de usos de suelo y su repercusión en las necesidades de transporte, porque "la expansión de las actividades humanas a lo largo de los territorios urbanos crea necesidades de transporte que exigen una respuesta" (Gakemheimer, 2006). Se debe considerar, además, que la oferta de transporte condiciona la demanda de suelo a través de las condiciones de accesibilidad (véase Torrens, 2000). Las sociedades más modernas demandan que esta planificación urbana, además de facilitar la movilidad, tenga cada vez más en cuenta los objetivos de reducción de emisiones contaminantes y de efecto invernadero; y limite, en lo posible, la agresión medioambiental, especialmente al entorno no urbano. Por ejemplo, reducir la dispersión urbana y potenciar la generación de suelos con usos mixtos es una estrategia que genera menos ocupación de suelo no urbano, facilita que los desplazamientos sean más cortos y favorece modos de transporte menos

contaminantes, más baratos y más compatibles con el entorno social urbano. Las publicaciones científicas de indicadores medioambientales y económicos facilitan el análisis de estrategias y la toma de decisiones, como el impacto ambiental de la movilidad metropolitana que se analiza en el artículo Perez-Lopez et al. (2021) para el caso de la comunidad universitaria.

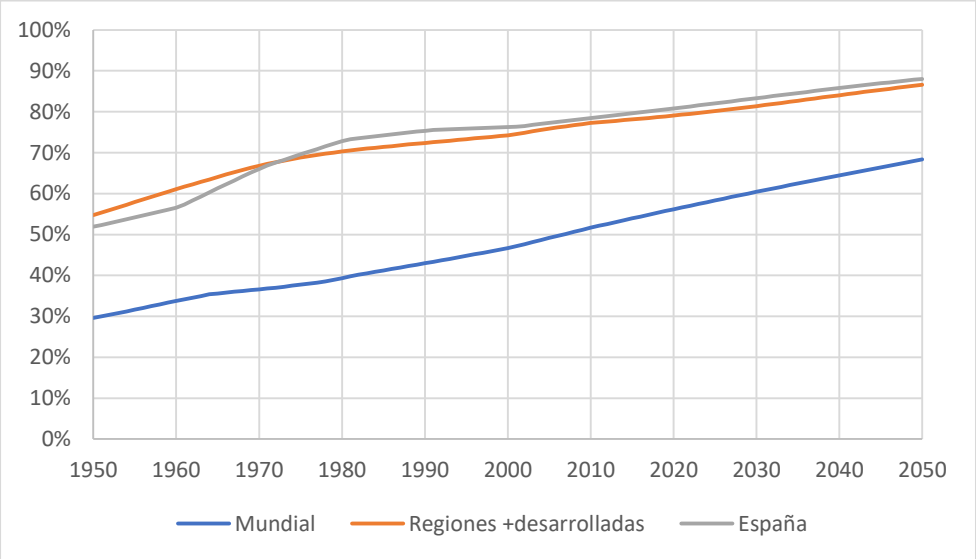

Figura 1. 2. Gráfico de secuencia de la serie anual 1950-2050 del porcentaje de población urbana respecto al total. Elaboración propia a partir del WUP (2018) de Naciones Unidas.

El marco conceptual más ampliamente utilizado en este ámbito es el de los modelos de interacción entre el sistema de usos del suelo y el sistema de transporte (LUTI). Estos modelos permiten simular matemáticamente la interacción entre los diferentes fenómenos. Por tanto, los modelos LUTI permiten pronosticar la evolución del sistema urbano, tanto el actual como el de escenarios que simulan la implantación de nuevas políticas o cambios exógenos. Los modelos LUTI pueden ayudar a responder cuestiones como la de cómo se distribuirá geográficamente el crecimiento de la población en un área urbana determinada, de qué depende la elección de una localización concreta para la vivienda familiar o el centro de trabajo o bajo qué circunstancias los proyectos de transporte apoyan el desarrollo de usos del suelo (Martínez, 2000). Las dos primeras de estas cuestiones se responden con uno de los submodelos LUTI, el modelo de elección de la localización residencial. Tal y como se detalla en el siguiente capítulo, este modelo de reparto espacial de la residencia tiene un enfoque desagregado, que se centra en el estudio de la elección individual. En la literatura de modelos de elección de localización residencial hay planteamientos basados en la maximización de la entropía. De todas formas, el planteamiento más extendido actualmente en contexto LUTI es el econométrico desagregado de los modelos de elección discreta (Pagliara et al., 2010), derivados de la teoría económica de maximización de la utilidad aleatoria. Este planteamiento integra mejor la fuerte naturaleza económica y social que tiene esta decisión, tal y como se analiza en el segundo capítulo de esta tesis.

Las alternativas de elección del planteamiento microeconomético son de tipo multinomial. En el caso de los modelos de elección de localización residencial las

alternativas son además de naturaleza geográfica, por lo que se denominan modelos de elección espacial. La naturaleza espacial de las alternativas hace que estos modelos de elección presenten ciertas características específicas, como son el elevado número de alternativas y la presencia de dependencia espacial entre ellas. Pero en muchas aplicaciones en contexto LUTI no se ha tenido en cuenta la correlación especial entre alternativas, o al menos no de forma específica. En las aplicaciones más recientes sí se especifican modelos que suponen la presencia de correlación entre alternativas. En la literatura de los modelos de elección discreta se han planteado, hasta la fecha, dos enfoques que son compatibles con alternativas de naturaleza espacial. Un enfoque, surgido a partir del modelo nested logit, se basa en el uso de estructuras de agrupación de alternativas, llamadas nidos. Los nidos deben ser diseñados por el analista a partir del conocimiento del contexto de aplicación, para recoger la correlación entre las alternativas de elección. Este primer enfoque no es específico de la correlación espacial entre alternativas. El segundo enfoque, surgido a partir del modelo spatially correlated logit, sí es específico de la correlación espacial entre las alternativas y utiliza métricas espaciales para recogerla. Esta tesis postula que ambos enfoques son compatibles y que su integración ofrece un potencial de mejora sobre los modelos actuales, especialmente a la hora de utilizar la información espacial de las alternativas para mejorar su capacidad explicativa y predictiva.

1.2. Objetivos

El objetivo principal de esta tesis es la propuesta de avances en los modelos de elección discreta con correlación espacial, que mejoren su capacidad explicativa y predictiva, especialmente en contexto de predicción de la demanda de usos de suelo y transporte. En concreto, esta tesis se plantea el objetivo de proponer un nuevo modelo de elección discreta con correlación espacial, que sea compatible con la elección de la localización residencial en un contexto LUTI, y que mejore la capacidad explicativa y predictiva de los modelos actuales compatibles con este contexto.

Para lograr este objetivo, esta tesis plantea la hipótesis de que los dos enfoques actuales de modelización en este campo, que fueron descritos en el apartado anterior, son compatibles. Además, esta tesis postula que se puede diseñar un modelo que integre ambos enfoques, que sea compatible con un contexto de elección de la localización residencial y que mejore la capacidad explicativa y predictiva de los modelos actuales compatibles con este contexto. Para lograr este objetivo general, se han perseguido los siguientes objetivos específicos:

- Analizar el estado del conocimiento de los modelos de elección discreta, con especial hincapié en la compatibilidad con correlación espacial entre alternativas de elección.
- Proponer una nueva métrica de la correlación espacial entre alternativas de elección, que mejore las existentes para la elección de la localización residencial en determinados contextos urbanos caracterizados por zonificaciones irregulares o con tamaños de zona muy diferentes.
- Proponer un nuevo modelo de elección discreta compatible con la elección de la localización residencial en un contexto urbano, que integre los dos enfoques

actuales en este campo, y que, al menos en este contexto, mejore la capacidad explicativa y predictiva de los modelos actuales.

- Proponer un marco metodológico que permita comparar empíricamente el modelo propuesto con los de los modelos actuales. Poner en práctica este marco de comparación en una aplicación real de la métrica y del modelo propuesto y de sus alternativas actuales.

1.3. Estructura de la tesis

El contenido de esta tesis está organizado en cinco capítulos. El primero es la presente introducción. El segundo capítulo aborda el marco teórico y econométrico, profundizando en el conocimiento de los modelos de predicción de la demanda urbana de usos del suelo y transporte en los modelos LUTI y en los modelos de elección de localización residencial. En este capítulo también se analizan los modelos de elección discreta y su compatibilidad con elecciones entre alternativas de naturaleza espacial. En concreto, en este capítulo se analiza el enfoque para modelizar la correlación entre alternativas de elección basado en agrupaciones de alternativas, conocidas como nidos. Además, en este capítulo se aplican, en un contexto empírico real, todos los modelos compatibles con elección espacial que utilizan el enfoque que se analiza a lo largo del capítulo, y se propone una metodología para compararlos.

En el tercer capítulo se analizan los modelos con enfoque de correlación espacial actuales y se propone una generalización parsimoniosa de este tipo de modelos. También se propone una métrica espacial de las alternativas que sea apropiada para zonificaciones basadas en áreas administrativas que tienen diferentes tamaños y formas irregulares, hecho muy habitual en contextos urbanos, como el utilizado en esta tesis para aplicar los modelos. En el capítulo se incluye la aplicación de los modelos con el enfoque analizado a lo largo del capítulo, así como los modelos de la generalización propuesta, incluyendo el especificado con la métrica propuesta.

En el cuarto capítulo se formula y analiza un nuevo modelo de elección discreta, propuesto en esta tesis para modelización de elección espacial. El modelo propuesto se aplica en el contexto empírico y se compara con los modelos analizados en los capítulos anteriores. En el quinto capítulo se presentan las conclusiones de esta tesis y las líneas futuras de investigación. El capítulo Bibliografía recoge el detalle de las citas científicas incluídas a lo largo de los capítulos anteriores.

A lo largo del proceso de desarrollo de esta tesis se han elaborado varias publicaciones relacionadas con ella, como el artículo Perez-Lopez et al. (2021) citado en el apartado Motivación. El último capítulo, Anexos, recoge las siguientes publicaciones directamente derivadas de esta tesis (en la versión restringida sólo se incluye la tercera publicación por estar cedidos los derechos de las dos primeras a un editor):

- Pérez-López J-B y Orro A (2016) Residential location choice models with spatial correlation. In: Dell'Olio L, Cordera R y Ibeas A (eds) Land Use - Transport Interaction Models. The TRANSPACE model. Santander, GIST: 114–150.
- Pérez-López J-B, Novales M, Varela-Garcia F-A y Orro A (2020) Residential location econometric choice modeling with irregular zoning: common border

- Perez-Lopez J-B, Novales M y Orro A (2022) Spatially correlated nested logit model for spatial location choice. Transportation Research Part B: Methodological, 161 (2022): 1-12. https://doi.org/10.1016/j.trb.2022.05.007.

Parte de los contenidos de esta tesis están reflejados en esas publicaciones. Al final de la tesis se indican también otras publicaciones en las que ha participado el autor relacionadas con la tesis doctoral. Parte del desarrollo de esta tesis se ha llevado a cabo dentro de los proyectos TRANSPACE (TRA2012–37659, financiado por el Ministerio de Economía y Competitividad) e IMPROVEBUS (RTI2018-097924-B-I00, financiado por la Agencia Estatal de Investigación-MCIN y FEDER).

1.4. Aportaciones

En esta tesis se parte de una revisión de la literatura para llevar a cabo nuevos desarrollos teóricos y experimentales. La principal aportación consiste en el desarrollo de un nuevo modelo de elección discreta, compatible con elección entre alternativas espaciales, que se ha mostrado eficiente en contexto LUTI. Se incluye el desarrollo de sus formulaciones de correlación y elasticidad. Otra aportación importante de esta tesis es la propuesta de una métrica espacial que se ha mostrado eficiente en los modelos logit con correlación espacial aplicados en un contexto empírico urbano con una zonificación basada en áreas administrativas, especialmente si tienen diferentes tamaños y formas irregulares. También se ha propuesto en esta tesis un planteamiento de la generalización del modelo spatially correlated logit, que es parsimonioso y que se ha mostrado eficiente en la aplicación empírica realizada en esta tesis. Otra aportación de esta tesis es la deducción de la función generatriz GEV del modelo paired generalized nested logit, que es un modelo importante en la revisión bibliográfica realizada en esta tesis. Hasta donde conoce el autor de esta tesis, esta formulación no había sido deducida en la literatura. Además, en esta tesis se propone un marco metodológico para comparar la capacidad explicativa y predictiva de los modelos aplicados, incluyendo la propuesta de una medida de capacidad de predicción mediante medias geométricas.

2. Modelos de elección espacial para localización residencial

En este capítulo se analiza el marco teórico y econométrico de los modelos de demanda urbana de usos del suelo y transporte, en contexto LUTI y aplicado a la elección de localización residencial. Además, se analizan los modelos de elección discreta y su compatibilidad con elecciones entre alternativas de naturaleza espacial. En concreto, se analiza el primer enfoque utilizado para recoger correlación entre alternativas de elección en este tipo de modelos, basado en agrupaciones jerárquicas de las alternativas que deben ser diseñadas por el analista. Por último, en este capítulo se aplican todos los modelos con este enfoque presentes en la literatua y se propone una metodología para comparar su capacidad explicativa y predictiva. Esta propuesta ha sido publicada en el capítulo de libro Pérez-López y Orro (2016).

2.1. Marco teórico y econométrico

2.1.1. Modelos de predicción de usos de suelo y transporte

En muchos ámbitos científicos y profesionales, los estudios de demanda utilizan cada vez más habitualmente los modelos matemáticos de predicción del comportamiento individual ante elecciones de tipo multinomial. Este es el planteamiento que se utiliza en los modelos de interacción entre usos del suelo y transporte (land-use transport interaction, LUTI). Este enfoque es el más ampliamente utilizado en la actualidad en este campo.

Los modelos LUTI se basan en la distribución geográfica de la población y sus actividades, así como en las redes de transporte que las comunican. Hansen (1959) hizo el planteamiento pionero de que las decisiones de desplazamiento y localización se determinan la una a la otra. Los estudios posteriores en este campo ponen de manifiesto el reconocimiento generalizado de la noción de interacción entre usos del suelo y transporte. Acheampong y Silva (2015) analizan las seis décadas de investigaciones en este campo desde la publicación de Hansen. Estas investigaciones se centran en comprender y predecir las elecciones de localización de las zonas de residencia y trabajo, y su interacción con los patrones de transporte derivados de las actividades diarias, incluyendo la elección del modo de transporte y la ruta. Los modelos LUTI permiten una representación abstracta del funcionamiento del sistema urbano o regional. Una vez que el modelo ha sido calibrado frente a un escenario conocido, puede ser usado para hacer predicciones sobre el estado futuro del sistema. Por lo tanto, los modelos LUTI pueden ser usados como una herramienta de ayuda a la toma de decisiones en el campo de la planificación urbana y el transporte. La arquitectura de los modelos LUTI se basa en una estructura espacial urbana y dos componentes principales: el subsistema de usos del suelo y el subsistema de transporte (Torrens, 2000). La figura 2.1 representa el marco conceptual de los modelos LUTI, formado por los dos subsistemas y su interacción. El subsistema de usos del suelo influye en el subsistema de transporte debido a la necesidad de movilidad de personas y el transporte de mercancías (Wegener, 1994). La población localizada en un área concreta demanda transporte para mercancías y para realizar actividades. Las mercancías y las actividades pueden estar localizadas en diferentes áreas del sistema territorial, que se comunican a través de las redes de transporte. El subsistema de usos del suelo se relaciona con el de transporte a través del concepto de generación de viajes. Recíprocamente, el subsistema de

transporte influye en el modelo de elección de la localización residencial a través de la accesibilidad. La idea de accesibilidad fue introducida por Hansen (1959), entendida como "el potencial de oportunidades de interacción entre zonas". Bhat et al. (2002) describen cinco tipos principales de medidas de accesibilidad, como son las medidas de utilidad, las medidas de tiempo o espacio, las medidas de gravedad, la separación espacial y las oportunidades acumuladas. Geurs y van Wee (2004) interpretan la accesibilidad como la facilidad con la que una persona puede acceder a una actividad en otro lugar utilizando un sistema de transporte.

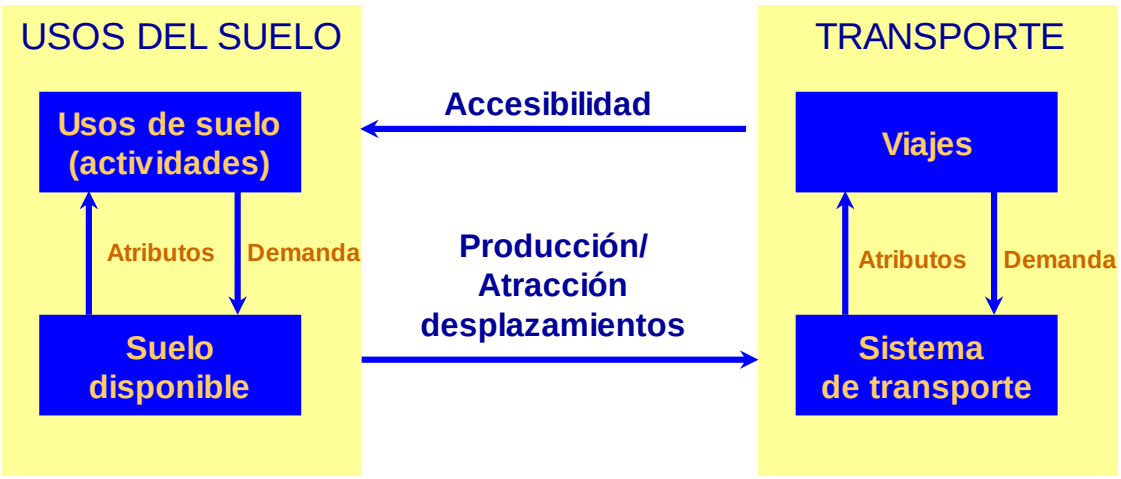

Figura 2.1. Estructura básica de un modelo LUTI. Adaptado de "Integrating Transportation and Land Use Planning: Addressing the Requirements of Federal Legislation and Rule Making". Louden, W. et al. TRB Paper Nº 971022. Enero, 1997 y memoria del proyecto TRANSPACE.

Para predecir la demanda de transporte, el subsistema de transporte de los modelos LUTI se centra en comprender el patrón de viaje. En la literatura se pueden encontrar dos enfoques principales para modelizar la demanda de transporte, el modelo de cuatro etapas y los modelos basados en actividades. A partir de la década de los cincuenta del siglo XX, el enfoque del modelo de cuatro etapas fue el habitual para predecir la demanda de transporte y para evaluar el funcionamiento de los sistemas de transporte y de los proyectos de infraestructuras de transporte a gran escala (McNally, 2000). Las cuatro etapas de este enfoque son la etapa de generación de viajes, la de distribución de viajes, la de elección modal y la de asignación de ruta. La unidad de análisis con este enfoque es el viaje, considerando como tal un desplazamiento de un origen a un destino con un motivo, que puede emplear diversos modos de transporte. La unidad espacial en la que ocurren los viajes son zonas de análisis de transporte, que se definen a partir de características de usos del suelo, administrativas, demográficas y socio-económicas (Fox, 1995; Martínez et al., 2007). El enfoque basado en actividades para los modelos de predicción de la demanda de transporte cobra fuerza, como alternativa al modelo de cuatro etapas, a partir de la década de los noventa del siglo XX. El principio fundamental de este enfoque es que la demanda de transporte se deriva de que las personas necesitan el transporte para realizar actividades interconectadas (McNally y Rindt, 2007). Acheampong y Silva (2015) recopilan diversas aplicaciones y técnicas de modelización con el enfoque de actividades, como son CEMDAP (Bhat et al., 2004), MORPC (PB Consult, 2005), FAMOS (Pendyala et al., 2005), NYMTC (Vovsha y Chiao, 2008), SFCTA (Outwater y Charlton, 2008) y SACSIM (Bradley et al., 2009). La mayoría de

los modelos LUTI existentes en la práctica emplean el modelo de cuatro etapas para el submodelo de transporte, especialmente los de enfoque de interación espacial y basados en la teoría de la utilidad (Acheampong y Silva, 2015).

El subsistema de usos del suelo representa la localización de la población y de sus actividades. Las principales actividades son las que se realizan en el hogar o residencia familiar y en el centro de trabajo. Adicionalmente, se consideran actividades auxiliares, como la formación, las actividades relacionadas con los cuidados, las compras o el ocio. Una de las principales áreas de investigación en los modelos LUTI ha sido comprender el comportamiento a largo plazo de las decisiones de localización (y re-localización) de la residencia y el centro de trabajo, así como su interdependencia (Acheampong y Silva, 2015). Según la teoría económica de maximización de la utilidad, en igualdad de condiciones, se elegirá la localización residencial que minimice los costes de transporte a la localización del centro de trabajo. El planteamiento habitual en el submódulo de usos del suelo de los modelos LUTI se basa en el enfoque económico, y supone que la localización del centro de trabajo es predeterminada o exógena a la elección de la localización residencial (Wadell, 1993; Wadell et al., 2007). Algunos modelos plantean que las elecciones de localización de residencia y centro de trabajo se deciden conjuntamente (por ejemplo, Boschmann, 2011), aunque este enfoque presenta un reto de multi-dimensionalidad (Wadell et al., 2007).

2.1.2. Modelos de elección de la localización residencial

Los modelos de elección de la localización residencial, en un planteamiento de interacción entre usos del suelo y transporte, tienen como objetivo predecir dónde las personas eligen la localización geográfica de su residencia familiar y/o comprender cómo lo hacen. Estos modelos son un tema de interés no solo en ingeniería civil, sino también en otros ámbitos científicos y profesionales como son el urbanismo, la economía urbana, la sociología, la psicología o la geografía (Guo y Bhat, 2004). Este interés se debe, en gran medida, a que el área residencial ocupa en torno a dos tercios del suelo urbano. Además, la elección de la localización residencial es un factor muy importante para la calidad de vida de la ciudadanía. El individuo decisor puede ser tanto una persona, como una familia, organización o algún tipo de entidad de decisores. Para cualquiera de ellos las características de su residencia son muy importantes, ya que determinan, por ejemplo, el precio de compra o alquiler, que en la mayoría de los casos es uno de los conceptos más importantes de las finanzas familiares. Además, las características de una residencia influyen decisivamente en la disponibilidad de infraestructuras de servicios para el individuo decisor y su familia, que es un factor clave en sus relaciones personales y en el tipo de la actividad social que pueden realizar. Por ejemplo, los servicios de transporte disponibles influyen en el tiempo y dinero dedicado al transporte, e indirectamente, influyen también en el tiempo y presupuesto disponible para realizar otras actividades. La localización de la zona en la que se ubica es una de las características esenciales de una residencia familiar. Por ejemplo, los servicios de transporte disponibles son muy diferentes de unas zonas urbanas a otras. Además, el tipo de vivienda que se busca no está disponible en todas las zonas residenciales (Guo y Bhat, 2007).

Pinjari y Bhat (2011) consideran que la elección de la localización residencial impacta directamente en la estructura espacial y define el entorno de actividades – transporte disponible para cada domicilio o persona. Para Schirmer et al. (2014), la elección de la localización residencial no solo influye en el sistema de transporte, también activa otras dinámicas urbanas. Los autores hacen una revisión del enfoque de la modelización de la elección de la localización residencial. Los avances pioneros en modelización de usos del suelo se basaron en gran medida en las ideas de von Thunen (1826) y de Hansen (1959). Posteriores trabajos se basaron en la interacción espacial (ver Lowry, 1964), utilizando la aproximación teórica de la maximización de la entropía. Las generaciones más recientes de modelos de elección de la localización residencial utilizan un enfoque teórico económico. El motivo principal para elegir este enfoque es la enorme importancia que suelen tener los aspectos económicos en esta decisión, que ya hemos visto que es crítica para la mayor parte de las familias.

En la teoría económica hay dos enfoques econométricos para modelizar la demanda residencial de cada zona, así como otros tipos de elecciones, el enfoque agregado de la econometría clásica y el desagregado de la microeconometría (ver Koppelman y Bhat, 2006). El enfoque agregado modela directamente la cuota de mercado de decisores que escogen cada alternativa. Lo hace mediante una función de características de las alternativas y de atributos socio-demográficos de categorías de decisores. Este enfoque agregado es el enfoque de la econometría clásica, que utiliza principalmente modelos de regresión lineal. Por otro lado, el enfoque desagregado postula que el comportamiento agregado es el resultado de numerosas decisiones individuales. Este enfoque modela primero la elección individual de cada decisor, como función de las características individuales de las alternativas, de atributos socio-demográficos de cada individuo y de otros factores que puedan influir en la elección. Los resultados obtenidos en el modelo individual se pueden agregar posteriormente, para establecer el reparto de la demanda total entre las alternativas de elección. Siguiendo este enfoque desagregado, Alonso (1960 y 1964) considera que la localización residencial se elige porque maximiza la utilidad, en función de ciertas características económicas y sociales. McFadden (1978) avanzó en este enfoque teórico económico desagregado para la modelización de la localización residencial. El autor estableció el marco de los modelos de elección discreta, derivados de la teoría económica de maximización de la utilidad aleatoria (RUM; Thurstone, 1927; McFadden, 1974).

Una fortaleza del marco de los modelos de elección discreta es su capacidad para cuantificar el impacto de diferentes características de la localización y de su interacción con características de la residencia. Koppelman y Bhat (2006) consideran que el enfoque desagregado es de naturaleza causal, y que explica los motivos por los que el individuo toma la decisión, dadas sus circunstancias. Este enfoque utiliza microdatos para la estimación del modelo. Los microdatos proporcionan mayor variación en el comportamiento de interés y sus factores determinantes que los datos agregados, por tanto, los modelos que los utilizan son más fiables por unidad de dato que los que utilizan datos agregados. En base a estas ventajas, los autores consideran que el enfoque desagregado de los modelos de elección discreta es más apropiado que el agregado a la hora de modelizar y predecir el comportamiento de toma de decisiones de individuos, por los siguientes motivos: 1) captura más eficientemente el comportamiento de elección y proporciona predicciones más precisas; 2) tiene un comportamiento más

eficiente a la hora de transferirlo en el tiempo o el espacio; 3) es más adecuado que el agregado para el análisis de políticas proactivas; 4) está menos vinculado a los datos de estimación; 5) es más probable que incluya las variables analizadas. Pagliara et al. (2010) afirman que los modelos de elección discreta son el marco matemático más utilizado en los modelos de predicción de elección de localización residencial en el contexto LUTI. Estos modelos también se han utilizado extensamente en otros aspectos del transporte, especialmente en modelos de prognosis (Ben-Akiva et al., 2002), como los modelos de elección de modo de transporte (Koppelman y Bath, 2006; Anta et al., 2016) o destino (Bhat et al., 1998; Train, 1998). También se utiliza en muchos otros campos, como la elección de marca en marketing (Kalyanam y Putler, 1997; Bucklin et al., 1995) o para capturar el comportamiento de compra (Takahashi, 2019).

Tras el pionero trabajo de McFadden (1978), se realizaron otros estudios acerca de elección de la localización residencial con el enfoque desagregado de los modelos de elección discreta, como los estudios de Weisbrod et al. (1980), Vyvere et al. (1980) y Anas (1982). Pagliara et al. (2010) y Schirmer et al. (2014) revisan un amplio espectro de estudios de este tipo más recientes. Los estudios revisados por los autores se aplican a diferentes tipos de áreas espaciales a lo largo del mundo, como ciudades, áreas metropolitanas, regiones o países. Además, los estudios utilizan diferentes zonificaciones del espacio de las alternativas de elección.

En análisis espacial, las localizaciones del área geográfica en estudio pueden ser tratadas con un planteamiento continuo o discreto del espacio (Pagliara et al., 2010). En los modelos de elección espacial, las alternativas de elección pueden ser un gran número de unidades espaciales muy pequeñas, que en el límite pueden simular un continuo. En el caso de los modelos de elección de la localización residencial, estas localizaciones pueden ser viviendas unifamiliares, edificios o unidades residenciales. Schirmer et al. (2014) revisan diversos estudios con este planteamiento, como es el caso del estudio de Axhausen et al. (2004), aplicado a la ciudad de Karlsrue en Alemania; el estudio de Kim et al. (2005), aplicado a Oxfordshire en UK; el estudio de Bürgle (2006), aplicado al área de Gran Zúrich en Suiza; el estudio de Habib y Miller (2009), aplicado al área de Gran Toronto en Canadá; los estudios de Lee y Waddell (2010), aplicados al estrecho de Puget en EE.UU.; el estudio de Zhou y Kockleman (2008), aplicado a la ciudad de Austin en EE.UU.; y el estudio de Belart (2011), aplicado al cantón Zúrich en Suiza.

El tratamiento discreto del espacio facilita la integración del submodelo de usos del suelo con el de transporte en contexto LUTI. Una opción es que el tratamiento discreto se base en el diseño de mallas o celdas espaciales, pero este tratamiento discreto del espacio es muy poco frecuente en modelización de la elección de la localización residencial en contexto LUTI. El principal motivo es la dificultad para obtener datos de los atributos de alternativas de elección. Schirmer et al. (2014) solo incluyen el caso del estudio de Waddell (2006), aplicado a la región del estrecho de Puget en EE.UU. El tratamiento del espacio más habitual en modelos de elección de la localización residencial para contexto LUTI se basa en la zonificación del territorio. Pueden emplearse las zonas de análisis de transporte o, en general, zonificaciones basadas en barrios o en algún tipo de áreas administrativas (como distritos o secciones censales, ayuntamientos, etc.). Esta zonificación del espacio tiene la ventaja de que se puede aplicar en la práctica con cierta facilidad, porque es habitual que existan datos estadísticos de las áreas administrativas, que en muchos casos son públicos. Esta

zonificación del espacio de alternativas es la que aplicamos en esta tesis. Schirmer et al. (2014) incluyen diversos estudios de este tipo, como el estudio de Ben-Akiva y Bowman (1998), aplicado a la ciudad de Boston en EE.UU.; el estudio de Srour et al. (2002), aplicado al condado de Dallas en EE.UU.; los estudios de de Palma et al. (2005 y 2007), aplicados al área de Gran París en Francia; el estudio de Zondang y Pieters (2005), aplicado a los Países Bajos; el estudio de Andrew y Meen (2006), aplicado a Londres y Sudeste de Inglaterra; el estudio de Guo y Bhat (2007), aplicado al área de la Bahía de San Francisco en EE.UU.; el estudio de Chen et al. (2008), aplicado a la región del estrecho de Puget en EE.UU.; los estudios de Pinjari et al. (2009 y 2011), aplicados a la Bahía de San Francisco (EE.UU.) y el estudio de Zolfaghari et al. (2012), aplicado al área de Gran Londres (Reino Unido).

El analista no puede incluir todas las variables explicativas en la función de utilidad observada, por la dificultad para encontrar datos de muchas de ellas y para evitar multicolinealidad. La mitad de los estudios revisados por Schirmer et al. (2014) utilizan veinte o menos variables explicativas, y dos tercios de ellos utilizan quince o menos variables explicativas. Las variables explicativas pueden ser de diferentes tipos. En el caso de que las alternativas sean unidades residenciales se pueden incluir variables relativas al precio o valor de las viviendas, su tipología, tamaño y otras características, como la disponibilidad de vistas o el número de plazas de garaje. Las zonas de localización pueden caracterizarse mediante atributos del entorno construido, como la cantidad de espacio verde, las redes de transporte o la densidad de viviendas en la zona; atributos de puntos de interés, como colegios o instalaciones deportivas; y atributos socio-económicos, como la densidad de población y su distribución por cuestiones étnicas, económicas o laborales.

2.1.3. Modelos de elección discreta

El planteamiento de esta tesis para modelizar la elección de la localización residencial, en contexto de predicción de la demanda urbana de usos de suelo y transporte, es el enfoque desagregado de los modelos de elección discreta multinomiales, con una zonificación del espacio en estudio a partir de áreas administrativas. Las alternativas de estos modelos de elección son los elementos de la zonificación, que forman un conjunto discreto y exhaustivo de zonas espaciales mutuamente excluyentes, que denotamos $\{y_1, \dots, y_A\}$. Esta partición del espacio puede considerarse el dominio de la variable que recoge las elecciones individuales, que es de naturaleza cualitativa nominal y espacial, que denotamos Y. La regla de decisión no se puede plantear como una regresión lineal, porque los estimadores estadísticos son sesgados e inconsistentes en esta situación. El planteamiento de los modelos de elección individual del enfoque desagregado se basa en modelizar la probabilidad de que el decisor elija cada una de las alternativas, que denotamos $P_i = P(Y = y_i), i = 1, \dots, A$.

Thurstone (1927) originalmente derivó la probabilidad de elección en términos de estímulos psicológicos. Marschak (1960) interpretó los estímulos como utilidad, y derivó una probabilidad de elección consistente con la teoría económica de la maximización de la utilidad, tal y como se explica a continuación (Train, 2009). En esta teoría económica, un individuo perfectamente informado y racional se enfrenta a la elección entre A alternativas excluyentes. El decisor obtendría un cierto nivel de utilidad o provecho de

cada alternativa $y_j, j \in \langle 1, \dots, A \rangle$, que es una variable que se supone continua y que denotamos U_j. El decisor siempre elije la alternativa que le reporta más utilidad, lo que define el modelo de comportamiento de elección multinominal (2.1).

$$Y = y_i, \forall i \in \{1, \dots, A\} \Leftrightarrow \left\{U_i = \max_{j \in \{1,\dots,A\}} U_j\right\}, \forall i \in \{1, \dots, A\} \Leftrightarrow$$

$$\Leftrightarrow \{U_i \geq U_j, \forall j \in \{1, \dots, A\}, j \neq i\}, \forall i \in \{1, \dots, A\} \quad (2.1)$$

El decisor conoce la utilidad de cada alternativa, pero el analista no. Lo que sí puede hacer el analista es obtener observaciones de ciertas variables. Por un lado, el analista puede obtener observaciones de algunas características de las alternativas que considera el decisor. Por ejemplo, en el caso de modelos de elección de la localización residencial en contexto LUTI, el analista puede obtener observaciones del tiempo dedicado por los decisores al viaje entre la zona de residencia y el centro de trabajo. Por otro lado, el analista puede obtener observaciones de algunos de los atributos socio-económicos de los decisores, como puede ser su edad, su nivel de formación o sus ingresos familiares. Por último, el analista puede obtener observaciones de otras características del contexto de elección, como el motivo de viaje en el caso de modelos de elección modal en transporte. Por tanto, el analista puede especificar una función paramétrica, llamada utilidad observada o sistemática. Cuando los regresores de esta función son específicos de cada alternativa, se dice que la utilidad observada es condicional, y la denotamos $V_i, i \in \{1, \dots, A\}$. Las observaciones recopiladas se pueden utilizar para estimar los parámetros desconocidos de la utilidad sistemática de cada alternativa. En contexto LUTI, la expresión habitualmente elegida para la función de utilidad observada es la de una función lineal en los parámetros, aunque se han propuesto funciones no lineales más flexibles, como la utilización de transformaciones Box-Cox de las variables explicativas (Gaudry y Wills, 1978; Hensher y Johnson, 1981: 186-191; Orro, 2006; Orro et al., 2010). Las funciones de utilidad observada de cada alternativa con estructura lineal se pueden escribir como se muestra en la fórmula (2.2), donde V_i es la función de utilidad observada de la i-ésima alternativa, β es el vector de coeficientes y X_i el vector de regresores de la i-ésima alternativa.

$$V_i = \beta' X_i, \forall i \epsilon \{1, \dots, A\} \quad (2.2)$$

En los modelos de elección discreta con alternativas que tengan características particulares a nivel individual se pueden especificar una serie de constantes específicas para cada alternativa. Estas constantes capturan el efecto de la media de todos los factores que no son observados por las variables explicativas (Ben-Akiva y Bierlaire, 1999:5-34). Los modelos de elección discreta derivados de RUM pueden recoger variaciones sistemáticas en las preferencias o gustos de los decisores, mediante la incorporación de atributos individuales en la utilidad observada. En el caso de funciones de utilidad observada lineales, los atributos individuales se pueden incluir en modo aditivo y/o multiplicativo. En el caso aditivo representan variaciones sistemáticas en las constantes específicas de alternativa, y en el caso multiplicativo, representan variaciones sistemáticas de los coeficientes de los regresores con los que interaccionan. Por tanto, los regresores de la función de utilidad observada pueden corresponder a variables explicativas, funciones de ellas (por ejemplo, potencias, logaritmos o tasas) o de sus interacciones.

Llamamos término de error a la diferencia entre la utilidad percibida por el individuo y su valor observado, que es desconocida por serlo la utilidad. En expresiones condicionales de la utilidad, el término de error de cada alternativa, que denotamos $\varepsilon_i, i \in \langle 1, \dots, A \rangle$, es su componente aleatoria. Para cada alternativa, esta componente contiene los factores que influyen en la utilidad del decisor pero que no fueron incluidos por el analista en la utilidad observada, además de otros sesgos aleatorios, como la diferencia entre el valor de cada atributo percibido por el individuo y el considerado por el analista o las diferencias en los gustos de los individuos en especificaciones con parámetros fijos. Por este motivo, también se le denomina error aleatorio o perturbación. El esquema de la utilidad de cada alternativa en un modelo RUM establece que la utilidad es una variable aleatoria con el esquema aditivo que se muestra en la ecuación (2.3). Los parámetros desconocidos de la utilidad observada deben ser estimados a partir de muestras aleatorias de elecciones individuales. Habitualmente, la estimación de los parámetros desconocidos se realiza conjuntamente mediante el método estadístico de máxima verosimilitud. Para ello, es necesario que previamente se haya derivado la expresión de la probabilidad de elección de cada alternativa del modelo.

$$U_i = V_i + \varepsilon_i, \forall i \epsilon \{1, \dots, A\} \tag{2.3}$$

La probabilidad de que un decisor n elija cada i-ésima alternativa, P_{ni}, denominada probabilidad de elección de la alternativa, se muestra en la ecuación (2.4). La expresión de la probabilidad de elección de cada alternativa en un modelo RUM se deriva de la probabilidad de que sea la alternativa con mayor utilidad para el decisor, que se expresa y desarrolla matemáticamente en (2.5). La expresión resultante es una función de distribución del vector de errores aleatorios $\varepsilon_n := (\varepsilon_{n1}, \dots, \varepsilon_{nA})$. Por tanto, la probabilidad de elección se calcula resolviendo una integral $(A-1)$-dimensional, tal y como muestra la ecuación (2.6), en la que el integrando es el producto entre la función de densidad conjunta del vector de errores aletorios $f(\varepsilon_n)$ y la función indicadora de la expresión $(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} + V_{nj}, \forall j \neq i)$. La función indicadora $I(\cdot)$, toma el valor 1 cuando la expresión entre paréntesis es cierta y 0 en caso contrario. Como el analista desconoce $f(\varepsilon_n)$, para resolver esta integral se utiliza un planteamiento de inferencia estadística paramétrica. Este planteamiento consiste en suponer que la distribución de probabilidad conjunta de los errores aleatorios es un modelo de distribución de probabilidad conocido.

$$P_{ni} = P(Y_n = y_{ni}), \forall i \epsilon \{1, \dots, A\} \tag{2.4}$$

$$\begin{aligned} P_{ni} &= P\left(U_i = \max_j U_j\right) = P\big(U_{ni} \geq U_{nj}, \forall j \neq i\big) = \\ &= P\big(V_{ni} + \varepsilon_{ni} \geq V_{nj} + \varepsilon_{nj}, \forall j \neq i\big) = \\ &= P\big(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}, \forall j \neq i\big), \forall i \epsilon \{1, \dots, A\} \end{aligned} \tag{2.5}$$

$$P_{ni} = \int_{\varepsilon} I\big(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} + V_{nj}, \forall j \neq i\big) f(\varepsilon_n) d\varepsilon_n, \forall i \epsilon \{1, \dots, A\} \tag{2.6}$$

El modelo estadístico elegido para la distribución de probabilidad conjunta del vector de errores aleatorios, ε_n, determina la forma de $f(\varepsilon_n)$, lo que establece a su vez la forma de la integral de la probabilidad de elección de cada alternativa que tiene el modelo de elección discreta, lo que define su tipología. Por tanto, cada modelo estadístico define un tipo diferente de modelo de elección discreta. En unos casos, la integral se puede

resolver analíticamente. Cuando sucede esto se dice que la probabilidad de elección del modelo tiene una expresión matemática cerrada. Cuando la integral no tiene solución analítica, solo se puede calcular un valor aproximado de la probabilidad de elección, mediante integración numérica y simulación estadística del modelo elegido.

El modelo elegido para la distribución de probabilidad conjunta del vector de errores aleatorios puede ser paramétrico. Por tanto, el tipo de modelo de elección discreta derivado de la teoría de maximización de la utilidad aleatoria que se ha definido puede depender de una serie de parámetros. El valor que tienen los parámetros en cada contexto empírico se estima mediante algún método estadístico, utilizando datos de muestra del contexto empírico de aplicación. En concreto, se estiman conjuntamente mediante el método estadístico de máxima verosimilitud, que denominamos máxima verosimilitud simulada cuando la probabilidad de elección no tiene una estructura matemática cerrada. Este método permite la estimación puntual de parámetros, a partir de muestras aleatorias de las que se conoce la distribución de probabilidad paramétrica conjunta. En particular, permite la estimación conjunta de todos los parámetros desconocidos de modelos de elección discreta de la teoría de maximización de la utilidad aleatoria. Este método se realiza en dos etapas. En la primera etapa se desarrolla la llamada función de verosimilitud del modelo, consistente en la densidad conjunta de una muestra aleatoria observada como función de los parámetros, y que se recoge en la fórmula (2.7), donde β es el vector de parámetros desconocidos del modelo, N_{est} es el tamaño de la muestra de estimación, que se supone aleatoria simple, $P_{ni}(\beta)$ es la probabilidad de que el individuo n de la muestra elija la alternativa i-ésima si el vector de parámetros del modelo toma el valor β, y δ_{ni} es la función indicadora, que tiene el valor 1 si el individuo n de la muestra elije la i-ésima alternativa y 0 en caso contrario. En la segunda etapa del método de máxima verosimilitud se calcula el valor del vector de parámetros que maximiza la función de verosimilitud, que equivale a maximizar el logaritmo natural de la función de verosimilitud (ver fórmula 2.8), debido a la naturaleza monótona creciente de la transformación logarítmica y que facilita la derivación durante el proceso de optimización.

$$L(\beta) := \prod_{i=1}^{A} \prod_{n=1}^{N_{est}} \left(P_{ni}(\beta)\right)^{\delta_{ni}} \tag{2.7}$$

$$LL(\beta) = \ln[L(\beta)] = \sum_{i=1}^{A} \sum_{n=1}^{N_{est}} \delta_{ni} \ln[P_{ni}(\beta)] \tag{2.8}$$

Los principales planteamientos para elegir el modelo de distribución de probabilidad conjunta del vector de errores aleatorios se basan en dos teoremas de la inferencia estadística, que dan lugar a los dos principales tipos de modelos de elección discreta derivados de la teoría de maximización de la utilidad aleatoria, los modelos probit y logit. Los modelos probit utilizan el planteamiento clásico de la inferencia estadística, que se basa en el teorema central del límite. Este teorema postula que, bajo ciertas condiciones, el efecto agregado de diferentes factores aleatorios independientes tiene una distribución de probabilidad asintótica Gaussiana. El error aleatorio de cada alternativa representa el efecto agregado de muchos factores omitidos, cada uno con un impacto relativamente pequeño en el valor de cada alternativa. Por este motivo, el planteamiento de los modelos probit consiste en suponer que el vector de errores aleatorios tiene una distribución de probabilidad conjunta asintótica normal multivariante. El modelo multinomial probit (Daganzo, 1979), supone que el vector de errores aleatorios tiene una distribución de probabilidad conjunta asintótica $N(\vec{0}, \Omega)$,

donde Ω es la matriz de varianzas-covarianzas del vector de errores aleatorios de las alternativas. Como la función de distribución de una variable aleatoria normal multivariante no tiene solución analítica, la probabilidad de elección de los modelos probit no es matemáticamente cerrada. La matriz de varianzas-covarianzas contará con $\mathrm{A}(\mathrm{A}+1)/2$ parámetros desconocidos que es necesario estimar, donde A es el número de alternativas de elección. La complejidad matemática de este modelo dificulta su estimación e interpretación, lo que puede haber limitado la extensión de su uso. Dado que la complejidad aumenta potencialmente con el número de alternativas, este enfoque no suele ser viable para acomodar la correlación entre las alternativas espaciales. Algunos ejemplos son los estudios de Bolduc (1992), Garrido y Mahmassani (2000) y Schnier y Felthoven (2011). Para considerar la correlación espacial entre alternativas, la mayoría de las aplicaciones que se basan en el modelo Probit limitan el número de unidades espaciales posibles. Por ejemplo, Schnier y Felthoven (2011) consideran hasta 24 resultados posibles, mientras que Garrido y Mahmassani (2000) consideran hasta 41.

Los modelos logit son modelos de elección discreta derivados de RUM que utilizan un planteamiento alternativo para la elección de la distribución conjunta de los errores aleatorios. La probabilidad de elección de los modelos de elección discreta se obtiene a partir de un extremo de la distribución de probabilidad conjunta de los errores aleatorios, máximo o mínimo según se especifique, como se vio en la fórmula (2.1). El teorema de valores extremos de la inferencia estadística postula que, bajo ciertas condiciones, el máximo de variables Gaussianas tiene una distribución de probabilidad asintótica valor extremo. En base a este teorema, los modelos logit plantean que el vector de errores aleatorios tiene una distribución de probabilidad conjunta asintótica valor extremo de algún tipo. La distribución valor extremo tiene tres parámetros: el parámetro de localización y el de forma con dominio $\mathbb{R}$ y el parámetro de escala con dominio $\mathbb{R}^{+}$. Esta distribución es combinación de tres tipos de distribuciones: Gumbel o tipo I, Fréchet o tipo II y Weibull o tipo III (Johnson y Kotz, 1970). Bajo ciertos supuestos, las probabilidades de elección de estos modelos tienen una expresión matemática cerrada. Además, las distribuciones de probabilidad que se generan en los modelos de elección discreta de tipo logit tienen formas funcionales bastante semejantes a la normal. El enfoque logit es el que más extensamente se ha utilizado en la práctica para los modelos de elección discreta derivados de la teoría de maximización de la utilidad. Hay diferentes tipos de modelos logit. En los capítulos siguientes de esta tesis se estudian los principales, focalizando en los que se adaptan a los objetivos de la tesis.

Los modelos de elección discreta también pueden considerar variaciones aleatorias en los gustos o preferencias de los individuos decisores, que se pueden implementar especificando algunos de los coeficientes de la función de utilidad observada como variables aleatorias dentro de la población, en lugar de como parámetros desconocidos. Los modelos especificados con este planteamiento se denominan modelos mixtos con coeficientes aleatorios. El enfoque paramétrico de los modelos mixtos con coeficientes aleatorios consiste en suponer que los coeficientes aleatorios se distribuyen en la población según algún modelo de probabilidad paramétrico conocido. De esta forma, el modelo mixto con coeficientes aleatorios incrementa el número de parámetros desconocidos según el número de coeficientes que se hayan especificado como variables aleatorias, y el número de parámetros del modelo de distribución de

probabilidad de cada uno. Los modelos de distribución de probabilidad que se han aplicado más extensamente en este planteamiento son el normal, lognormal, triangular y uniforme. El más habitualmente utilizado es el normal, que es el que supondremos en esta investigación. En este caso, dado que la distribución Gaussiana tiene dos parámetros, la media y la desviación típica, cada coeficiente especificado como variable aleatoria incrementa en uno el número de parámetros desconocidos del modelo.

Para medir la influencia individual de cada regresor en la probabilidad de elección de cada alternativa de un modelo de elección discreta derivado de RUM se utilizan las elasticidades del modelo, dado que la función de densidad del modelo de distribución de probabilidad Gumbel tiene una expresión de tipo exponencial. Dado el m-ésimo regresor de la i-ésima alternativa de un modelo, X_{im}, la elasticidad directa de X_{im}, $E_{X_{im}}^{P_i}$, mide la variación porcentual esperada en la probabilidad de elección de esa alternativa, por cada aumento de un punto porcentual del regresor. Por tanto, $E_{X_{im}}^{P_j}$ es la derivada del logaritmo natural de P_i respecto a la derivada del logaritmo natural de X_{im}, que se puede expresar multiplicando por X_{im} la derivada del logaritmo natural de P_i respecto a X_{im}. La elasticidad cruzada del mismo regresor con cualquier otra j-ésima alternativa, $E_{X_{im}}^{P_j}$, mide la variación porcentual esperada en la probabilidad de elección de la alternativa con la que se cruza, por cada aumento de un punto porcentual en el regresor. Por tanto, $E_{X_{im}}^{P_j}$ se puede expresar multiplicando por X_{im} la derivada del logaritmo natural de P_j respecto a X_{im}.

2.1.4. Modelos de elección espacial

Los modelos de elección espacial son un caso particular de los modelos de elección discreta multinomiales. En el contexto de predicción de la demanda urbana de usos de suelo y transporte, las elecciones de localización espacial más importantes se refieren a la elección de la localización residencial y, en menor medida, a la elección de la localización de actividades, como la del centro de trabajo. La elección entre alternativas de localización espacial también puede aparecer en otros tipos de modelos, como el destino de viaje o los modelos de elección de parada de embarque o desembarque del transporte público.

Weiss et al. (2019) afirman que la elección de la localización espacial generalmente se modeliza utilizando el enfoque econométrico basado en la teoría de maximización de la utilidad aleatoria, es decir, mediante modelos de elección discreta multinomial. Pero no todos los tipos son aplicables con alternativas espaciales. Por un lado, porque los modelos de elección espacial suelen tener muchas alternativas de elección, lo que imposibilita la aplicación de algunos tipos de modelos de elección discreta. Además, ha recibido varias críticas el uso extendido de modelos de elección discreta que no tienen en cuenta patrones de sustitución entre alternativas de localización, como el modelo logit multinomial que se presentará en el siguiente apartado (Chen et al., 2008). Es lógico suponer que hay patrones de sustitución entre diferentes alternativas de localización, por lo menos, dependiendo de la situación espacial de la misma (Hunt et al., 2004; Bhat y Guo, 2004). La dependencia espacial juega un papel clave en todos los fenómenos que involucran el espacio geográfico, como son los procesos sociales asociados al transporte y uso del suelo. A partir de la Primera Ley de Geografía de Tobler (Tobler, 1970), se

establece que todas las unidades espaciales exhiben algún grado de interdependencia espacial. Al estimar modelos estadísticos con datos geográficos, como en el caso de los modelos de transporte y uso del suelo, esta interdependencia espacial toma la forma de correlación espacial entre los datos. El tratamiento de la correlación espacial en los modelos discretos multinomiales, como es el caso de los modelos de elección espacial que nos ocupan en esta tesis, difiere de forma sustancial de su tratamiento en modelos con un término de error aleatorio por observación, como son los modelos continuos (Cressie, 1993) o discretos binarios (LeSage, 2000; Ward y Gleditsch, 2002).

Como pone de manifiesto Bahamonde-Birke (2021), dada la complejidad del tratamiento de la correlación espacial en modelos de elección discreta multinomial, esta no ha recibido el mismo nivel de atención en la literatura que correlaciones de tipo no-espacial, como la correlación entre los posibles resultados dadas sus propias características (por ejemplo, la independencia entre alternativas irrelevantes), la correlación entre diferentes niveles de decisión (por ejemplo, la elección simultánea de modo de transporte y destino del viaje) o la correlación entre las respuestas proporcionadas por el mismo individuo (por ejemplo, datos de panel, pseudo-panel o preferencias declaradas). La complejidad del tratamiento de la correlación espacial en los modelos de elección discreta multinomial no proviene de su tratamiento teórico, sino de su aplicación a casos reales. Desde un punto de vista teórico existen diversos modelos que pueden incorporar la correlación espacial. Por ejemplo, el modelo Probit totalmente especificado o algunas especificaciones del modelo mixed logit, que serán descritos en la siguiente sección de esta tesis, pueden capturar completamente la correlación espacial entre los diferentes resultados. Pero la aplicación de estos modelos a casos reales de elección espacial tiene unos requerimientos de disponibilidad de software, de datos y computacionales que complican, o directamente impiden, su utilización. La naturaleza de estas dificultades será explicada en la siguiente sección de esta tesis, durante la descripción de cada uno de estos modelos derivados de la teoría de maximización de la utilidad aleatoria. Consecuentemente, la correlación espacial termina siendo completamente ignorada en muchas aplicaciones (Sener et al., 2011) o como mucho, tratada de forma excesivamente simplificada.

En modelos de elección discreta multinomial, Bahamonde-Birke (2021) diferencia la correlación espacial inducida a nivel de observaciones y a nivel de alternativas de elección. La correlación espacial entre diferentes observaciones se asemeja a un proceso kriging de modelos continuos en inferencia estadística espacial. Por ejemplo, en un modelo de distribución del espacio urbano en el que la variable dependiente sea el uso que se le da a cada unidad espacial (residencial, comercial, zona verde, u otros), es lógico suponer que unas zonas tengan mayor probabilidad de tener el mismo uso que otras, por algún tipo de dependencia espacial. Podría tratarse, por ejemplo, de las zonas más próximas o contiguas en alguna dirección. Los modelos de elección de la localización residencial no presentan este nivel de correlación espacial. Por el contrario, cuando las alternativas de elección son unidades espaciales, es muy probable que aparezca dependencia espacial entre ellas. Esta correlación espacial entre alternativas se refiere a las preferencias de sustitución por parte del tomador de decisiones. Cuando un decisor tiene una valoración elevada de una zona, es probable que las zonas con mayor relación espacial (por ejemplo, las más próximas) también tengan valoraciones elevadas y que, a igualdad de características observadas, sean mejores sustitutas que otras zonas con

menor relación espacial. Las preferencias se deben a elementos espaciales no observados de la utilidad. Es lógico suponer que esta dependencia sucede en los modelos de elección de la localización residencial que nos ocupan en esta tesis. El alto número de alternativas de elección que suelen presentar los modelos de elección espacial impide o dificulta el uso de algunos enfoques para capturar esta correlación. Ya hemos visto que esto sucede con el modelo probit, si no se incluyen restricciones en la estructura de correlación de los errores aleatorios de este modelo. También sucede con el modelo mixed logit, que utiliza una estructura de componentes de error que permite patrones flexibles de correlación entre alternativas (Train, 2009: 143-145). Pero este enfoque suele ser inviable para la correlación entre alternativas espaciales, porque dichas correlaciones requieren especificar tantos componentes de error como pares de alternativas correlacionadas, que suelen ser demasiados para el proceso de estimación.

2.1.5. Endogeneidad

La endogeneidad es un tema relevante en los modelos de elección discreta y, más concretamente, en los modelos de elección de la localización residencial. Esta endogeneidad en la elección de la localización espacial fue el tema de la tesis doctoral de Guevara-Cue (2005), supervisada por Moshe Ben-Akiva. La endogeneidad ocurre cuando existe correlación entre el término de error y alguna variable independiente de la función de utilidad observada. Siguiendo a Guevara (2010) y Guevara y Ben-Akiva (2012), en modelos de elección de la localización residencial, la endogeneidad generalmente se debe a la omisión de algunos de los atributos que pueden influir en la elección de la ubicación residencial de un hogar. Como es probable que otros atributos de una vivienda estén correlacionados con el precio, el término de error estará correlacionado con el precio observado y, por lo tanto, el modelo sufrirá de endogeneidad. Esta mala especificación supondrá que el impacto del precio en el proceso de elección no se establecerá correctamente, y los estimadores de los parámetros del modelo pueden ser inconsistentes. Cada vivienda es casi única, y no es posible que el investigador tenga en cuenta todos sus atributos. Los atributos omitidos como la calidad de la construcción, el diseño, el estado, las vistas, la ubicación dentro del edificio o el mobiliario estarán correlacionados con el precio y también con las alternativas, lo que puede causar esta endogeneidad.

Este problema se ha tenido en cuenta en varias investigaciones y existen algunos métodos para abordarlo, como el método de función de control o la función de control actualizada (para el modelo de transporte estratégico, véase Guerrero et al., 2021). Sin embargo, hasta donde sabemos, las investigaciones que consideran la endogeneidad en los modelos de elección discreta para las elecciones de ubicación espacial se refieren a modelos donde las alternativas son la unidad de vivienda específica para vivir. Esta situación puede presentarse, por ejemplo, en enfoques de microsimulación. En esos casos, la endogeneidad es casi inevitable y su principal causa es la omisión de atributos de la vivienda que se correlacionan con el precio (Guevara y Ben-Akiva, 2012). El método de función de control requiere encontrar variables instrumentales que estén correlacionadas con el precio pero que no estén correlacionadas con el término de error. Siguiendo la idea de tomar como variables instrumentales el precio promedio observado del mismo producto en zonas adyacentes, Guevara y Ben-Akiva (2006) utilizan el precio promedio de otras unidades de vivienda ubicadas en el mismo municipio. Más

detalladamente, las variables instrumentales que emplearon para resolver la endogeneidad en el precio se construyen como un promedio de los precios de otras viviendas próximas con atributos observados (diferentes al precio) similares (Guevara, 2010). Sus resultados muestran que se trata de una variable instrumental apropiada.

El autor de esta tesis no ha encontrado ningún estudio de modelización de la elección entre zonas residenciales que haya considerado la existencia de un problema de endogeneidad. Aunque varios artículos (Guevara, 2010 p. 22; Guevara y Ben-Akiva, 2012; Guevara, 2015 o Guevara y Polanco, 2016) parecen señalar que Bhat y Guo (2004) informan que los coeficientes estimados del precio de la vivienda no pueden ser significativos, o incluso positivos, cuando no se abordó la endogeneidad. En dicho trabajo, Bhat y Guo encontraron un coeficiente de precio no significativo (así se referencia en Guevara y Ben-Akiva, 2006), pero consideran que esto puede ser consecuencia de la resolución utilizada para representar la ubicación en este estudio. No hay referencia sobre endogeneidad en Bhat y Guo (2004). Por otro lado, en un artículo reciente (Gopalakrishnan et al., 2020), los autores abordan el problema de la endogeneidad en el precio, lo que resulta en un coeficiente de precio estadísticamente igual a cero. Pero dijeron que "si bien no se puede descartar que puedan existir otras fuentes de endogeneidad para otras variables, no hay una razón previa para pensar que eso podría ser relevante, y los resultados del modelo no corregido parecen respaldar esa afirmación".

2.2. Modelos logit

En este apartado se estudian los modelos logit más extensamente presentes en la literatura. Algunos de estos modelos consideran correlaciones entre alternativas, pero no tienen en cuenta la correlación espacial entre alternativas, o al menos no la tienen en cuenta de forma específica.

2.2.1. Modelo multinomial logit

El modelo de elección discreta derivado de RUM más extendido y simple es el modelo multinomial logit (MNL; McFadden, 1974; Domencich y McFadden, 1975). Este modelo logit supone que los errores aleatorios de la utilidad son independientes entre alternativas y entre individuos, con distribuciones de probabilidad marginal Gumbel idénticas. El modelo de distribución de probabilidad Gumbel es, como ya se ha señalado, la distribución valor extremo de tipo I, que corresponde a la distribución valor extremo con parámetro de forma igual a cero ($\xi = 0$). La fórmula (2.9) muestra la función de distribución de probabilidad Gumbel(λ, η), donde $\lambda \in \mathbb{R}$ es el parámetro de localización y $\eta > 0$ es el parámetro de escala. Esta distribución es unimodal asimétrica con moda λ, tiene media $\lambda + \gamma\eta$, donde γ es la constante de Euler-Mascheroni; y la desviación típica es $\eta\pi/\sqrt{6}$. Sin pérdida de generalidad, el modelo multinomial logit se puede plantear suponiendo que la distribución de probabilidad marginal de los errores aleatorios es Gumbel típica (Train, 2009:34; Abbe et al., 2007). La distribución Gumbel típica se refiere a la que tiene parámetro de localización de valor cero ($\lambda = 0$) y el parámetro de escala de valor uno ($\eta = 1$), es decir, Gumbel(0,1), de forma que se normaliza la escala de la utilidad. En esta investigación aplicamos esta tipificación a todos los modelos logit que se analizan y proponen. Por tanto, la función de distribución

de probabilidad marginal de los errores aleatorios del modelo multinomial logit es la que se muestra en la fórmula (2.10). Derivando esta expresión se obtiene la función de densidad marginal de los errores aleatorios del modelo multinomial logit que se muestra en la fórmula (2.11). De igual modo, se puede deducir que los errores aleatorios del modelo multinomial logit son unimodales con valor de la moda igual a 0, la media tiene como valor la constante de Euler-Mascheroni y la desviación típica se muestra en la fórmula (2.12). Por tanto, las varianzas de los errores aleatorios del modelo multinomial logit tienen igual valor, es decir, son homocedásticos.

$$F(y) = exp\{-e^{-(y-\lambda)/\eta}\} \tag{2.9}$$

$$F_{\varepsilon_i}(y) = exp\{-e^{-y}\}, \forall i \epsilon \{1, \dots, A\} \tag{2.10}$$

$$f_{\varepsilon_i}(y) = e^{-y} exp\{-e^{-y}\}, \forall i \epsilon \{1, \dots, A\} \tag{2.11}$$

$$\sigma = \frac{\pi}{\sqrt{6}} \tag{2.12}$$

Además de homocedásticos, los errores aleatorios de este modelo son incorrelados entre alternativas y entre observaciones. La fórmula (2.13) muestra la matriz de varianzas-covarianzas de este modelo, que presenta una estructura escalar, donde I_A es la matriz identidad de orden igual al número de alternativas. Pero el modelo multinomial logit no recoge variaciones aleatorias en estas preferencias o gustos. Gracias a las restricciones que establecen los supuestos del modelo multinomial logit, la probabilidad de elección de las alternativas de este modelo tiene la estructura matemática cerrada, que se conoce como probabilidad logit y se muestra en la fórmula (2.14).

$$\Omega = \sigma^2 I_A \tag{2.13}$$

$$P_i = \frac{e^{V_i}}{\sum_{j=1}^{A} e^{V_j}}, \forall i \epsilon \{1, \dots, A\} \tag{2.14}$$

Considerando una utilidad observada lineal en los parámetros, si llamamos β_m al coeficiente del m-ésimo regresor de la i-ésima alternativa, X_{im}, la elasticidad directa de X_{im}, $E_{X_{im}}^{P_i}$, mide la variación porcentual esperada en P_i ante un incremento de un punto porcentual de X_{im}. En el caso del modelo multinomial logit, la elasticidad directa de un regresor X_{im} se calcula multiplicando el regresor por la derivada del logaritmo natural de la expresión (2.25) respecto a él, cuyo resultado se muestra en la fórmula (2.15). Como puede observarse, la expresión de la elasticidad directa es la misma en todas las alternativas. Esto se debe a la independencia e igual distribución de probabilidad de la utilidad aleatoria de todas las alternativas.

$$E_{X_{im}}^{P_i} = \frac{d \ln P_i}{d \ln X_{im}} = \frac{d \ln P_i}{d X_{im}} X_{im} = (1 - P_i)\beta_m X_{im}, \forall i \epsilon \{1, \dots, A\} \tag{2.15}$$

Considerando de nuevo una utilidad observada lineal en los parámetros, y β_m el coeficiente de X_{im}, la elasticidad cruzada de X_{im} en la j-ésima alternativa, $E_{X_{im}}^{P_j}$, mide la variación porcentual esperada en P_j ante un incremento de un punto porcentual de X_{im}. En el caso de un modelo multinomial logit con función de utilidad observada lineal en parámetros, $E_{X_{im}}^{P_j}$ se calcula multiplicando X_{im} por la derivada de P_j, cuyo resultado se muestra en la fórmula (2.16). Al igual que sucede con la correlación directa, la expresión de la elasticidada cruzada es igual en todos los pares de alternativas, debido a la

independencia de los errores aleatorios de la utilidad entre pares de alternativas y a la igualdad de distribución de probabilidad de la utilidad aleatoria de todas las alternativas.

$$E_{X_{im}}^{P_j} = \frac{d \ln P_j}{d \ln X_{im}} = \frac{d \ln P_j}{d X_{im}} X_{im} = -P_i \beta_m X_{im}, \forall i, j \epsilon \{1, \dots, A\} \quad (2.16)$$

En el contexto de los modelos LUTI, los datos de las muestras de estimación suelen ser de tipo corte transversal, utilizando encuestas de preferencias reveladas. En este contexto, se puede suponer incorrelación entre observaciones, tanto de tipo temporal como individual (esta correlacióin es debida a diferentes respuestas de un mismo individuo, que es habitual en encuestas de preferencias declaradas, pero no es razonable suponerla en las de preferencias reveladas). Sin embargo, en modelización de elecciones espaciales las alternativas son áreas geográficas, entre las que es muy difícil justificar incorrelación. Las alternativas que tengan mayor similitud espacial pueden ser consideradas preferentemente como sustitutas, existiendo por tanto correlación espacial entre alternativas. Las similitudes son debidas a elementos espaciales no observados de la utilidad (Hunt et al., 2004). Por tanto, en modelización de elecciones espaciales es necesario considerar modelos más flexibles que el multinomial logit, que permitan al menos incorporar correlación entre alternativas. Adicionalmente, pueden considerarse variaciones en los gustos de los decisores, esto es, heterogeneidad en la componente no observada de la utilidad.

2.2.2. Modelo mixed logit

Train (2009: 46) describe las limitaciones del modelo multinomial logit en el contexto de relajación de dos de los supuestos de su definición: incorrelación y homocedasticidad. Hay diferentes enfoques RUM para incorporar correlación entre alternativas y variaciones en los gustos de los decisores en modelos de elección discreta. Las clases de modelos RUM con estructuras de error generales lo hacen en una única estructura, como son los modelos mixed logit y el modelo multinomial probit. El modelo mixed logit más sencillo es el mixed multinomial logit (MMNL). El modelo MMNL es muy flexible, y puede aproximar cualquier modelo RUM (McFadden y Train, 2000), incluido el modelo multinomial probit. La probabilidad de elección de una alternativa i en un modelo MMNL se recoge en la fórmula (2.17), donde $P_i(\beta)$ es una función mixta (se le llama así en literatura estadística a la media ponderada de muchas funciones) de la probabilidad logit evaluada para diferentes valores del vector de parámetros β, que se recoge en (2.18), y $f(\beta)$ es la función de densidad conjunta del vector de parámetros β, que llamamos distribución mixta. Las ponderaciones de la función mixta se obtienen a partir de la distribución mixta. La distribución mixta puede ser tanto discreta como continua. En el caso discreto, el modelo logit se conoce como modelo de clase latente, y ha sido utilizado tanto en transporte (ver Bhat, 1997; Greene y Henser, 2003) como en psicología, marketing y otras disciplinas (ver ejemplos en Kamakura y Russell, 1989; Chintagunta et al., 1991). En los modelos de clase latente, β podrá tomar un número finito de posibles valores. La probabilidad de elección de un modelo de clase latente se recoge en la fórmula (2.19), donde $\{b_1, \dots, b_M\}$ son los M posibles valores que puede tomar β y s_m es la cuota de población en cada segmento m, que se calcula con la fórmula (2.20). Se pueden estimar en cada segmento conjuntamente los valores posibles de β y la probabilidad de cada uno. La especificación mixed multinomial logit de clase latente

es útil cuando se pueden identificar M segmentos de la población, cada uno con sus propias preferencias o comportamiento.

$$P_i = \int P_i(\beta) f(\beta) d\beta \,, \forall i \epsilon \{1, \dots, A\} \tag{2.17}$$

$$P_i(\beta) = \frac{e^{V_i(\beta)}}{\sum_{j=1}^{A} e^{V_j(\beta)}}, \forall i \epsilon \{1, \dots, A\} \tag{2.18}$$

$$P_i = \sum_{m=1}^{M} P(\beta = b_m)\, P_i(b_m), \forall i \epsilon \{1, \dots, A\} \tag{2.19}$$

$$s_m = P(\beta = b_m), \forall m = 1, \dots, M \tag{2.20}$$

De todas formas, se han aplicado más extensamente especificaciones continuas del modelo mixed multinomial logit. La fórmula (2.21) recoge la expresión de la probabilidad de elección de un modelo mixed multinomial logit con distribución mixta normal, donde Φ es la función de densidad multivariante normal con los siguientes parámetros, vector de medias b y matriz de varianzas-covarianzas Ω. En el caso de la distribución mixta normal, al igual que sucede en la mayoría de modelos de distribución de probabilidad, la probabilidad de elección no tiene una expresión matemática cerrada.

$$P_i = \int P_i(\beta) \Phi(\beta | b, \Omega) d\beta, \forall i \epsilon \{1, \dots, A\} \tag{2.21}$$

El modelo mixed multinomial logit puede ser derivado bajo diferentes especificaciones de la función de utilidad, que son formalmente equivalentes (Train, 2009: 134-137). La especificación del modelo mixed multinomial logit mediante una estructura de componentes de error permite patrones flexibles de correlación entre alternativas. Este enfoque no suele ser viable para acomodar la correlación entre las alternativas espaciales, porque ya hemos dicho que hay que especificar tantas componentes de error como el número de pares de alternativas correladas, que suele ser un número excesivamente alto para el proceso de estimación. Si no se incluyen restricciones en la estructura de correlación de los errores aleatorios de este modelo, el número de parámetros a estimar en la matriz de varianzas-covarianzas de los errores aleatorios puede ser tan alto que haga inviable el proceso de estimación de este modelo.

La derivación más sencilla y extensamente aplicada del modelo mixed multinomial logit se basa en coeficientes aleatorios (Train, 2009: 137). En esta especificación, ciertos coeficientes de la función de densidad son variables aleatorias, que se pueden interpretar como las variaciones en las preferencias dentro de la población, según alguna distribución de probabilidad supuesta por el analista. Para un análisis de la repercusión de una u otra especificación del modelo mixed multinomial logit en la correlación y la heteroscedasticidad, puede consultarse Cherchi y Ortúzar (2004). La distribución mixta será determinada por la que se suponga para los coeficientes aleatorios. En esta investigación supondremos que es normal, tal y como ya hemos explicado.

Las especificaciones mixed multinomial logit de coeficientes aleatorios pueden superponerse a modelos logit más flexibles que el multinomial logit, que denominamos núcleo logit de un modelo mixed logit. Cualquier modelo logit con probabilidades de elección de cada alternativa de naturaleza paramétrica también se puede especificar

con una estructura mixed logit permitiendo que los parámetros sean aleatorios, e integrando la función sobre la distribución de parámetros (Greene, 2001). En estas especificaciones, $P_i(\beta)$ es la probabilidad de elección del núcleo logit que se haya escogido, que preferentemente tendrá una estructura matemática cerrada. Los núcleos logit pueden incorporar correlación entre alternativas. Por tanto, esta especificación de los modelos mixed logit permite construir modelos logit que incluyen explícitamente la correlación entre alternativas y las variaciones en los gustos. Si el núcleo logit es compatible con un modelo de elección de localización espacial, este hecho no se ve afectado por la incorporación de coeficientes aleatorios. Por todo esto, el planteamiento de esta investigación es un modelo mixed logit con un núcleo que tenga una expresión matemáticamente cerrada y sea compatible con modelización espacial en contexto LUTI. El núcleo logit que se elige en este contexto es una extensión del modelo multinomial logit, que permita un número elevado de alternativas e incorpore correlación entre ellas, al menos de naturaleza espacial.

2.2.3. Modelo nested logit

El modelo hierarchical o nested logit (NL; Williams, 1977; Daly y Zachary, 1978; McFadden, 1978) extiende el modelo multinomial logit para permitir ciertas estructuras de correlación entre alternativas. Al igual que el modelo multinomial logit, los términos de error aleatorio de la utilidad de las alternativas del modelo hierarchical logit son homocedásticos y tienen distribución de probabilidad marginal Gumbel. Sin embargo, a diferencia del modelo logit multinomial, el modelo hierarchical logit permite ciertas estructuras de correlación entre errores aleatorios de la utilidad, con distribución conjunta valor extremo. Este modelo utiliza agrupamientos jerárquicos de las alternativas para recoger la correlación entre alternativas. Los agrupamientos de alternativas, llamados nidos, deben ser diseñados por el analista. Para el diseño de los nidos, el analista debe utilizar variables no incorporadas en la función de utilidad. Por ejemplo, en contexto de modelización de la elección de localización residencial urbana estas variables pueden representar el atractivo de la zona para el decisor, debido al prestigio de la zona, el tipo de arquitectura imperante, sus vistas o la disponibilidad de servicios, como pueden ser los de transporte, escolar, ocio o empleo. El modelo hierarchical logit es compatible con RUM siempre que los llamados parámetros de disimilitud de los nidos cumplan la propiedad (2.22). Los parámetros de disimilitud modulan el valor de la correlación entre los pares de alternativas. El valor de la correlación entre los errores aleatorios de la utilidad de dos alternativas i, j, se calcula mediante la fórmula (2.23) si ambas alternativas pertenecen a un mismo nido N_k, y tiene valor cero si pertenecen a diferentes nidos (y utilizando la normalización asumida en todos los modelos logit). La estructura de la matriz de varianzas-covarianzas resultante es una matriz diagonal por bloques, uno por nido, a diferencia de la estructura escalar del modelo multinomial logit. Cada bloque correspondiente al nido N_k contará con unos en la diagonal y el valor $(1 - \mu_k^2)$ en el resto, todos ellos multiplicados por el mismo valor σ^2 de la fórmula (2.7), que es la diagonal de la matriz de varianzas-covarianzas del modelo multinomial logit (Álvarez-Daziano y Munizaga, 2002; Carrasco y Ortúzar, 2002). Si una alternativa es incorrelada con el resto de alternativas, formará un nido al que solo pertenece ella. En este caso también se dice que pertenece al nido raíz. Los parámetros de disimilitud toman el valor uno en las alternativas del nido raíz. Es muy habitual utilizar una estructura de nidos de dos niveles, es decir, los nidos no tienen sub-nidos

jerarquizados con ellos. En esta tesis nos centraremos en el hierarchical logit de dos niveles, al que nos referiremos como nested logit.

$$0 < \mu_k \leq 1, \forall \mathrm{k} = 1, \dots, \mathrm{M} \tag{2.22}$$

$$Corr\left(\varepsilon_i, \varepsilon_j\right) = (1 - \mu_k^2), \forall i, j \epsilon N_k, \forall k \in \{1, \dots, M\} \tag{2.23}$$

La fórmula (2.24) muestra la expresión de la probabilidad de elección de las alternativas pertenecientes a cualquier nido N_k que no sea el raíz en un modelo nested logit. La expresión de la probabilidad de elección de las alternativas del nido raíz es la misma que en el modelo multinomial logit (2.14). Esta expresión se puede descomponer según la fórmula (2.25), donde $P_{i|k}$ (2.26) es la probabilidad condicional de elección de la i-ésima alternativa si se selecciona el nido N_k, y P_k (2.27) es la probabilidad de elegir ese nido. Las ecuaciones (2.26) y (2.27) son modificaciones de la de Papola (2004) para facilitar la comparación entre el modelo nested logit con los modelos logit que se presentarán y propondrán en los siguientes capítulos de esta tesis.

$$P_i = \frac{e^{V_i/\mu_k}\left(\sum_{j \epsilon Nido_k} e^{V_j/\mu_k}\right)^{\mu_k - 1}}{\sum_{l=1}^{M}\left(\sum_{j \epsilon Nido_l} e^{V_j/\mu_l}\right)^{\mu_l}}, \forall i \epsilon \{1, \dots, A\}, i \epsilon N_k \tag{2.24}$$

$$P_i = P_{i|k} \cdot P_k, \forall i \epsilon \{1, \dots, A\}, i \epsilon N_k \tag{2.25}$$

$$P_{i|k} = \frac{\left(e^{V_i}\right)^{1/\mu_k}}{\sum_{j \epsilon N_k}\left(e^{V_j}\right)^{1/\mu_k}}, \forall i \epsilon \{1, \dots, A\}, \forall k \epsilon \{1, \dots, M\} \tag{2.26}$$

$$P_k = \frac{\left(\sum_{j \epsilon N_k}\left(e^{V_j}\right)^{1/\mu_k}\right)^{\mu_k}}{\sum_{l=1}^{M}\left(\sum_{r \epsilon N_l}(e^{V_r})^{1/\mu_l}\right)^{\mu_l}}, \forall k \epsilon \{1, \dots, M\} \tag{2.27}$$

A diferencia del modelo multinomial logit, las elasticidades en el modelo nested logit no tienen la misma expresión para todas las alternativas. Dado que las alternativas del nido raíz son incorreladas con el resto de alternativas, la expresión de la elasticidad directa y de la elasticidad cruzada de las alternativas del nido raíz coincide con la del modelo multinomial, que se muestra en las fórmulas (2.15) y (2.16), respectivamente. La fórmula (2.28) muestra la expresión de la elasticidad directa del modelo nested logit de alternativas que pertenecen a un nido diferente del nido raíz. La fórmula (2.29) muestra la expresión de la elasticidad cruzada de pares de alternativas que pertenecen a un mismo nido. Esta expresión se ha obtenido a partir de la propuesta por Papola (2004), aunque con una expresión que facilite su comparación con los modelos logit que se presentarán y propondrán en los siguientes capítulos de esta tesis. En el caso de pares de alternativas pertenecientes a nidos diferentes, la expresión de la elasticidad cruzada coincide con la de las alternativas del nido raíz, es decir, la del modelo multinomial logit.

$$E_{X_{im}}^{P_i} = \left[(1 - P_i) + (\mu_k^{-1} - 1)\left(1 - P_{i|k}\right)\right]\beta_m X_{im}, \forall i \epsilon N_k, k \epsilon \{1, \dots, M\} \tag{2.28}$$

$$E_{X_{im}}^{P_j} = -\left[P_i + (\mu_k^{-1} - 1)P_{i|k}\right]\beta_m X_{im}, \forall i, j \epsilon N_k, k \epsilon \{1, \dots, M\} \tag{2.29}$$

El aumento en el número de nidos tiene el efecto positivo de que incrementa la flexibilidad del modelo nested logit para medir la correlación entre alternativas. Pero tiene el efecto negativo de que incrementa la complejidad del modelo, pues aumenta el número de parámetros de disimilitud, que son desconocidos y tendrán que ser estimados. El modelo nested logit es compatible con la modelización de elección de

localización espacial, siempre que el analista diseñe una estructura de nidos que tenga un número viable de nidos. La ventaja del modelo nested logit, frente al resto modelos de elección discreta derivados de RUM que incorporan correlación espacial que se han presentado por ahora en esta tesis, radica en que el enfoque de nidos le permite al analista incorporar estructuras de correlación entre alternativas relativamente complejas, con un número de parámetros desconocidos mucho menor. Además, la eficacia del modelo nested logit para recoger la correlación entre alternativas depende de la capacidad del analista para diseñar los nidos. El modelo nested logit tiene el hándicap de necesitar que el analista diseñe la estructura de nidos.

El modelo nested logit se puede especificar con la restricción de que todos los parámetros de disimilitud tengan el mismo valor, que llamaremos restricted nested logit (RNL). Esta parsimoniosa especificación del modelo nested logit tiene un solo parámetro desconocido adicional al modelo multinomial logit. Al igual que en la especificación no restringida, en el restricted nested logit los pares de alternativas de diferentes nidos son incorreladas. Pero en el modelo restricted nested logit, los pares de alternativas de un mismo nido son igualmente correladas, sea cual sea el nido al que ambas pertenecen. La fórmula (2.30) muestra el valor de la correlación entre pares de alternativas de un mismo nido, siendo μ el parámetro de disimilitud de todos los nidos. Desafortunadamente, esta especificación solo será eficiente en contextos de aplicación donde todos los pares de alternativas estén igualmente correladas.

$$Corr(\varepsilon_i, \varepsilon_j) = (1 - \mu^2), \forall i, j \epsilon N_k, \forall k \in \{1, \dots, M\} \quad (2.30)$$

Tanto el modelo nested logit como su especificación restricted nested logit pueden incorporar variación en los gustos de los decisores si se incorporan como núcleo de una especificación mixed logit con coeficientes aleatorios (M-NL y MRNL, respectivamente).

2.3. Aplicación y comparación de modelos de elección espacial

En esta sección, al igual que en los dos próximos capítulos, será necesario especificar, estimar, validar y comparar modelos de elección discreta de tipo logit para modelizar la elección espacial en contexto de planificación urbana y del transporte. En concreto, se pretende validar su uso para la elección de localización espacial en contexto LUTI urbano. Para ello, los modelos que se evalúan en esta tesis, tanto los descritos en la literatura como las nuevas propuestas que se recogen en esta tesis, se aplican en un mismo caso real. En este apartado se propone una metodología para validar y comparar los resultados de estimación obtenidos con cada modelo especificado.

2.3.1. Estimación y aceptación

Los parámetros desconocidos de los modelos de tipo logit se estiman conjuntamente mediante el método de máxima verosimilitud, utilizando datos de muestra recopilados para este fin. En esta investigación todos los modelos se estiman con ayuda de Biogeme (Bierlaire, 2003). Biogeme es un programa informático de software libre y código abierto específico para la estimación de modelos paramétricos mediante máxima verosimilitud, con especial énfasis en los modelos de elección discreta. En todas las estimaciones se utilizó el mismo algoritmo de optimización, el DONLP2 (Spellucci, 1993).

En las aplicaciones que se realizan en este capítulo y los dos siguientes solo se consideran aceptables las especificaciones de los modelos en las que todos los parámetros estimados sean significativos, y en las que los signos de los valores obtenidos en la estimación de los parámetros desconocidos sean coherentes con los esperados teóricamente. Esta exigencia favorece el objetivo principal de las aplicaciones que se realizan en esta tesis, que no es otro que comparar su capacidad explicativa y predictiva. La significación de cada parámetro se comprueba mediante el contraste de Wald de su coeficiente estimado. El contraste de Wald es semejante al t-test de la regresión lineal. En ambos casos se trata de un contraste bilateral de nulidad del cociente entre el coeficiente estimado y su error típico. La diferencia radica en que el estadístico del contraste de Wald tiene distribución asintótica normal estándar. El nivel de significación que se utiliza en esta investigación para este contraste es 0,05, y es el mismo que se utiliza en el resto de contrastes que se apliquen en esta tesis. Los únicos parámetros desconocidos en el modelo logit multinomial son los coeficientes de la función de utilidad observada. Consideramos que un regresor de la función de utilidad observada especificada por el analista es relevante cuando su respectivo coeficiente es significativo según el contraste de Wald. El resto de modelos logit que se evalúan en esta tesis están anidados con el modelo multinomial logit, y presentan parámetros estructurales desconocidos adicionales a los coeficientes de la función de utilidad observada.

2.3.2. Bondad de ajuste

Las técnicas estadísticas de bondad de ajuste de los modelos estimados mediante máxima verosimilitud utilizan estadísticos y contrastes basados en el valor de la verosimilitud final (L). L es el valor de la función de verosimilitud de la muestra de estimación que se obtiene al finalizar el algoritmo de optimización, cuando se alcanza la convergencia. L se calcula según la fórmula (2.31), donde $N_{Est.}$ es el tamaño de la muestra de estimación, y $\hat{P}_{n,Est.}$ es el valor de la probabilidad de elección de la alternativa elegida por cada decisor n de esa muestra (que podemos denominar como alternativa correcta), calculada con la expresión de la probabilidad de elección del modelo que incluye el valor de los parámetros desconocidos que se obtuvo en la estimación con la misma muestra.

Los estadísticos de bondad de ajuste que utilizamos es esta investigación se obtienen de dos formas diferentes a partir de L. En la primera forma, los estadísticos de bondad de ajuste se obtienen directamente de L mediante una función matemática creciente. Los parámetros que maximizan L también maximizan funciones crecientes de ella. El primer estadístico de este tipo es el logaritmo natural de L (LL). Este estadístico es el que habitualmente utilizan los algoritmos de optimización utilizados en el proceso de estimación por máxima verosimilitud. Por este motivo suele ser el valor de salida de los procesos de optimización, como sucede en el programa utilizado en esta investigación. El segundo estadístico de este tipo, que llamamos Fitting Geometric (FG), es la media geométrica (GM) de las probabilidades de elección $\left(\hat{P}_{1,Est.}, \dots, \hat{P}_{N_{est.},est.}\right)$. A la hora de establecer una metodología para comparar la capacidad explicativa y predictiva de los modelos aplicados, se ha definido en esta tesis un factor de ajuste basado en la media geométrica, en lugar de la media artimética que se usa más habitualmente, porque presenta mejores propiedades con datos del tipo de las probabilidades. Además, tal y

como se deduce en (2.32), FG es una función creciente de L y LL que se puede calcular directamente a partir de ellas. Este estadístico es semejante al Fitting Factor (de Luca y Cantarella, 2009), pero utilizando la media geométrica en lugar de la aritmética, porque presenta mejores propiedades con datos del tipo de las probabilidades.

$$L := \prod_{n=1}^{N_{est.}} \hat{P}_{n,est.} \tag{2.31}$$

$$FG = GM\left(\hat{P}_{1,est.}, \dots, \hat{P}_{N_{est.},est.}\right) := \sqrt[N_{Est.}]{\prod_{n=1}^{N_{est.}} \hat{P}_{n,est.}} = \sqrt[N_{est.}]{L} = exp\left\{\frac{\ln L}{N_{est.}}\right\} =$$

$$= exp\left\{\frac{LL}{N_{est.}}\right\} \tag{2.32}$$

La segunda forma de obtener estadísticos de bondad de ajuste a partir de L se basa en el cociente entre LL y la log-verosimilitud de la muestra de estimación que se obtiene con algún modelo de referencia. En esta investigación se utilizan tres estadísticos de este tipo, en los que se usa como modelo de referencia el modelo Nulo (modelo que solo tiene término de error, con función de utilidad observada con valor constante igual a cero): el likelihood ratio index (LRI) de McFadden (1974) formulado en (2.33); el adjusted LRI (ALRI) de Horowitz (1983) formulado en (2.34), donde p es el número de parámetros estimados; y el Akaike likelihood ratio index basado en el Akaike information criterion (AIC; Ben-Akiva y Swait, 1986) formulado en (2.35). Los dos últimos penalizan el número de parámetros que se hayan estimado, y coinciden con ρ^2 si los modelos comparados tienen igual número de parámetros. Los modelos que penalizan el número de parámetros estimados permiten comparar modelos que estiman diferente número de parámetros. AIC penaliza más la incorporación de parámetros que $\bar{\rho}_H^2$, lo que favorece modelos más parsimoniosos. Aunque la fórmula es parecida, ambos estadísticos tienen justificaciones diferentes. Dado que los modelos mixtos se estiman en esta investigación mediante máxima verosimilitud simulada, para calcular los estadísticos de bondad de ajuste se actuará de forma análoga, pero utilizando la verosimilitud simulada (SL) en lugar de L.

$$\rho^2 = 1 - \frac{LL(Modelo)}{LL(Nulo)} \tag{2.33}$$

$$\bar{\rho}_H^2 = 1 - \frac{LL(Modelo) - \frac{p}{2}}{LL(Nulo)} \tag{2.34}$$

$$AIC = 1 - \frac{LL(Modelo) - \mathrm{p}}{LL(Nulo)} \tag{2.35}$$

Si se quiere comparar la bondad de ajuste de dos modelos que estiman los mismos parámetros se puede utilizar cualquiera de los estadísticos de bondad de ajuste anteriores. Con todos ellos se obtiene la misma conclusión. Si los dos modelos que se quieren comparar estiman diferentes parámetros, pero están anidados (uno de los modelos se puede obtener mediante restricciones lineales de los parámetros del otro), entonces para comparar la bondad de ajuste utilizamos el contrate de la razón de verosimilitudes (LRT). En este capítulo utilizamos este contraste para comparar el modelo NL con su especificación RNL, ambos con el modelo MNL y todos los modelos estimados con el modelo Nulo. Para un nivel de significación dado, el LRT contrasta si el modelo con mayor número de parámetros tiene significativamente la misma verosimilitud que el modelo con el que está anidado, frente a la hipótesis alternativa unilateral de que dicha verosimilitud es mayor. El cociente entre ambas verosimilitudes

no tiene una distribución conocida, pero la transformación que se muestra en la fórmula (2.36), llamada devianza o estadístico Wilks del $Modelo_2$ respecto al $Modelo_1$, tiene una distribución asintótica $\chi^2_{(p_2-p_1)}$ (Wilks, 1938), donde p_i es el número de parámetros estimados en cada uno de los dos modelos que se comparan, L_i es la verosimilitud final de cada modelo y LL_i es el logaritmo neperiano de L_i. El LRT del $Modelo_2$ respecto al $Modelo_1$, $LRT(Modelo_2, Modelo_1)$, es significativo cuando la devianza es mayor que el cuantil $\chi^2_{(p_1-p_2),\alpha}$, donde α es el nivel de significación dado, que en esta investigación será 0,05.

$$LRT(Modelo_2, Modelo_1) = -2log\frac{L_2}{L_1} = -2(LL_2 - LL_1) \tag{2.36}$$

Los modelos con un número de parámetros desconocidos diferente se pueden comparar, aunque no estén anidados. Para ello se utilizan los estadísticos de bondad de ajuste global que penalizan el número de parámetros estimados. En esta investigación se utilizan el ALRI de Horowitz y el AIC.

2.3.3. Validación

Las técnicas estadísticas de validación permiten comparar la capacidad predictiva de los modelos estimados con muestras de test diferentes a las de estimación. El tipo de validación que realizamos en esta tesis es lo que Parady et al. (2021) denominan validación interna. Las muestras de entrenamiento y test se obtendrán mediante particiones aleatorias de la muestra recopilada. Para ello, en esta investigación utilizamos dos técnicas de validación cruzada: CV-4 y CV-10. En ambos casos se mide la precisión de las predicciones de un modelo mediante un estadístico. En esta investigación utilizamos el estadístico que llamamos Predicting Geometric (PG). Este estadístico es la media geométrica de $\left(\hat{P}_{1,test.}, \dots, \hat{P}_{N_{test},test}\right)$, tal y como se recoge en la fórmula (2.37), donde N_{test} es el tamaño de la muestra de test, y $\hat{P}_{n,test}$ es el valor de la probabilidad de elección de la alternativa que elige cada encuestado n de la muestra de test, calculada con la expresión de la probabilidad de elección del modelo que incluye el valor de los parámetros desconocidos que se obtuvo en la estimación con la muestra de entrenamiento. El estadístico habitualmente utilizado con este fin se basa en la media aritmética (Başar y Bhat, 2004; de Luca y Cantarella, 2009; Martínez-Pardo et al., 2020), en lugar de la media geométrica que proponemos en esta tesis, por los mismos motivos que los explicados en el caso FG para GoF.

$$PG = GM\left(\hat{P}_{1,test.}, \dots, \hat{P}_{N_{test},test}\right) = \sqrt[N_{test}]{\prod_{n=1}^{N_{test}} \hat{P}_{n,test}} =$$

$$= \prod_{n=1}^{N_{test}} \sqrt[N_{test}]{\hat{P}_{n,test}} \tag{2.37}$$

Los procesos CV-4 y CV-10 son procesos de Validación Cruzada. En estos procesos se realiza una partición aleatoria de la muestra en 4 y 10 grupos respectivamente. Cada una de las submuestras aleatorias generará una iteración de *split validation* (ver Parady et al., 2021), en la que una submuestra se emplea como muestra de test y las restantes como muestras de entrenamiento. En las tablas de resultados se presentará el valor medio de los 4 o 10 valores de PG obtenidos, respectivamente. Esta media también se calculará mediante la media geométrica (por los mismos motivos explicados

anteriormente), utilizando la fórmula (2.38), donde PG_k es la Predicting Geometric obtenida en cada una de las K iteraciones que se realizan en un CV de K grupos.

$$PG\text{-}CV\text{-}K = \sqrt[K]{\prod_{k=1}^{K} PG_k} = \prod_{k=1}^{K} \sqrt[K]{PG_k}\,, K \in \{4{,}10\} \tag{2.38}$$

2.3.4. Tabla de resultados

Para cada modelo estimado se muestra una tabla que contiene resultados de la estimación de parámetros, de los contrastes y estadísticos de bondad de ajuste y de los estadísticos de validación ($Val.$). Para cada parámetro estimado, la tabla de resultados incluye el valor estimado (Valor), el error típico de estimación (SE), el estadístico del contraste de Wald, así como el p-valor y una codificación de la significatividad (Sig.) de este contraste. La codificación que se utiliza en esta investigación para mostrar la significación de un contraste es la siguiente: si el contraste es significativo, se muestra el código "**" si lo es con el nivel de significación 0,01, y "*" si solo lo es con el nivel de significación 0,05; si el contraste no es significativo se muestra el código ".".

En los coeficientes fijos estimados de la función de utilidad observada, las tablas de resultados también incluyen el valor del coeficiente estandarizado (SC) y su valor relativo (RI). Los coeficientes estandarizados permiten comparar la influencia relativa de cada predictor, aunque estos estén medidos en diferentes escalas. Incluso permite comparar la influencia relativa entre predictores continuos y cualitativos dicotómicos o politómicos. Un predictor tiene mayor influencia relativa en la decisión cuanto mayor sea el valor absoluto de su coeficiente tipificado. El signo del coeficiente estandarizado tiene la misma interpretación que el del coeficiente estimado. Hay diferentes estadísticos que tipifican los coeficientes fijos estimados (ver Menard, 2004 y 2011). En esta investigación utilizamos el estadístico de tipo partially standardized logistic regression coefficients propuesto por Menard (1995) y Agresti (1996:129). Este estadístico tipifica cada coeficiente fijo estimado b_m mediante la fórmula (2.39), donde $SD(X_m)$ es la desviación típica muestral de su correspondiente regresor X_m. En el caso de regresores cualitativos, se utiliza una codificación numérica de los regresores. Los coeficientes así tipificados se interpretan en base a desviaciones típicas del regresor, en lugar de hacerlo en base a sus unidades de medida. Para medir la influencia relativa de cada predictor, se utiliza su influencia relativa (RI), que es el peso relativo de su coeficiente estandarizado, tal y como refleja la fórmula (2.40), donde s es el número de coeficientes estimados de la función de utilidad observada.

$$\dot{b}_m = b_m \cdot SD(X_m) \tag{2.39}$$

$$RI(X_m) = \frac{\dot{b}_m}{\sum_{r=1}^{s} \dot{b}_r} \tag{2.40}$$

Para comparar los modelos aceptados se analiza su bondad de ajuste (GoF) y su validación. La bondad de ajuste permite comparar la capacidad de explicación de los modelos (Hilbe, 2009) y la validación permite comparar la capacidad predictiva de los modelos (Parady et al., 2021). Con respecto a la bondad de ajuste, las tablas incluyen los estadísticos LL, o el logaritmo neperiano de SL (SLL) en el caso de modelos mixtos, FG, LRI de McFadden (ρ^2), ALRI de Horowitz ($\bar{\rho}_H^2$) y el AIC. En los modelos que aniden algún otro modelo estimado y aceptado, la tabla de resultados incluye la devianza (W) del LRT con cada uno, junto con el p-valor y la codificación de la significación del contraste. La

codificación que se utiliza es la descrita anteriormente para el contraste de Wald. Con respecto a la validación, la tabla de resultados de cada modelo incluirá el PG obtenido en los dos procesos de validación cruzada realizados: *PG-CV-4* y *PG-CV-10*.

2.4. Aplicación a un caso real

En este apartado, se describe la aplicación empírica a un caso real de los modelos de elección discreta multinomial, derivados de la teoría económica de maximización de la utilidad aleatoria, que se describieron en las subsecciones anteriores y que se mostraron compatibles con la modelización de elección espacial. En concreto, se han aplicado los modelos multinomial logit y nested logit (incluyendo la especificación restringida), así como de sus correspondientes especificaciones mixed logit con coeficientes aleatorios. Estos modelos se han aplicado según la metodología descrita en el apartado anterior, con la finalidad de comparar su capacidad explicativa y predictiva en el contexto empírico de aplicación. Esta misma metodología se utilizará para comparar los resultados obtenidos con estos modelos con los que se obtengan en los modelos con correlación espacial de los dos siguientes capítulos. La aplicación real corresponde a una modelización de elección de la localización residencial en la ciudad de Santander. La figura 2.2 muestra la ciudad de Santander, situada en la costa norte de España, y que es capital de la comunidad autónoma de Cantabria.

Figura 2.2. Ortofoto de la ciudad de Santander. Memoria del proyecto INTERLAND.

2.4.1. Zonificación y muestra

En la aplicación de un modelo de elección espacial con enfoque zonal, la primera tarea consiste en diseñar la zonificación de las alternativas. En modelización de la elección de localización residencial, la zonificación consiste en diseñar una partición del área geográfica total en estudio. Las zonas diseñadas contienen las localizaciones residenciales disponibles, que vienen determinadas por la necesidad de información para la planificación urbana.

La zonificación de la ciudad de Santander y la muestra de datos empleados para la aplicación empírica parten de desarrollos realizados en el marco de los proyectos de investigación INTERLAND (Ibeas et al., 2013) y TRANSPACE (Dell'Olio et al., 2016; Pérez-López y Orro, 2016). Dado que el planteamiento de esta tesis es metodológico, se ha considerado adecuado tomar una base de datos y una zonificación en las que ya había resultados publicados para comparar con los nuevos desarrollos. Esta base de datos fue recopilada en 2008 por el Grupo de Investigación en Sistemas de Transporte de Santander (Ibeas et al., 2013). La zonificación que se utiliza en esta aplicación empírica se basa en las secciones censales de la ciudad de Santander que se muestran en la figura 2.3.

Figura 2.3. Secciones censales de la ciudad de Santander. Memoria del proyecto INTERLAND.

La ciudad de Santander albergaba una población de 182.302 habitantes en el momento de la recogida de los datos, donde más del 70% se concentra en el sector servicios, incluso si se incluye el resto de municipios del área metropolitana. La zonificación de esta investigación está formada por 26 alternativas que agrupan las secciones censales de la ciudad. La población se distribuye bastante homogéneamente entre las alternativas, con un promedio de 7011 habitantes en cada una y un coeficiente de variación de 0,2. La mediana de las superficies de las alternativas es de 0,33 km^2 y su densidad poblacional media es de 212 habitantes por km^2. Estas áreas geográficas que constituyen las alternativas tienen formas muy irregulares y elevada variabilidad de tamaños, tal y como se muestra en la figura 2.4. La codificación de colores en esta figura representa los terciles del número de individuos en la muestra que han elegido esa área por kilómetro cuadrado, que facilita identificar patrones de decisión menos dependientes de la zonificación.

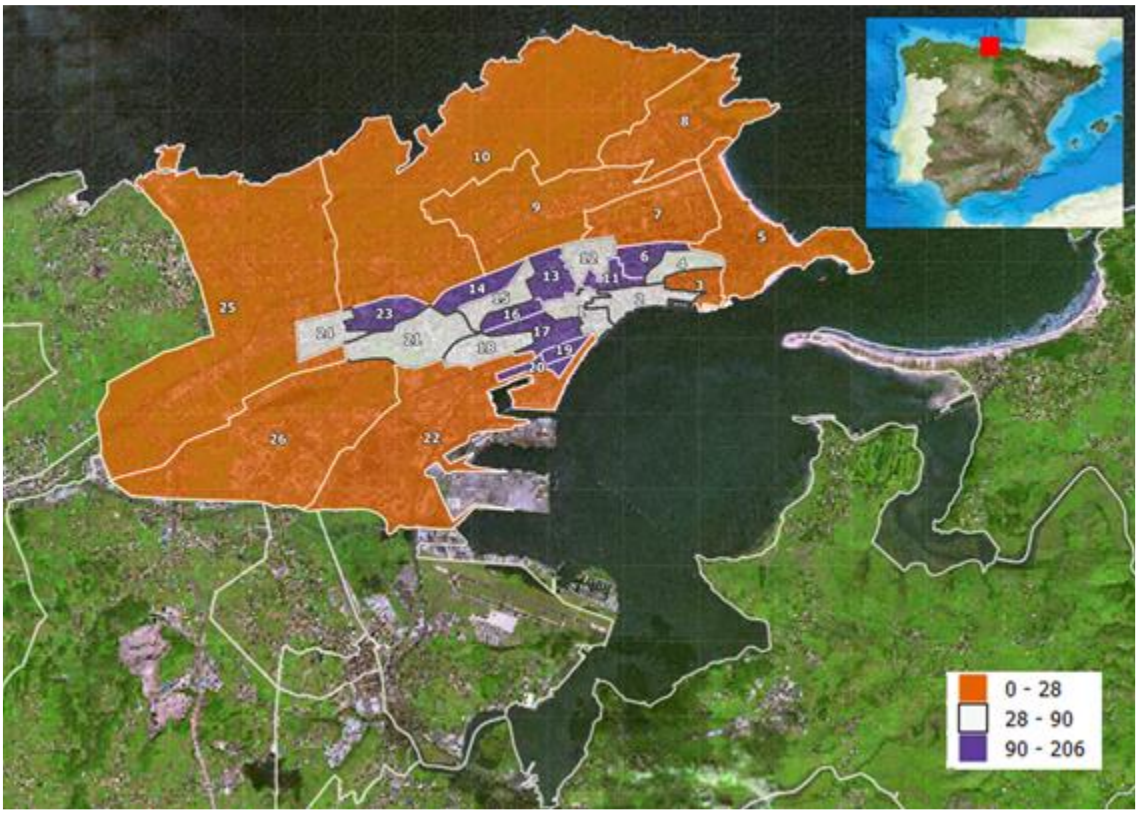

Figura 2.4. Zonificación de alternativas.

Una vez diseñada la zonificación de alternativas, la segunda tarea de la aplicación empírica consiste en diseñar los datos de muestra para estimación y validación. Como mínimo será necesaria una muestra, que se utiliza para estimar el valor de los parámetros desconocidos de cada modelo de elección utilizado en el contexto empírico de aplicación. Además, esta muestra permite comparar los modelos en el contexto empírico mediante el análisis de los resultados de bondad de ajuste. A la hora de validar el modelo, es recomendable utilizar muestras diferentes a la de estimación, aunque, como ya se ha señalado, existen técnicas de validación que utilizan la muestra de estimación.

El diseño muestral incluye la elección de las variables explicativas y el procedimiento de recopilación de cada una. La elección de las variables explicativas se basa en la teoría y estudios empíricos previos en modelización de la elección de localización residencial que se recogen en la literatura. La información que se obtenga de cada variable explicativa elegida debe incluir metadatos referidos al tipo de relación que se espera que tenga con la variable de elección.

En el enfoque micro-econométrico de los modelos de elección discreta basado en RUM, los datos de la muestra deben contener la decisión del individuo y las variables explicativas de la misma a nivel individual. En el caso de los modelos condicionales, como son los que se estiman en esta investigación, la muestra debe contener el valor específico de cada variable explicativa en cada alternativa de elección. Por tanto, además de la variable que contiene las elecciones, la muestra de datos contiene variables explicativas de la elección que pueden ser de tres tipos. El primer tipo de variables explicativas es el de las variables de nivel de alternativa, que son invariantes entre individuos. Las variables de este tipo recogen características socio-económicas y

demográficas de cada alternativa. Los datos de las variables de este tipo generalmente se recopilan a partir de información pública de las áreas administrativas que las componen. El segundo tipo de variables explicativas es el de las variables de nivel de alternativa que son específicas de cada individuo decisor. Estas variables se calculan habitualmente a partir de datos recopilados mediante encuestas individuales. El tercer tipo de variables explicativas, es el de las variables socio-económicas del individuo decisor, que se recopilan en las encuestas realizadas.

La muestra contiene 534 elecciones individuales, de decisores que residen y trabajan en la ciudad de Santander. Las variables de la muestra son: la elección del decisor, un atributo socio-económico del individuo y el valor de 9 variables explicativas de la elección en cada una de las 26 alternativas. La tabla 2.1 recoge la distribución de frecuencias de las elecciones del decisor en la muestra, correspondiente a la zona de residencia del decisor que se ha recopilado mediante encuesta. La tabla 2.2 recoge el nombre, descripción, tipo y estadísticos descriptivos del atributo individual y de las nueve variables explicativas incluidas en la muestra de datos. Los estadísticos descriptivos que recogen para cada variable cuantitativa son media y desviación típica muestral. En el caso de las tres variables cualitativas (dicotómicas), recoge la distribución de frecuencias.

Elección	Nº decisores	Elección	Nº decisores	Elección	Nº decisores
1	23	10	4	19	19
2	22	11	16	20	26
3	6	12	9	21	27
4	19	13	42	22	4
5	11	14	40	23	30
6	39	15	21	24	24
7	31	16	30	25	1
8	12	17	34	26	11
9	3	18	30		

Tabla 2.1. Distribución de frecuencias de las elecciones de los decisores en la muestra.

Las variables explicativas son de los tres tipos descritos en la metodología. Siete de ellas recogen características socio-económicas y demográficas de cada alternativa, siendo por tanto invariantes entre individuos. Las variables de este tipo son la accesibilidad al empleo en la zona residencial (AC), el número de residentes extra-comunitarios en la zona residencial (FO, en miles de personas), el número de viviendas en la zona residencial (HO), una variable dicotómica que indica si la zona residencial tiene una consideración de prestigio (PS), el precio medio de las viviendas de la zona residencial (PR, en millones de Euros), el número de centros de formación en un radio de un kilómetro del centroide de la zona residencial (SC) y el tiempo medio de espera en las paradas de transporte público de la zona residencial (WT, en minutos). La accesibilidad es un indicador gravitatorio de tipo Hansen, que mide la accesibilidad al empleo en una zona origen del desplazamiento. Por tanto, es una medida del potencial de

oportunidades de empleo de la zona. La accesiblidad de cada zona i se ha calculado teniendo en cuenta la posible naturaleza multicentro en las áreas urbanas, utilizando la expresión de Coppola y Nuzzolo (2011) que se recoge en la fórmula (2.41); donde JO es una variable que mide el número de empleos en la zona de destino de cada desplazamiento, CO es una variable que mide el coste de cada desplazamiento (que se calculó usando un modelo de transporte suponiendo que el desplazamiento se hizo en coche con la congestión de la hora punta de la mañana) y α_1, α_2 son dos parámetros desconocidos, cuyos valores estimados son 0,26 y -0,12, respectivamente (véase Ibeas et al., 2013).

$$AC(i) = \sum_j e^{\alpha_2 CO(i,\tau_j)} JO(\tau_j)^{\alpha_1} \quad (2.41)$$

Nombre	Descripción	Tipo	Promedio/ Distribución	Desviación típica
AC	Accesibilidad al empleo en la zona de residencia.	Alternativa	29,34	7,42
JT	Tiempo de viaje en minutos entre la zona de residencia y la de empleo.	Alternativa específica de individuo	7,57	3,93
FO	Número de extranjeros extracomunitarios en la zona de residencia (en miles de personas).	Alternativa	0,461	0,224
IN	Atributo dicotómico que indica si la zona de residencia es la misma que la de trabajo	Alternativa específica de individuo	NO: 93,45% SÍ: 6,55%	
HO	Logaritmo natural del número de viviendas en la zona de residencia (en miles de viviendas).	Alternativa	2,651	0,567
PS	Atributo dicotómico que indica que la zona de residencia se considera (subjetivamente) especialmente prestigiosa.	Alternativa	NO: 95,51% SÍ: 4,49%	
PR	Precio medio de las viviendas de la zona de residencia (en millones de Euros).	Alternativa	0,28761	0,12670
SC	Número de centros de educación primaria o secundaria a una distancia de no más de un km del centroide de la zona de residencia.	Alternativa	2,22	1,70
WT	Tiempo medio de espera en las paradas de transporte público de la zona de residencia (en minutos).	Alternativa	10,51	0,78
H	Atributo dicotómico que indica si los ingresos mensuales netos de la familia del decisor se pueden considerar altos (más de 2500 €).	Individual	NO: 76,97% SÍ: 23,03%	

Tabla 2.2. Variables explicativas de la muestra de datos (basada en Ibeas et al., 2013).

Dos de las variables explicativas incluidas en la muestra recogen información de cada alternativa específica del individuo. Los datos de estas variables se han calculado a partir de la información recopilada en la encuesta y un modelo de transporte. Las variables de este tipo son: el tiempo que tardaría el individuo decisor en desplazarse a su centro de trabajo en caso de residir en cada una de las alternativas de zona de residencia (JT, en minutos), y una variable dicotómica que indica si la zona de residencia coincidiría con la del centro de trabajo en caso de residir en cada una de las alternativas de zona de residencia (IN). Por último, la muestra cuenta con un atributo dicotómico del decisor calculado a partir de la información recopilada mediante encuesta. El atributo indica si el individuo tiene unos ingresos familiares altos o no. Se ha considerado que sí los tiene cuando los ingresos mensuales netos de la familia superen los 2500 euros (H).

2.4.2. Función de utilidad observada

En los modelos de elección de la localización residencial suele haber limitaciones a la hora de especificar las constantes específicas de alternativa en las funciones de utilidad observada lineales. Por un lado, en los modelos de elección espacial es habitual un alto número de alternativas. Si se especifican las constantes específicas de muchas de las alternativas, pueden surgir problemas de identificación. Por otro lado, en los modelos zonales las alternativas son de tipo agregado, por lo que las variables específicas de alternativa presentan características idénticas para todos los individuos, lo que puede provocar confusión entre los valores genéricos de las variables zonales y las estimaciones de las constantes específicas de alternativa (Ibeas et al., 2013). En esta tesis consideramos alternativas de tipo zonal. Por todo lo anterior, en los modelos de elección de la localización residencial que se aplican en esta tesis, la utilidad observada que se especifica será de tipo condicional, con expresiones lineales sin constates específicas para cada alternativa.

Una vez elegida la forma funcional de la utilidad observada de un modelo logit, el siguiente paso consiste en elegir sus regresores, cumpliendo con los requisitos descritos en la metodología de estimación y comparación de modelos de elección espacial descrita en el apartado 2.2. En esta tesis se eligen los regresores mediante el siguiente proceso de paso atrás y adelante. El proceso comienza con la estimación de un modelo saturado. El modelo saturado que consideramos especifica las nueve variables explicativas presentes en la muestra, que son todas las variables de la muestra salvo el factor de nivel de ingresos altos. La variable que indica el número de viviendas en la zona se incorpora a la función de utilidad observada a través de su logaritmo natural. Además, el modelo saturado incluye las ocho interacciones de cada una de estas variables con el factor de nivel alto de ingresos, todas salvo la interacción con la variable que indica si la zona de residencia es la misma que la del trabajo, por ser dicotómica. De esta forma, el modelo saturado incluye 17 regresores. Una especificación estimada se considera válida cuando todos los parámetros estimados son válidos, según los criterios descritos en la metodología del apartado 2.2. Si alguno de los regresores del modelo saturado no cumple los requisitos de validez descritos en la metodología, el siguiente paso de este proceso consiste en estimar el modelo que queda tras eliminar el regresor que presente un mayor p-valor en el test de Wald de su correspondiente coeficiente. Este procedimiento continúa eliminando regresores no válidos. Si en algún paso del proceso un nuevo regresor se muestra no válido, el proceso comienza de nuevo a partir del

modelo saturado, pero eliminando únicamente este último regresor. Este proceso continúa hasta que la especificación estimada sea válida. La especificación de la función de utilidad observada válida que finalmente se elige es la que presente mejores valores de bondad de ajuste según los criterios de la metodología.

En el caso de que se consideren variaciones individuales en las preferencias o gustos de los decisores a través de una especificación logit mixto con coeficientes aleatorios, el último paso de la especificación de la función de utilidad observada consiste en elegir los coeficientes que se especificarán como variables aleatorias, así como el modelo de distribución de probabilidad que se supondrá para cada uno. En esta tesis se eligen los coeficientes aleatorios mediante un proceso hacia delante o forward. En la primera fase de este proceso se estiman todos los modelos correspondientes a la función de utilidad observada elegida anteriormente, pero con uno de los coeficientes especificado como una variable aleatoria con distribución de probabilidad Gaussiana. Como ya se ha indicado, dado que el modelo de probabilidad Gaussiano tiene los parámetros media y desviación típica, en los modelos logit mixto con coeficientes aleatorios el número de parámetros desconocidos se incrementa en un número igual al de coeficientes especificados como aleatorios. Las especificaciones logit mixto con coeficientes aleatorios no tienen una estructura cerrada. Por este motivo, los modelos logit mixto no se pueden estimar mediante máxima verosimilitud. En su lugar, como ya se ha indicado, se utilizará máxima verosimilitud simulada. En todas las estimaciones de modelos logit mixto con coeficientes aleatorios que se realizan en esta tesis utilizaremos 1000 extracciones de simulación. Entre las especificaciones válidas, se repite el proceso, pero especificando dos de los coeficientes como variables aleatorias con distribución de probabilidad Gaussiana. En caso de que alguna de estas especificaciones sea válida, el proceso continuaría con tres variables aleatorias. El proceso se detiene cuando ninguna de las especificaciones estimadas sea válida. De entre las especificaciones válidas, se elige la que presente mejores resultados de bondad de ajuste según los criterios descritos en la metodología.

Modelo multinomial logit

En este apartado, se especifica el modelo multinomial logit aplicado al contexto empírico real de la ciudad de Santander. Los parámetros desconocidos de este modelo son los coeficientes de la función de utilidad observada. Por tanto, el proceso consiste en la selección de los regresores de la función de utilidad observada, siguiendo el procedimiento stepwise descrito anteriormente.

El primer paso de este proceso de selección de regresores comienza con la estimación de la especificación del modelo multinomial logit saturado con los 17 regresores, que denominamos MNL-0. La tabla 2.3 muestra los resultados de estimación y bondad de ajuste de esta especificación. MNL-0 no es válido porque presenta regresores no relevantes. En concreto, 10 de los regresores se muestran no relevantes con un nivel de significación del 5%.

MNL-0						
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**
Estimación	β_{AC}	0,0107	0,00963	1,11	0,27	·
	$\beta_{AC \cdot H}$	-0,0160	0,0181	-0,88	0,38	·
	β_{JT}	-0,0453	0,0237	-1,92	0,06	·
	$\beta_{JT \cdot H}$	0,00584	0,0417	0,14	0,89	·
	β_{FO}	-0,830	0,424	-1,96	0,05	*
	$\beta_{FO \cdot H}$	-1,58	0,995	-1,58	0,11	·
	β_{IN}	0,313	0,229	1,37	0,17	·
	β_{HO}	1,41	0,343	4,11	0,00	**
	$\beta_{HO \cdot H}$	1,49	0,820	1,81	0,07	·
	β_{PS}	-0,984	0,301	-3,27	0,00	**
	$\beta_{PS \cdot H}$	1,95	0,552	3,52	0,00	**
	β_{PR}	-1,33	0,641	-2,07	0,04	*
	$\beta_{PR \cdot H}$	-0,594	1,10	-0,54	0,59	·
	β_{SC}	-0,0926	0,0451	-2,05	0,04	*
	$\beta_{SC \cdot H}$	0,246	0,0844	2,92	0,00	**
	β_{WT}	-0,140	0,0811	-1,73	0,08	·
	$\beta_{WT \cdot H}$	0,0203	0,187	0,11	0,91	·
		Nº par. estimados		17		
Bondad de ajuste		LL		-1660,571		
		FG		0,0446		
		ρ^2		0,0456		
		$\bar{\rho}_H^2$		0,0407		
		AIC		0,0358		
		$LRT(Nulo)$		158,51	0,00	**

Tabla 2.3. Resultados de estimación y bondad de ajuste de la primera iteración del proceso stepwise MNL.

Partiendo de la especificación MNL-0, y tras 70 etapas del proceso stepwise, se elige la especificación del modelo multinomial logit que denominamos MNL. La tabla 2.4 muestra los resultados de estimación, bondad de ajuste y validación de MNL. Esta especificación del modelo multinomial logit tiene seis regresores: tiempo de viaje al centro de trabajo (JT), número de extra-comunitarios residentes en la zona (FO), logaritmo neperiano del número de viviendas disponibles en la zona (HO), precio medio de las viviendas en la zona (PR) y las interacciones de nivel alto de ingresos (H) con el prestigio de la zona ($PS \cdot H$) y con el nº de centros de formación próximos al centroide de la zona ($SC \cdot H$). Los regresores son relevantes globalmente, pues el contraste de la razón de verosimilitudes de MNL respecto al modelo nulo es significativo. Además, cada regresor es relevante individualmente, utilizando el contraste de Wald, y todos tienen los signos esperados teóricamente. Respecto a la endogeneidad, las alternativas no son viviendas específicas sino áreas, y la variable de la función de utilidad es el precio medio

de las viviendas en el área. En esa situación, no se espera endogeneidad debido a atributos omitidos de una vivienda específica que están correlacionados con el precio. Los resultados no muestran indicios de endogeneidad, porque el precio tiene el signo esperado teóricamente, aunque no se puede descartar y es recomendable analizar detenidamente este tema en modelos estimados para el análisis de políticas. Las mismas consideraciones sobre la endogeneidad pueden aplicarse a los restantes modelos estimados en esta tesis doctoral. Por tanto, esta especificación es válida. JT, FO y PR presentan una relación inversa con la utilidad del decisor, HO, $PS \cdot H$ y $SC \cdot H$ presentan una relación directa con la utilidad del decisor. Utilizando los coeficientes estandarizados de este modelo, se deduce que la variable explicativa más influyente en la decisión es JT, un 26% del total, seguida con bastante diferencia de HO, $SC \cdot H$ y PR.

MNL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**RI**
Estimación	β_{JT}	-0,114	0,0291	-3,92	0,00	**	-0,448	26%
	β_{FO}	-0,870	0,307	-2,84	0,00	**	-0,195	11%
	β_{HO}	1,49	0,265	5,63	0,00	**	0,339	20%
	β_{PR}	-2,16	0,429	-5,03	0,00	**	-0,274	16%
	$\beta_{PS \cdot H}$	1,25	0,272	4,61	0,00	**	0,108	6%
	$\beta_{SC \cdot H}$	0,224	0,0516	4,33	0,00	**	0,336	20%
		Nº par. est.		6				
Bondad de ajuste		LL		-1667,968				
		FG		0,0440				
		ρ^2		0,0413				
		$\bar{\rho}_H^2$		0,0396				
		AIC		0,0379				
		LRT(Nulo)		143,71	0,00	**		
Val.		*PG-CV-4*		0,04348				
		PG-CV-10		0,04340				

Tabla 2.4. Resultados de estimación, bondad de ajuste y validación del modelo multinomial logit: MNL.

Mixed multinomial logit

En este apartado, se especifica el modelo mixto con coeficientes aleatorios que tiene de núcleo el modelo MNL estimado en el apartado anterior. La especificación se realiza mediante el proceso forward descrito anteriormente. En la primera fase del proceso forward, se han estimado seis especificaciones, cada una con 5 coeficientes fijos (todos los de MNL salvo uno), y uno aleatorio según una distribución normal de parámetros desconocidos. De las seis estimaciones, solo una de ellas presenta resultados válidos, según el criterio descrito anteriormente. La especificación válida es la que considera aleatorio el coeficiente de la interacción $SC \cdot H$. En la segunda fase del proceso forward, se han estimado las cinco especificaciones, cada una con 4 coeficientes fijos, y dos aleatorios según una distribución normal de parámetros desconocidos, siendo uno de

ellos siempre el correspondiente a la interacción $SC \cdot H$. Ninguna de estas 5 estimaciones presenta resultados válidos. Por tanto, se detiene el proceso forward y se selecciona la especificación mixed multinomial logit válida de la primera fase del proceso, que denominamos MMNL.

La tabla 2.5 muestra los resultados de estimación, bondad de ajuste y validación de MMNL, que tiene núcleo GEV MNL y el coeficiente $SC \cdot H$ aleatorio con distribución Gaussiana de parámetros desconocidos. Esta especificación es válida, porque todos los parámetros estimados son significativos y con signos correctos. Los resultados no muestran indicios de endogeneidad, porque el precio tiene el signo esperado teóricamente. El modelo MMNL obtiene resultados ligeramente superiores a su núcleo MNL, tanto en capacidad explicativa como en capacidad predictiva. Este resultado confirma que, en este modelo, la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. Los dos indicadores de bondad de ajuste penalizados por el número de parámetros estimados son ligeramente superiores en el caso del modelo MMNL respecto al MNL. El $\bar{\rho}_H^2$ de MMNL es 0,0399 frente a 0,0396 del MNL. El AIC de MMNL es 0,03786 frente a 0,03785 en el MNL. Los dos indicadores de validación también son ligeramente superiores en el modelo MMNL. El *PG-CV-4* del modelo MMNL es 0,04352 frente a 0,04350 en el modelo MNL. El *PG-CV-10* del modelo MMNL es 0,04343 frente a 0,04340 en el modelo MNL.

MMNL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**RI**
Estimación	β_{JT}	-0,118	0,0293	-4,03	0,00	**	-0,464	34%
	β_{FO}	-0,914	0,309	-2,95	0,00	**	-0,205	15%
	β_{HO}	1,49	0,266	5,62	0,00	**	0,339	25%
	β_{PR}	-2,12	0,433	-4,89	0,00	**	-0,269	20%
	$\beta_{PS \cdot H}$	1,12	0,287	3,91	0,00	**	0,097	7%
	$E(\beta_{SC \cdot H})$	0,16	0,0766	2,09	0,04	*		
	$\sigma(\beta_{SC \cdot H})$	0,289	0,127	2,27	0,02	*		
		Nº par. est.		7				
Bondad de ajuste		LL		-1666,953				
		FG		0,0441				
		ρ^2		0,0419				
		$\bar{\rho}_H^2$		0,0399				
		AIC		0,0379				
		LRT(Nulo)		145,74	0,00	**		
Val.		*PG-CV-4*		0,04352				
		PG-CV-10		0,04343				

Tabla 2.5. Resultados de estimación, bondad de ajuste y validación de la especificación mixed multinomial logit: MMNL.

2.4.3. Estructura de nidos de alternativas

Antes de estimar un modelo NL es necesario diseñar la estructura de nidos que agrupe las alternativas. El diseño consiste en la elección del número de nidos y de cuál es el nido al que pertenece cada una. El objetivo del diseño es que las alternativas de un mismo nido tengan la mayor correlación posible entre ellas, y que sean incorreladas con las alternativas de otros nidos. La figura 2.5 representa la estructura de nidos que se utiliza en esta tesis y que se diseñó para el modelo nested logit (Ibeas et al., 2013). En este capítulo se utiliza en la estimación de las diferentes especificaciones de ese modelo, pero en el capítulo cuarto se utiliza también para estimar el nuevo modelo que se propone. La estructura está formada por tres nidos (A, B y C), quedando dos alternativas en el nido raíz, es decir, incorreladas entre ellas y con el resto de alternativas. Puede observarse que esta estructura de nidos tiene una fuerte componente espacial.

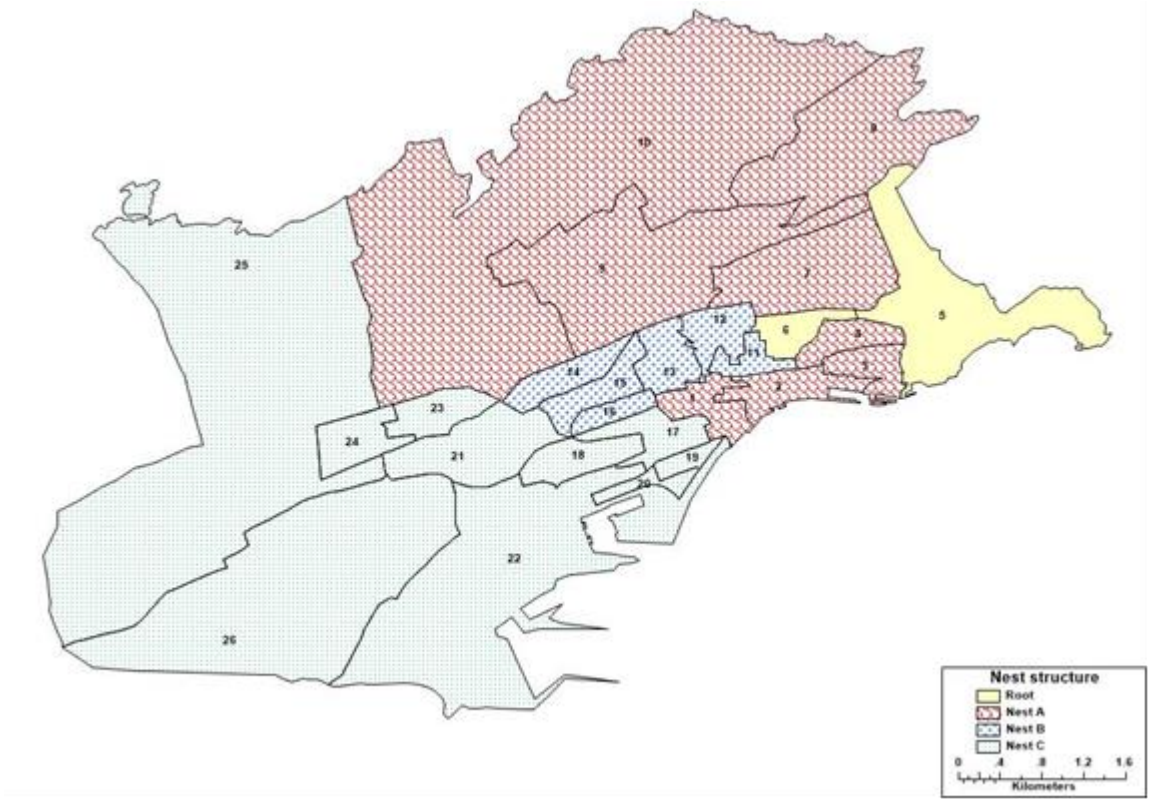

Figura 2.5. Estructura de los nidos de alternativas. Elaboración propia a partir de la zonificación de Ibeas et al. (2013).

Modelo nested logit

En este capítulo se especifica y estima el modelo nested logit aplicado al contexto empírico real de la ciudad de Santander, y utilizando la estructura de nidos de alternativas descrita anteriormente. Se ha realizado un nuevo proceso stepwise de especificación de la función de utilidad observada, utilizando el modelo nested logit. La tabla 2.6 muestra los resultados de estimación y bondad de ajuste de la primera iteración del proceso.

NL-0						
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**
Estimación	β_{AC}	0,00544	0,00864	0,63	0,53	·
	$\beta_{AC \cdot H}$	-0,0139	0,0162	-0,86	0,39	·
	β_{JT}	-0,0891	0,0361	-2,47	0,01	**
	$\beta_{JT \cdot H}$	-0,00125	0,0654	-0,02	0,98	·
	β_{FO}	-0,805	0,395	-2,04	0,04	*
	$\beta_{FO \cdot H}$	-1,35	0,863	-1,56	0,12	·
	β_{IN}	0,216	0,192	1,13	0,26	·
	β_{HO}	1,31	0,351	3,73	0,00	**
	$\beta_{HO \cdot H}$	1,29	0,728	1,78	0,08	·
	β_{PS}	-0,666	0,273	-2,44	0,01	**
	$\beta_{PS \cdot H}$	1,76	0,491	3,58	0,00	**
	β_{PR}	-1,56	0,604	-2,58	0,01	**
	$\beta_{PR \cdot H}$	-0,658	0,987	-0,67	0,51	·
	β_{SC}	-0,0527	0,0392	-1,35	0,18	·
	$\beta_{SC \cdot H}$	0,214	0,0729	2,93	0,00	**
	β_{WT}	-0,0949	0,0722	-1,31	0,19	·
	$\beta_{WT \cdot H}$	0,0443	0,163	0,27	0,79	·
	$\mu_A{}^{-1}$	1,24	0,153	8,09	0,00	**
	$\mu_B{}^{-1}$	1,27	0,140	9,08	0,00	**
	$\mu_C{}^{-1}$	1,10	0,0977	11,27	0,00	**
		Nº par. est.		20		
Bondad de ajuste		LL		-1654,730		
		FG		0,0451		
		ρ^2		0,0489		
		$\bar{\rho}_H^2$		0,0432		
		AIC		0,0374		
		$LRT(Nulo)$		170,19	0,00	**

Tabla 2.6. Resultados de estimación y bondad de ajuste de la primera iteración del proceso stepwise NL.

El proceso stepwise con el modelo nested logit obtiene los mismos seis regresores que utilizando el modelo multinomial logit. La tabla 2.7 muestra los resultados de estimación, bondad de ajuste y validación obtenidos con modelo nested logit especificado con la misma función de utilidad que el modelo MNL. La especificación es válida, porque todos los parámetros estimados en este modelo se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad. El orden de la influencia relativa de los regresores en la elección de localización residencial, que se obtiene a partir de los valores de los coeficientes tipificados, es el mismo que el obtenido en MNL.

El modelo NL se muestra más apropiado que el modelo MNL en este contexto, porque presenta mayor capacidad explicativa y predictiva que el modelo MNL. El modelo NL está anidado con el modelo MNL, pues adicionalmente estima los parámetros de disimilitud de los nidos. El modelo NL presenta mejores resultados de bondad de ajuste que el modelo MNL, porque el contraste de la razón de verosimilitudes del modelo NL respecto al modelo MNL es significativo. Además, el modelo NL también presenta mejores resultados de validación que el modelo MNL, tanto en la validación cruzada de cuatro grupos como en la de diez. Esto confirma la presencia de correlación entre alternativas de elección en este contexto empírico, y que la estructura de nidos diseñada en el modelo NL es capaz de recuperar una parte significativa de ella.

NL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,104	0,0271	-3,83	0,00	**	-0,409	24%
	β_{FO}	-1,00	0,300	-3,34	0,01	**	-0,224	13%
	β_{HO}	1,55	0,305	5,10	0,00	**	0,353	21%
	β_{PR}	-2,17	0,426	-5,08	0,00	**	-0,275	16%
	$\beta_{PS \cdot H}$	1,22	0,262	4,65	0,00	**	0,105	6%
	$\beta_{SC \cdot H}$	0,210	0,0460	4,57	0,00	**	0,315	19%
	${\mu_A}^{-1}$	1,25	0,147	8,55	0,00	**		
	${\mu_B}^{-1}$	1,26	0,128	9,81	0,00	**		
	${\mu_C}^{-1}$	1,05	0,089	11,84	0,00	**		
		Nº par. est.		9				
Bondad de ajuste		LL		-1661,270				
		FG		0,0446				
		ρ^2		0,0452				
		$\bar{\rho}_H^2$		0,0426				
		AIC		0,03998				
		$LRT(Nulo)$		157,11	0,00	**		
		$LRT(MNL)$		13,40	0,00	**		
		$LRT(RNL)$		13,55	0,01	**		
Val.		*PG-CV-4*		0,043899				
		PG-CV-10		0,043802				

Tabla 2.7. Resultados de estimación, bondad de ajuste y validación de la aplicación del modelo nested logit: NL.

El modelo NL también se muestra más apropiado que la especificación mixta MMNL en este contexto, porque significativamente presenta mayor capacidad explicativa y predictiva que el modelo MMNL. El modelo NL presenta mejores resultados de bondad de ajuste que el modelo MMNL en los dos indicadores de bondad de ajuste que penalizan el número de parámetros estimados, $\bar{\rho}_H^2$ y AIC. Además, el modelo NL también presenta mejores resultados de validación que el modelo MMNL, tanto en la validación cruzada de cuatro grupos como en la de diez. Esto confirma que la estructura de

correlación descrita por el diseño de nidos del modelo NL es más eficiente que la incorporación de variaciones en los gustos al modelo MNL mediante la estructura de coeficientes aleatorios elegida.

La tabla 2.8 muestra los resultados obtenidos con la especificación restricted nested logit del modelo NL anterior, que denominamos RNL. Este modelo tiene un parámetro adicional al modelo MNL, con el que está anidado. Además, el modelo NL está anidado con el modelo RNL, que es más sencillo porque reduce el número de parámetros desconocidos del modelo NL en una cifra igual al del número de nidos menos uno. La flexibilidad del modelo RNL respecto al modelo MNL compensará el parámetro adicional cuando exista correlación entre los pares de alternativas de un mismo nido, y tenga una intensidad semejante en todos ellos. La flexibilidad del modelo NL respecto al RNL compensará los parámetros desconocidos adicionales siempre que la correlación entre pares de alternativas de un mismo nido tenga diferente intensidad en cada nido. El modelo RNL especificado en este capítulo es válido, porque todos los parámetros estimados son significativos, los signos coinciden con los esperados teóricamente y no muestra indicios de endogeneidad. Pero este modelo no mejora los resultados de bondad de ajuste de MNL y tampoco respecto al modelo NL (véase en la tabla 2.7 el contraste de la razón de verosimilitudes de NL respecto a RNL). En base a estos resultados, las alternativas de un mismo nido son correladas, pero la intensidad no es la misma en todos los nidos.

RNL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,12	0,0308	-3,89	0,00	**	-0,413	26%
	β_{FO}	-0,894	0,315	-2,84	0,01	**	-0,174	11%
	β_{HO}	1,52	0,277	5,49	0,00	**	0,298	19%
	β_{PR}	-2,31	0,45	-5,13	0,00	**	-0,271	17%
	$\beta_{PS \cdot H}$	1,25	0,284	4,39	0,00	**	0,098	6%
	$\beta_{SC \cdot H}$	0,228	0,0527	4,33	0,00	**	0,307	20%
	μ^{-1}	1,14	0,0985	8,87	0,00	**		
		Nº par. est.		7				
Bondad de ajuste		LL		-1.667,207				
		FG		0,0441				
		ρ^2		0,0417				
		$\bar{\rho}_H^2$		0,0397				
		AIC		0,0377				
		$LRT(Nulo)$		145,23	0,00	**		
		$LRT(MNL)$		1,52	0,22	·		
Val.		*PG-CV-4*		0,04352				
		PG-CV-10		0,04343				

Tabla 2.8. Resultados de estimación, bondad de ajuste y validación de la aplicación del modelo restricted nested logit: RNL.

Mixed nested logit

Se ha repetido el proceso forward de selección de coeficientes aleatorios, en este caso con el modelo NL, y se ha obtenido la misma estructura mixta de coeficientes aleatorios que en el caso MMNL. La tabla 2.9 presenta los resultados de estimación, bondad de ajuste y validación de esta especificación mixed logit, que llamamos M-NL, que tiene núcleo GEV NL y la estructura de coeficientes mixtos descrita anteriormente.

	M-NL							
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,108	0,0273	-3,95	0,00	**	-0,425	31%
	β_{FO}	-1,04	0,303	-3,42	0,00	**	-0,233	17%
	β_{HO}	1,52	0,304	5,00	0,00	**	0,346	25%
	β_{PR}	-2,07	0,429	-4,81	0,00	**	-0,262	19%
	$\beta_{PS\cdot H}$	1,11	0,272	4,07	0,00	**	0,096	7%
	$E(\beta_{SC\cdot H})$	0,149	0,0679	2,19	0,03	*		
	$\sigma(\beta_{SC\cdot H})$	-0,261	0,112	-2,33	0,02	*		
	$\mu_A{}^{-1}$	1,28	0,148	8,62	0,00	**		
	$\mu_B{}^{-1}$	1,24	0,128	9,73	0,00	**		
	$\mu_C{}^{-1}$	1,05	0,0885	11,88	0,00	**		
		Nº par. est.		10				
Bondad de ajuste		SLL		-1660,111				
		FG		0,0447				
		ρ^2		0,0458				
		$\bar{\rho}_H^2$		0,0429				
		AIC		0,0401				
		$LRT(Nulo)$		159,43	0,00	**		
		$LRT(MMNL)$		13,68	0,00	**		
		$LRT(MRNL)$		10,76	0,00	**		
Val.		*PG-CV-4*		0,043886				
		PG-CV-10		0,043798				

Tabla 2.9. Resultados de estimación, bondad de ajuste y validación de la aplicación del modelo mixed nested logit: M-NL.

El modelo es válido, pues todos los parámetros estimados en este modelo se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad. La especificación M-NL presenta mejores resultados de bondad de ajuste que NL en los dos indicadores de bondad de ajuste que penalizan el número de parámetros estimados, $\bar{\rho}_H^2$ y AIC. Sin embargo, los resultados de validación cruzada son ligeramente superiores en su núcleo GEV. El valor de PG-CV-4 en M-NL es 0,043886 frente a 0,043899 en NL, y el valor de PG-CV-10 en M-NL es 0,043798 frente a 0,043802 en NL. La especificación mixta del modelo nested logit se muestra más

apropiada que la especificación mixta del modelo multinomial logit, pues M-NL presenta mayor capacidad explicativa y predictiva que MMNL.

En la tabla 2.10 se muestran los resultados de estimación, bondad de ajuste y validación de la especificación mixta de RNL, que llamamos MRNL, utilizando la misma estructura de coeficientes aleatorios que en los casos anteriores: MMNL y M-NL. Esta especificación MRNL tiene un comportamiento semejante al de su núcleo RNL. MRNL es una especificación válida, porque todos los parámetros estimados se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad. Esta especificación mejora la capacidad explicativa y predictiva de su núcleo GEV. Por otro lado, MRNL no mejora el ajuste de MMNL, pues el contrate de la razón de verosimilitudes respecto a él no es significativo. Por último, M-RNL mejora el ajuste de MRNL, porque el contraste de la razón de verosimilitudes de M-NL respecto a MRNL sí es significativo, tal y como se recoge en la tabla 2.9.

	MRNL							
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,124	0,031	-4,00	0,00	**	-0,425	34%
	β_{FO}	-0,933	0,317	-2,94	0,00	**	-0,182	14%
	β_{HO}	1,51	0,277	5,46	0,00	**	0,298	24%
	β_{PR}	-2,26	0,451	-5,02	0,00	**	-0,266	21%
	$\beta_{PS \cdot H}$	1,14	0,299	3,81	0,00	**	0,088	7%
	$E(\beta_{SC \cdot H})$	0,166	0,0774	2,14	0,03	*		
	$\sigma(\beta_{SC \cdot H})$	0,299	0,135	2,22	0,03	*		
	μ^{-1}	1,15	0,0967	8,99	0,00	**		
		Nº par. est.		8				
Bondad de ajuste		SLL		-1666,274				
		FG		0,0441				
		ρ^2		0,0423				
		$\bar{\rho}_H^2$		0,0400				
		AIC		0,0377				
		$LRT(Nulo)$		147,10	0,00	**		
		$LRT(MMNL)$		1,36	0,24	·		
Val.		*PG-CV-4*		0,043886				
		PG-CV-10		0,043798				

Tabla 2.10. Resultados de estimación, bondad de ajuste y validación de la aplicación del modelo mixed restricted nested logit: MRNL.

2.5. Resumen y conclusiones

En este capítulo se analizan los modelos de interacción entre usos de suelo y transporte para la predicción de la demanda urbana de usos de suelo y transporte. En este contexto, se analiza la problemática de la modelización de elección espacial, focalizando

en el caso concreto de la elección de la localización residencial. El enfoque que se considera más apropiado es el econométrico desagregado. Este enfoque se basa en el marco de los modelos de elección discreta multinomial, derivados de la teoría económica de maximización de la utilidad aleatoria.

No todos los modelos de elección discreta son compatibles con elecciones espaciales, por la presencia de dependencia espacial entre alternativas o por el elevado número de alternativas de elección. En este capítulo se analizan estas problemáticas, y otras como la endogeneidad. El modelo multinomial logit es el primer planteamiento, pero supone incorrelación entre alternativas, que no se justifica en modelización de elección espacial, aunque este enfoque todavía está presente en muchas aplicaciones. Además, se describe un primer enfoque para la modelización de elección espacial, el del modelo nested logit. Este modelo es el más extendido en presencia de correlación entre alternativas, que se incorpora al modelo mediante una estructura de nidos de alternativas diseñada por el analista. Ambos modelos son compatibles con especificaciones mixtas con coeficientes aleatorios que permiten incorporar variaciones en los gustos de los individuos decisores.

Los dos modelos que se han descrito en este capítulo, multinomial logit (MNL) y nested logit (NL), se han aplicado a un mismo caso real de elección de la localización residencial, en la ciudad de Santander. Los regresores de la función de utilidad observada se eligieron en ambos casos siguiendo el procedimiento descrito en este capítulo. En ambos modelos se obtuvo la misma función de utilidad observada. Los criterios de estimación y comparación de los modelos se han descrito en este capítulo, y son los mismos para todos los modelos que se estiman en esta tesis. En el caso del modelo nested logit, se estimó también una especificación que denominamos restricted nested logit (RNL), que supone que los pares de alternativas de un mismo nido son igualmente correladas, sea cual sea el nido al que ambas pertenecen.

La figura 2.6 compara visualmente los valores obtenidos en los indicadores de bondad de ajuste (los que penalizan el número de parámetros estimados) y validación de todos los núcleos logit estimados en este capítulo. La especificación RNL no mejora significativamente el ajuste del modelo MNL, pero el modelo NL muestra mayor capacidad explicativa y predictiva que los modelos MNL y RNL. Esto pone de manifiesto la presencia de correlación entre alternativas de elección en este contexto empírico, y que la estructura de nidos diseñada para esta aplicación es capaz de representarla de forma significativamente eficiente. Además, la mayor capacidad explicativa y predictiva del modelo nested logit respecto a su especificación restricted nested logit pone de manifiesto que, en este contexto empírico, la correlación entre pares de alternativas de un mismo nido no es necesariamente igual en todos los nidos.

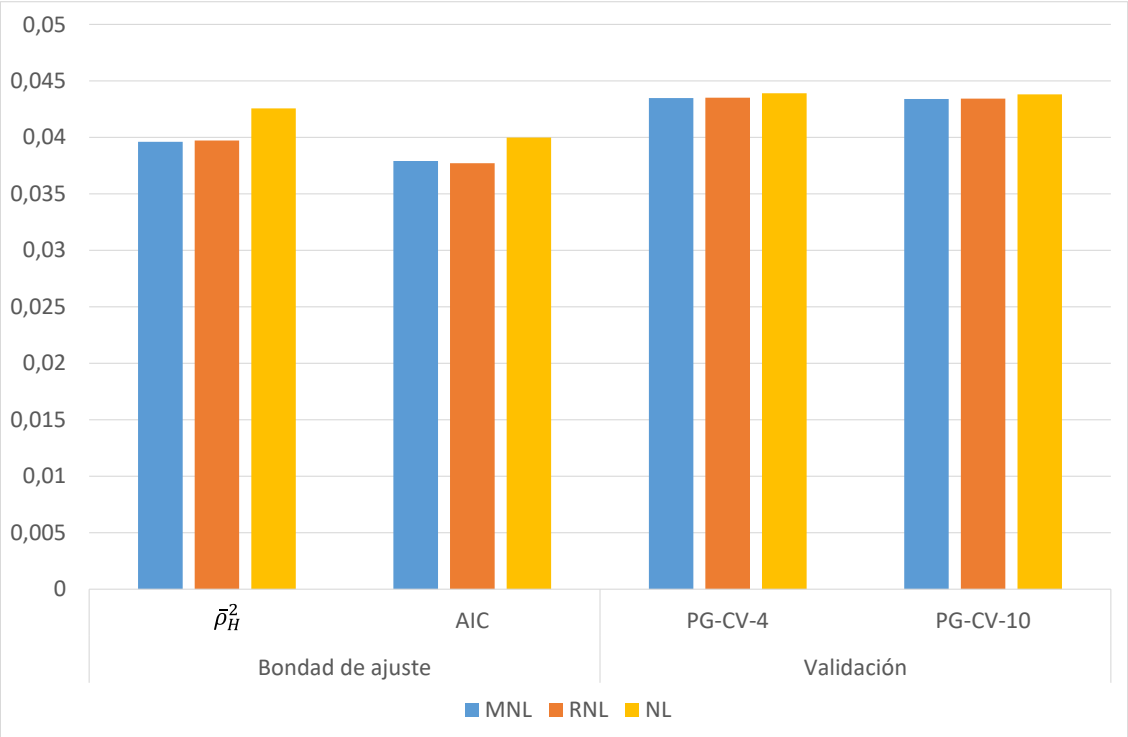

Figura 2.6. Estadísticos de bondad de ajuste y validación cruzada de los núcleos GEV estimados en este capítulo.

Las correspondientes especificaciones mixtas con coeficientes aleatorios de ambos modelos se aplican al mismo contexto empírico. La elección de los coeficientes aleatorios se realizó mediante el procedimiento descrito en esta capítulo en el modelo multinomial logit y en el nested logit, y de nuevo se obtuvo la misma conclusión en ambos. Todas las especificaciones mixtas mejoran la capacidad explicativa y predictiva de sus núcleos logit. Este resultado confirma que, en este modelo, la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. La comparación entre las especificaciones mixtas obtiene conclusiones semejantes a las obtenidas con sus respectivos núcleos logit, tal y como muestra la figura 2.7. La especificación MRNL no mejora significativamente el ajuste de MMNL, pero M-NL mejora la capacidad explicativa y predictiva de ambos.

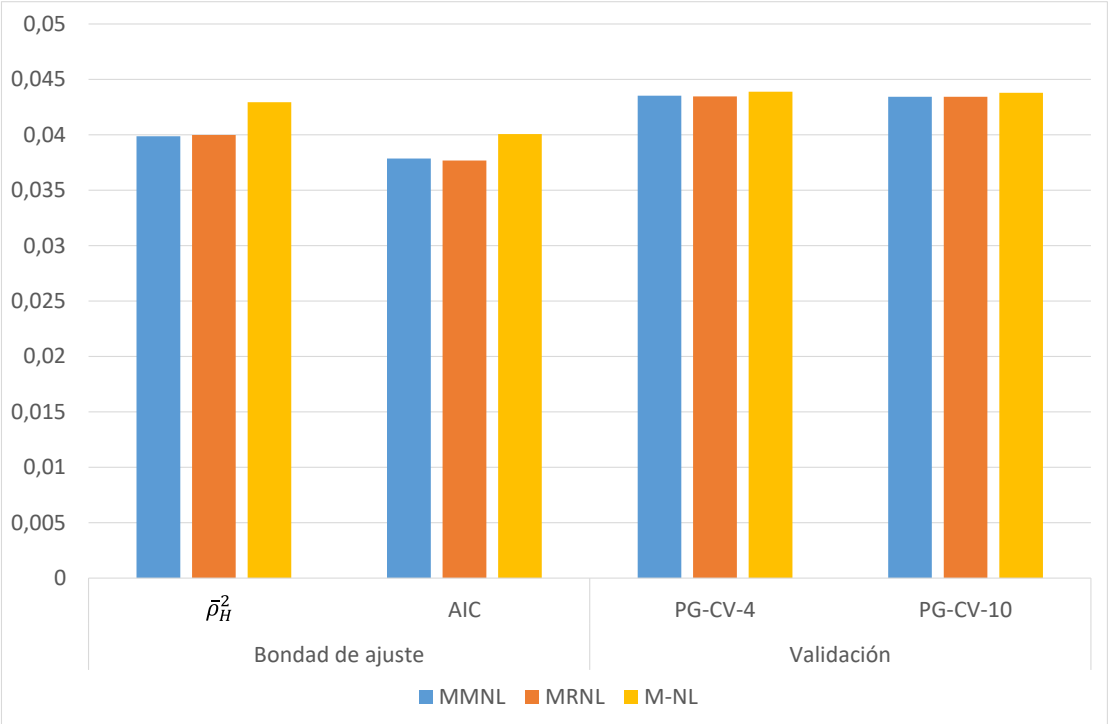

Figura 2.7. Estadísticos de bondad de ajuste y validación cruzada de las especificaciones mixtas estimadas en este capítulo.

3. Nuevas métricas para modelos de elección discreta con correlación espacial

En este capítulo se estudia un segundo enfoque de los modelos de elección discreta para incorporar correlación entre alternativas de elección, que sea compatible con modelización de elección espacial. Este enfoque no requiere el diseño de la estructura de nidos de alternativas que sí precisa el enfoque del modelo nested logit, descrito en el capítulo anterior.

El enfoque analizado permite incorporar correlación espacial entre alternativas utilizando la información espacial de las mismas. El planteamiento inicial con este enfoque, y el más sencillo, es un modelo de la familia paired generalized nested logit que utiliza la información espacial dicotómica de la contigüidad espacial o no de las alternativas. Posteriores generalizaciones de este planteamiento proponen utilizar información espacial más compleja, que requiere del uso de sistemas de información geográfica (GIS), pero que recoja con mayor precisión la correlación espacial entre las alternativas. En esta tesis proponemos una generalización que, siendo más sencilla que otras anteriores, se muestra suficientemente eficiente. Además, se propone una nueva métrica espacial que sea eficiente en aplicaciones con zonificaciones de diferentes tamaños y formas irregulares. Los modelos analizados se aplican al caso real de elección de la localización residencial en la ciudad de Santander, que se presentó en el capítulo anterior. Los resultados se comparan con la misma metodología propuesta también en el capítulo anterior. Esta propuesta ha sido publicada en el artículo Pérez-López et al. (2020).

3.1. Modelos generalized nested logit

McFadden (1978) generalizó los modelos logit con la clase de modelos generalized extreme value (GEV). Al igual que el modelo multinomial logit, los errores aleatorios de muchos de los modelos GEV son homocedásticos (una excepción sería el heteroscedastic logit; Steckel y Vanhonacker ,1988; Bhat, 1995 y Recker, 1995; véase Train, 2009:92), con igual distribución marginal Gumbel. Pero a diferencia del modelo multinomial logit, los errores aleatorios de los modelos GEV pueden ser mutuamente dependientes, al igual que sucede con el modelo hierarchical logit. La distribución conjunta del vector de errores aleatorios se obtiene mediante la fórmula (3.1). McFadden (1978) demostró que esta función es una distribución multivariante valor extremo, siempre que la función $G(x_1, \dots, x_A)$, llamada función generatriz, cumpla las propiedades que se describen a continuación, establecidas por McFadden (1978) y revisadas por Ben-Akiva y Francois (1983):

- La función generatriz es diferenciable no-negativa definida en $(\mathbb{R}^+)^A$, siendo A el número de alternativas del modelo.
- La función generatriz GEV es homogénea de grado $\xi > 0$, es decir, cumple la propiedad (3.2).
- La función generatriz es no acotada según la propiedad (3.3).
- Las derivadas parciales cruzadas de la función generatriz de orden impar son no-negativas, y las de orden par son no-positivas, es decir, cumple la propiedad (3.4).

$$F_{\varepsilon_1,\ldots,\varepsilon_A}(x_1,\ldots,x_A) = exp\{-G(e^{-x_1},\ldots,e^{-x_A})\}, x_i > 0, \forall i = 1,\ldots,A \quad (3.1)$$

$$G(\alpha y) = \alpha^{\xi} G(y), \forall y \in (\mathbb{R}^+)^A \quad (3.2)$$

$$\lim_{y_i \to +\infty} G(y_1,\ldots,y_i,\ldots,y_A) = +\infty, \forall y_i > 0, i = 1,\ldots,A \quad (3.3)$$

$$(-1)^k \frac{\partial^k G}{\partial y_{i1},\ldots,\partial y_{ik}}(y) \leq 0, \forall y \in (\mathbb{R}^+)^A \quad (3.4)$$

La probabilidad de elección de cada alternativa de un modelo GEV se calcula mediante la fórmula (3.5), donde $y_i = e^{V_i}$. Dada la homogeneidad de la función generatriz y utilizando el teorema de Euler (Abbe et al., 2007), la probabilidad de elección de cada alternativa se puede escribir según la fórmula (3.6), donde $G_i(y_1,\ldots,y_A) = \partial G(y_1,\ldots,y_A)/\partial y_i$.

$$P_i = \frac{y_i \cdot \frac{\partial G(y_1,\ldots,y_A)}{\partial y_i}}{G(y_1,\ldots,y_A)}, \forall i = 1,\ldots,A \quad (3.5)$$

$$P_i = \frac{e^{V_i + \ln G_i(y_1,\ldots,y_A)}}{\sum_{j=1}^{A} e^{V_j + \ln G_j(y_1,\ldots,y_A)}}, \forall\, i = 1,\ldots,A \quad (3.6)$$

McFadden (1978) demostró que los modelos GEV son consistentes con la teoría económica de maximización de la utilidad aleatoria. Cada función generatriz elegida determina un modelo GEV diferente. El modelo multinomial logit es GEV, y su función generatriz se muestra en la fórmula (3.7). Los modelos GEV permiten incorporar restricciones en la matriz de covarianzas que, si son relativamente simples, mantienen una estructura matemática cerrada para la probabilidad de elección. Por tanto, los modelos GEV permiten especificar familias de modelos logit más flexibles que el modelo multinomial logit. Estos modelos relajan la hipótesis de incorrelación entre alternativas, pero que pueden mantener estructuras matemáticas cerradas de las probabilidades de elección. Al igual que el resto de modelos logit, los modelos GEV pueden actuar como núcleo de especificaciones mixed logit con coeficientes aleatorios (MGEV; Bhat y Guo, 2004; Hess et al., 2005), que permiten especificar variaciones aleatorias en los gustos o preferencias de los individuos decisores.

$$G(e^{V_1},\ldots,e^{V_A}) = \sum_{i=1}^{A} e^{V_i} \quad (3.7)$$

Se han presentado diferentes modelos GEV. Los más importantes serán analizados en detalle a lo largo de esta tesis, aunque hay otros utilizados con menor frecuencia como es el modelo ordered GEV (Small, 1987) o el modelo product differentiation logit (Bresnahan et al., 1997). El enfoque de nidos de alternativas diseñados por el analista, propuesto en el modelo hierarchical logit para especificar estructuras de correlación entre alternativas, se puede trasladar a la formulación GEV. La fórmula (3.8) muestra la función generatriz del modelo nested logit o modelo hierarchical logit de dos niveles, en la que $(\mu_1,\ldots,\mu_M)$ es el vector de parámetros de disimilitud de la estructura de nidos definida por el analista. Los nidos pueden recoger tanto correlación espacial no observada, como correlación debida a variables no observadas de naturaleza no espacial.

$$G(e^{V_1}, \dots, e^{V_A}) = \sum_{k=1}^{M} \left(\sum_{i \epsilon Nest_k} (e^{V_i})^{1/\mu_k} \right)^{\mu_k} \quad (3.8)$$

Las extensiones GEV del modelo nested logit se basan en el modelo cross-nested logit (CNL; Small, 1987; Vovsha, 1997; Ben-Akiva y Bierlaire, 1999:5-34; Papola, 2004). Los modelos cross-nested logit permiten que las alternativas puedan pertenecer a más de un nido. Por este motivo incorporan los llamados parámetros de asignación. Estos parámetros son adicionales a los parámetros de disimilitud, ya presentes en el modelo nested logit. Los parámetros de asignación se interpretan como el nivel de pertenencia de cada alternativa a cada nido (Abbe et al., 2007). El número de parámetros de asignación de un modelo cross-nested logit es $A \cdot M$, siendo A el número de alternativas del modelo y M el número de nidos diseñados por el analista.

Wen y Koppelman (2001) hicieron un planteamiento GEV semejante al modelo cross-nested logit denominado generalized nested logit (GNL). Al igual que sucede en el modelo CNL, los modelos GNL permiten que las alternativas puedan pertenecer a más de un nido. Por este motivo también incorporan los parámetros de asignación, además de los parámetros de disimilitud. Los autores del modelo GNL incluyen la restricción de que los parámetros de asignación de cada alternativa sumen la unidad, como refleja la ecuación (3.9) (véase Abbe et al. 2007 para analizar propuestas de normalización en otras formulaciones cross-nested logit). Esta normalización permite que los parámetros de asignación de cada alternativa representen la proporción de pertenencia a cada nido. En esta tesis consideramos la propuesta GNL como una subfamilia del CNL, que se obtiene al añadir esta normalización a la formulación de Ben-Akiva y Bierlaire (1999:5-34).

$$\sum_{k=1}^{M} \alpha_{ik} = 1 \,, \forall i = 1, \dots, A \quad (3.9)$$

La fórmula (3.10) muestra la función generatriz GEV del modelo GNL, donde $\alpha_{ik} \geq 0$ es el parámetro de asignación de cada i-ésima alternativa al nido k-ésimo. Los parámetros de asignación tienen valor cero cuando la alternativa no pertenece al nido. Los modelos GNL están anidados con el modelo nested logit de dos niveles que utiliza los mismos parámetros de disimilitud, pues colapsa en él si cada alternativa pertenece a un solo nido con parámetro de asignación de valor uno. El modelo GNL es consistente con la teoría de maximización de la utilidad aleatoria, si los parámetros de disimilitud cumplen la propiedad (3.11). El nido raíz tiene parámetros de disimilitud y asignación de valor uno. El cálculo de la correlación no observada se realiza a partir de la función de distribución conjunta. En los modelos generalized nested logit esta función no tiene solución analítica, por lo que se requiere el uso de integración numérica. Papola (2004) propuso una aproximación a este cálculo, que se muestra en la fórmula (3.12). A partir de esta aproximación puede concluirse que la correlación no observada entre pares de alternativas de modelos GNL se modula mediante todos los parámetros estructurales, los de asignación y los de disimilitud.

$$G(e^{V_1}, \dots, e^{V_A}) = \sum_{k=1}^{M} \left(\sum_{i \epsilon Nest_k} (\alpha_{ik} \cdot e^{V_i})^{1/\mu_k} \right)^{\mu_k} \quad (3.10)$$

$$0 < \mu_k \leq 1, \forall k = 1, \dots, M \quad (3.11)$$

$$\widehat{Corr}(\varepsilon_i, \varepsilon_j) = \sum_{k=1}^{K} \sqrt{\alpha_{ik} \cdot \alpha_{j\mathrm{k}}} (1 - \mu_k^2), \ \forall i, j \in \{1, \dots, A\} \quad (3.12)$$

A partir de la fórmula (3.5) se calcula la expresión de la probabilidad de elección de las alternativas en un modelo GNL que se muestra en la fórmula (3.13). La probabilidad de elección de las alternativas que solo pertenecen al nido raíz coincide con la del modelo multinomial logit. La probabilidad de elección de una alternativa en el modelo GNL se puede expresar según la fórmula (3.14), como la suma de los productos de la probabilidad de elección de la alternativa si se elige cada nido (3.15) multiplicada por la probabilidad de elegir ese nido (3.16). La fórmula (3.17) muestra la elasticidad directa del modelo GNL. Puede comprobarse fácilmente que, en las alternativas que solo pertenecen al nido raíz, esta expresión coincide con la del modelo multinomial logit (2.26). La fórmula (3.18) muestra la elasticidad cruzada de un par de alternativas. Puede comprobarse fácilmente que, en los nidos donde no coincidan ambas alternativas, el sumando tomará el valor cero. Si ambas alternativas pertenecen al nido raíz o no coinciden en ningún nido, la expresión de la elasticidad cruzada del modelo GNL coincide con la del modelo multinomial logit (2.27).

$$P_i = \frac{\sum_{k=1}^{M}\left[\left(\alpha_{ik}e^{V_i}\right)^{1/\mu_k}\left(\sum_{j\epsilon Nest_k}\left(\alpha_{jk}\cdot e^{V_j}\right)^{1/\mu_k}\right)^{\mu_k-1}\right]}{\sum_{m=1}^{M}\left(\sum_{l\epsilon Nest_m}\left(\alpha_{lm}\cdot e^{V_l}\right)^{1/\mu_m}\right)^{\mu_m}} \tag{3.13}$$

$$P_i = \sum_{k=1}^{M} P_{i|k}\cdot P_k\,,\forall i\epsilon\{1,\dots,A\} \tag{3.14}$$

$$P_{i|k} = \frac{\left(\alpha_{ik}e^{V_i}\right)^{1/\mu_k}}{\sum_{l\epsilon Nest_m}\left(\alpha_{lm}\cdot e^{V_l}\right)^{1/\mu_m}} \tag{3.15}$$

$$P_k = \frac{\left(\sum_{j\epsilon Nest_k}\left(\alpha_{jk}\cdot e^{V_j}\right)^{1/\mu_k}\right)^{\mu_k}}{\sum_{m=1}^{M}\left(\sum_{l\epsilon Nest_m}\left(\alpha_{lm}\cdot e^{V_l}\right)^{1/\mu_m}\right)^{\mu_m}} \tag{3.16}$$

$$E_{X_{im}}^{P_i} = \frac{\sum_{k=1}^{M} P_{i|k}P_k\left[(1-P_i)+\left(\frac{1}{\mu_k}-1\right)(1-P_{i|k})\right]}{P_i}\beta_m X_{im}\forall i\in\{1,\dots,A\} \tag{3.17}$$

$$E_{X_{im}}^{P_j} = -\left[P_i + \frac{\sum_{k=1}^{M}\left(\frac{1}{\mu_k}-1\right)P_{i|k}P_kP_{j|k}}{P_j}\right]\beta_m X_{im} \tag{3.18}$$

El modelo GNL es más flexible que el modelo nested logit, aunque a costa de incrementar significativamente el número de parámetros estructurales. Por ejemplo, en un modelo de dos niveles con 26 alternativas repartidas en tres nidos, el modelo nested logit tiene tres parámetros estructurales adicionales a los que tendría un modelo multinomial logit, correspondientes a los tres parámetros de disimilitud. Sin embargo, el modelo GNL correspondiente tiene 81 parámetros estructurales adicionales a los que tendría un modelo multinomial logit, los tres de disimilitud más 26·3=78 de asignación. En el caso de que sean 100 alternativas y el mismo número de nidos, en el modelo nested logit seguiría teniendo 3 parámetros estructurales adicionales a los que tendría un modelo multinomial logit, pero en el GNL habría ahora 303 parámetros estructurales adicionales a los que tendría un modelo multinomial logit. Dado el elevado número de alternativas habitual en los modelos de elección espacial, los modelos GNL no suelen ser viables en este contexto. El problema es que el elevado número de parámetros desconocidos haría inviable la estimación del modelo. La única posibilidad consiste en encontrar mecanismos que permitan, o bien calcular una parte de los parámetros

estructurales sin necesidad de estimarlos, o bien especificar restricciones que reduzcan su número (Abbe et al., 2007).

El modelo paired combinatorial logit (PCL; Chu, 1981 y 1989; Koppelman y Wen, 2000) es un modelo GEV que utiliza parcialmente el enfoque de nidos de alternativas del modelo nested logit, pero tiene la ventaja de que utiliza una estructura de nidos no diseñada por el analista. Esta estructura de nidos consiste en considerar un nido para cada par de alternativas. La función generatriz GEV del modelo PCL se muestra en la fórmula (3.19), utilizando parámetros de disimilitud μ_{ij} en lugar de los parámetros de similitud $\sigma_{ij} = 1 - \mu_{ij}$ planteados por los autores. Sustituyendo esta función generatriz GEV en la ecuación (3.5) se obtiene la probabilidad de elección de cada alternativa en este modelo (3.20). El modelo PCL está anidado con el modelo multinomial logit, pues colapsa en él cuando los parámetros de disimilitud de todos los pares de alternativas tienen valor uno.

$$G(e^{V_1}, \dots, e^{V_A}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \left((e^{V_i})^{1/\mu_{ij}} + (e^{V_j})^{1/\mu_{ij}} \right)^{\mu_{ij}} \tag{3.19}$$

$$P_i = \frac{\sum_{j \neq i} (e^{V_i})^{1/\mu_{ij}} \left((e^{V_i})^{1/\mu_{ij}} + (e^{V_j})^{1/\mu_{ij}} \right)^{\mu_{ij}-1}}{\sum_{k=1}^{A-1} \sum_{l=k+1}^{A} \left((e^{V_k})^{1/\mu_{kl}} + (e^{V_l})^{1/\mu_{kl}} \right)^{\mu_{kl}}}, \forall i = 1, \dots, A \tag{3.20}$$

Los parámetros estructurales del modelo PCL adicionales a los que tendría un modelo multinomial logit son los parámetros de disimilitud. El número de parámetros de disimilitud del modelo PCL es igual al número de pares de alternativas, que es $A(A-1)/2$, donde A es el número de alternativas del modelo, tal y como se demuestra en (3.21), sin más que sumar los $A-1$ primeros términos de la serie aritmética $(A-i)$. Por tanto, el modelo PCL tiene la desventaja de que, cuando el número de alternativas es alto, el número de parámetros estructurales se incrementa de forma potencial, hasta el punto de que puede hacer inviable la estimación de todos los parámetros. Por ejemplo, un modelo con 26 alternativas tendrá 325 parámetros de disimilitud, y uno con 100 alternativas tendrá 4950 parámetros de disimilitud. Por tanto, siguiendo el mismo razonamiento descrito para el modelo GNL, el modelo PCL tampoco es viable para modelos de elección espacial, a no ser que se encuentren mecanismos que permitan o bien calcular una parte de los parámetros de disimilitud sin necesidad de estimarlos, o bien especificar restricciones que reduzcan su número.

$$\sum_{i=1}^{A-1} \sum_{j=i+1}^{A} 1 = \sum_{i=1}^{A-1} (A-i) = (A-1) \frac{(A-1)+(A-[A-1])}{2} = \frac{A(A-1)}{2} \tag{3.21}$$

Wen y Koppelman (2001) combinaron los modelos PCL y GNL para formular un nuevo modelo GEV de la familia GNL que llamaron paired generalized nested logit (PGNL). El modelo PGNL tiene la misma estructura de nidos del modelo PCL, es decir, la formada por todos los pares de alternativas. Además, el modelo PGNL incrementa la flexibilidad del modelo PCL porque añade los parámetros de asignación a cada nido del modelo GNL. En concreto, el modelo PGNL incorpora dos parámetros de asignación por cada par de nidos adicionales al parámetro de disimilitud que ya utiliza el modelo PCL. Dado que en la literatura no se ha encontrado una función generatriz GEV para el modelo PGNL, en esta tesis se ha deducido la expresión que recoge la fórmula (3.22), donde $\alpha_{i,ij}$ y $\alpha_{j,ij}$ son los parámetros de asignación de las alternativas i y j al nido formado por ambas

alternativas, respectivamente, y μ_{ij} es el parámetro de disimilitud de ese nido. Esta expresión fue desarrollada a partir de la expresión de probabilidad recogida por los autores, y permite encuadrar el modelo PGNL en el marco de los modelos GEV.

$$G(e^{V_1}, \dots, e^{V_A}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \left[\left(\alpha_{i,ij} e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij} e^{V_j}\right)^{1/\mu_{ij}} \right]^{\mu_{ij}} \quad (3.22)$$

El modelo PGNL mantiene la ventaja del modelo PCL de no necesitar que el analista diseñe una estructura de nidos. Además, el modelo PGNL es más flexible que el modelo PCL, porque incorpora los parámetros de asignación que permiten modelar el grado de pertenencia de cada alternativa a cada nido. La desventaja del modelo PGNL es que todo esto lo consigue a costa de incrementar aún más el número de parámetros estructurales del modelo PCL. En concreto, el modelo PGNL tiene el triple de parámetros estructurales (adicionales a la función de utilidad observada) que el modelo PCL. Por tanto, un modelo PGNL con A alternativas tiene A·(A-1)·3/2 parámetros estructurales, que corresponden a $A \cdot (A-1)/2$ parámetros de disimilitud y $A \cdot (A-1)$ parámetros de asignación. Además, es fácil observar que el número de parámetros estructurales del modelo PGNL es mucho mayor que en el modelo PCL y que en cualquier especificación GNL basada en una estructura de nidos elegida por el analista.

En el ejemplo de un modelo con 26 alternativas, el modelo PGNL tiene 975 parámetros estructurales adicionales a los que tendría el modelo multinomial logit, frente a los 325 parámetros de disimilitud del modelo PCL, los 81 del modelo GNL y los tres del modelo NL (suponiendo en los dos últimos modelos que el analista elige una estructura de tres nidos). En el caso de que sean 100 alternativas, el modelo PGNL tiene 14850 parámetros estructurales adicionales a los que tendría el modelo multinomial logit, frente a los 4950 del PCL, los 302 del GNL y los tres del NL (suponiendo en los dos últimos modelos que el analista elige una estructura de tres nidos). Por tanto, siguiendo el mismo razonamiento descrito para los modelos GNL y PCL, el modelo PGNL tampoco es viable para modelos de elección espacial, a no ser que se encuentren mecanismos que permitan o bien calcular una parte de los parámetros de disimilitud sin necesidad de estimarlos, o bien especificar restricciones que reduzcan su número.

3.2. Modelo spatially correlated logit

Bhat y Guo (2004) proponen un nuevo modelo logit, el spatially correlated logit (SCL), que permite incorporar correlación espacial entre alternativas sin necesidad de especificar una estructura de nidos de alternativas como la del modelo nested logit o el GNL. Este nuevo modelo GEV utiliza la función generatriz (3.23), que tiene $A(A-1)+1$ parámetros estructurales adicionales al modelo multinomial logit, los $A(A-1)$ parámetros de asignación, $\alpha_{i,ij}$, dos por cada alternativa; y el parámetro de disimilitud, $0 < \mu \leq 1$. Pero el único parámetro estructural desconocido adicional al modelo multinomial logit es el de disimilitud.

$$G(e^{V_1}, \dots, e^{V_A}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \left(\left(\alpha_{i,ij} e^{V_i}\right)^{1/\mu} + \left(\alpha_{j,ij} e^{V_j}\right)^{1/\mu} \right)^{\mu} \quad (3.23)$$

Este modelo incorpora la correlación espacial entre alternativas a través de los $A(A-1)$ parámetros estructurales adicionales al de disimilitud, dos para cada par de alternativas. Estos parámetros estructurales, denominados parámetros de asignación, no necesitan ser estimados, sino que se calculan previamente al proceso de estimación

a partir de cierta información espacial de los pares de alternativas. En concreto, los parámetros de asignación se calculan a partir de la fórmula (3.24), donde la variable espacial dicotómica ω_{ij} toma el valor 1 cuando ambas alternativas i, j son contiguas (es decir, tienen alguna parte de la frontera en común), y 0 en caso contrario. Puede comprobarse fácilmente que los parámetros de asignación así definidos cumplen la condición $\sum_{j\neq i}\alpha_{i,ij}, \forall i = 1, \dots, A$. Esto es, que la asignación total de cada alternativa con el resto de alternativas (sean contiguas o no) es la unidad. Esta propiedad es coherente con la propiedad (3.9) de los modelos GNL. A partir de ella, puede comprobarse fácilmente que el modelo SCL está anidado con el modelo multinomial logit, pues colapsa en él al añadir la restricción de que el parámetro de disimilitud tiene valor uno.

$$\alpha_{i,ij} = \frac{\omega_{ij}}{\sum_{l=1}^{A}\omega_{il}}, \forall i, j \epsilon \{1, \dots, A\} \tag{3.24}$$

La expresión de la probabilidad de elección de cada alternativa en el modelo SCL se muestra en la fórmula (3.25), que al igual que en el resto de modelos GEV, se deriva de su función generatriz (3.23) aplicando la fórmula (3.5). La fórmula (3.26) muestra la elasticidad directa del modelo SCL, y la fórmula (3.27) su elasticidad cruzada.

$$P_i = \sum_{\substack{j=1 \\ j\neq i}}^{A}\left[\frac{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu}}{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu}+\left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu}}\frac{\left(\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu}+\left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu}\right)^{\mu}}{\sum_{k=1}^{A-1}\sum_{l=k+1}^{A}\left(\left(\alpha_{k,kl}e^{V_k}\right)^{1/\mu}+\left(\alpha_{l,kl}e^{V_l}\right)^{1/\mu}\right)^{\mu}}\right],$$

$$\forall i \epsilon \{1, \dots, A\} \tag{3.25}$$

$$E_{X_{im}}^{P_i} = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i|ij}P_{ij}\left[(1-P_i)+\left(\frac{1}{\mu}-1\right)(1-P_{i|ij})\right]}{P_i}\beta_m X_{im} \tag{3.26}$$

$$E_{X_{im}}^{P_j} = -\left[P_i + \left(\frac{1}{\mu}-1\right)\frac{P_{i|ij}P_{ij}P_{j|ij}}{P_j}\right]\beta_m X_{im} \tag{3.27}$$

El modelo SCL es un modelo logit orientado a la modelización de elección espacial, porque incorpora correlación entre alternativas y el número de parámetros estructurales es independiente del número de alternativas. La ventaja de este modelo frente al modelo multinomial logit es que es más flexible y permite incorporar correlación entre alternativas. La ventaja del modelo SCL frente al modelo nested logit es que no necesita que el analista diseñe una estructura de nidos para recoger la correlación espacial entre alternativas. Además, el modelo SCL es más sencillo que el modelo nested logit, porque tiene un solo parámetro estructural desconocido, mientras que el modelo nested logit tiene un número de parámetros estructurales igual al número de nidos de la estructura diseñada por el analista. En el caso de un modelo nested logit con un único nido, tendría el mismo número de parámetros estructurales desconocidos y no necesita que el analista diseñe una estructura de nidos, pero tiene la desventaja de que no incorpora la correlación espacial entre alternativas que sí capturan los parámetros del modelo SCL.

Aunque los autores no lo especifican, en esta tesis proponemos que el modelo SCL se puede plantear como una especificación del modelo PGNL, que combina las dos estrategias para reducir el número de parámetros estructurales desconocidos que explicamos en las subsecciones anteriores: calcular una parte de los parámetros estructurales sin necesidad de estimarlos, y especificar restricciones que reduzcan su

número. Por un lado, si consideramos en el modelo PGNL la restricción de que todos los parámetros de disimilitud sean iguales, tendríamos el único parámetro de disimilitud del modelo SCL. Por otro lado, si se calculan los parámetros de disimilitud a partir de la información espacial de las alternativas tal y como muestra la fórmula (3.24) del modelo SCL, tendríamos los parámetros de asignación del modelo PGNL, pero ninguno de ellos sería desconocido. Por tanto, esta estrategia evita la necesidad de estimar la totalidad de los $A(A-1)$ parámetros de asignación del modelo PGNL. Teniendo en cuenta ambas estrategias, el número de parámetros estructurales desconocidos se reduce de los $A(A-1)3/2$ del modelo PGNL, a solo uno en esta especificación denominada SCL. Se puede comprobar fácilmente que la fórmula (3.23) y las fórmulas (3.25) a (3.27) se pueden deducir de las correspondientes del modelo PGNL sin más que aplicar la restricción de que todos los parámetros de asignación sean iguales a μ.

Pero el modelo SCL tiene carencias. Por un lado, el único parámetro estructural que se estima con los datos de la muestra, el parámetro de disimilitud, es el mismo para todos los pares de alternativas. La eficiencia del modelo SCL para recoger la correlación espacial entre alternativas mediante los parámetros de asignación depende de si la dicotomía de vecindad, combinada con igual disimilitud de todos los pares de alternativas, refleja de forma eficaz la correlación entre alternativas. Parece razonable pensar que la dicotomía de vecindad pueda reflejar bastante eficientemente la correlación entre alternativas espaciales con formas regulares. Se trataría, por ejemplo, de la correlación existente en el caso de zonificaciones basadas en mallas de celdas o, al menos, en espacios urbanos que hayan sido planificados con formas regulares. Pero en muchos contextos de modelización de elección de localización espacial, la zonificación de las alternativas suele basarse en áreas administrativas. En concreto, en el caso de modelización de la elección de localización residencial urbana, lo más habitual es utilizar zonificaciones de este tipo. El motivo es la disponibilidad de fuentes de datos públicas de las áreas administrativas y, por tanto, de las alternativas consistentes en estas áreas o agrupaciones de ellas.

Las áreas administrativas pueden tener formas muy irregulares. En concreto, en los centros urbanos es muy habitual esta situación, especialmente si se trata zonas históricas, como es habitual en Europa. En aplicaciones empíricas con este tipo de zonificación, la estructura de correlación espacial entre alternativas que es capaz de recoger el modelo SCL es reducida. La figura 3.1 muestra un ejemplo teórico de zonificación irregular en el que la contigüidad dicotómica del modelo SCL tiene el mismo valor en los pares de alternativas 2-1 y 2-3. Sin embargo, no parece razonable deducir la misma dependencia espacial entre ambos pares de alternativas.

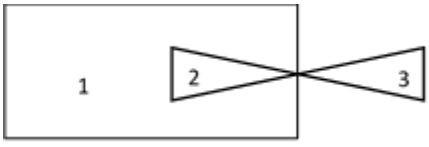

Figura 3.1 Ejemplo teórico de zonificación irregular.

Al igual que el resto de modelo GEV, el modelo SCL puede especificarse como núcleo de un modelo mixed logit con coeficientes aleatorios, denominado mixed spatially correlated logit (MSCL). Bhat y Guo (2004) aplicaron una especificación MSCL a un caso de elección de la localización residencial en contexto LUTI. Los resultados obtenidos fueron comparados con los obtenidos con el modelo multinomial logit especificado con

la misma función de utilidad observada. El área espacial del estudio cubre una parte del condado de Dallas, en Texas (USA), que incluye las ciudades de University Park, Highland Park y Dallas y cubre 98 de las 383 Transport Analysis and Processing zonas de este condado. Los datos de muestra se basaron principalmente en la 1996 Dallas–Fort Worth metropolitan area household activity survey. El modelo mixto SCL obtuvo un mejor ajuste que el MNL.

Sener et al. (2011) propusieron una generalización del modelo SCL, los modelos generalized spatially correlated logit (GSCL). Esta familia de modelos permite definir modelos GEV con correlación espacial que recogen estructuras de correlación entre alternativas más complejas que el modelo SCL. La diferencia de los modelos GSCL respecto al modelo SCL radica únicamente en que los parámetros de asignación de esta nueva familia de modelos se calculan según la fórmula (3.28), donde z_{ij} son atributos que caracterizan la relación entre cada par de alternativas i, j ($z_{ii} = 0, \forall i$), y ϕ' es un vector de parámetros desconocidos que es necesario estimar. Los parámetros de asignación así calculados también cumplen las propiedades (3.29) y (3.30). Se puede comprobar fácilmente que los modelos GSCL no están necesariamente anidados con el modelo SCL. Por ejemplo, Sener et al. (2011) proponen el modelo de la familia GSCL denominado distance-based spatially correlated logit, en el que el vector z_{ij} utiliza el logaritmo neperiano de la distancia euclídea entre pares de alternativas d_{ij}, según la fórmula (3.31).

$$\alpha_{i,ij} = \frac{e^{\phi' z_{ij}}}{\sum_k e^{\phi' z_{ik}}} \tag{3.28}$$

$$0 < \alpha_{i,ij} < 1, \forall\, i, j \tag{3.29}$$

$$\sum_j \alpha_{i,ij} = 1, \forall\, i, j \tag{3.30}$$

$$\alpha_{i,ij} = \frac{e^{\phi' \ln d_{ij}}}{\sum_k e^{\phi' \ln d_{ik}}} = \frac{{d_{ij}}^{\phi'}}{\sum_k {d_{ik}}^{\phi'}} \tag{3.31}$$

Los modelos GSCL pueden tener ventajas frente al modelo SCL. Los modelos de esta familia son más flexibles que el modelo SCL, porque pueden tener más parámetros estructurales desconocidos que el modelo SCL, recogidos en el vector ϕ'. Los autores señalan que estos parámetros adicionales pueden incorporar información espacial y no espacial. Por un lado, los modelos GSCL pueden incorporar estructuras de correlación espacial entre alternativas más complejas que el modelo SCL, a través de la elección de los atributos z_{ij}, como es el caso del modelo distance-based spatially correlated logit. Por otro lado, estos modelos GSCL también pueden incorporar correlación no-espacial entre alternativas a través de la definición de los parámetros de ϕ'. Pero los modelos GSCL también pueden tener desventajas frente al modelo SCL. La formulación de los parámetros de asignación a partir de valores exponenciales (3.28) puede ser innecesariamente compleja, y no permite los valores cero y uno. Esto puede limitar la eficacia del modelo de esta familia que se especifique. Además, la aparición de los parámetros desconocidos adicionales a los del modelo SCL, ϕ', no deja clara ni su naturaleza ni su formulación. Al igual que el modelo SCL, los modelos de la familia generalized spatially correlated logit son especificaciones PGNL en las que todos los pares de alternativas tienen igual parámetro de disimilitud.

3.3. Modelos SCL-b

En esta tesis proponemos una extensión del modelo SCL diferente de GSCL, que denominamos SCL-based (SCL-b), publicada en Pérez-López et al. (2020). Al igual que el modelo SCL, los modelos SCL-b son especificaciones del modelo PGNL, con la restricción de que todos los parámetros de disimilitud son iguales. A diferencia del modelo SCL, los parámetros de asignación de cada modelo SCL-b son una normalización de los valores de una métrica espacial entre pares de alternativas, f. Los parámetros de asignación de los modelos SCL-b se calculan según la fórmula (3.32), donde $f(i,j)$ es el valor de una métrica espacial f para cada par de alternativas i,j. Los valores de la métrica espacial cumplen que $f(i,j) \geq 0\ \forall i,j$ y $f(i,i) = 0, \forall i$.

$$\alpha_{i,ij} = \frac{f(i,j)}{\sum_{l=1}^{A} f(i,l)}, \forall i,j \epsilon \{1,\dots,A\} \tag{3.32}$$

La métrica espacial f que se elija debe conseguir que los valores $f(i,j)$ que se obtengan representen lo mejor posible la dependencia espacial entre ambas alternativas en el contexto empírico de aplicación. Los valores de la métrica espacial entre pares de alternativas (así como los parámetros de asignación) tendrán una relación directa con la correlación espacial que haya entre ellas. Los valores de la métrica espacial se pueden representar matricialmente y los parámetros de asignación también.

La matriz de valores de la métrica $W = \big(f(i,j)\big)$ es una matriz simétrica de dimensión igual al número de alternativas y con ceros en la diagonal. Los valores de los parámetros de asignación se calculan a partir de la matriz W, y también se pueden representar matricialmente. La matriz de parámetros de asignación $\Lambda = \big(\alpha_{i,ij}\big)$ es una matriz cuadrada con la misma dimensión que W pero no simétrica, que se calcula normalizando por filas de la matriz W, tal y como muestra la ecuación (3.32). Esta representación matricial permite establecer una analogía con la autocorrelación espacial en los modelos de regresión lineal. La matriz W tiene una función análoga a la matriz de ponderaciones o pesos espaciales del contexto de autocorrelación espacial (se habla de matriz de interacciones espaciales o de contigüidades cuando únicamente recoge la variable dicotómica de contigüidad). El estudio de la dependencia espacial en modelos de regresión lineal ha tenido un tratamiento más extenso que para los modelos de elección discreta (Fleming, 2004). Dada su analogía, a la hora de elegir la métrica espacial para un modelo SCL-b se pueden utilizar las investigaciones de las matrices de pesos espaciales del contexto de autocorrelación espacial (véase Chasco, 2003 para una revisión de las especificaciones de esta matriz). Por ejemplo, la métrica booleana de contigüidad del modelo SCL y la métrica espacial basada en el modelo de gravitación de GDSCL fueron estudiadas en el contexto de autocorrelación espacial en Anselin (1980, 1988). La variable espacial del modelo distance-based spatially correlated logit de la familia GSCL y la métrica espacial del modelo BSCL que proponemos en este capítulo fueron estudiados en el contexto de autocorrelación espacial (véase Cliff y Ord, 1981: 14-16). Para un estudio más profundo de las matrices de pesos espaciales en contexto de autocorrelacion espacial se pueden consultar Stetzer (1982) o Florax y Rey (1995).

En la mayor parte de las métricas que se elijan para el cálculo de los valores $f(i,j)$ será necesario utilizar un sistema de información geográfica, que deberá contar con cartografía de las áreas espaciales correspondientes a las alternativas, y la capacidad de cálculo espacial necesario para la métrica elegida. El modelo SCL es un caso particular

SCL-b, sin más que definir la métrica espacial dada por la expresión dicotómica de la contigüidad entre alternativas, es decir, 1 si dos alternativas tienen borde en común y 0 en caso contrario. La matriz W del SCL es booleana. Esta métrica espacial quizás sea la única que no necesita utilizar un sistema de información geográfica. De todas formas, el modelo SCL no está anidado con el resto de modelos SCL-b.

La expresión de la elasticidad directa y cruzada de los modelos SCL-b coincide con la del modelo SCL, con la salvedad de que los parámetros de asignación se calculan de forma diferente en cada caso.

La ventaja de los modelos SCL-b frente al SCL es que pueden incorporar estructuras de correlación espacial entre alternativas más complejas que este último, sin incrementar el número de parámetros estructurales desconocidos. Como desventaja frente al modelo SCL, los modelos SCL-b tienen la necesidad de elegir y calcular una métrica espacial adecuada al contexto de aplicación. La ventaja de los modelos SCL-b, frente a los modelos generalized spatially correlated logit, radica que no incrementan el número de parámetros estructurales desconocidos del modelo SCL, que tienen una definición más sencilla y totalmente cerrada de los parámetros de asignación, y que además sí permiten los valores extremos 0 y 1. La desventaja de los modelos SCL-b, frente a los modelos generalized spatially correlated logit, se encuentra en que son menos flexibles y no pueden incorporar correlación no-espacial entre alternativas, porque carecen de los parámetros ϕ'.

Al igual que el resto de modelos GEV vistos anteriormente, los modelos SCL-b pueden actuar como núcleo de un modelo MGEV para acomodar heterogeneidad en los gustos de los decisores.

3.3.1. Gravitational-distance spatially correlated logit

A la hora de elegir una métrica espacial para un modelo SCL-b, parece lógico suponer que la correlación espacial entre pares de alternativas tiene una dependencia inversa con la distancia entre ellas. Este razonamiento fue utilizado por ejemplo en el modelo distance-based spatially correlated logit descrito anteriormente. Anselin (1980) propone una matriz de pesos espaciales que utiliza la distancia del modelo de gravitación, consistente en la inversa de la distancia al cuadrado. En esta investigación proponemos utilizar esta distancia como métrica espacial para definir un modelo SCL-b (presentada inicialmente en Pérez-López y Orro, 2016). En concreto, se trataría de la métrica espacial consistente en la inversa de la distancia cuadrática entre los centroides de pares de alternativas, que se recoge en la fórmula (3.33), donde d_{ij} es la distancia entre los centroides de dos alternativas i, j. La sencillez de esta métrica permite que se puedan calcular valores aproximados de la matriz W con el uso de la escala del mapa, siempre que se hayan calculado los centroides de las zonas. De todas formas, para obtener precisión en los cálculos de W es más apropiado el uso de un sistema de información geográfica.

$$f(i,j) = {d_{ij}}^{-2}, \forall i,j \epsilon \{1, \dots, A\} \tag{3.33}$$

El modelo SCL-b que denominamos gravitational-distance spatially correlated logit (GDSCL), utiliza esta métrica espacial. Esta métrica puede ser eficiente a la hora de recoger la correlación espacial entre alternativas en contextos empíricos que cuenten

con alternativas de formas regulares. Por ejemplo, las zonificaciones basadas en mallas tienen formas regulares. Estas se pueden encontrar en ciertos entornos urbanos, especialmente fuera de las zonas históricas. La eficiencia se debe a que el centroide es un buen representante de zonas de este tipo.

Al igual que sucedía con la contigüidad del modelo SCL, la métrica gravitational-distance no parece que pueda ser eficiente en zonificaciones basadas en áreas administrativas, que en muchas ocasiones tienen formas irregulares, porque en este contexto, los centroides de las alternativas no representan bien su área. Por este motivo, el modelo gravitational-distance spatially correlated logit puede no ser apropiado para el contexto de modelización de la elección de localización residencial, si se aplica a una zonificación basada en áreas administrativas de formas irregulares, como tampoco lo es el modelo SCL.

En el ejemplo de zonificación irregular de la figura 3.1, se puede deducir que dependiendo de la forma y tamaño de las alternativas 1 y 3, el valor de la métrica espacial entre las alternativas 1-2 podría ser mayor que entre 1-3. Denominamos MGDSCL a la especificación MGEV del modelo GDSCL.

3.3.2. Common-border spatially correlated logit

A la hora de elegir un modelo SCL-b para modelizar elecciones de localización espacial, en este capítulo proponemos el modelo common border spatially correlated logit (BSCL), que utiliza como métrica espacial la longitud de frontera común entre pares de alternativas contiguas, a la que denominamos λ_{ij}, como se muestra en (3.34). Esta métrica espacial ya ha sido utilizada en el contexto de autocorrelación (véase Cliff y Ord, 1981:14-16). Se puede comprobar, sin más que utilizar la fórmula (3.32), que el parámetro de asignación de una alternativa respecto a otra en el modelo BSCL es la proporción de perímetro que comparten ambas alternativas respecto del perímetro total de la primera de las alternativas.

$$f(i,j) = \lambda_{ij}, \forall i,j \epsilon \{1, \dots, A\} \tag{3.34}$$

Parece lógico suponer que el planteamiento de esta métrica es más robusto en contexto de modelización de la elección de localización residencial que los de los modelos SCL-b anteriores. Por un lado, el modelo BSCL parece más eficiente que el modelo SCL para recoger la correlación espacial entre alternativas en este contexto empírico, ya que el modelo BSCL permite calcular un nivel de contigüidad entre alternativas frente a la especificación binaria del modelo SCL. También parece más eficiente el modelo BSCL que el GDSCL, debido a que la forma y tamaño de las alternativas influye decisivamente en la capacidad de su centroide para representarla. Denominamos MBSCL a la especificación MGEV del modelo BSCL.

3.4. Aplicación a un caso real

En este apartado se presentan y analizan los resultados de la aplicación empírica de los modelos descritos en este capítulo en el mismo caso de la ciudad de Santander empleado en el capítulo 2.

El primer paso es comprobar si es necesario un modelo con correlación espacial. En esta investigación utilizaremos el modelo SCL para hacer esta comprobación. El criterio que

se propone es considerar la existencia de correlación espacial entre alternativas, y necesidad de utilizar un modelo que la incluya, cuando el coeficiente de disimilitud del modelo SCL se muestre significativamente distinto de 0 y de 1, y el modelo SCL presente una bondad de ajuste significativamente mayor que el MNL (con el que está anidado). Si estas condiciones no se cumplen, para evitar descartar casos en los que la contigüidad entre alternativas no refleje en absoluto la dependencia espacial de las alternativas, se realiza la misma comprobación utilizando el resto de modelos SCL-b que se consideren apropiados para el contexto empírico de la aplicación.

En caso de considerar la presencia de correlación espacial entre alternativas, es necesario elegir el modelo SCL-b más apropiado para el contexto empírico de aplicación. Se escoge el modelo SCL-b que obtenga mejores resultados de bondad de ajuste y validación, de entre los que se hayan estimado y aceptado. Antes de estimar un modelo SCL-b es necesario calcular sus parámetros de asignación mediante la métrica espacial que lo define. Para ello, primero se calculan los valores de la métrica espacial para cada par de alternativas de la zonificación, normalmente con ayuda de un sistema de información geográfica (GIS). En esta investigación utilizamos QGIS (2018), que es un sistema de información geográfica de software libre y código abierto. A partir de ellos se calculan los parámetros de asignación de cada par de alternativas de la zonificación con la fórmula (3.32).

El objetivo principal de esta aplicación es comprobar empíricamente si el modelo BSCL se muestra más apropiado para este contexto que el resto de modelos logit con correlación espacial presentados en esta investigación. La estimación y comparación de los modelos se realizará según la metodología descrita en el apartado 2.3. En el resto del capítulo se describe una aplicación empírica de los modelos analizados en esta investigación a un contexto de modelización de la elección de localización residencial urbana con zonificación basada en áreas administrativas.

3.4.1. Modelo spatially correlated logit

Para estimar el modelo SCL es necesario calcular previamente los parámetros de asignación correspondientes al contexto empírico de aplicación. Estos parámetros se calculan mediante la fórmula (3.24), a partir de la variable dicotómica que indica la contigüidad o no entre las áreas geográficas de los pares de alternativas. Los valores de la variable contigüidad se han calculado a partir de la figura 2.4, sin necesidad de la ayuda de un GIS, aunque puede haber ocasiones en las que sea necesario contar con la ayuda de uno para estar seguros. Por ejemplo, entre las zonas 21 y 10 puede ser necesario confirmar si ambas zonas tienen frontera en común, aunque sea pequeña. En este caso se ha confirmado la contigüidad entre estas dos zonas con ayuda del sistema de información geográfica QGIS (2018). La tabla 3.1 muestra la representación matricial de los valores obtenidos de la variable de la variable de contigüidad entre zonas (W).

W	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1		1									1	1	1			1	1		1			1				
2	1		1	1							1															
3		1		1	1																					
4		1	1		1	1					1															
5			1	1		1	1	1	1																	
6				1	1		1				1	1														
7					1	1			1			1														
8					1				1	1																
9							1	1		1		1	1	1	1											
10								1	1					1							1		1	1	1	
11	1	1		1		1						1														
12	1					1	1		1		1		1													
13	1								1			1		1	1	1										
14									1	1			1		1						1					
15									1				1	1		1	1	1			1					
16	1												1		1		1	1								
17	1															1		1	1	1		1				
18															1	1	1				1	1				
19	1																1			1		1				
20																	1		1			1				
21										1				1	1			1				1	1	1		1
22	1																1	1	1	1	1					1
23										1											1			1		
24										1											1		1		1	
25										1											1			1		1
26																					1	1			1	

Tabla 3.1. Matriz de contigüidades en la zonificación de la ciudad de Santander.

Una situación semejante a la del ejemplo teórico expuesto en el apartado 3.2 la podemos observar en esta zonificación de la ciudad de Santander. La Figura 3.2 es una ampliación de la figura 2.4 en torno a la zona 21. En la imagen puede observarse que la variable dicotómica de contigüidad tiene el mismo valor para el par de zonas 21 y 10, que para los pares de zonas 21 y 23 o 21 y 18. De esta forma, un modelo SCL especificado con esta variable asigna la misma correlación espacial entre las alternativas 21 y 10 que entre las alternativas 21 y 23 o entre las alternativas 21 y 18. Sin embargo, atendiendo a la configuración espacial de estas alternativas, no parece razonable pensar que esta estructura de correlación espacial represente de forma eficaz la dependencia espacial entre esas alternativas.

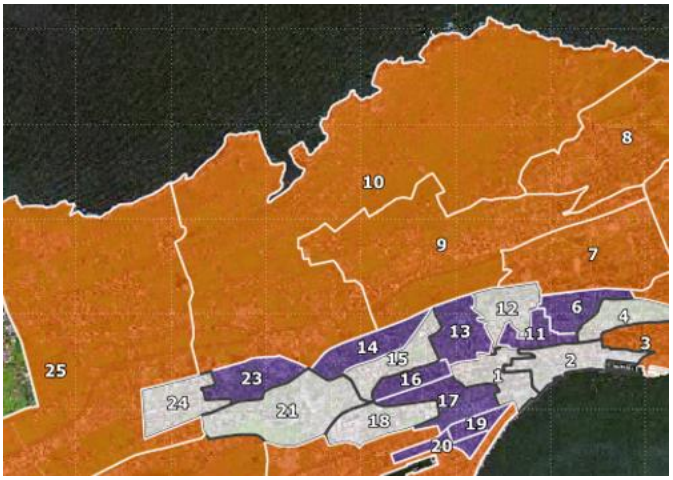

Figura 3.2. Ampliación de la de zonificación de la Figura 2.4 en torno a la alternativa 21.

En el caso del modelo SCL, se ha repetido el proceso stepwise de selección de regresores de la función de utilidad observada, descrito en la metodología del capítulo anterior. La tabla 3.2 muestra los resultados de estimación y bondad de ajuste de la primera iteración del proceso, que denominamos SCL-0. Esta especificación no es válida, porque presenta once regresores no relevantes.

SCL-0						
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**
Estimación	β_{AC}	0,0132	0,0139	0,96	0,34	·
	$\beta_{AC \cdot H}$	-0,021	0,0249	-0,84	0,40	·
	β_{JT}	-0,156	0,0656	-2,39	0,02	*
	$\beta_{JT \cdot H}$	0,000861	0,102	0,01	0,99	·
	β_{FO}	-0,931	0,541	-1,72	0,09	·
	$\beta_{FO \cdot H}$	-2,34	1,36	-1,72	0,09	·
	β_{IN}	0,259	0,28	0,92	0,36	·
	β_{HO}	1,61	0,453	3,56	0,00	**
	$\beta_{HO \cdot H}$	2,05	1,1	1,87	0,06	·
	β_{PS}	-1,06	0,38	-2,79	0,01	**
	$\beta_{PS \cdot H}$	2,39	0,762	3,13	0,00	**
	β_{PR}	-1,36	0,806	-1,68	0,09	·
	$\beta_{PR \cdot H}$	-1,22	1,48	-0,82	0,41	·
	β_{SC}	-0,12	0,0601	-1,99	0,05	*
	$\beta_{SC \cdot H}$	0,311	0,115	2,7	0,01	**
	β_{WT}	-0,186	0,107	-1,74	0,08	·
	$\beta_{WT \cdot H}$	-0,00552	0,226	-0,02	0,98	·
	μ^{-1}	1,53	0,155	4,21	0,00	**
		Nº par. est.		18		
Bondad de ajuste		LL		-1657,350		
		FG		0,0449		
		ρ^2		0,0474		
		$\bar{\rho}_H^2$		0,0422		
		AIC		0,0371		
		$LRT(Nulo)$		164,95	0,00	**

Tabla 3.2. Resultados de estimación y bondad de ajuste de la primera iteración del proceso stepwise SCL.

El proceso stepwise de selección de regresores de la función de utilidad observada da como resultado la misma especificación que se obtuvo en el capítulo anterior, con los modelos multinomial logit y nested logit. La tabla 3.3 muestra los resultados de estimación, bondad de ajuste y validación del modelo spatially correlated logit con esta especificación de la función de utilidad, que denominamos SCL. El modelo SCL es válido,

porque todos los parámetros estimados se muestran significativos, y con los signos esperados teóricamente. Respecto a la endogeneidad, y al igual que se explicó en los modelos estimados anteriormente, las alternativas no son viviendas específicas sino áreas, y la variable de la función de utilidad es el precio medio de las viviendas en el área. En esa situación, no se espera endogeneidad debido a atributos omitidos de una vivienda específica que están correlacionados con su precio. Los resultados no muestran indicios de endogeneidad, porque el precio tiene el signo esperado teóricamente, aunque no se puede descartar y es recomendable analizar detenidamente este tema en modelos estimados para el análisis de políticas.

SCL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**RI**
Estimación	β_{JT}	-0,174	0,0517	-3,37	0,00	**	-0,684	31%
	β_{FO}	-1,03	0,417	-2,47	0,00	**	-0,231	10%
	β_{HO}	1,93	0,398	4,85	0,00	**	0,439	20%
	β_{PR}	-2,50	0,561	-4,46	0,00	**	-0,317	14%
	$\beta_{PS \cdot H}$	1,56	0,382	4,08	0,00	**	0,135	6%
	$\beta_{SC \cdot H}$	0,274	0,0716	3,83	0,00	**	0,411	19%
	μ^{-1}	1,55	0,127	5,08	0,00	**		
		Nº par. estimados		7				
Bondad de ajuste		LL		-1665,940				
		FG		0,0442				
		ρ^2		0,0425				
		$\bar{\rho}_H^2$		0,0405				
		AIC		0,0384				
		$LRT(Nulo)$		147,77	0,00	**		
		$LRT(MNL)$		4,06	0,04	*		
Val.		*PG-CV-4*		0,04353				
		PG-CV-10		0,04349				

Tabla 3.3. Resultados de estimación, bondad de ajuste y validación del modelo spatially correlated logit: SCL.

El modelo SCL presenta mayor capacidad explicativa y predictiva que el modelo MNL estimado en el capítulo anterior bajo las mismas condiciones. Por un lado, el modelo SCL tiene una mayor bondad de ajuste, pues el contraste de la razón de verosimilitudes es significativo respecto al modelo MNL, con el que está anidado. Por otro lado, el modelo SCL supera ambos indicadores de validación cruzada del modelo MNL. Esto pone de manifiesto que las alternativas de elección del contexto empírico tienen correlación espacial, y que la métrica de contigüidad del modelo SCL es capaz de capturar una parte significativa de ella. Dado que la especificación restricted nested logit no mejora la capacidad explicativa y predictiva del modelo MNL, el hecho de mejorar el modelo MNL ya dará por hecho que también lo hace con RNL.

Pero el modelo NL estimado en el capítulo anterior mejora la capacidad explicativa y predictiva de SCL. Tanto los dos indicadores de bondad de ajuste que penalizan el número de parámetros, $\bar{\rho}_H^2$ y AIC, como los indicadores de validación cruzada de la precisión en predicción, *PG-CV-4* y *PG-CV-10,* son superiores en NL a los de SCL. Esto pone de manifiesto que la estructura de nidos diseñada por el analista para la aplicación es capaz de incorporar mejor la correlación entre alternativas, presente en el contexto empírico, que la correlación espacial entre alternativas que recoge la métrica de SCL.

Mixed spatially correlated logit

Se ha repetido el proceso forward de selección de coeficientes aleatorios, en este caso con el modelo SCL, y se ha obtenido la misma estructura mixta de coeficientes aleatorios que en el caso MMNL y M-NL. Esta estructura mixta, que denominamos MSCL, tiene todos los coeficientes fijos, salvo el correspondiente a la interacción SC·H, que se supone aleatoria según una distribución normal de parámetros desconocidos. Los resultados no muestran indicios de endogeneidad. La tabla 3.4 muestra los resultados de estimación, bondad de ajuste y validación de la especificación mixta de coeficientes aleatorios y núcleo SCL.

	MSCL							
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,182	0,0535	-3,41	0,00	**	-0,716	39%
	β_{FO}	-1,07	0,424	-2,53	0,00	**	-0,240	13%
	β_{HO}	1,94	0,402	4,83	0,00	**	0,442	24%
	β_{PR}	-2,43	0,567	-4,29	0,00	**	-0,308	17%
	$\beta_{PS\cdot H}$	1,41	0,403	3,51	0,00	**	0,122	7%
	$E(\beta_{SC\cdot H})$	0,206	0,101	2,06	0,04	*		
	$\sigma(\beta_{SC\cdot H})$	0,421	0,193	2,18	0,03	*		
	μ^{-1}	1,58	0,0885	11,88	0,00	**		
		Nº par. est.		8				
Bondad de ajuste		SLL		-1664,796				
		FG		0,0442				
		ρ^2		0,0430				
		$\bar{\rho}_H^2$		0,0408				
		AIC		0,0385				
		$LRT(Nulo)$		150,06	0,00	**		
		$LRT(MNL)$		6,34	0,04	*		
Val.		*PG-CV-4*		0,04355				
		PG-CV-10		0,04352				

Tabla 3.4. Resultados de estimación, bondad de ajuste y validación del modelo mixed spatially correlated logit: MSCL

La especificación MSCL mejora la capacidad explicativa y predictiva de su núcleo GEV, que implica que la incorporación de variaciones en los gustos de los individuos decisores en uno de los regresores compensa el incremento de un parámetro desconocido. Por un lado, los dos indicadores de bondad de ajuste que penalizan el número de parámetros desconocidos, $\bar{\rho}_H^2$ y AIC, son mayores en MSCL que en su núcleo SCL. Por otro lado, los dos indicadores de validación cruzada son mayores en MSCL que en SCL. El *PG-CV-4* de MSCL es 0,04355 frente a 0,04353 en SCL. El *PG-CV-10* de MSCL es 0,04352 frente a 0,04349 en SCL.

La especificación mixta con núcleo SCL mejora la capacidad explicativa y predictiva de la que tiene núcleo MNL (por tanto, también la que tiene núcleo RNL), con la que está anidada; al igual que sucedía con los respectivos núcleos GEV. Por un lado, mejora la bondad de ajuste, porque el contraste de la razón de verosimilitudes respecto a MMNL es significativo. Además, los dos indicadores de validación cruzada son mayores en MSCL que en MMNL. El *PG-CV-4* de MSCL es 0,04355 frente a 0,04352 en MMNL. El *PG-CV-10* de MSCL es 0,04352 frente a 0,04343 en MMNL.

Sin embargo, la especificación mixta M-NL mejora la capacidad explicativa y predictiva de MSCL, de nuevo, de igual manera que sucede con los respectivos núcleos GEV. Tanto los indicadores de bondad de ajuste como los de validación cruzada de la precisión en predicción son mejores en el modelo M-NL que en el MSCL.

3.4.2. Gravitational-distance spatially correlated logit

La métrica espacial del modelo gravitational-distance spatially correlated logit se basa en la distancia entre los centroides de los pares de alternativas. La figura 3.3 muestra, para el caso de aplicación de Santander, el centroide de cada área geográfica de la zonificación de alternativas descrita en el capítulo anterior.

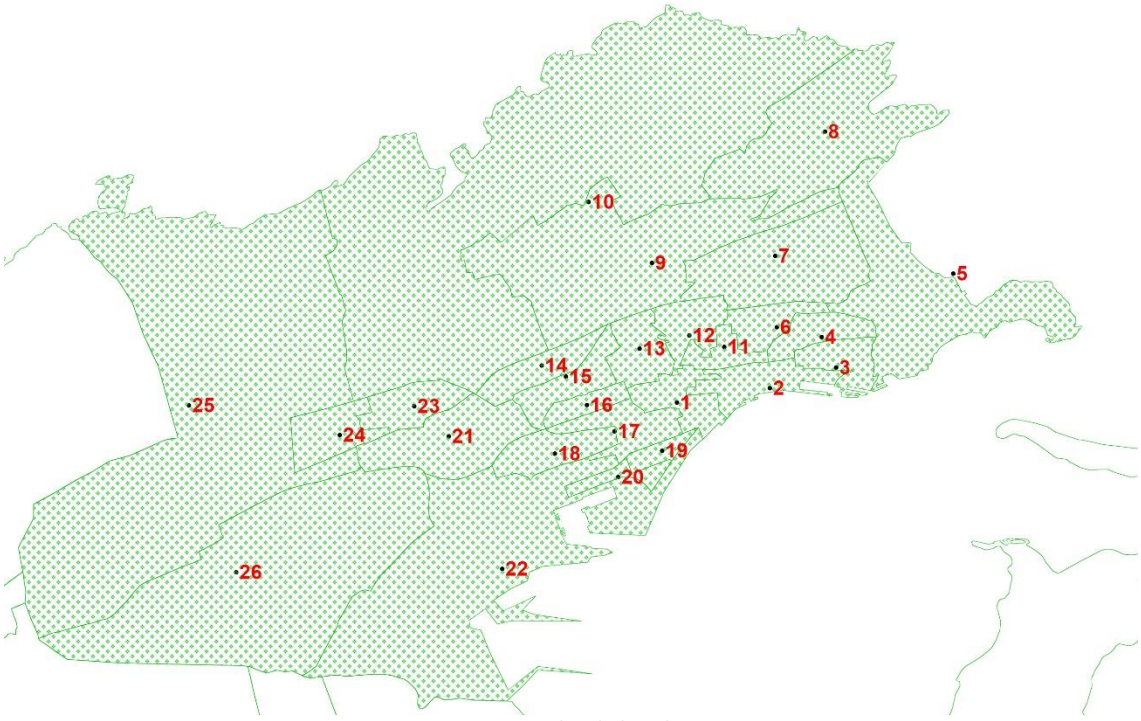

Figura 3.3. Centroides de las alternativas.

En esta imagen se intuye que la distancia euclídea entre centroides de pares de alternativas puede no recoger la correlación espacial entre ellas, debido a las formas irregulares de muchas de las alternativas. Para comprobarlo, se pueden observar los valores de la métrica espacial gravitational-distance para cada par de alternativas de la zonificación, calculados con ayuda del sistema de información geográfica QGIS (2018), que se muestran matricialmente (W) en la tabla 3.5. Por ejemplo, la zona 25 es contigua a la zona 10, sin embargo, la distancia euclídea a su centroide es mayor que muchas otras zonas con las que no tiene frontera común, como las zonas 14 a 18; e incluso mayor que con la zona 20, a pesar de que entre ellas se interponen las zonas 24, 21, 18 y 22. Puede resaltarse también que hay centroides que están ubicados fuera de sus zonas, como en la zona 10.

W	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
01		0,90	1,56	1,52	2,92	1,20	1,69	2,96	1,36	2,10	0,70	0,65	0,63	1,35	1,09	0,86	0,66	1,27	0,48	0,91	2,21	2,31	2,52	3,25	4,68	4,53
02	0,90		0,67	0,70	2,07	0,59	1,27	2,51	1,65	2,49	0,59	0,93	1,31	2,20	1,96	1,76	1,55	2,16	1,19	1,69	3,11	3,10	3,42	4,15	5,58	5,42
03	1,56	0,67		0,33	1,44	0,69	1,22	2,26	2,03	2,86	1,09	1,44	1,89	2,82	2,59	2,42	2,21	2,82	1,85	2,34	3,77	3,74	4,07	4,80	6,22	6,08
04	1,52	0,70	0,33		1,40	0,44	0,89	1,97	1,78	2,58	0,94	1,27	1,75	2,70	2,48	2,34	2,18	2,79	1,88	2,37	3,70	3,79	3,96	4,72	6,11	6,05
05	2,92	2,07	1,44	1,40		1,77	1,72	1,83	2,89	3,56	2,31	2,60	3,09	4,04	3,84	3,73	3,58	4,19	3,27	3,76	5,08	5,17	5,32	6,08	7,44	7,45
06	1,20	0,59	0,69	0,44	1,77		0,68	1,93	1,35	2,17	0,54	0,84	1,33	2,28	2,08	1,97	1,85	2,45	1,61	2,09	3,31	3,51	3,56	4,32	5,69	5,69
07	1,69	1,27	1,22	0,89	1,72	0,68		1,28	1,18	1,86	1,00	1,12	1,57	2,47	2,31	2,30	2,28	2,84	2,16	2,60	3,57	3,98	3,75	4,51	5,80	5,99
08	2,96	2,51	2,26	1,97	1,83	1,93	1,28		2,08	2,37	2,28	2,35	2,74	3,52	3,42	3,47	3,51	4,03	3,43	3,86	4,64	5,21	4,74	5,49	6,64	7,05
09	1,36	1,65	2,03	1,78	2,89	1,35	1,18	2,08		0,84	1,06	0,78	0,83	1,44	1,37	1,50	1,65	2,05	1,80	2,08	2,56	3,26	2,66	3,42	4,65	4,97
10	2,10	2,49	2,86	2,58	3,56	2,17	1,86	2,37	0,84		1,90	1,60	1,49	1,63	1,69	1,94	2,21	2,43	2,48	2,65	2,61	3,61	2,57	3,26	4,30	4,90
11	0,70	0,59	1,09	0,94	2,31	0,54	1,00	2,28	1,06	1,90		0,35	0,81	1,76	1,54	1,43	1,33	1,92	1,16	1,61	2,78	3,01	3,03	3,78	5,17	5,16
12	0,65	0,93	1,44	1,27	2,60	0,84	1,12	2,35	0,78	1,60	0,35		0,49	1,44	1,24	1,19	1,17	1,71	1,13	1,52	2,50	2,87	2,72	3,48	4,85	4,90
13	0,63	1,31	1,89	1,75	3,09	1,33	1,57	2,74	0,83	1,49	0,81	0,49		0,95	0,75	0,74	0,83	1,29	1,00	1,25	2,01	2,49	2,23	2,99	4,36	4,42
14	1,35	2,20	2,82	2,70	4,04	2,28	2,47	3,52	1,44	1,63	1,76	1,44	0,95		0,26	0,58	0,94	0,85	1,41	1,30	1,12	1,99	1,29	2,05	3,41	3,54
15	1,09	1,96	2,59	2,48	3,84	2,08	2,31	3,42	1,37	1,69	1,54	1,24	0,75	0,26		0,34	0,70	0,74	1,16	1,08	1,26	1,94	1,48	2,24	3,63	3,68
16	0,86	1,76	2,42	2,34	3,73	1,97	2,30	3,47	1,50	1,94	1,43	1,19	0,74	0,58	0,34		0,37	0,56	0,84	0,75	1,36	1,77	1,66	2,39	3,82	3,73
17	0,66	1,55	2,21	2,18	3,58	1,85	2,28	3,51	1,65	2,21	1,33	1,17	0,83	0,94	0,70	0,37		0,61	0,49	0,44	1,59	1,70	1,94	2,63	4,09	3,87
18	1,27	2,16	2,82	2,79	4,19	2,45	2,84	4,03	2,05	2,43	1,92	1,71	1,29	0,85	0,74	0,56	0,61		1,03	0,65	1,03	1,22	1,42	2,07	3,54	3,26
19	0,48	1,19	1,85	1,88	3,27	1,61	2,16	3,43	1,80	2,48	1,16	1,13	1,00	1,41	1,16	0,84	0,49	1,03		0,49	2,05	1,91	2,42	3,10	4,56	4,25
20	0,91	1,69	2,34	2,37	3,76	2,09	2,60	3,86	2,08	2,65	1,61	1,52	1,25	1,30	1,08	0,75	0,44	0,65	0,49		1,67	1,42	2,07	2,70	4,18	3,78
21	2,21	3,11	3,77	3,70	5,08	3,31	3,57	4,64	2,56	2,61	2,78	2,50	2,01	1,12	1,26	1,36	1,59	1,03	2,05	1,67		1,37	0,44	1,05	2,51	2,42
22	2,31	3,10	3,74	3,79	5,17	3,51	3,98	5,21	3,26	3,61	3,01	2,87	2,49	1,99	1,94	1,77	1,70	1,22	1,91	1,42	1,37		1,77	2,02	3,39	2,55
23	2,52	3,42	4,07	3,96	5,32	3,56	3,75	4,74	2,66	2,57	3,03	2,72	2,23	1,29	1,48	1,66	1,94	1,42	2,42	2,07	0,44	1,77		0,76	2,16	2,34
24	3,25	4,15	4,80	4,72	6,08	4,32	4,51	5,49	3,42	3,26	3,78	3,48	2,99	2,05	2,24	2,39	2,63	2,07	3,10	2,70	1,05	2,02	0,76		1,48	1,65
25	4,68	5,58	6,22	6,11	7,44	5,69	5,80	6,64	4,65	4,30	5,17	4,85	4,36	3,41	3,63	3,82	4,09	3,54	4,56	4,18	2,51	3,39	2,16	1,48		1,66
26	4,53	5,42	6,08	6,05	7,45	5,69	5,99	7,05	4,97	4,90	5,16	4,90	4,42	3,54	3,68	3,73	3,87	3,26	4,25	3,78	2,42	2,55	2,34	1,65	1,66	

Tabla 3.5. Matriz de valores de la métrica espacial gravitational-distance de la zonificación de la ciudad de Santander.

El modelo gravitational-distance spatially correlated logit se ha aplicado especificado con la misma función de utilidad observada que el resto de modelos GEV. La tabla 3.6 muestra los resultados de estimación, bondad de ajuste y validación. La verosimilitud que se obtiene es la misma que con el modelo MNL, a pesar de contar con un parámetro estructural adicional, el parámetro de disimilitud. Esto se debe a que el valor estimado de este parámetro es igual a uno, por lo que el modelo GDSCL estimado colapsa con el modelo MNL. Esto no se debe a la falta de correlación espacial entre alternativas, cuya existencia ya se comprobó con el modelo SCL, sino a la incapacidad de esta métrica espacial para recoger dicha correlación en este contexto empírico concreto. Esto confirma empíricamente, al menos en el contexto de la aplicación, la hipótesis planteada anteriormente, relativa a que la métrica espacial gravitational-distance no es eficiente con zonificaciones que contengan alternativas de formas irregulares. Al colapsar con el modelo MNL, no es necesario que realicemos las comparaciones del resto de modelos de esta tesis con el modelo GDSCL.

GDSCL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,114	0,0291	-3,92	0,00	**	-0,448	30%
	β_{FO}	-0,870	0,307	-2,84	0,00	**	-0,195	11%
	β_{HO}	1,49	0,265	5,63	0,00	**	0,339	20%
	β_{PR}	-2,16	0,429	-5,03	0,00	**	-0,274	14%
	$\beta_{PS \cdot H}$	1,25	0,272	4,61	0,00	**	0,108	6%
	$\beta_{SC \cdot H}$	0,224	0,0516	4,33	0,00	**	0,336	19%
	μ^{-1}	1	0,000	1,73E+08	0,00	**		
		Nº pár. est.		7				
Bondad de ajuste		LL		-1667,968				
		FG		0,0440				
		ρ^2		0,0413				
		$\bar{\rho}_H^2$		0,0393				
		AIC		0,0373				
		$LRT(Nulo)$		143,71	0,00	**		
		$LRT(MNL)$		0	1,00	·		
Val.		*PG-CV-4*		0,0435				
		PG-CV-10		0,0434				

Tabla 3.6. Resultados de estimación, bondad de ajuste y validación del modelo gravitational-distance spatially correlated logit: GDSCL.

Mixed gravitational-distance spatially correlated logit

La tabla 3.7 muestra los resultados de la especificación mixta del modelo GDSCL con la misma estructura de coeficientes aleatorios diseñada con el modelo SCL, que denominamos MGDSCL. Este modelo tampoco mejora los resultados de bondad de ajuste del modelo MMNL, y colapsa en él, pues el parámetro de disimilitud no se muestra significativo. Por este motivo tampoco realizamos las comparaciones del resto de modelos de esta tesis con el modelo MGDSCL.

MGDSCL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,118	0,0293	-4,03	0,00	**	-0,464	34%
	β_{FO}	-0,915	0,309	-2,96	0,00	**	-0,205	15%
	β_{HO}	1,49	0,266	5,62	0,00	**	0,339	25%
	β_{PR}	-2,12	0,433	-4,89	0,00	**	-0,269	20%
	$\beta_{PS \cdot H}$	1,12	0,287	3,90	0,00	**	0,097	7%
	$E(\beta_{SC \cdot H})$	0,159	0,0765	2,08	0,04	*		
	$\sigma(\beta_{SC \cdot H})$	0,291	0,126	-2,31	0,02	*		
	μ^{-1}	1,00	1.80e+308	0,00	1,00	·		
		Nº pár. est.		8				
Bondad de ajuste		SLL		-1666,905				
		FG		0,0441				
		ρ^2		0,0419				
		$\bar{\rho}_H^2$		0,0396				
		AIC		0,0373				
		$LRT(Nulo)$		145,84	0,00	**		
		$LRT(MNL)$		2,13	0,35	·		
Val.		*PG-CV-4*		0,0435				
		PG-CV-10		0,0434				

Tabla 3.7. Resultados de estimación, bondad de ajuste y validación del modelo mixed gravitational distance spatially correlated logit: MGDSCL

3.4.3. Common-border spatially correlated logit

La métrica espacial common-border se basa en la longitud de la frontera común entre pares de alternativas. Esta medida mantiene el enfoque del modelo SCL de considerar incorrelación espacial entre zonas no contiguas. En el caso de que dos zonas sean contiguas, esta métrica permite establecer un grado de contigüidad. Para ello, se utiliza la proporción de longitud de frontera común entre zonas contiguas, que se incorpora al modelo como parámetro de asignación, para modelizar la correlación espacial entre pares de alternativas. En el caso de zonificaciones con alternativas que tengan diferentes tamaños y formas irregulares, esta métrica representa más eficientemente la correlación espacial entre alternativas contiguas. La tabla 3.8 muestra matricialmente los valores de esta métrica obtenidos con la ayuda del sistema de información geográfica QGIS (2018).

Por ejemplo, expusimos en el sub-apartado 3.4.1 que el modelo SCL asigna la misma correlación espacial al par de zonas 21-10 que a los pares 21-23 o 21-18, que no parece coherente con la posible dependencia espacial que se intuye a la vista de la figura 3.2. Sin embargo, la métrica common-border asigna valores que sí parecen razonbles, pues asigna una valor 23 veces mayor al par 21-18 que al par 21-10, y 37 veces superior en el caso del par 21-23.

W	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
01		1298									288	215	493			158	840		162			123				
02	1298		930	245							649															
03		930		870	623																					
04		245	870		585	795					104															
05			623	585		100	1343	609	158																	
06				795	100		1029				702	206														
07					1343	1029			2228			585														
08					609				3063	2471																
09					158		2228	3063		4939		246	454	642												
10								2471	4939					899							26		1081	136	2188	
11	288	649		104		702						1075														
12	215					206	585		246		1075		962													
13	493								454			962			693	240										
14									642	899					1245						400					
15													693	1245		950					444					
16	158												240		950		888									
17	840															888		1161	573	118		466				
18																	1161				605	1101				
19	162																573			554		390				
20																	118		554			1930				
21										26				400	444			605				711	954	364	172	657
22	123																466	1101	390	1930	711					2545
23										1081											954			800		
24										136											364		800		1810	
25										2188											172			1810		3822
26																					657	2545			3822	

Tabla 3.8. Matriz de valores de la métrica espacial common-border de la zonificación de la ciudad de Santander.

La tabla 3.9 muestra los resultados de la aplicación del modelo common-border spatially correlated logit, que llamamos BSCL, en el contexto empírico de la ciudad de Santander. Se ha utilizado la misma función de utilidad observada que en el resto de modelos GEV estimados. El modelo es válido, porque todos los parámetros estimados de este modelo se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad. Los coeficientes tipificados de este modelo indican la misma ordenación de influencia de los regresores que en los modelos estimados anteriormente.

El modelo BSCL mejora la capacidad explicativa y predictiva del modelo SCL, y por tanto también las de los modelos MNL y RNL. Por un lado, el modelo BSCL mejora la bondad de ajuste del modelo SCL. Con el mismo número de parámetros, se obtiene mayor verosimilitud. Por tanto, todos los indicadores de bondad de ajuste son mayores en BSCL que en SCL. Además, los dos indicadores de validación cruzada son también mayores en el modelo BSCL que en el SCL. Esto demuestra que, en este contexto empírico, la mayor flexibilidad del modelo BSCL es capaz de representar mejor la correlación espacial entre alternativas, y sin necesidad de parámetros desconocidos adicionales.

La comparación entre los modelos BSCL y NL no es concluyente en este contexto empírico. Por un lado, el modelo NL presenta mayor capacidad predictiva que el modelo BSCL, pues los dos indicadores de validación cruzada del modelo NL son superiores a los del modelo BSCL. Sin embargo, respecto a la bondad de ajuste, el resultado no es concluyente. Ambos modelos no están anidados. Los dos modelos sí están anidados con el modelo MNL, y ambos mejoran significativamente la bondad de ajuste del modelo MNL mediante el contraste de la razón de verosimilitudes. Por tanto, es necesario comparar los indicadores de bondad de ajuste que penalizan el número de parámetros desconocidos. El modelo NL obtiene un mayor valor en $\bar{\rho}_H^2$ que BSCL. Sin embargo, el valor del AIC del modelo NL es 0,03998, que es ligeramente inferior al 0,04003 del modelo BSCL. Esto es posible porque el indicador AIC penaliza más que $\bar{\rho}_H^2$ el número de parámetros desconocidos del modelo que se estima.

BSCL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,18	0,0518	-3,48	0,00	**	-0,708	30%
	β_{FO}	-1,12	0,447	-2,50	0,01	**	-0,251	11%
	β_{HO}	2,05	0,417	4,90	0,00	**	0,467	20%
	β_{PR}	-2,63	0,596	-4,42	0,00	**	-0,333	14%
	$\beta_{PS \cdot H}$	1,58	0,396	3,98	0,00	**	0,136	6%
	$\beta_{SC \cdot H}$	0,302	0,0752	4,01	0,00	**	0,453	19%
	μ^{-1}	1,74	0,0979	5,88	0,00	**		
		Nº pár. est.		7				
Bondad de ajuste		LL		-1,663,183				
		FG		0,0444				
		ρ^2		0,0441				
		$\bar{\rho}_H^2$		0,0420				
		AIC		0,04003				
		$LRT(Nulo)$		153,28	0,00	**		
		$LRT(MNL)$		9,57	0,00	**		
Val.		*PG-CV-4*		0,04384				
		PG-CV-10		0,04369				

Tabla 3.9. Resultados de estimación, bondad de ajuste y validación del modelo common-border spatially correlated logit: BSCL

Mixed common-border spatially correlated logit

Hemos llamado MBSCL a la aplicación de la especificación mixta con coeficientes aleatorios y núcleo BSCL. La estructura de coeficientes aleatorios es la misma que en el resto de modelos mixtos estimados en esta tesis. En la tabla 3.10 se muestran los resultados de la aplicación del modelo MBSCL al contexto empírico de la ciudad de Santander. El modelo es válido, porque todos los parámetros estimados se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad.

El modelo MBSCL mejora la capacidad explicativa y predictiva de su núcleo GEV. Este resultado confirma que, en este modelo, la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. Los dos indicadores de bondad de ajuste que penalizan el número de parámetros desconocidos, y AIC, son mayores en el modelo MBSCL que en BSCL. Lo mismo sucede con los dos indicadores de validación, el valor de *PG-CV-4* en MBSCL es 0,04389 frente a 0,04384 en BSCL y el valor de *PG-CV-10* en MBSCL es 0,04375 frente a 0,04370 en BSCL.

El modelo MBSCL mejora la capacidad explicativa y predictiva del modelo MSCL y, por tanto, también de los modelos MNL y RNL. MBSCL y MSCL estiman los mismos parámetros, pero MBSCL obtiene una mayor verosimilitud y, por tanto, también mejora

la bondad de ajuste de MSCL. Además, los dos indicadores de validación cruzada del modelo MBSCL son mayores que los del modelo BSCL.

A la hora de comparar los modelos MBSCL y M-NL, se obtiene que ninguno es concluyentemente mejor con todos los indicadores, al igual que con sus núcleos GEV. En este caso, ninguno presenta mayor capacidad explicativa ni predictiva que el otro. En ambos casos, un indicador es superior en un modelo y el otro indicador es superior en el otro modelo. Respecto a los indicadores de bondad de ajuste, el $\bar{\rho}_H^2$ del modelo M-NL es superior al del modelo MBSCL, pero con el AIC sucede lo contrario. Respecto a los indicadores de la validación cruzada sucede algo semejante. El *PG-CV-4* del modelo MBSCL es ligeramente superior al del modelo NL (0,043889 frente a 0,043886), pero el *PG-CV-10* del modelo MBSCL es inferior al del NL.

MBSCL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,192	0,0546	-3,52	0,00	**	-0,755	39%
	β_{FO}	-1,16	0,459	-2,54	0,01	**	-0,260	13%
	β_{HO}	2,06	0,425	4,85	0,00	**	0,469	24%
	β_{PR}	-2,57	0,608	3,42	0,00	**	-0,326	17%
	$\beta_{PS\cdot H}$	1,47	0,428	-4,23	0,00	**	0,127	7%
	$E(\beta_{SC\cdot H})$	0,228	0,109	2,08	0,04	*		
	$\sigma(\beta_{SC\cdot H})$	0,480	0,209	2,30	0,02	*		
	μ^{-1}	1,80	0,0975	5,70	0,00	**		
		Nº pár. est.		8				
Bondad de ajuste		SLL		-1.661,816				
		FG		0,0445				
		ρ^2		0,0448				
		$\bar{\rho}_H^2$		0,0425				
		AIC		0,0402				
		$LRT(Nulo)$		156,02	0,00	**		
		$LRT(MNL)$		12,33	0,00	**		
Val.		*PG-CV-4*		0,043889				
		PG-CV-10		0,043746				

Tabla 3.10. Resultados de estimación, bondad de ajuste y validación del modelo mixed common-border spatially correlated logit: MBSCL.

3.5. Resumen y conclusiones

El enfoque para incorporar correlación entre alternativas espaciales en modelos de elección discreta que se describe en este capítulo se basa en generalizaciones del modelo spatially correlated logit (SCL). En este capítulo se propone la generalización SCL-b, que utiliza un único parámetro estructural adicional al modelo MNL. Los modelos

SCL-b utilizan una métrica espacial entre alternativas para recoger la correlación espacial. La métrica se debe elegir en función de las características espaciales del contexto empírico de aplicación. En este capítulo se propone la métrica espacial common-border, basada en la proporción de frontera común a cada par de alternativas, para las zonificaciones basadas en áreas administrativas que tengan formas irregulares.

Para situar el modelo SCL en la literatura de los modelos de elección discreta, en este capítulo se describe la familia de modelos logit generalized extreme value (GEV). Dentro de la familia de modelos GEV, se estudian especialmente generalizaciones del modelo nested logit. Las generalizaciones cross-nested logit o generalized nested logit aumentan la flexibilidad del modelo nested logit, al permitir que las alternativas de elección pertenezcan a más de un nido. Para ello, incorporan los denominados parámetros de asignación, que ponderan la pertenencia de cada alternativa a cada nido. Debido al elevado número de alternativas habitual en los modelos de elección espacial, estas generalizaciones del modelo nested logit no son compatibles con modelización de elección espacial, salvo que se pueda reducir el número de parámetros estructurales desconocidos. El modelo paired combinatorial logit considera un nido por cada par de alternativas, aunque sin utilizar parámetros de asignación. Este planteamiento le permite evitar la necesidad de diseñar la estructura de nidos, pero también requiere de un número muy alto de parámetros estructurales. Por tanto, este modelo tampoco es compatible con modelización de elección espacial, salvo que se pueda reducir el número de parámetros estructurales desconocidos. El modelo paired generalized nested logit combina este último modelo con generalized nested logit, lo que aumenta la flexibilidad de ambos y evita la necesidad de diseñar la estructura de nidos de alternativas. Pero es el modelo con más parámetros estructurales desconocidos. Por este motivo, es inviable en modelización de elección espacial si no se reduce drásticamente su número mediante restricciones y/o se calculan previamente mediante algún procedimiento.

En este capítulo exponemos que el modelo spatially correlated logit es una especificación del modelo PGNL, en la que se toman dos caminos para reducir el número de parámetros estructurales desconocidos a solo uno. Por un lado, se incluye la restricción de que todos los nidos de pares de alternativas tengan el mismo parámetro de disimilitud. Esta restricción reduce el número de parámetros de disimilitud a uno solo. Por otro lado, se calculan los parámetros de asignación antes de la estimación del modelo, utilizando la información espacial dicotómica de si el par de alternativas son contiguas o no. Este cálculo permite que no haya parámetros de asignación desconocidos. El modelo SCL y los modelos SCL-b son modelos GEV y, por tanto, compatibles con especificaciones mixtas con coeficientes aleatorios. Estas especificaciones permiten incorporar en el modelo variaciones en los gustos de los individuos decisores.

Por último, en este capítulo se aplican, al mismo caso real que en el capítulo anterior, el modelo SCL y las especificaciones SCL-b que utilizan la métrica gravitational-distance y common-border, que denominamos GDSCL y BSCL respectivamente. También se aplican a todos los modelos GEV sus correspondientes especificaciones MGEV, utilizando en todos ellos la misma estructura de coeficientes aleatorios que se utilizó en el capítulo anterior.

Los tres núcleos GEV se comparan entre ellos y con los modelos estimados en el capítulo anterior, multinomial logit (MNL), nested logit (NL) y restricted nested logit (RNL). Los modelos SCL y BSCL presentan mejores resultados de capacidad explicativa y predictiva que el modelo MNL, lo que demuestra la presencia de correlación espacial entre alternativas, que ambos modelos son capaces de capturar. El modelo GDSCL no mejora los resultados del modelo MNL y colapsa en él, seguramente debido a que la métrica espacial gravitational-distance no es eficiente en una zonificación con alternativas de formas irregulares, como la que se utiliza en esta tesis. Se comprueba de esta forma que la eficiencia de los modelos SCL-b dependerá en gran medida de la elección de métrica espacial para cada contexto empírico. Como futura línea de investigación sería recomendable estudiar posibles recomendaciones para la elección de métrica espacial en diferentes circunstancias empíricas. También se pone de manifiesto que la métrica espacial common-border se muestra eficiente en el caso de zonificaciones con alternativas de formas irregulares, pues el modelo BSCL mejora la capacidad explicativa y predictiva del modelo SCL, y se presenta como la especificación SCL-b con mejores resultados en este contexto, tal y como refleja la figura 3.4.

Los modelos BSCL y SCL mejoran los resultados de la especificación RNL del modelo nested logit, utilizando el mismo número de parámetros desconocidos. El modelo NL, con tres nidos, mejora los resultados del modelo SCL. Sin embargo, el modelo NL no mejora de forma concluyente los resultados del modelo BSCL, porque el modelo NL obtiene mejores resultados en validación cruzada de la precisión en predicción que el modelo BSCL, pero no lo mejora de forma concluyente en bondad de ajuste, ya que el modelo NL obtiene mayor resultado en AIC, pero menor en $\bar{\rho}_H^2$. Esto se debe a que el primer indicador de bondad de ajuste penaliza menos el número de parámetros desconocidos que el segundo. Estos resultados pueden ser debidos, por un lado, a la posible presencia de correlación entre alternativas que no sea de naturaleza espacial, que podría recoger la estructura de nidos del modelo NL, pero no lo pueden hacer los modelos SCL-b; aunque en este caso no parece la causa más probable, porque la estructura de nidos tiene una fuerte naturaleza espacial. Por otro lado, también ponen de manifiesto que un modelo SCL-b con una métrica apropiada al contexto empírico, como es en este caso el modelo BSCL, es capaz de capturar de forma más eficiente que el modelo NL la correlación espacial entre alternativas.

Por tanto, el enfoque basado en nidos descrito en el capítulo anterior y el enfoque basado en métricas espaciales descrito en este capítulo, son compatibles con modelización de elección espacial. Que uno u otro sea el más eficiente, puede depender de la habilidad del analista para diseñar la estructura de nidos o de la elección de la métrica espacial más apropiada al contexto empírico de aplicación.

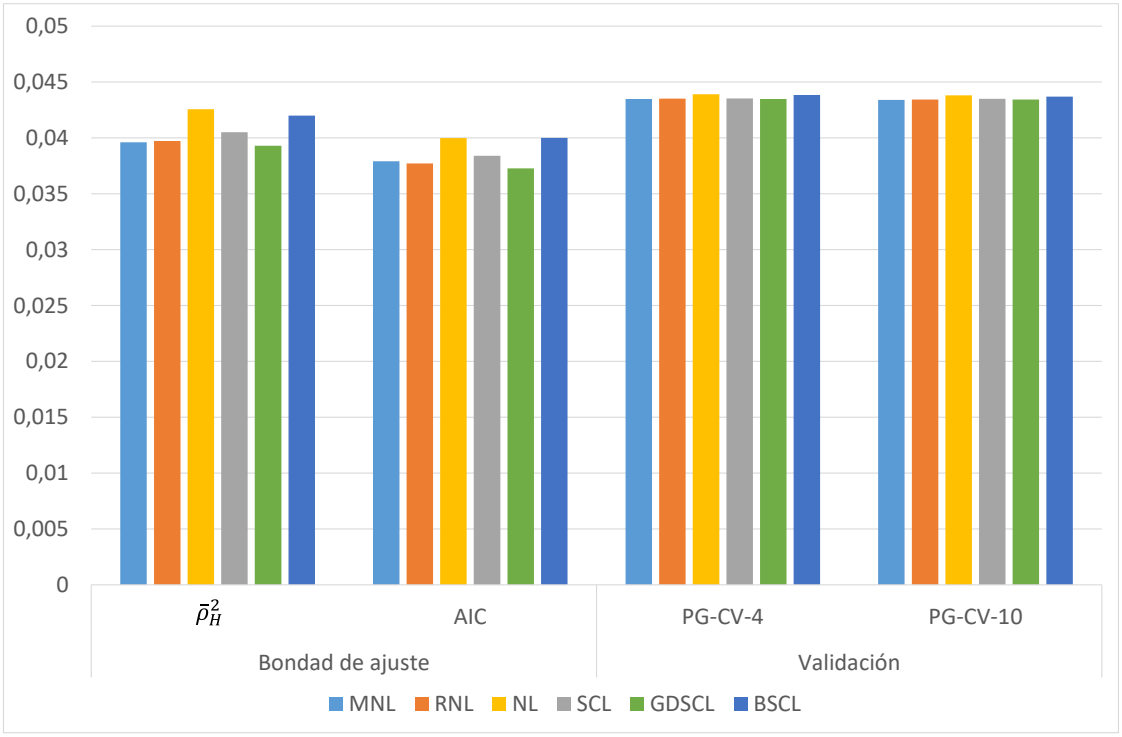

Figura 3.4. Estadísticos de bondad de ajuste y validación cruzada de los núcleos GEV estimados en los capítulos 2 y 3.

Todas las especificaciones mixtas mejoran la capacidad explicativa y predictiva de sus núcleos. Este resultado confirma, en este modelo, que la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. La comparación entre las especificaciones MGEV obtiene las mismas conclusiones relativas que con sus respectivos núcleos GEV, tal y como muestra la figura 3.5. MBSCL es la especificación mixta SCL-b que mejores resultados obtiene. M-NL también mejora a MSCL, pero no es mejor que MBSCL en todos los indicadores, por lo que no es concluyentemente mejor.

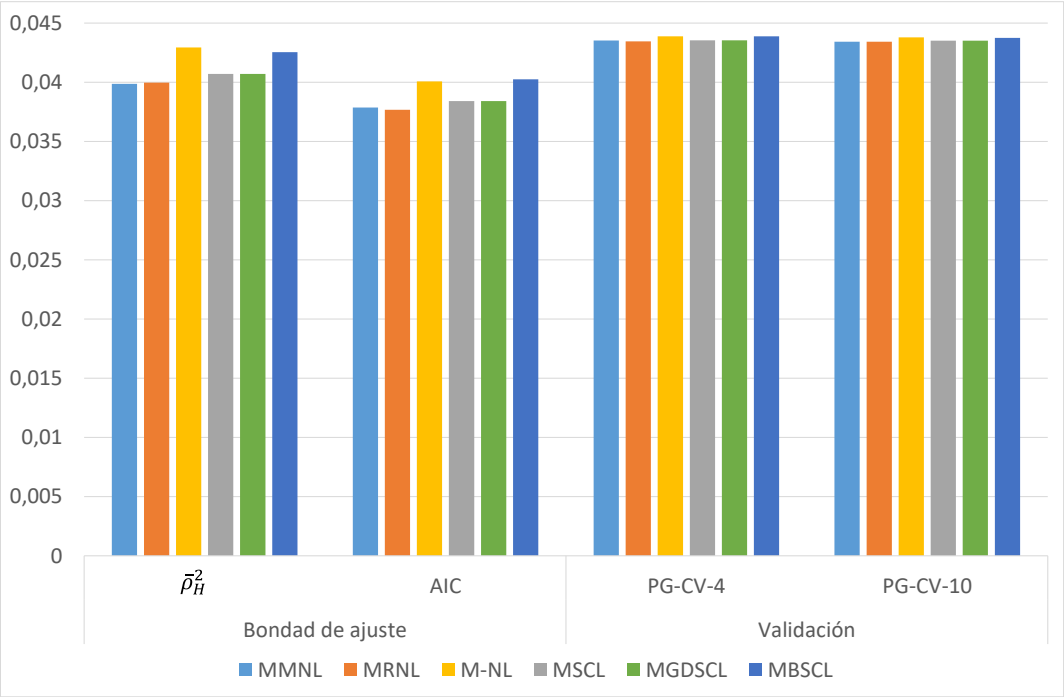

Figura 3.5. Estadísticos de bondad de ajuste y validación cruzada de las especificaciones mixtas estimadas en los capítulos 2 y 3.

4. Integración de los enfoques de correlación espacial en modelos de elección discreta: el modelo spatially correlated nested logit

En los dos capítulos anteriores de esta tesis se han analizado los dos enfoques actuales para incorporar correlación espacial entre alternativas en los modelos de elección discreta con alternativas de naturaleza espacial: el modelo nested logit (NL) basado en estructuras de nidos de alternativas diseñados por el analista, y los modelos logit con correlación espacial basados en generalizaciones del modelo spatially correlated logit (SCL). Los modelos de ambos enfoques forman parte de la familia de modelos logit generalized extreme value (GEV), por lo que tienen estructura matemática cerrada. Esta investigación postula que ambos enfoques son compatibles, y que su combinación puede mejorar los resultados de los modelos especificados con dichos enfoques. En este capítulo se propone un nuevo modelo GEV que combina ambos enfoques, que denominamos spatially correlated nested logit (SCNL). A lo largo de este capítulo se formula y analiza el nuevo modelo, así como su especificación mixta con coeficientes aleatorios, que permite incorporar variaciones en los gustos de los individuos decisores. El modelo propuesto y su especificación mixta se aplican al mismo contexto empírico de los dos capítulos anteriores, y los resultados obtenidos se comparan con los de los modelos de los dos enfoques actuales. Esta propuesta ha sido publicada en el artículo Perez-Lopez et al. (2022).

4.1. Modelo spatially correlated nested logit

La formulación del modelo NL de dos niveles se puede realizar a partir de la función generatriz GEV que se recoge en la fórmula (3.8), donde μ_k es el parámetro de disimilitud del k-ésimo nido de un total de M nidos, que se estima con los datos de muestra. La fórmula (2.24) recoge la probabilidad de elección de cada alternativa i de un nido N_k que se deriva de la función generatriz, que tiene una estructura matemática cerrada. El nuevo modelo SCNL que proponemos en esta investigación flexibiliza el modelo NL, sin añadir parámetros desconocidos al proceso de estimación, y manteniendo la compatibilidad con la modelización de elección espacial. En el modelo SCNL propuesto, al igual que en el modelo NL, las alternativas pertenecen a un solo nido, son incorreladas con las alternativas de nidos diferentes y correladas con las del mismo nido. Pero en el modelo NL, los pares de alternativas de un mismo nido tienen el mismo valor de correlación, que se calcula con la fórmula (2.23). La flexibilidad que aporta el modelo SCNL frente al modelo NL, se debe a que el nuevo modelo SCNL permite modelar la correlación espacial entre las alternativas de un mismo nido. Esto se hace a partir de una métrica de la correlación espacial de las alternativas, siguiendo el enfoque de los modelos logit con correlación espacial. Por tanto, en contexto de modelización de elección espacial, y dada una estructura de M nidos diseñada por el analista, el modelo SCNL propuesto en esta investigación permite superponer una métrica espacial. Esta métrica espacial incorpora al modelo una estructura de correlación espacial entre alternativas dentro de cada nido, que debe reflejar la dependencia espacial entre ellas.

El modelo SCNL es un modelo GEV con la función generatriz que se muestra en la fórmula (4.1). En concreto, el modelo SCNL es una especificación GNL, que será consistente con la maximización de la utilidad aleatoria siempre que $\mu_1, \dots, \mu_M \in (0, 1]$,

como todo modelo GNL (Wen y Koppelman, 2001). Los parámetros de disimilitud de cada par de alternativas, μ_{ij}, se obtienen según la fórmula (4.2), donde $\delta_k(i,j)$ es una función indicadora, que toma valor 1 si ambas alternativas (i,j) pertenecen a un mismo nido N_k, y nula en caso contrario. De este modo, los parámetros de disimilitud de pares de alternativas de un mismo nido N_k son iguales, y su valor μ_k se estima con los datos de muestra. En pares de alternativas de nidos diferentes este parámetro vale 1. Los parámetros de asignación de cada par de alternativas de un modelo SCNL se obtienen mediante la métrica espacial que se haya elegido, y se calculan de la misma forma que en el modelo SCL o cualquiera de sus generalizaciones que utilice la misma métrica. Por ejemplo, en el caso del modelo SCL, los parámetros de asignación se calculan mediante la fórmula (3.24), y en el caso de los modelos SCL-b, estos parámetros se obtienen mediante la fórmula (3.32). En esta tesis supondremos que el modelo SCNL se ha especificado a partir de un modelo SCL-b.

$$G(e^{V_1},\dots,e^{V_A}) = \sum_{i=1}^{A-1}\sum_{j=i+1}^{A}\left[\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}} \quad (4.1)$$

$$\mu_{ij} := \sum_{k=1}^{M}\mu_k\delta_k(i,j) + \prod_{k=1}^{M}[1-\delta_k(i,j)], \forall i,j\epsilon\{1,\dots,A\}, j\neq i \quad (4.2)$$

En el segundo capítulo de esta tesis se explicó que, en modelos de elección espacial, es habitual que la función de utilidad observada no incluya un conjunto completo de constantes específicas de alternativa, debido al elevado número de alternativas que suelen incluir este tipo de aplicaciones. En concreto, la aplicación a un caso real que se realiza en esta tesis no incluye constantes específicas de alternativa. Este tipo de especificaciones de la función de utilidad observada puede provocar que el modelo esté artificialmente sesgado, debido a que las esperanzas de los errores aleatorios no sean constantes entre alternativas. En los modelos cross-nested logit (CNL), y en concreto en los modelos GNL como el SCNL, es suficiente con que los parámetros de asignación estén normalizados a la unidad (Abbe et al., 2007). En la ecuación (4.3) se demuestra que en el modelo SCNL se cumple esta normalización. Por tanto, el modelo SCNL se puede especificar con una función de utilidad observada que no incluya un conjunto completo de constantes específicas de alternativa. El modelo así especificado no estará artificialmente sesgado por este motivo.

$$\sum_{j=1}^{A}\alpha_{i,ij} = \sum_{j=1}^{A}\frac{f(i,j)}{\sum_{l=1}^{A}f(i,l)} = \frac{\sum_{j=1}^{A}f(i,j)}{\sum_{l=1}^{A}f(i,l)} = 1, \forall i\epsilon\{1,\dots,A\} \quad (4.3)$$

Teorema. Los parámetros de asignación del modelo SCNL son independientes de la unidad de medida utilizada en la métrica espacial. Demostración:

> Sea f la métrica espacial del modelo y $\alpha_{i,ij}$ el parámetro de disimilitud de cada alternativa i con cada una de las restantes alternativas j, $i,j\epsilon\{1,\dots,A\}$. Si ahora tenemos la misma métrica espacial pero medida con otra unidad de medida, f', entonces hay un número no nulo $a\epsilon\mathbb{R}$, tal que $f'(i,j) = a\cdot f(i,j)$, para todo $i,j\in\{1,\dots,A\}$. El parámetro de disimilitud calculado ahora con la nueva unidad de medida es:

$$\alpha'_{i,ij} := \frac{f'(i,j)}{\sum_{l=1}^{M}f'(i,l)} = \frac{a\cdot f(i,j)}{\sum_{k=1}^{M}a\cdot f(i,l)} = \frac{a}{a}\cdot\frac{f(i,j)}{\sum_{k=1}^{M}f(i,l)} := \alpha_{i,ij}$$

En un modelo SCNL, la función de probabilidad de cada alternativa de elección se muestra en las fórmulas (4.4), (4.5) y (4.6), que como sucede en todos los modelos GEV,

tienen una estructura matemática cerrada. Los parámetros desconocidos del modelo se estiman mediante máxima verosimilitud.

$$P_i := \sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i|ij} P_{ij}\,, \forall i \epsilon \{1, \dots, A\} \tag{4.4}$$

$$P_{i|ij} = \frac{\left(\alpha_{i,ij} e^{V_i}\right)^{1/\mu_{ij}}}{\left(\alpha_{i,ij} e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij} e^{V_j}\right)^{1/\mu_{ij}}} \tag{4.5}$$

$$P_{ij} = \frac{\left(\left(\alpha_{i,ij} e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij} e^{V_j}\right)^{1/\mu_{ij}}\right)^{\mu_{ij}}}{\sum_{r=1}^{A-1} \sum_{l=r+1}^{A} \left(\left(\alpha_{r,rl} e^{V_r}\right)^{1/\mu_{rl}} + \left(\alpha_{l,rl} e^{V_l}\right)^{1/\mu_{rl}}\right)^{\mu_{rl}}} \tag{4.6}$$

La función de distribución valor extremo del vector de errores aleatorios de las ecuaciones de utilidad de un modelo SCNL $(\varepsilon_1, \dots, \varepsilon_A)$ se recoge en la fórmula (4.7). Esta fórmula se ha deducido a partir de la expresión del modelo SCL (Bhat y Guo, 2004) incorporando los parámetros μ_{ij} diferentes en cada nido de la estructura NL. La función de distribución marginal de cada error aleatorio ε_i es univariante de valor extremo. Como se comprueba en la fórmula (4.8), la función de distribución de este modelo es Gumbel estándar si los parámetros de asignación de cada alternativa están normalizados a la unidad, como se ha visto que se cumple en el caso SCNL.

$$F(\varepsilon_1, \dots, \varepsilon_A) = \exp\left\{-\sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \left[\left(\alpha_{i,ij} e^{-\varepsilon_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij} e^{-\varepsilon_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}}\right\} \tag{4.7}$$

$$F(\varepsilon_i) = \exp\left(-\sum_{\substack{j=1 \\ j\neq i}}^{A} \alpha_{i,ij}\, e^{-\varepsilon_i}\right) = \exp(-e^{-\varepsilon_i}), \forall i \epsilon \{1, \dots, A\} \tag{4.8}$$

La correlación no observada entre cada par de alternativas (i, j) se calcula mediante integración numérica, a partir de la función de distribución marginal bivariante de los errores aleatorios de las alternativas (Abbe et al., 2007), que se recoge en la fórmula (4.9); aunque también se puede utilizar la formulación propuesta por Papola (2004) para obtener un valor aproximado de forma más simplificada. La fórmula (4.10) es la adaptación de esta propuesta al modelo SCNL. Teniendo en cuenta la formulación de los parámetros de disimilitud del modelo SCNL (4.2), es fácil comprobar que el valor que se obtiene con esta aproximación en los pares de alternativas de distintos nidos es cero, y el que se obtiene en los pares de alternativas de un mismo nido N_k es $\sqrt{\alpha_{i,ij} \cdot \alpha_{j,ij}}(1 - \mu_k^2)$.

$$H\left(\varepsilon_i, \varepsilon_j\right) = \exp\left\{-\left[\left(1 - \alpha_{i,ij}\right) e^{-\varepsilon_i} + \left(1 - \alpha_{j,ij}\right) e^{-\varepsilon_j}\right] - \left[\left(\alpha_{i,ij} e^{-\varepsilon_i}\right)^{1/\mu_{ij}} + \right.\right.$$
$$\left.\left. + \left(\alpha_{j,ij} e^{-\varepsilon_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}}\right\} \forall i, j \epsilon \{1, \dots, A\}, j \neq i \tag{4.9}$$

$$\widehat{Corr}\left(\varepsilon_i, \varepsilon_j\right) = \sqrt{\alpha_{i,ij} \cdot \alpha_{j,ij}}(1 - {\mu_{ij}}^2), \forall i, j \epsilon \{1, \dots, A\} \tag{4.10}$$

Dado que las alternativas del nido raíz son incorreladas con el resto de alternativas, y al igual que sucede con el modelo nested logit, la expresión de la elasticidad directa y de la elasticidad cruzada de las alternativas del nido raíz coincide con la del modelo multinomial logit. Considerando de nuevo una utilidad observada lineal en los parámetros, si llamamos β_m al coeficiente del m-ésimo regresor de la i-ésima

alternativa, X_{im}, la tabla 4.1 recoge la expresión de la elasticidad directa de X_{im} en los modelos MNL, NL, SCL-b y SCNL.

Modelo	Elasticidad directa
SCNL	$\frac{\sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i\|ij}P_{ij}\left[(1-P_i)+\left(\mu_{ij}^{-1}-1\right)\left(1-P_{i\|ij}\right)\right]}{P_i}\beta_m X_{im}$ • Si i es una alternativa del nido raíz: $(1-P_i)\beta_m X_{im}$ • Si i es una alternativa de un nido N_k, $k\epsilon\{1,\dots,M\}$ $\frac{\sum_{j\notin N_k} P_{i\|ij}P_{ij}\,(1-P_i)+\sum_{\substack{j\in N_k \\ j\neq i}} P_{i\|ij}P_{ij}\left[(1-P_i)+(\mu_k^{-1}-1)\left(1-P_{i\|ij}\right)\right]}{P_i}\beta_m X_{im}$
SCL-based	$\frac{\sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i\|ij}P_{ij}\left[(1-P_i)+(\mu^{-1}-1)\left(1-P_{i\|ij}\right)\right]}{P_i}\beta_m X_{im}$
NL	• Si i es una alternativa del nido raíz: $(1-P_i)\beta_m X_{im}$ • Si i es una alternativa de un nido N_k, $k\epsilon\{1,\dots,M\}$ $\left[(1-P_i)+(\mu_k^{-1}-1)\left(1-P_{i\|k}\right)\right]\beta_m X_{im}, \forall i\epsilon\{1,\dots,A\}$
MNL	$(1-P_i)\beta_m X_{im}$

Tabla 4.1. Elasticidad directa de cada alternativa iϵ{1,…,A}.

En las mismas condiciones, la tabla 4.2 recoge la elasticidad cruzada de X_{im} en la j-ésima alternativa de los mismos modelos. En pares de alternativas de diferentes nidos, la fórmula de la elasticidad cruzada del modelo SCNL es la misma que en el modelo NL. En pares de alternativas de un mismo nido, ambas formulaciones son similares. Sin embargo, esta similitud es engañosa. A diferencia del modelo NL, la probabilidad condicional del modelo SCNL depende del efecto implícito de los parámetros de asignación. Las elasticidades SCNL son equivalentes a las de los modelos SCL-b en el caso de que todas las alternativas estén en el mismo nido. En comparación con los modelos SCL-b, el cálculo de la elasticidad directa en el modelo SCNL depende de los parámetros de disimilitud del nido. Si la métrica espacial se basa en la contigüidad, como la del modelo SCL o BSCL, la elasticidad cruzada de las alternativas no contiguas para la especificación basada en SCNL y SCL-b es igual a la de MNL.

Modelo	Elasticidad cruzada
SCNL	$-\left[P_i + \left(\mu_{ij}^{-1} - 1\right)\frac{P_{i\|ij}P_{ij}P_{j\|ij}}{P_j}\right]\beta_m X_{im}$ • Si i, j no están en el mismo nido: $-P_i\beta_m X_{im}$ • Si i, j están en el mismo nido N_k, $k\epsilon\{1, \dots, M\}$ $-\left[P_i + (\mu_k^{-1} - 1)\frac{P_{i\|ij}P_{ij}P_{j\|ij}}{P_j}\right]\beta_m X_{im}$
SCL-based	$-\left[P_i + (\mu^{-1} - 1)\frac{P_{i\|ij}P_{ij}P_{j\|ij}}{P_j}\right]\beta_m X_{im}$
NL	• Si i, j no están en el mismo nido: $-P_i\beta_m X_{im}$ • Si i,j están en el mismo nido N_k, $k\epsilon\{1, \dots, M\}^*$ $-\left[P_i + (\mu_k^{-1} - 1)P_{i\|k}\right]\beta_m X_{im}$ (*) Formulación modificada de Papola (2004) para facilitar la comparación
MNL	$-P_i\beta_m X_{im}$

Tabla 4.2. Elasticidad cruzada de cada par de alternativas i,j∈{1,…,A},j≠i.

El modelo SCNL requiere un diseño y un proceso de estimación más complejos que los modelos NL y SCL-b. En cuanto al diseño, en comparación con NL, el modelo SCNL requiere seleccionar una métrica espacial adecuada a la zonificación del área geográfica, y calcular los valores de la métrica en la zonificación normalmente utilizando un sistema de información geográfica; con respecto a los modelos basados en SCL, SCNL requiere diseñar una estructura de nidos de alternativas apropiada para el contexto empírico de aplicación. En cuanto a la estimación, aunque el modelo SCNL tiene los mismos parámetros desconocidos que el modelo NL, su proceso de estimación es más complejo, porque el modelo SCNL es una especificación reducida del modelo PGNL; en comparación con los modelos SCL-b, la presencia de la estructura de nidos aumenta el número de parámetros desconocidos en una cifra igual al número de nidos menos uno. Cuando un modelo SCNL usa métricas espaciales como la contigüidad SCL original o la métrica BSCL, la cantidad de parámetros de SCNL es la misma que la de NL. El modelo SCNL es compatible con las especificaciones basadas en SCL que requieren la estimación de parámetros adicionales, como en el caso de los modelos GSCL.

El modelo SCNL propuesto es una especificación GEV y, por tanto, es compatible con una especificación mixta de coeficientes aleatorios, que denominamos mixed SCNL (MSCNL), que permite incorporar al modelo variaciones en los gustos de los individuos decisores.

4.2. Aplicación a un caso real

En este apartado se aplica el modelo spatially correlated nested logit propuesto en esta tesis al mismo caso real que en los capítulos anteriores. El objetivo principal de esta aplicación es evaluar empíricamente la capacidad explicativa y predictiva del modelo SCNL propuesto, y compararla con el resto de modelos compatibles con modelización de elección espacial, que se analizaron y aplicaron en capítulos anteriores de esta tesis. Para ello, el modelo SCNL se especifica con la función de utilidad que se diseñó en los modelos de los dos enfoques descritos en los dos capítulos anteriores. Además, la estructura de nidos que se utiliza en este capítulo es la misma que se utilizó en el segundo capítulo, en los modelos que utilizan el enfoque de nidos, con el fin de evitar sesgar la comparación. Esta estructura de nidos de alternativas fue diseñada para el modelo NL, y tiene una fuerte componente espacial para capturar tanto los patrones de correlación espacial entre las alternativas como los resultantes de características no espaciales. Por lo tanto, este diseño probará la capacidad del modelo SCNL para capturar correlaciones espaciales entre alternativas que aún no han sido identificadas por la estructura de nidos, lo que servirá para verificar la capacidad del modelo SCNL para complementar el modelo NL. Respecto a la métrica de la correlación espacial entre alternativas, el modelo SCNL se especifica con la métrica common-border, que en el tercer capítulo de esta tesis se demostró empíricamente como la más apropiada para el contexto de aplicación, que denominamos common-border spatially correlated nested logit (BSCNL). También se especificará el modelo mixto BSCNL con la estructura de coeficientes aleatorios elegida en los modelos de ambos enfoques, que denominamos mixed common-border spatially correlated nested logit (MBSCNL). Esta especificación mixta se comparará con el resto de especificaciones MGEV estimadas en los capítulos anteriores de esta tesis.

4.2.1. Spatially correlated nested logit

La tabla 4.3 muestra los resultados de estimación, bondad de ajuste y validación del modelo BSCNL, propuesto en esta investigación para modelización de elección espacial en contextos empíricos como el de este caso. El modelo es aceptable, porque todos los parámetros estimados en este modelo se muestran significativos y con los signos esperados teóricamente. Los resultados no muestran indicios de endogeneidad. El orden de la influencia relativa de los regresores en la elección de localización residencial, en función de los valores de los coeficientes tipificados, es el mismo que en el resto de modelos logit estimados en el capítulo anterior.

Respecto al modelo multinomial logit, el modelo BSCNL mejora su capacidad explicativa y predictiva. Por un lado, mejora significativamente el ajuste del modelo MNL, con el que está anidado, ya que el contraste de la razón de verosimilitudes del modelo BSCNL es significativo respecto al modelo MNL. Por otro lado, el modelo BSCNL también mejora los resultados de validación cruzada con 4 grupos y 10 grupos del modelo MNL. Por tanto, el modelo BSCNL demuestra empíricamente que su mayor flexibilidad es capaz de recoger la correlación entre alternativas.

Respecto al enfoque de nidos para recoger la correlación entre alternativas, el modelo BSCNL mejora la capacidad explicativa y predictiva de las dos especificaciones del modelo nested logit. En la comparación con el modelo RNL, el modelo SCNL obtiene

mejores resultados con los dos indicadores de bondad de ajuste que penalizan el número de parámetros desconocidos, $\bar{\rho}_H^2$ y AIC. En la comparación con el modelo NL, que tiene los mismos parámetros desconocidos que el modelo BSCNL, se puede comprobar que todos los estadísticos de bondad de ajuste son superiores en el caso del modelo BSCNL. Por otro lado, el modelo BSCNL también mejora los resultados de validación cruzada con 4 grupos y 10 grupos de ambas especificaciones nested logit. Por tanto, el modelo BSCNL demuestra empíricamente que es capaz de recoger la correlación entre alternativas más eficientemente que el modelo NL, sin necesidad de parámetros desconocidos adicionales.

BSCNL								
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,104	0,0271	-3,85	0,00	**	-0,409	27%
	β_{FO}	-0,892	0,284	-3,14	0,00	**	-0,200	13%
	β_{HO}	1,29	0,283	4,57	0,00	**	0,294	20%
	β_{PR}	-1,99	0,399	-4,99	0,00	**	-0,252	16,8%
	$\beta_{PS\cdot H}$	1,02	0,26	3,91	0,00	**	0,088	6%
	$\beta_{SC\cdot H}$	0,173	0,0469	3,69	0,00	**	0,259	17,3%
	${\mu_A}^{-1}$	3,11	1,42	2,19	0,03	*		
	${\mu_B}^{-1}$	2,27	0,773	2,94	0,00	**		
	${\mu_C}^{-1}$	1,49	0,43	3,48	0,00	**		
		Nº par. est.		9				
Bondad de ajuste		LL		-1659,038				
		FG		0,0447				
		ρ^2		0,0464				
		$\bar{\rho}_H^2$		0,0438				
		AIC		0,0413				
		$LRT(Nulo)$		161,57	0,00	**		
		$LRT(MNL)$		17,86	0,00	**		
Val.		*PG-CV-4*		0,0442				
		PG-CV-10		0,0440				

Tabla 4.3. Resultados de estimación, bondad de ajuste y validación del modelo common-border spatially correlated nested logit: BSCNL.

Respecto al enfoque de correlación, el modelo BSCNL mejora la capacidad explicativa y predictiva de las tres especificaciones basadas en el modelo spatially correlated logit. Los modelos SCL y GDSCL utilizan métricas de correlación espacial entre alternativas que, como se demostró en el capítulo anterior, no son tan eficientes en el contexto empírico de aplicación como la métrica common-border propuesta en esta tesis. El modelo BSCNL no está anidado con estos dos modelos. Por un lado, el modelo BSCNL obtiene mejores resultados que SCL y GDSCL, con los dos indicadores de bondad de ajuste que penalizan el número de parámetros desconocidos, $\bar{\rho}_H^2$ y AIC. Por otro lado,

el BSCNL también mejora los resultados de validación cruzada con 4 grupos y 10 grupos de ambas especificaciones SCL-b. El modelo BSCNL utiliza la misma métrica que el modelo BSCL, y también mejora su capacidad explicativa y predictiva. Por tanto, el modelo BSCNL demuestra empíricamente que es capaz de incorporar al modelo tanto la correlación entre alternativas que recoge la estructura de nidos, como la correlación espacial entre alternativas que recoge la métrica de correlación espacial entre alternativas. De esta forma, el modelo BSCNL es el modelo GEV más apropiado para modelización de elección espacial en el contexto empírico de aplicación.

4.2.2. Mixed spatially correlated nested logit

La tabla 4.4 muestra los resultados de aplicación del modelo MBSCNL con la especificación de coeficientes aleatorios diseñada en el capítulo anterior, en la que el coeficiente correspondiente a la interacción $SC \cdot H$ es el único que se supone aleatorio, según una distribución normal. Todos los parámetros estimados en este modelo se muestran significativos, salvo el parámetro de disimilitud del nido A. Este parámetro no es significativo con un nivel de significación del 5%, aunque sí con 10%. De todas formas, se mantiene esta especificación, para que al comparar el modelo MBSCNL con el resto de modelos MGEV estimados en esta tesis, todos los modelos tengan la misma estructura de nidos. Los resultados no muestran indicios de endogeneidad. El orden de la influencia relativa de los regresores en la elección de localización residencial, en función de los valores de los coeficientes fijos tipificados, es el mismo que en el resto de modelos MGEV estimados hasta el momento.

La especificación mixta del modelo BSCNL muestra la mayor capacidad explicativa y predictiva de todos los modelos estimados en esta tesis. Por un lado, mejora la capacidad explicativa y predictiva de su núcleo GEV. Este resultado confirma que, en este modelo, la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. Por un lado, la especificación mixta MBSCNL presenta mejores resultados de bondad de ajuste que BSCNL en los dos indicadores de bondad de ajuste que penalizan el número de parámetros estimados, $\bar{\rho}_H^2$ y AIC. Además, aunque ambos modelos obtienen el mismo resultado en la validación cruzada de 4 grupos, en la de 10 grupos la especificación MBSCNL obtiene un resultado superior a su núcleo GEV.

Además, la especificación mixta MBSCNL mejora la capacidad explicativa y predictiva de todos los modelos mixtos estimados en los capítulos anteriores de esta tesis y, por tanto, también de sus núcleos GEV. Respecto a MMNL, con quien está anidado, mejora significativamente su bondad de ajuste, pues el contraste de la razón de verosimilitudes es significativo (ver tabla 4.4). También mejora los resultados obtenidos en la validación cruzada (ver tabla 2.5). El modelo MBSCNL también mejora la bondad de ajuste de las especificaciones mixtas MRNL y M-NL, pues tanto $\bar{\rho}_H^2$ como AIC son superiores, y sus resultados de validación cruzada (ver tablas 2.9 y 2.10). También presenta las mismas mejoras con respecto al MBSCL, con el que comparte la métrica de correlación espacial entre alternativas, common-border, propuesta en esta tesis (ver tabla 3.10).

	MBSCNL							
	Parámetro	**Valor**	**SE**	**Wald**	**p-valor**	**Sig.**	**SC**	**IR**
Estimación	β_{JT}	-0,108	0,027	-3,94	0,00	**	-0,425	34%
	β_{FO}	-0,924	0,288	-3,20	0,00	**	-0,207	16%
	β_{HO}	1,29	0,282	4,57	0,00	**	0,294	23%
	β_{PR}	-1,98	0,402	-4,92	0,00	**	-0,251	20%
	$\beta_{PS \cdot H}$	0,932	0,267	3,49	0,00	**	0,080	6%
	$E(\beta_{SC \cdot H})$	0,131	0,064	2,04	0,04	*		
	$\sigma(\beta_{SC \cdot H})$	0,276	0,118	2,33	0,02	*		
	$\mu_A{}^{-1}$	3,53	2,03	1,74	0,08	·		
	$\mu_B{}^{-1}$	2,43	0,954	2,55	0,01	**		
	$\mu_C{}^{-1}$	1,49	0,434	3,44	0,00	**		
		Nº par. est.		10				
Bondad de ajuste		SLL		-1.657,860				
		FG		0,0448				
		ρ^2		0,0471				
		$\bar{\rho}_H^2$		0,0442				
		AIC		0,0414				
		$LRT(Nulo)$		163,93	0,00	**		
		$LRT(MMNL)$		18,19	0,00	**		
Val.		*PG-CV-4*		0,0442				
		PG-CV-10		0,0441				

Tabla 4.4. Resultados de estimación, bondad de ajuste y validación del modelo mixed common-border spatially correlated nested logit: MBSCNL.

4.3. Resumen y conclusiones

En este capítulo se propone un nuevo modelo de elección discreta, compatible con modelización de elección espacial. El nuevo modelo combina los dos enfoques actuales, descritos en los capítulos segundo y tercero de esta tesis, respectivamente. Un enfoque utiliza nidos de alternativas diseñados por el analista para recoger la correlación entre alternativas. El otro enfoque utiliza métricas espaciales para recoger la correlación espacial entre alternativas. Esta tesis postula que ambos enfoques son compatibles. El modelo propuesto, spatially correlated nested logit, combina de forma eficiente ambos enfoques, y mejora la capacidad explicativa y predictiva de los modelos de ambos enfoques especificados en las mismas condiciones. Esta tesis formula y analiza el nuevo modelo.

El modelo propuesto se aplica en el mismo contexto empírico que el resto de modelos estimados en esta tesis, en la ciudad de Santander, especificado bajo las condiciones elegidas en los capítulos anteriores para el resto de modelos: función de utilidad observada, estructura de nidos de alternativas y métrica espacial common-border. El modelo aplicado, BSCNL, mejora la capacidad explicativa y predictiva del resto de

modelos estimados en esta tesis. La figura 4.1 permite apreciar que BSCNL obtiene mejores resultados en todos los estadísticos de bondad de ajuste y validación cruzada.

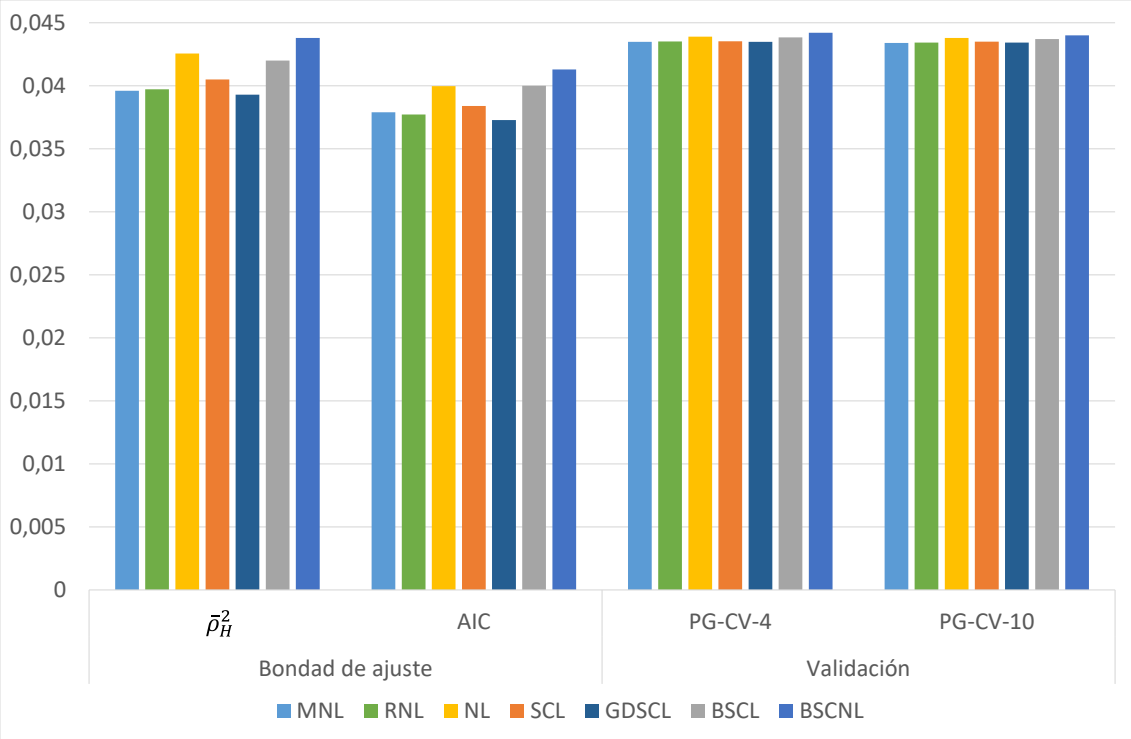

Figura 4.1. Estadísticos de bondad de ajuste y validación cruzada de los núcleos GEV estimados en la tesis.

Todas las especificaciones mixtas mejoran la capacidad explicativa y predictiva de sus núcleos. Este resultado confirma, en este modelo, que la incorporación de variaciones en los gustos del individuo decisor en uno de los regresores compensa el incremento de un parámetro desconocido. La comparación entre las especificaciones MGEV obtiene las mismas conclusiones relativas que con sus respectivos núcleos GEV, tal y como muestra la figura 4.2. El modelo MBSCNL es la especificación mixta que mejores resultados obtiene en los estadísticos de bondad de ajuste y validación cruzada.

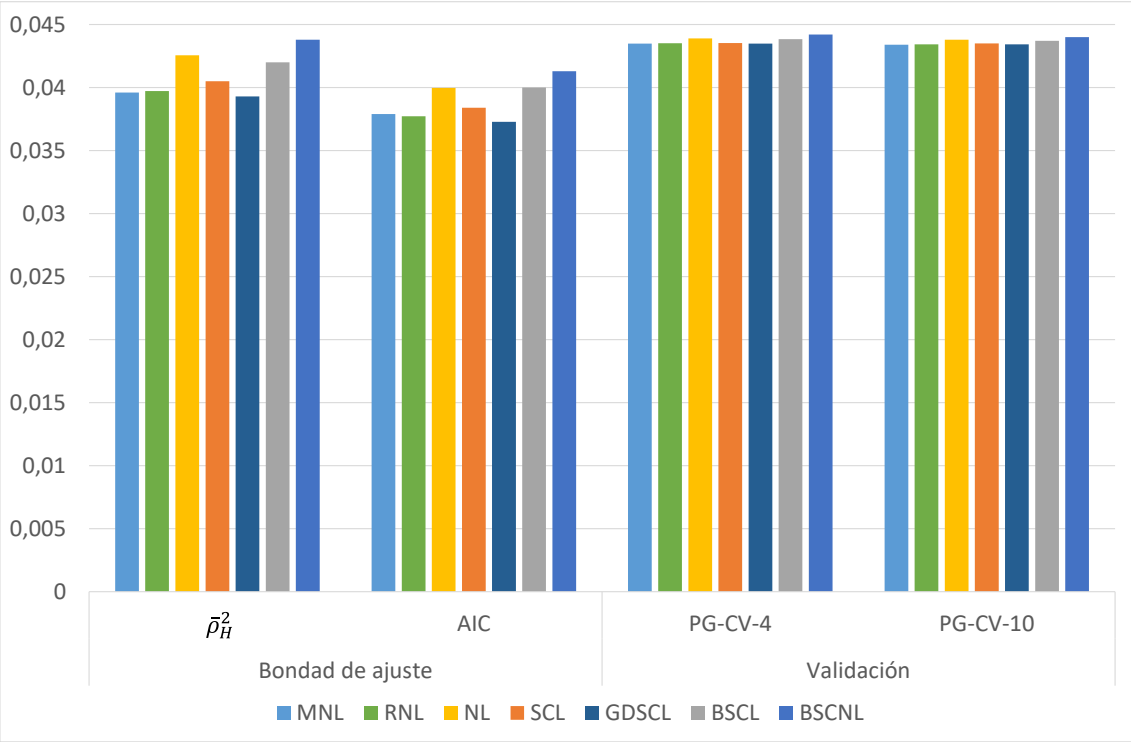

Figura 4.2. Estadísticos de bondad de ajuste y validación cruzada de las especificaciones mixtas estimadas en la tesis

5. Conclusiones y líneas de investigación futura

5.1. Conclusiones

Esta tesis se ha centrado en el estudio de los modelos de elección en los que las alternativas son áreas geográficas, poniendo el foco en el caso de elección de la localización de la residencia familiar en el contexto de los modelos de interacción de usos del suelo y transporte (LUTI). Los modelos LUTI son el marco conceptual más ampliamente utilizado actualmente para abordar la predicción de demanda de usos de suelo y transporte, en economía urbana y planificación del transporte.

El planteamiento más extendido para modelizar la elección de la localización residencial en contexto LUTI es el econométrico desagregado de los modelos de elección discreta. Estos modelos, derivados de la teoría económica de maximización de la utilidad aleatoria, son preferidos debido a la fuerte naturaleza económica y social que tiene esta elección. Pero los modelos de elección espacial presentan características específicas, principalmente el elevado número de alternativas y la presencia de dependencia espacial entre ellas, de manera que no todos los modelos de elección discreta son viables para modelizar elecciones de este tipo. En la literatura de los modelos de elección discreta se han planteado dos enfoques compatibles, el enfoque de nidos de alternativas, que surge a partir del modelo nested logit (NL) y que no es específico de elecciones espaciales; y el enfoque de correlación espacial, que surge a partir del modelo spatially correlated logit (SCL) y que sí es específico de elecciones espaciales.

En el capítulo segundo se analizan y comparan empíricamente los modelos con el primer enfoque. En el capítulo tercero se hace lo mismo con los modelos con el segundo enfoque. Además, en este capítulo se propone una nueva familia de modelos con este enfoque, a partir de la generalización del modelo SCL (SCL-b). A diferencia de otras generalizaciones de este tipo, como la denominada generalized spatially correlated logit (GSCL), SCL-b añade un solo parámetro estructural a la función de utilidad observada. Además, frente a las métricas empleadas hasta el momento en este tipo de modelos, basadas en la contigüidad entre zonas o en la distancia entre centroides, en este capítulo se propone una nueva métrica que se basa en la proporción de la frontera que comparten cada par de zonas. Esta métrica se considera especialmente adecuada en contextos urbanos con zonificaciones de formas irregulares o diferentes tamaños, donde puede mejorar los resultados obtenidos con las métricas utilizadas anteriormente. Este tipo de zonificaciones son muy habituales en la práctica, pues frecuentemente se utilizan áreas administrativas como base para la recopilación de datos, y es común que las áreas administrativas de muchas ciudades tengan formas irregulares y diferentes tamaños, principalmente en regiones con una larga historia urbana, como es el caso de Europa. Con el planteamiento de la métrica propuesta se evitan efectos no deseados, como por ejemplo, que dos zonas que apenas están en contacto tengan la misma correlación espacial que dos zonas que comparten gran parte de su frontera. Se corrigen también efectos debidos al tamaño variable de las zonas, especialmente presentes con métricas basadas en la distancia entre centroides. Para constatar esa posibilidad de mejora, en la tesis se aplica la métrica propuesta al caso real de la ciudad de Santander, con una zonificación basada en áreas administrativas, que presenta alternativas de diferentes tamaños y formas irregulares, y utilizando la

generalización SCL-b propuesta. Los resultados obtenidos con la nueva métrica presentan mayor capacidad explicativa y predictiva que los obtenidos con el resto de métricas. En este capítulo también se deduce la función generatriz del modelo paired generalized nested logit (PGNL), que es un modelo importante en la literatura científica. Hasta donde conoce el autor de esta tesis, esta formulación no había sido deducida en la misma.

En el capítulo cuarto de esta tesis se postula que los dos enfoques para modelos de elección discreta entre alternativas de naturaleza espacial son compatibles, y que su integración ofrece un potencial de mejora sobre ellos. La principal aportación de esta tesis, y que se recoge en ese capítulo, es la propuesta de un nuevo modelo de elección discreta espacial que integra ambos enfoques, denominado spatially correlated nested logit (SCNL). El modelo SCNL incorpora la correlación espacial y no espacial existente entre alternativas espaciales, sin necesidad de añadir parámetros desconocidos adicionales. Las alternativas de diferentes nidos son incorreladas en el nuevo modelo, al igual que sucede en el modelo NL. Pero SCNL, adicionalmente a NL, modela la correlación espacial de las alternativas de un mismo nido mediante el uso de una métrica espacial. El modelo SCNL es una especificación generalized extreme value (GEV) que presenta una estructura matemática cerrada, y es compatible con una especificación mixta de coeficientes aleatorios que le permita incorporar variaciones en los gustos de los individuos decisores. A lo largo del capítulo se deducen las fórmulas del nuevo modelo correspondientes a la probabilidad de elección, la función de distribución del vector de errores aleatorios de las ecuaciones de utilidad, la función de distribución marginal del error aleatorio de cada ecuación de utilidad, la función de distribución marginal bivariante de los errores aleatorios de cada par de alternativas, así como una aproximación de la correlación entre ellos. En este capítulo también se deduce la fórmula de la función generatriz GEV y de las elasticidades directa y cruzada, que se comparan con la de los modelos de los dos enfoques anteriores.

El modelo SCNL requiere un diseño y un proceso de estimación más complejos que los modelos NL y SCL-b. En cuanto al diseño, en comparación con NL, el modelo SCNL requiere seleccionar una métrica espacial adecuada a la zonificación del área geográfica, y calcular los valores de la métrica en la zonificación, normalmente utilizando un sistema de información geográfica; con respecto a los modelos basados en SCL, SCNL requiere diseñar una estructura de nidos de alternativas apropiada para el contexto empírico de aplicación. En cuanto a la estimación, aunque el modelo SCNL tiene los mismos parámetros desconocidos que el modelo NL, su proceso de estimación es más complejo, porque el modelo SCNL es una especificación reducida del modelo PGNL; en comparación con los modelos SCL-b, la presencia de la estructura de nidos aumenta el número de parámetros desconocidos en una cifra igual al número de nidos menos uno. Cuando un modelo SCNL usa métricas espaciales como la contigüidad SCL original o la métrica BSCL, la cantidad de parámetros desconocidos de SCNL es la misma que la de NL. El modelo SCNL es compatible con otros modelos del enfoque basado en SCL, aunque requieran la estimación de parámetros adicionales, como en el caso de los modelos GSCL. Los resultados obtenidos en este capítulo demuestran la mayor capacidad explicativa y predictiva del modelo desarrollado para este caso de estudio.

5.2. Líneas de investigación futura

El trabajo realizado en esta tesis y sus aportaciones han dejado abiertas diferentes líneas de investigación futura, que se describen a continuación:

- Estudiar posibles generalizaciones del modelo spatially correlated nested logit (SCNL). Por ejemplo, con interacciones de estructuras de nidos no-espaciales, en las que se permita la correlación espacial entre alternativas de nidos distintos.
- Estudiar nuevas métricas espaciales y elaborar un manual de recomendaciones para diferentes tipos de zonificación.
- Desarrollar software específico para la estimación del modelo SCNL y de su especificación mixta.
- Estudiar la casuística de los modelos de elección discreta con correlación espacial entre alternativas de elección y entre observaciones simultáneamente. Hasta donde sabe el autor, ningún estudio reportado en la literatura la ha considerado.
- Estudiar la casuística de los modelos de elección discreta con correlación inducida por variables explicativas correlacionadas espacialmente. Hasta donde sabe el autor, ningún estudio reportado en la literatura la ha considerado.
- Aplicación de las propuestas a nuevos casos reales con características diferentes.

5.3. Listado de publicaciones realizadas

En este apartado se recopilan las publicaciones del autor de esta tesis en transportes y economía urbana, que se realizan durante el periodo de elaboración de la misma.

5.3.1. Artículos científicos con indexación Journal of Citation Reports (JCR)

Perez-Lopez J-B, Novales M y Orro A (2022) Spatially correlated nested logit model for spatial location choice. Transportation Research Part B: Methodological 161 (2022): 1-12. https://doi.org/10.1016/j.trb.2022.05.007.

Perez-Lopez J-B, Orro A y Novales M (2021) Environmental impact of mobility in higher-education institutions: the case of the ecological footprint at the University of A Coruña (Spain). Sustainability, 13(11), 6190. https://doi.org/10.3390/su13116190.

Novales M, Orro A, Pérez-López J-B, Feal J y Bugarín M R (2021) Increasing Boarding Lost Time at Regular Bus Stops during Rainy Conditions: A Case Study. Journal of Public Transportation, 23(1), 4. https://doi.org/10.5038/2375-0901.23.1.4.

Pérez-López J-B, Novales M, Varela-Garcia F-A y Orro A (2020) Residential location econometric choice modeling with irregular zoning: common border spatial correlation metric. Networks and Spatial Economics 20, 785–802. https://doi.org/10.1007/s11067-020-09495-5.

Orro A, Novales M, Monteagudo Á, Pérez-López, J-B y Bugarín M R (2020) Impact on city bus transit services of the COVID–19 lockdown and return to the new normal: The case of A Coruña (Spain). Sustainability, 12(17), 7206. https://doi.org/10.3390/su12177206.

Longarela-Ares Á, Calvo-Silvosa A y Pérez-López J B (2020) The influence of economic barriers and drivers on energy efficiency investments in maritime shipping, from the

perspective of the principal-agent problem. Sustainability, 12(19), 7943. https://doi.org/10.3390/su12197943.

Anta J, Pérez-López J-B, Martínez-Pardo A, Novales M y Orro A (2016) Influence of the weather on mode choice in corridors with time-varying congestion: a mixed data study. Transportation, 43(2), 337-355. https://doi.org/10.1007/s11116-015-9578-1.

5.3.2. Capítulos de libro

Pérez-López J-B y Orro A (2016) Residential location choice models with spatial correlation. In: Dell'Olio L, Cordera R y Ibeas A (eds) Land Use - Transport Interaction Models. The TRANSPACE model. Santander: GIST, pp. 114–150.

Bibliografía

Abbe E, Bierlaire M y Toledo T (2007) Normalization and correlation of cross-nested logit models. Transportation Research Part B: Methodological, 41: 795–808.

Acheampong R A y Silva E A (2015) Land use–transport interaction modeling: A review of the literature and future research directions. Journal of Transport and Land use, 8(3): 11-38.

Agresti A (1996) An Introduction to Categorical Data Analysis. New York: Wiley.

Alonso W (1960). A theory of the urban land market. Bobbs-Merrill Company, College Division.

Alonso W (1964) Location and Land Use. Cambridge: Harvard University Press.

Álvarez-Daziano R y Munizaga M A (2002) Modelación flexible de elecciones discretas: una revisión crítica. Actas del XI Congreso Panamericano de Ingeniería de Tránsito y Transporte. Noviembre 2002, Quito, Ecuador.

Andrew M y Meen G (2006) Population structure and location choice: A study of london and south east england. Papers in Regional Science, 85(3): 401–419.

Anselin L (1980) Estimation methods for spatial autoregressive structures. Regional Science Dissertation and Monograph Series (Ithaca, NY).

Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic.

Anta J, Pérez-López J-B, Martínez-Pardo A, Novales M y Orro A (2016) Influence of the weather on mode choice in corridors with time-varying congestion: a mixed data study. Transportation, 43(2): 337-355.

Axhausen K, Scott D, König D y Jürgens C (2004) Locations, commitments and activity spaces. In M. Schreckenberg and R. Selten, eds., Human Behaviour and Traffic Networks: 205–230. Berlin: Springer.

Bahamonde-Birke F J (2021) A brief discussion on the treatment of spatial correlation in multinomial discrete models. Journal of Transport and Land Use, 14(1): 521-535.

Başar G y Bhat C (2004) A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. Transportation Research Part B: Methodological, 38: 889–904.

Belart B C (2011) Wohnstandortwahl im Grossraum Zürich (Master's thesis, IVT, ETH Zürich).

Ben-Akiva M y Bierlaire M (1999) Discrete choice methods and their applications to short-term travel decisions. In Hall R (ed) Handbook of Transportation Science.

Ben-Akiva M y Bowman J L (1998) Integration of an activity-based model system and a residential location model. Urban Studies, 35: 1131–1153.

Ben-Akiva M y Francois B (1983) Mu-homogenous generalized extreme value model. Working Paper. Department of Civil Engineering, MIT, Cambridge, MA.

Ben-Akiva M y Swait J (1986) The Akaike likelihood ratio index. Transportation Science, 20(2): 133-136.

Ben-Akiva M, McFadden D, Train K, Walker J, Bhat C, Bierlaire M, Bolduc D, Boersch-Supan A, Brownstone D, Bunch D S y Daly A (2002) Hybrid choice models: Progress and challenges. Marketing Letters 13(3): 163-175.

Bhat C R y Guo J (2004) A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transportation Research Part B: Methodological, 38(2): 147–168.

Bhat C R, Govindarajan A y Pulugurta V (1998) Disaggregate attraction-end choice modeling: formulation and empirical analysis. Transportation Research Record, 1645: 60-68.

Bhat C R, Guo J Y, Srinivasan S y Sivakumar A (2004) Comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record, 1894: 57–66.

Bhat C R, Handy S L, Kockelman K, Mahmassani S L, Gopal A, Srour I M y Weston L (2002) Development of an Urban Accessibility Index: Formulations, Aggregation, and Application. Working Paper Report 7-4938-4, University of Texas Austin, Austin.

Bhat, C R (1995). A heteroscedastic extreme value model of intercity mode choice. Transportation Research Part B: Methodological, 29: 471-483.

Bhat, C R (1997) An endogenous segmentation mode choice model with an application to intercity travel. Transportation Science, 31: 34-48.

Bierlaire M (2003) BIOGEME: a free package for the estimation of discrete choice models. Ascona, Switzerland, 3rd Swiss transportation research conference.

Bolduc D (1992) Generalized autoregressive errors: The multinomial probit model. Transportation Research Part B: Methodological, 26: 155-170.

Boschmann, E E (2011) Job access, location decision, and the working poor: A qualitative study in the Columbus, Ohio, metropolitan area. Geoforum, 42: 671–682.

Bradley M, Bowman J y Griesenbeck B (2009) SACSIM: an applied activity-based model system with fine-level spatial and temporal resolution. Journal of Choice Modelling 3(1): 5-31.

Bresnahan T F, Stern S y Trajtenberg M (1997) Market segmentation and the sources of rents from innovation: personal computers in the late 1980s. RAND Journal of Economics 28: 17–44.

Bucklin R E, Gupta S y Han S (1995) A brand's eye view of response segmentation in consumer brand choice behavior. Journal of Marketing Research, 32(1): 66-74.

Bürgle M (2006) Residential location choice model of the Greater Zurich area. In STRC, ed., 6th Swiss Transport Research Conference. Ascona. URL http://www.strc.ch/conferences/2006/Buergle_STRC_2006.pdf.

Carrasco J A y Ortúzar J de D (2002) Review and assessment of the nested logit model. Transport Reviews, 22(2): 197-218.

Chasco C (2003) Econometría espacial aplicada a la predicciónextrapolación de datos microterritoriales. Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid.

Chen J, Chen C y Timmermans H J P (2008) Accessibility trade-offs in household residential location decisions. Transportation Research Record, 2077: 71–79.

Cherchi E y de Dios Ortuzar J (2010) Can mixed logit reveal the actual data generating process? Some implications for environmental assessment. Transportation Research Part D: Transport and Environment, 15(7): 428-442.

Chintagunta P, Jain D y Vilcassim N (1991) Investigating heterogeneity in brand preference in logit models for panel data. Journal of Marketing Research, 28: 417-428.

Chu C (1981) Structural issues and sources of bias in residential location and travel mode choice models. Unpublished Ph.D. Dissertation. Department of Civil Engineering, Northwestern University, USA.

Chu C (1989) A paired combinatorial logit model for travel demand analysis. Proceedings of the Fifth World Conference on Transportation Research, 4(Ventura, CA): 295-309.

Cliff A y Ord J (1981) Spatial Processes: Models and Applications. London: Pion.

Coppola P y Nuzzolo A (2011) Changing accessibility, dwelling price and the spatial distribution of socio-economic activities. Research in transportation economics, 31(1): 63-71.

Cressie N (1993) Statistics for spatial data (Fourth ed.). Hoboken, NJ: John Wiley and Sons.

Daganzo C (1979) Multinomial Probit: The Theory and its Application to Demand Forecasting. Academic Press, New York.

Daly A J y Zachary S (1978) Improved multiple choice models. In: Hensher D. A. y Dalvi MQ (eds), Determinants of Travel Choice (Westmead: Saxon House): 335-357.

de Luca S y Cantarella G E (2009) Validation and comparison of choice models. In: Saleh, W., Sammer, G. (Eds.), Travel Demand Management and Road User Pricing: Success, Failure and Feasibility. Ashgate publications: 37–58.

de Palma A, Motamedi K, Picard N, y Waddell P (2005) A model of residential location choice with endogenoushousing prices and traffic for the París region. European Transport, 31: 67–82.

de Palma A, Picard N y Waddell P (2007) Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater Paris region. Journal of Urban Economics, 62(2): 204–230.

Dell'Olio L, Cordera R y Ibeas A (eds) Alonso A, Alonso B, Barreda R, Comi A, Coppola R, González E, Monzón A, Moura J, Nogués S, Nuzzolo A, Orro A, Papa E, Perez-Lopez J-B, Reques P, Sañudo R y Wang Y (2016) Land use - transport interaction models. The TRANSPACE model. 1st edn. Santander: GIST.

Domencich T A y McFadden D (1975) Urban travel demand: A behavioral analysis. American Elsevier, New York.

Fleming M M (2004) Techniques for estimating spatially dependent discrete choice models. In: Anselin L, Florax R J G M y Rey S J (eds) Advances in spatial econometrics. Springer: 145-168.

Florax R J G M y REY S (1995) The impacts of misspecified spatial interaction in linear regression models. En New directions in spatial econometrics. Ed. Springer: 111-135.

Fox M (1995) Transport planning and the human activity approach. Journal of Transport Geography 3: 105–116.

Gakemheimer R (2006) Transporte y uso del suelo en los países en vías de desarrollo: planificar en medio de la controversia. In Madrid, España: I Congreso Internacional sobre Desarrollo Humano.

Garrido R A y Mahmassani H S (2000) Forecasting freight transportation demand with the space-time multinomial probit model. Transportation Research Part B: Methodological, 34: 403-418.

Gaudry M J y Wills M J (1978) Estimating the functional form of travel demand models. Transportation Research, 12(4): 257–289.

Geurs K T y van Wee B (2004) Accesibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport Geography, 12: 127-140.

Gopalakrishnan R, Guevara A y Ben-Akiva M (2020) Combining multiple imputation and control function methods to deal with missing data and endogeneity in discrete-choice models. Transportation Research Part B: Methodological, 142: 45-57.

Greene W (2001) Fixed and random effects in nonlinear models. Working Paper, Stern School of Business, New York University.

Greene W y Hensher D A (2003) A latent class model for discrete choice analysis: contrasts with mixed logit. Transportation Research Part B: Methodological, 37(8): 681-698.

Guerrero T E, Guevara C A, Cherchi E y Ortúzar J D D (2021). Forecasting with strategic transport models corrected for endogeneity. Transportmetrica A, Transport Science: 1-28.

Guevara-Cue C A (2005) Addressing endogeneity in residential location models (Doctoral dissertation, Massachusetts Institute of Technology).

Guevara-Cue C A (2010) Endogeneity and Sampling of Alternatives in Spatial Choice Models (Doctoral dissertation, Ph. D. Thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA).

Guevara C A (2015) Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transportation Research Part A: Policy and Practice, 82: 240-254.

Guevara C A y Ben-Akiva M (2006) Endogeneity in residential location choice models. Transportation Research Record, 1977(1): 60-66.

Guevara C A y Ben-Akiva M (2012) Change of Scale and Forecasting with the Control-Function Method in Logit Models. Transportation Science, 46(3): 425–437.

Guevara C A and Polanco D (2016) Correcting for Endogeneity due to Omitted Attributes in Discrete-Choice Models: the Multiple Indicator Solution. Transportmetrica A: Transport Science, 12: 458–478.Guo J Y y Bhat C R (2004) Modifiable areal units: problem or perception in modeling of residential location choice? Transportation Research Record, 1898: 138–147.

Guo J Y y Bhat C R (2007) Operationalizing the concept of neighborhood: Application to residential location choice analysis. Journal of Transport Geography, 15: 31-45.

Habib K M N y Miller E J (2009) Reference-dependent residential location choice model within a relocation context. Transportation Research Record, 2133: 92–99.

Hansen W G (1959) How accessibility shapes land use. Journal of the American Institute of Planners, 25(2): 73–76.

Hensher D A y Johnson L W (1981) Applied discrete-choice modelling. Cross Helm, London, 1981.A

Hess S, Bierlaire M y Polak J (2005) Capturing taste heterogeneity and correlation structure with Mixed GEV models. In: Scarpa R y Alberini A (eds), Applications of Simulation Methods in Environmental and Resource Economics, Springer Publisher, Dordrecht, The Netherlands, chapter 4: 55-76.

Hilbe J M (2009) Logistic regression models. CRC press. ISBN 978-1-138-10671-0.

Horowitz J (1983) Statistical comparison of non-nested probabilistic discrete choice models. Transportation Science, 17: 319-350.

Hunt L M, Boots B y Kanaroglou P S (2004) Spatial choice modelling: new opportunities to incorporate space into substitution patterns. Progress in Human Geography, 28(6): 746-766.

Ibeas A, Cordera R, Dell'Olio L y Coppola P (2013) Modeling the spatial interactions between workplace and residential location. Transportation Research A: Policy and Practice, 49: 110–122

Johnson N y Kotz S (1970) Distributions in Statistics: Constinuous Univariate Distributions. John Wiley. New York. Chapter 21.

Kalyanam K y Putler D S (1997) "Incorporating demographic variables in brand choice models: an indivisible alternatives framework," Marketing Science, 16: 166-181.

Kamakura, W A y Russell A (1989) A probabilistic choice model for market segmentation and elasticity structure. Journal of Marketing Research, 26: 379-390.

Kim J H, Pagliara F y Preston J (2005) The intention to move and residential location choice behavior. Urban Studies, 42: 1621–1636.

Koppelman F S y Bhat C H (2006) A self-instructing course in mode choice modeling: multinomial and nested logit models.

Koppelman F S y Wen C H (2000) The paired combinatorial logit model: properties, estimation and application. Transportation Research Part B: Methodological, 34(2): 75-89.

Lee B H Y y Waddell P (2010) Residential mobility and location choice: A nested logit model with sampling of alternatives. Transportation, 37: 587–601.

LeSage J P (2000) Bayesian estimation of limited dependent variable spatial autoregressive models. Geographical Analysis, 32(1): 19–35.

Longarela-Ares Á, Calvo-Silvosa A y Pérez-López J B (2020) The influence of economic barriers and drivers on energy efficiency investments in maritime shipping, from the perspective of the principal-agent problem. Sustainability, 12(19),): 7943.

Lowry I S (1964) A model of metropolis. Technical report, RAND Europe.

Marschak, J. (1960). Binary-choice constraints and random utility indicators. In K. Arrow, ed., Stanford Symposium on Mathematical Methods in the Social Science, Stanford University Press, Standford, CA: 312-329.

Martínez F J (2000) Towards a land-use and transport interaction framework. Handbook of Transport Modelling. Eds Henser D A, Button K J. Elsevier Science: 145-164.

Martínez L M, Viegas J M y Silva E A (2007) Zoning decisions in transport planning and their impact on the precision of results. Transportation Research Record, 1994: 58–65.

Martínez-Pardo A, Orro A y Garcia-Alonso L (2020) Analysis of port choice: a methodological proposal adjusted with public data. Transportation Research Part A: Policy and Practice, 136: 178-193.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P (ed). Frontiers in Econometrics, Academic Press, New York: 105-142.

McFadden D (1978) Modelling the choice of residential location. In: Karlqvist A, Jundqvist L, Snickars F y Weibull J (eds) Spatial interaction theory and planning models. North Holland. Amsterdam: 75–96.

McFadden D y Train K (2000) Mixed MNL models for discrete response. Journal of Applied Econometrics, 15(5): 447-470.

McNally M G (2000) The activity approach. In Handbook of Transport Modeling, edited by D. A. Hensher, and K. J. Button. Oxford: Pergamon.

McNally M G y Rindt C (2007) The Activity-Based Approach. URL: http://escholarship.org/uc/item/86h7f5v0.

Menard S (1995) Applied Logistic Regression Analysis. Thousand Oaks, CA: Sage: 48.

Menard S (2004) Six Approaches to Calculating Standardized Logistic Regression Coefficients. The American Statistician, 58(3): 218-23.

Menard S (2011) Standards for standardized logistic regression coefficients. Social Forces, 89(4): 1409-1428.

Novales M, Orro A, Pérez-López J B, Feal J y Bugarín M R (2021) Increasing Boarding Lost Time at Regular Bus Stops during Rainy Conditions: A Case Study. Journal of Public Transportation, 23(1): 4.

Orro A (2006) Modelos de elección discreta en transportes con coeficientes aleatorios. Cátedra abertis.

Orro A, Novales M y Benitez F G (2010) Box-Cox Mixed Logit Model for Travel Behavior Analysis. In AIP Conference Proceedings 1281(1): 679-682. American Institute of Physics.

Orro A, Novales M, Monteagudo Á, Pérez-López, J B y Bugarín M R (2020) Impact on city bus transit services of the COVID–19 lockdown and return to the new Normal: The case of A Coruña (Spain). Sustainability, 12(17): 7206.

Outwater M L y Charlton B (2008) The San Francisco model in practice validation, testing, and application. Innovations in Travel Demand Modeling: Papers. Transportation Research Board Conference Proceedings 42 (2).

Pagliara F, Preston J y Simmonds D (Eds.) (2010). Residential location choice: Models and applications. Springer Science & Business Media.

Papola A (2004) Some developments on the cross-nested logit model. Transportation Research Part B: Methodological, 38: 833-851.

Parady G, Ory D y Walker J (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. Journal of Choice Modelling, 100257.

PB Consult (2005) The MORPC Travel Demand Model Validation and Final Report. Prepared for the Mid-Ohio Region Planning Commission.

Pendyala R M, Kitamura R, Kikuchi A, Yamamoto T y Fujji S (2005) FAMOS: Florida activity mobility simulator. Proceedings of the 84th Annual Meeting of the Transportation Research Board Washington, DC, January 9–13, 2005 (CD Rom).

Pérez-López J-B y Orro A (2016) Residential location choice models with spatial correlation. In: Dell'Olio L, Cordera R y Ibeas A (eds) Land Use - Transport Interaction Models. The TRANSPACE model. Santander, GIST: 114–150.

Pérez-López J-B, Novales M, Varela-Garcia F-A y Orro A (2020) Residential location econometric choice modeling with irregular zoning: common border spatial correlation metric. Networks and Spatial Economics, 20(3): 785-802.

Perez-Lopez J B, Orro y Novales (2021) Environmental impact of mobility in higher-education institutions: the case of the ecological footprint at the University of A Coruña (Spain). Sustainability, 13(11): 6190.

Perez-Lopez J-B, Novales M y Orro A (2022) Spatially correlated nested logit model for spatial location choice. Transportation Research Part B: Methodological, 161: 1-12.

Pinjari A R y Bhat C R (2011) Activity-based travel demand analysis. A Handbook of Transport Economics, 10: 213–248.

Pinjari A R, Bhat C R y Hensher D A (2009). Residential self-selection effects in an activity time-use behavior model. Transportation Research Part B: Methodological, 43(7): 729–748.

Pinjari A R, Pendyala R M, Bhat C R y Waddell P A (2011) Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. Transportation, 38(6): 933–958.

QGIS Development Team (2018) QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

Recker, W (1995). Discrete choice with an oddball alternative. Transportation Research Part B: Methodological, 29: 201-211.

Schirmer P M, Van Eggermond M A y Axhausen K W (2014) The role of location in residential location choice models: a review of literature. Journal of Transport and Land Use, 7(2): 3-21.

Schnier K E y Felthoven R G (2011) Accounting for spatial heterogeneity and autocorrelation in spatial discrete choice models: Implications for behavioral predictions. Land Economics, 87(3): 382–402.

Sener I N, Pendyala R M y Bhat C R (2011) Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. Journal of Transport Geography, 19: 294-303.

Small K A (1987) A discrete choice model for ordered alternatives. Econometrica, 55(2): 409–424.

Spellucci P (1993) DONLP2 Users Guide, Dept. of Mathematics, Technical University at Darmstadt, 64289 Darmstadt, Germany.

Srour I M, Kockelman K y Dunn P (2002) Accessibility Indices: Connection to Residential Land Prices and Location Choices. Transportation Research Record, 1805: 25–34.

Steckel, J H y Vanhonacker W R (1988). A heterogeneous conditional logit model of choice. Journal of Business & Economic Statistics, 6: 391-398.

Stetzer F (1982) Specifying weights in spatial forecasting models: the results of some experiments. Environmental and Planning A, 14: 571-584.

Takahashi K (2019) Local Relaxation of Constraints on Dissimilarity Parameters in the Generalized Nested Logit Model. International Journal of Japan Association for Management Systems, 11(1): 73-80.

Thurstone L (1927) A law of comparative judgement. Psycological Review, 34: 273-286.

Tobler W R (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(1): 234–240.

Torrens P M (2000) How land-use transportation models work. Centre for Advanced Spatial Analysis, London

Train K E (1998) “Recreation demand models with taste differences over people,” Land Economics, 74: 230-240.

Train K E (2009) Discrete Choice Methods with Simulation. Cambridge University Press, New York, New York, USA.

von Thunen J H (1826) Der isolierte Staat in beziehung auf Landwirtschaft und Nationalokonomie, Gustav Fisher, Stuttgart. English edition: The isolated state (trans: Wartenburg CM (1966), edited by Hall P). Oxford: Pergamon.

Vovsha P (1997) The cross-nested logit model: application to ode choice in the Tel-Aviv metropolitan area. Transportation Research Record, 1607: 6–15.

Vovsha P y Chiao K A (2008) Development of New York Metropolitan Transportation Council Tour-Based Model. Innovations in Travel Demand Modeling: Papers. Transportation Research Board Conference Proceedings, 42(2).

Vyvere Y, Oppewal H y Timmermans H (1998) The validity of hierarchical information integration choice experiments to model residential preference and choice. Geographical Analysis, 30(3): 254–272.

Waddell P (1993) Exogenous workplace choice in residential location models: Is the assumption valid? Geographical Analysis, 25: 65–82.

Waddell P (2006) Reconciling household residential location choices and neighborhood dynamics. WorkingPaper.

Waddell P, Bhat C R, Eluru N, Wang L y Pendyala R M (2007) Modeling interdependence in household residence and workplace choices. Transportation Research Record, 2003: 84–92.

Ward M D y Gleditsch K S (2002) Location, location, location: An MCMC approach to modeling the spatial context of war and peace. Political Analysis, 10(3): 244–260.

Wegener M (1994) Operational urban models state of the art. Journal of the American planning Association, 60(1): 17-29.

Weisbrod G, Ben-Akiva M y Lerman S R (1980) Tradeoffs in residential location decisions: transportation versus other factors. Transport Policy and Decision Making, 1(1).

Weiss A, Hasnine S y Habib K N (2019) A Comparative Study of Methods for Capturing Spatial Correlations in Location Choice through an Empirical Application on School Location Modelling. Conference: 98th Annual Meeting of TRB, January 13-17, 2019. At: Washington DC.

Wen C H y Koppelman F S (2001) The generalized nested logit model. Transportation Research Part B: Methodological, 35 (7): 627–641.

Wilks S S (1938) The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics, 9(1): 60-62.

Williams H C W L (1977) On the formation of travel demand models and economic evaluation measures of user benefit. Environment and Planning, 9(A): 285-344.

World Urbanization Prospects (2018) The United Nations, Department of Economic and Social Affairs, Population Division Revision, Online Edition.

Zolfaghari A, Sivakumar A y Polak J W (2012). Choice Set Pruning in Residential Location Choice Modelling: A Comparison of Sampling and Choice Set Generation Approaches in Greater London. Transportation Planning and Technology, 35(1): 87–106.

Zondag B y Pieters M (2005) Influence of accessibility on residential location choice. Transportation Research Record, 1902: 63–70.

Zhou B y Kockelman K (2008) Microsimulation of residential land development and household location choices: bidding for land in Austin, Texas. Transportation Research Record, 2077: 106–112.

Anexos

En este último capítulo se incluyen las publicaciones directamente derivadas de la tesis, que son las siguientes (en la versión restringida sólo se incluye la primera publicación por estar cedidos los derechos de las dos últimas a un editor):

- Perez-Lopez J-B, Novales M y Orro A (2022) Spatially correlated nested logit model for spatial location choice. Transportation Research Part B: Methodological, 161 (2022): 1-12. https://doi.org/10.1016/j.trb.2022.05.007.
- Pérez-López J-B, Novales M, Varela-Garcia F-A y Orro A (2020) Residential location econometric choice modeling with irregular zoning: common border spatial correlation metric. Networks and Spatial Economics 20: 785–802. https://doi.org/10.1007/s11067-020-09495-5.
- Pérez-López J-B y Orro A (2016) Residential location choice models with spatial correlation. In: Dell'Olio L, Cordera R y Ibeas A (eds) Land Use - Transport Interaction Models. The TRANSPACE model. Santander, GIST: 114–150.

Anexo A. Spatially correlated nested logit model for spatial location choice. Transportation Research Part B: Methodological, 161 (2022): 1-12

Transportation Research Part B 161 (2022) 1–12

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Spatially correlated nested logit model for spatial location choice

Jose-Benito Perez-Lopez [a,*], Margarita Novales [b], Alfonso Orro [b]

[a] Universidade da Coruña, Group of Railways and Transportation Engineering, Department of Economics, Facultad de Economía y Empresa, Campus de Elviña, 15071 A Coruña, Spain
[b] Universidade da Coruña, Group of Railways and Transportation Engineering, Department of Civil Engineering, ETS Ingenieros de Caminos, Canales y Puertos, Elviña, 15071 A Coruña, Spain

ARTICLE INFO

Keywords:
Discrete choice models
Transport land use interaction
Spatial models
Spatial correlation
Residential location choice
Microeconometric choice models

ABSTRACT

Residential location choice is a key component of the models for predicting land-use and transport demand in urban planning. In general, it requires to consider correlation between spatial alternatives. The approach of nested alternatives of the nested logit model has proved highly efficient in this context. This approach incorporates into the nested logit model both spatial and non-spatial correlations due to unobserved variables. The approach of metric extensions to the spatially correlated logit model specifies models for capturing spatial correlations between alternatives without having to design a nested structure. A model combining both approaches is proposed in this research. The spatially correlated nested logit model proposed herein models the correlation between alternatives of the nests of a nested logit model using a metric of spatial correlation between pairs of alternatives. The proposed model improves the properties of the nested logit model without the need of increasing the number of unknown parameters. Our model also improves the properties of a spatially correlated model with the same spatial metric. When needing to incorporate preference heterogeneity into the model, the proposed model is compatible with a mixed specification with random coefficients. The spatially correlated nested logit model was empirically applied to the real case of residential location choice in the city of Santander in Spain. In this empirical context, this model improved the explanatory and predictive power of the models that it combines.

1. Introduction

People are frequently faced with decisions requiring choosing between a discrete set of alternatives, such as decisions about purchasing, mode of transport and travel destinations, among others. As highlighted by Takahashi (2019), huge studies have been conducted to capture discrete purchasing behavior through discrete choice models. Spatial location choices, in which choice alternatives refer to geographical locations, are a key feature of advanced disaggregate models of travel and activity demand. Within these models, the most important aspects refer to residential location choice and to a lesser extent employment location choice. Spatial location choices can also appear in other types of models, such as travel destination or public transport boarding or alighting stop choice models.

In the models for predicting land-use and transportation demand in urban planning, currently, the most widely used approach consists of mathematical simulation models of the interaction between land-uses and transportation (LUTI, see Torrens, 2000). LUTI

* Corresponding author.
E-mail address: benito.perez@udc.es (J.-B. Perez-Lopez).

https://doi.org/10.1016/j.trb.2022.05.007
Received 11 June 2021; Received in revised form 28 April 2022; Accepted 8 May 2022

models use zoning of the spatial area under study and require a prediction model of the individual choice of the area of residence. The choice of area of residence is a crucial decision for many families for two key reasons. First, the area of residence will tremendously impact housing prices, whether for buying or renting. Second, the area of residence influences travel times to daily activities and the type of social life the family will have according to accessibility of services offered and their characteristics.

At a disaggregate level, Weiss et al. (2019) state that spatial location choice is typically modeled using the econometric approach based on random utility maximizing (RUM) hypothesis (Thurstone, 1927; McFadden, 1974). RUM discrete choice modeling is the most commonly used mathematical framework in prediction models of residential location choice in the LUTI context (Pagliara and Wilson, 2010). Various RUM models have been developed, but they are not all applicable with spatial alternatives. In these cases, some alternatives will be spatially correlated. Bahamonde-Birke (2021) presents a discussion of the different kinds of spatial correlation that affects multinomial discrete choice models and how they have been addressed in the discrete modeling literature. The modeling approaches considered in this paper deal with the spatial correlation among alternatives, that is common in transport and land-use models. This correlation refers to substitution preferences on the part of the decision maker and are due to unobserved spatial elements of the utility. As a result, the models that overlook correlation between alternatives, like multinomial logit, are not suitable in this context. In turn, spatial location choices usually entail a high number of alternatives, thereby preventing or hindering the use of some approaches to capture this correlation. Specifying a mixed logit model using an error component structure allows flexible patterns of correlation between alternatives (see Train, 2009). This approach is usually unfeasible for correlation between spatial alternatives because such correlations require specifying as many error components as pairs of correlated alternatives, which are usually too many for the estimation process. The same limitation occurs in the probit model (Daganzo, 1979). If constraints are not included in the correlation structure of the perturbations of this model, the number of parameters to estimate in the covariance matrix of the perturbations may be so high that the estimation process of this model becomes unfeasible (see applications of the probit model for spatial correlation in Bolduc, 1992; Garrido and Mahmassani, 2000).

The main goal of this research is to propose a new model for predicting land-use and transport demand in urban planning, based on RUM and focused in spatial location choice of residence. In this context, the choice alternatives are geographic areas. The new model must consider the spatial characteristics of these alternatives to improve the ability to explain and predict the behavior of decision-makers in comparison with other current RUM models. The new generalized extreme value (GEV) model proposed combines the two current GEV approaches compatible with spatial correlation between alternatives, by incorporating the spatially correlated logit approach into the nested logit model. This spatially correlated nested logit model considers correlation through pre-specified nests and uses spatial information on the alternatives, without the need of increasing the number of estimated parameters in relation to a nested logit approach. Thanks to this combination, the proposed model improves the explanatory and predictive power of the previous GEV models. This model is compatible with a mixed GEV specification, which makes it possible to incorporate variations in decision-makers' preferences. In the next section, we review the state of the art in GEV models compatible with spatial location choice modeling. In Section 3, we present the proposed GEV model. In Section 4, the new model is applied to an urban residential location choice empirical context and compared with the GEV models compatible with spatial location choice modeling analyzed in Section 2. Finally, Section 5 presents our conclusions.

2. GEV models with spatial correlation between alternatives

The most widespread and simple RUM-consistent discrete choice model is multinomial logit (MNL) (McFadden, 1974; Domencich and McFadden, 1975). The MNL model assumes that the stochastic components (ε_i) of the utility of alternative i (U_i) have a marginal type I extreme value distribution (Gumbel; Johnson and Kotz, 1970) independent and equally distributed. The MNL model assumes uncorrelation between alternatives and between observations, overlooking unobserved variations in preferences or tastes. The parameter of the perturbation scale is usually normalized to one; a similar approach has been used in all models considered in this article, without loss of generality (Abbe et al., 2007).

2.1. Nested logit

The hierarchical or nested logit model (NL; Williams, 1977; Daly and Zachary, 1978; McFadden, 1978) extends the MNL model to allow for specific structures of correlation between alternatives. The stochastic components of the NL model maintain homoscedasticity and have a joint extreme value distribution. This model clusters alternatives to assess the correlation between them. The clusters of alternatives, termed nests, must be designed by the analyst. To design the nests, the analyst must use variables not incorporated in the utility function. For example, in an urban residential location choice context, these variables may represent how attractive the area is to the decision maker for its prestige, prevailing architecture, views or accessibility of services, such as transport, schools, leisure or employment. In the NL model, each alternative belongs to a nest. The structure of the resulting variance-covariance matrix is a diagonal matrix by blocks, one per nest, unlike the scalar structure of the MNL model. The parameters $0 < \mu_k \leq 1$, termed dissimilarity parameters of each nest N_k, modulate the value of the correlation between pairs of alternatives. The correlation between the perturbations of two alternatives, $i$and$j$, is calculated using Eq. (1) if both alternatives belong to the same nest, N_k, and is null if they belong to different nests.

$$Corr(\varepsilon_i,\ \varepsilon_j) = \left(1-\mu_k^2\right),\ \forall i,j \in N_k,\ k \in \{1,\ \ldots,\ M\} \tag{1}$$

The NL model is compatible with spatial location choice modeling if the analyst designs a structure with a viable number of nests.

The increase in the number of nests increases not only the flexibility of the NL model in measuring the correlation between alternatives but also the number of dissimilarity parameters that the model will have to estimate. The handicap of the NL model lies in the need for the analyst to design the nest structure. Furthermore, the effectiveness of the NL model in collecting the correlation between alternatives will depend on the analyst's ability to design the nests. Eq. (2) shows the probability of each alternative i, where $P_{i|k}$ (3) is the conditional probability of the alternative i if nest N_k is selected and P_k (4) is the probability of choosing nest N_k. Eqs. (3) and (4) are modifications of that of Papola (2004) to facilitate the comparison between NL and the model proposed in this paper.

$$P_i = P_{i|k} \cdot P_k \tag{2}$$

$$P_{i|k} = \frac{\left(e^{V_i}\right)^{1/\mu_k}}{\sum_{j \in N_k} \left(e^{V_j}\right)^{1/\mu_k}} \tag{3}$$

$$P_k = \frac{\left(\sum_{j \in N_k} \left(e^{V_j}\right)^{1/\mu_k}\right)^{\mu_k}}{\sum_{l=1}^{M} \left(\sum_{r \in N_l} \left(e^{V_r}\right)^{1/\mu_l}\right)^{\mu_l}} \tag{4}$$

2.2. Spatially correlated logit

McFadden (1978) generalized the nest approach of the NL model in the class of generalized extreme value (GEV) models. The perturbations of the GEV models are homoscedastic, with a joint extreme value distribution. GEV models incorporate constraints in the covariance matrix from the nest structure, which, if relatively simple, maintain a closed structure. Probability is calculated based on a termed generating function $G(e^{V_1}, \ldots, e^{V_A})$, using Eq. (5). The generating function of a GEV model should meet a set of criteria established by McFadden (1978) and revised by Ben-Akiva and Francois (1983). Nest can assess both unobserved spatial correlation among alternatives and correlation due to unobserved non-spatial variables. As pointed out by Bahamonde-Birke (2021), GEV models cannot be used to capture spatial correlation among observations. GEV models can act as kernels of mixed logit specifications with random coefficients, termed mixed GEV (Bhat and Guo, 2004; Hess et al., 2005).

$$P_i = \frac{e^{V_i} \cdot \frac{\partial G\left(e^{V_1}, \ldots, e^{V_A}\right)}{\partial e^{V_i}}}{G\left(e^{V_1}, \ldots, e^{V_A}\right)}, \ \forall i \in \{1, \ldots, A\} \tag{5}$$

Both MNL and NL are GEV models. The generating function of MNL model is shown in Eq. (6) and that of the NL model in Eq. (7), where μ_k is the dissimilarity parameter of the nest N_k. GEV extensions of the NL model are based on cross-nested logit (CNL; Small, 1987; Vovsha, 1997; Ben-Akiva and Bierlaire, 1999; Papola, 2004), which Wen and Koppleman (2001) formulated as generalized NL (GNL). The generating function of GNL model is shown in Eq. (8), where $\alpha_{ik} \geq 0$ is the allocation parameter of alternative i to nest N_k for all M nests and A alternatives, with zero value when the alternative does not belong to the nest. In CNL or GNL models, the alternatives can belong to more than one nest. For this reason, they incorporate the allocation parameters, which are interpreted as the level of membership of each alternative to each nest (Abbe et al., 2007). These models are nested with the two-level NL model that uses the same dissimilarity parameters if each alternative belongs to a single nest with an allocation parameter value of one.

$$G\left(e^{V_1}, \ldots, e^{V_A}\right) = \sum_{i=1}^{A} e^{V_i} \tag{6}$$

$$G\left(e^{V_1}, \ldots, e^{V_A}\right) = \sum_{k=1}^{M} \left(\sum_{i \in Nest_k} \left(e^{V_i}\right)^{1/\mu_k} \right)^{\mu_k} \tag{7}$$

$$G\left(e^{V_1}, \cdots, e^{V_A}\right) = \sum_{k=1}^{M} \left(\sum_{i \in Nest_k} \left(\alpha_{ik} e^{V_i}\right)^{1/\mu_k} \right)^{\mu_k} \tag{8}$$

The GNL models include the constraint that the allocation parameters of each alternative add up to one, as expressed in Eq. (9) (see Abbe et al., 2007 to analyze normalization proposals in other CNL formulations). This normalization allows the allocation parameters of each alternative to represent the proportion of belonging to each nest.

$$\sum_{k=1}^{M} \alpha_{ik} = 1, \ \forall i = 1, \ \ldots, \ A \tag{9}$$

The unobserved correlation between pairs of alternatives of the CNL and GNL models is modulated by all structural parameters, that is, allocation and dissimilarity parameters. This correlation is calculated from the joint cumulative distribution function, by numerical integration. When the number of alternatives is high, the number of structural parameters of the GNL models increases considerably in relation to the NL model to the point that estimating all parameters is unfeasible. Under these conditions, it may be useful to calculate some parameters beforehand or incorporate constraints to reduce their number and then estimate the model only with the other parameters (Abbe et al., 2007).

The paired combinatorial logit (Chu, 1981; Chu, 1989; Koppelman and Wen, 2000) model proposes a GEV model with a nest structure not designed by the analyst but instead formed by each pair of alternatives. Therefore, this model has as many dissimilarity parameters as pairs of alternatives. Wen and Koppelman (2001) extended the paired combinatorial logit model with a GNL formulation termed paired generalized nested logit, which adds two allocation parameters for each pair of nests with respect to the paired combinatorial logit model.

The paired combinatorial logit and paired generalized nested logit models are not viable in the spatial location choice modeling context, except when previously calculating a significant number of structural parameters or incorporating constraints to reduce their number. As clearly shown, the number of structural parameters in the paired generalized nested logit model is much higher than in any other GNL specification based on a nest structure designed by the analyst. Using both possibilities, Bhat and Guo (2004) proposed, for the context of spatial location choice, a reduced specification of the paired generalized nested logit model, the spatially correlated logit (SCL) model, with a GEV generating function (10). On the one hand, the SCL model adds to paired generalized nested logit model the constraint that all pairs of contiguous alternatives have the same dissimilarity parameter $0 < \mu \leq 1$. On the other hand, the SCL model proposes that the paired generalized nested logit model allocation parameters be calculated before estimating the model, using data on the contiguity of the alternatives. These parameters are calculated using Eq. (11), where the value of the dichotomous spatial variable ω_{ij} is 1 when the alternatives i, j partly share the border, and 0 otherwise. Therefore, regardless of the number of alternatives, the SCL model requires estimating only one more parameter than the MNL model (with which the model is nested), the dissimilarity parameter.

$$G\left(e^{V_1},\ \ldots, e^{V_A}\ \right) = \sum_{i=1}^{A-1}\sum_{j=i+1}^{A}\left(\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu}\right)^{\mu} \tag{10}$$

$$\alpha_{i,ij} = \frac{\omega_{ij}}{\sum_{l=1}^{A}\omega_{il}},\ \forall i,j \in \{1,\ \cdots,\ A\} \tag{11}$$

The spatial approach of the SCL model was extended with new, spatially correlated GNL models. These SCL-based models use metrics of the spatial similarity of the alternatives to calculate allocation parameters between pairs of alternatives according to Eq. (12), where $f(i, j)$ is the value of a spatial metric f in each pair of alternatives i, j, whose values are non-negative, and which meets $f(i, i) = 0$, $\forall i$ (see Pérez-López et al., 2020).

$$\alpha_{i,ij} = \frac{f(i,\ j)}{\sum_{l=1}^{A} f(i,\ l)},\ \forall i,j \in \{1,\ \cdots,\ A\} \tag{12}$$

The distance-based SCL model (Sener et al., 2011) is an SCL-based model that uses a distance-based spatial metric. Both the contiguity of the alternatives and the distance-based metrics are efficient in a context of alternatives with a regular shape, such as some type of grid. However, residential location models commonly use zoning based on administrative areas, which tends to have irregular shapes, especially in cities with historic areas. Pérez-López et al. (2020) propose in this context an SCL-based model which uses the common border length between pairs of contiguous alternatives as a spatial metric to calculate allocation parameters (BSCL). Eq. (13) shows the probability of choosing each alternative i in BSCL model, where $P_{i|ij}$ (14) is the conditional probability of alternative i if the pair i, j is selected and P_{ij} (15) is the probability for the pair i, j.

$$P_i = \sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i|ij}\cdot P_{ij},\ \forall i \in \{1,\ \ldots,\ A\} \tag{13}$$

$$P_{i|ij} = \frac{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu}}{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu}} \tag{14}$$

$$P_{ij} = \frac{\left(\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu}\right)^{\mu}}{\sum_{r=1}^{A-1}\sum_{l=r+1}^{A}\left(\left(\alpha_{r,rl}e^{V_r}\right)^{1/\mu} + \left(\alpha_{l,rl}e^{V_l}\right)^{1/\mu}\right)^{\mu}} \tag{15}$$

3. Spatially correlated nested logit

In residential location choice context, alternatives are usually high in number and spatially correlated. The models with correlation between alternatives which we have considered viable or more appropriate for this context are GEV models with two different approaches. One approach is the NL model, with nested structures designed by the analyst for the application environment. The other approach corresponds to models based on spatially correlated logit model that use spatial correlation metrics between alternatives, which must be appropriate to the empirical context, such as the BSCL model, when the alternatives are built from irregularly shaped administrative geographic areas. This research postulates that both approaches are compatible and that their combination can improve the fit and predictive capability of the models specified with those approaches. The resulting GEV model has been termed spatially correlated nested logit (SCNL).

The SCNL model makes the NL model more flexible in the spatial location choice modeling context, without adding parameters to

the estimation process. The new model makes it possible to model the spatial correlation between alternatives of the same NL nest. The alternatives still belong to a single nest and are not correlated with alternatives from different nests. However, the pairs of alternatives in the same nest do not have the same correlation. The spatial correlation between pairs of alternatives of the same nest is modeled from a metric of the spatial correlation between alternatives, following the approach based on spatially correlated logit model.

The SCNL model proposed in this research has been formulated from a paired generalized nested logit specification, starting from a NL-type nest structure (each alternative belongs to a single nest) in a spatial location choice modeling context, and incorporating spatial correlation between the alternatives of the same nest. The GEV generating function of the SCNL model is Eq. (16). The allocation parameters of each pair of alternatives are calculated from a spatial metric between alternatives f, as shown in Eq. (12). The dissimilarity parameters of the pairs of alternatives are assessed using Eq. (17), where $\delta_k(i,j)$ is a Boolean function, which is 1 if both alternatives belong to the same nest, and null otherwise. Thus, the dissimilarity parameters of the pairs of alternatives of the same nest are equal, and their μ_k values are estimated with sample data, reaching the value 1 in pairs of different nest alternatives. As a GNL model, the condition that $\mu_1,\ldots, \mu_M \in (0, 1]$ ensures that the SCNL model is consistent with RUM (Wen and Koppelman, 2001).

$$G\left(e^{V_1},\ \ldots, e^{V_A}\ \right) = \sum_{i=1}^{A-1}\sum_{j=i+1}^{A}\left[\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}} \tag{16}$$

$$\mu_{ij} := \sum_{k=1}^{M}\mu_k\delta_k(i,j) + \prod_{k=1}^{M}[1-\delta_k(i,j)],\ \forall i,j \in \{1,\ \ldots,\ A\} \tag{17}$$

This SCNL model collapses on the SCL-based model specified with the same spatial metric when there is only one nest to which all alternatives belong. The allocation parameters of the SCNL model are independent of the unit of measure used in the spatial metric, as shown below.

Demonstration:

Let f the spatial metric of the model and $\alpha_{i,ij}$ the dissimilarity parameter of each alternative i with each other alternative j, $i, j \in \{1, \ldots, A\}$. If we now have the same spatial metric but measured with other metric unit, f', then there is a non-zero number $a \in \mathrm{R}$, such that $f'(i, j) = a \bullet f(i, j)$, for all $i, j \in \{1, \ldots, A\}$. The dissimilarity parameter calculated now with the new metric unit is:

$$\alpha'_{i,ij} := \frac{f'(i,j)}{\sum_{l=1}^{A} f'(i,l)} = \frac{a\cdot f(i,j)}{\sum_{l=1}^{A} a\cdot f(i,l)} = \frac{a}{a}\cdot\frac{f(i,j)}{\sum_{l=1}^{A} f(i,l)} = \alpha_{i,ij}$$

The spatial location choice models have a high number of alternatives and, for this reason, typically do not include a full set of alternative specific constants; therefore, the expectations of perturbations between alternatives would not be constant and therefore the model would be artificially biased. In the SCNL model, as in CNL models, normalizing the allocation parameters to one suffices to avoid this (Abbe et al., 2007). This normalization is demonstrated in Eq. (18).

$$\sum_{j=1}^{A}\alpha_{i,ij} = \sum_{j=1}^{A}\frac{f(i,\ j)}{\sum_{l=1}^{A} f(i,\ l)} = \frac{\sum_{j=1}^{A} f(i,\ j)}{\sum_{l=1}^{A} f(i,\ l)} = 1,\ \forall i \in \{1,\ \ldots,\ A\} \tag{18}$$

The probability function of the SCNL model (Eqs. (19), (20), and (21)) is the same as for paired generalized nested logit model (Wen and Koppelman, 2001), albeit with a different definition of the parameters μ_{ij}, and makes it possible to calculate the probability of each individual choosing the alternative i without integrations. The parameters of the model are estimated using maximum likelihood. The cumulative extreme-value distribution of the vector of perturbations of the utility equations of an SCNL model $(\varepsilon_1,\ldots, \varepsilon_A)$ is expressed in Eq. (22). The marginal cumulative distribution function of each perturbation ε_i is a univariant extreme value. As confirmed in Eq. (23), the function is the Gumbel standard if the allocation parameters of each alternative are normalized to one, a requirement met in the SCNL model.

$$P_i := \sum_{\substack{j=1\\ j\neq i}}^{A} P_{i|ij}P_{ij},\ \forall i \in \{1,\ \ldots,\ A\} \tag{19}$$

$$P_{i|ij} = \frac{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu_{ij}}}{\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu_{ij}}} \tag{20}$$

$$P_{ij} = \frac{\left(\left(\alpha_{i,ij}e^{V_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{V_j}\right)^{1/\mu_{ij}}\right)^{\mu_{ij}}}{\sum_{r=1}^{A-1}\sum_{l=r+1}^{A}\left(\left(\alpha_{r,rl}e^{V_r}\right)^{1/\mu_{rl}} + \left(\alpha_{l,rl}e^{V_l}\right)^{1/\mu_{rl}}\right)^{\mu_{rl}}} \tag{21}$$

$$F(\varepsilon_1,\ \ldots,\ \varepsilon_A) = \exp\left\{-\sum_{i=1}^{A-1}\sum_{j=i+1}^{A}\left[\left(\alpha_{i,ij}e^{-\varepsilon_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{-\varepsilon_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}}\right\} \tag{22}$$

Table 1
Direct elasticities of each alternative i∈{1,…,A}.

Model	Direct elasticity
SCNL	• If i is in root nest $(1 - P_i)\beta_m X_{im}$ • If i is in N_k nest, $k\epsilon\{1, \ldots, M\}$ $\frac{\sum_{\substack{j=1 \\ j \neq i}}^{A} P_{i\|ij} P_{ij}[(1 - P_i) + (\mu^{-1} - 1)(1 - P_{i\|ij})]}{P_i} \beta_m X_{im}$
SCL-based	
NL	• If i is in root nest $(1 - P_i)\beta_m X_{im}$ • If i is in N_k nest, $k\epsilon\{1, \ldots, M\}$ $[(1 - P_i) + (\mu_k^{-1} - 1)(1 - P_{i\|k})]\beta_m X_{im}, \ \forall i\epsilon\{1, \ldots, A\}$
MNL	$(1 - P_i)\beta_m X_{im}$

Table 2
Cross-elasticities of each pair of alternatives i,j∈{1,…,A},j≠i.

Model	Cross-elasticity
SCNL	$-\left[P_i + (\mu_{ij}^{-1} - 1)\frac{P_{i\|ij} P_{ij} P_{j\|ij}}{P_j}\right]\beta_m X_{im}$ • If i, j are not in the same nest $-P_i\beta_m X_{im}$ • If i, j are in N_k nest, $k \in \{1, \ldots, M\}$ $-\left[P_i + (\mu_k^{-1} - 1)\frac{P_{i\|ij} P_{ij} P_{j\|ij}}{P_j}\right]\beta_m X_{im}$
SCL-based	$-\left[P_i + (\mu^{-1} - 1)\frac{P_{i\|ij} P_{ij} P_{j\|ij}}{P_j}\right]\beta_m X_{im}$
NL	• If i, $j$are not in the same nest $-P_i\beta_m X_{im}$ • If i, j are in N_k nest, $k \in \{1, \ldots, M\}$* $-[P_i + (\mu_k^{-1} - 1)P_{i\|k}]\beta_m X_{im}$ (*) Modified from Papola (2004) for easier comparison
MNL	$-P_i\beta_m X_{im}$

$$F(\varepsilon_i) = \exp\left(-\sum_{\substack{j=1 \\ j\neq i}}^{A} \alpha_{i,ij} e^{-\varepsilon_i}\right) = \exp(-e^{-\varepsilon_i}), \ \forall i \in \{1, \ldots, A\} \tag{23}$$

The correlation between each pair of alternatives (i, j) is calculated by numerical integration from the marginal bivariant cumulative distribution function of the perturbations of the alternatives Abbe et al., 2007), which is expressed in Eq. (24). Eqs. (22) to ((24) have been deduced from the SCL equations (Bhat and Guo, 2004) by incorporating different μ_{ij} parameters in each nest of the NL structure.

$$H(\varepsilon_i, \varepsilon_j) = \exp\left\{-\left[(1-\alpha_{i,ij})e^{-\varepsilon_i} + (1-\alpha_{j,ij})e^{-\varepsilon_j}\right] - \left[\left(\alpha_{i,ij}e^{-\varepsilon_i}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij}e^{-\varepsilon_j}\right)^{1/\mu_{ij}}\right]^{\mu_{ij}}\right\} \forall i,j \in \{1, \ldots, A\}, \ j \neq i \tag{24}$$

From the approach proposed by Papola (2004), the unobserved correlation between alternatives can be approximated by Eq. (25). It is null when the alternatives are not in the same nest (like NL), and it is $\alpha_{i,ij}^{1/2}\alpha_{j,ij}^{1/2}(1-\mu_k^2)$ when both alternatives are in a N_k nest. In comparison with NL, the correlation between alternatives of the same nest is not constant and depends on the allocation parameters and, therefore, on the spatial metric used. With respect to SCL-based models, the correlation between alternatives depends on the nest to which both of them belong.

$$\widehat{Corr}(\varepsilon_i, \ \varepsilon_j) = \alpha_{i,ij}^{1/2}\alpha_{j,ij}^{1/2}\left(1-\mu_{ij}^2\right), \ \forall i,j \in \{1, \ldots, A\} \tag{25}$$

Considering a linear observed utility in the parameters, with coefficients β_m, the direct elasticity of the m-th regressor of alternative i, X_{im}, measures the expected percentage change in P_i for an increase of one percentage point of X_{im}. The cross-elasticity of X_{im} in P_j measures the expected percentage variation in P_j for an increase of one percentage point in X_{im}. Tables 1 and 2 show a comparison of direct and cross-elasticity of SCNL model with other GEV models.

Direct and cross-elasticity of the SCNL have the same formulation as MNL and NL models in the alternatives of the root nest. In comparison with SCL-based models, in the SCNL both formulations depend on the dissimilarity parameters of the nest for alternatives in the same nest. SCNL elasticities are equivalent to that of SCL-based models in the case that all alternatives are in the same nest. If spatial metric is based on contiguity, the cross-elasticity of non-contiguous alternatives for SCNL and SCL-based specification is equal

Table 3
Explanatory variables of the sample.

Name	Description	Type	Mean/ Distribution	Standard deviation
JT	Journey time in minutes between residential zone and employment zone.	Alternative –Specific of individual	7.57	3.93
FO	Number of non-EU foreigners in the residential zone (in thousands of people).	Alternative	0.461	0.224
HO	Natural log of the number of housing in the residential zone.	Alternative	7.858	0.228
PS	Dichotomous factor indicating that the residential zone has special prestige (subjective).	Alternative	NO: 95.51% YES: 4.49%	
PR	Average price of housing in the residential zone (in millions of €).	Alternative	0.28761	0.12670
SC	Number of primary and secondary education centers at a maximum distance of one km from residential zone centroid.	Alternative	2.22	1.70
WT	Average waiting time in minutes at public transport stops in the residential zone.	Alternative	10.51	0.78
H	Dichotomous factor indicating decider's high monthly net family incomes (more than 2500 €).	Individual	NO: 76.97% YES: 23.03%	

to that of MNL. SCNL and NL have the same cross-elasticity in alternatives from different nests (and equal to that of the MNL model). In alternatives belonging to the same nest, the second formulation in Table 2 of the cross-elasticity of the SCNL model looks quite similar to that of the NL model. However, this similarity is misleading. Unlike the NL model, the conditional probability of the SCNL model depends on the implicit effect of the assignment parameters.

The SCNL model requires a more complex design and estimation process than the NL and SCL-based models. Regarding the design, in comparison with NL, the SCNL model requires selecting a spatial metric appropriate to the zoning of the geographic area, and calculating the values of the metric in the zoning normally using a GIS; with respect to SCL-based models, SCNL requires designing a nested structure appropriate to the empirical context of application. Regarding estimation, although the SCNL model has the same unknown parameters as the NL model, its estimation process is more complex because the SCNL model is a reduced specification of the paired combinatorial nested logit model; in comparison with SCL-based models, the nested design increases the number of unknown parameters by a number equal to the number of nests minus one. When a SCNL model uses spatial metrics like the original SCL contiguity or the BSCL metric, the number of parameters of the SCNL is the same as that of the NL. The SCNL model is compatible with SCL-based specifications that require estimating additional parameters, as in the case of generalized spatially correlated logit (Sener et al., 2011).

SCNL model is compatible with a mixed GEV specification (MSCNL), which makes it possible to incorporate variations in decision-makers' preferences through an overlapping structure of random coefficients.

4. Empirical application of SCNL

This application of the SCNL model in the city of Santander (Spain) focuses on comparing the capacity to collect the spatial correlation between alternatives of this model, against the previous GEV models that are described in Section 2. To compare the explanatory power of the estimated models, we will use the statistical techniques of goodness-of-fit (GoF) (Hilbe, 2009). To compare how well the estimated models keep their predictive accuracy in a different sample, we will use statistical validation techniques as recommended by Parady et al. (2021). This application also shows the results of a proof of concept of MSCNL.

To avoid design bias, we use data and spatial elements designed to be applied with the models that are described in Section 2 (with the previous GEV approach) from research projects INTERLAND (see Ibeas et al., 2013) and TRANSPACE (Dell'Olio et al., 2016). The sample, the zoning and the nests structure are the same of Ibeas et al. (2013). Also, the spatial metric of the correlation between alternatives and the utility function, both the kernel GEV and its mixed specification are from Pérez-López et al. (2020).

Endogeneity has been established as a relevant issue in residential location choice models (Guevara and Ben-Akiva, 2006) and in other discrete choice models (Guevara and Ben-Akiva, 2012; Guerrero et al., 2021a; Guerrero et al., 2021b). When the alternatives are the specific dwelling to live, it is usually due to the omission of attributes of the dwelling that are correlated with the price and influence the choice. This misspecification will suppose that the impact of price in the choice process will not be correctly established and the estimators of the model parameters may be biased and inconsistent. It would be a serious problem for policy analysis. An indicator of the problem may be that the dwelling-unit price coefficient is non-significant, small or even positive. This problem can be addressed with the control function method (see Guevara and Ben-Akiva, 2012 for forecasting issues with that method and Guevara, 2015, for a critical assessment of several methods). This method requires to select adequate instrumental variables that are correlated with the price but are uncorrelated with the error term. For the kind of choice presented, those variables can be constructed as an average of the prices of other dwellings with similar observed attributes (other than price) and locating within certain vicinity (Guevara, 2010; Guevara and Ben-Akiva, 2012). In the application presented herein, the alternatives are not specific dwellings but areas, and the variable of the utility function is the mean price of dwellings in the area. In that situation, endogeneity due to omitted attributes of a specific dwelling that are correlated with price is not expected, although other sources of endogeneity cannot be discarded at all. As can be seen later, the results do not show indications of endogeneity, but it is advisable to carefully analyze this issue in models estimated for policy analysis (see Guerrero et al., 2021a; Guerrero et al., 2021b).

Fig. 1. Map of the scheme of the alternative residential zones in Santander.

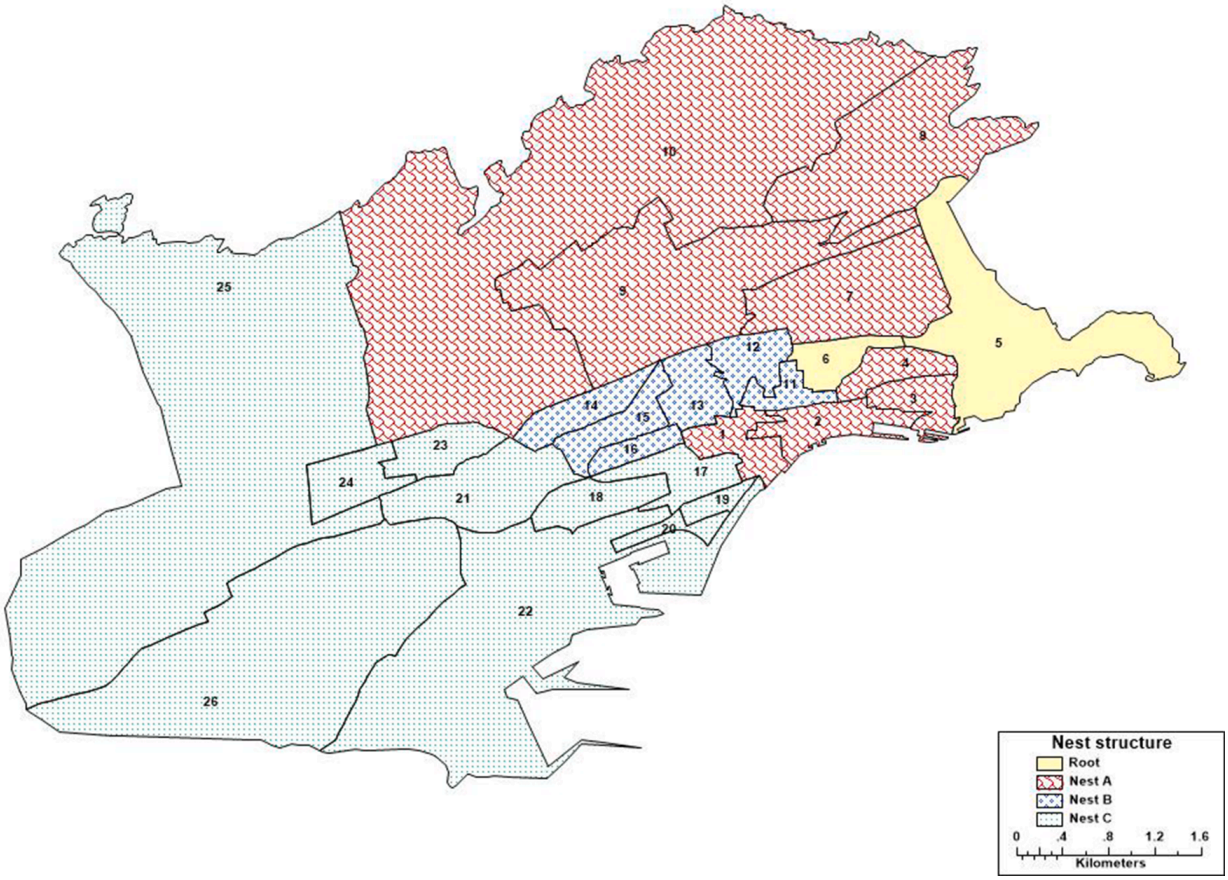

Fig. 2. Nest structure of alternatives.

4.1. Data and spatial elements

The sample contains 534 individual choices of deciders who live and work in the city of Santander. Table 3 includes descriptive statistics of the explanatory variables in the sample. The observed component of the utility function is a linear form which does not include alternative-specific constants. The regressors are the journey time (*JT*), the number of non-EU residents in the area (*FO*), the number of homes available in the area (*HO*) and the average house price in the area (*PR*), as well as the interactions between the high-

Table 4
Results of the estimation of the models NL, BSCL and BSCNL. LRT significant code: "**" if the test is significant at 1% significance level.

		BSCL			NL			BSCNL		
	Parameter	Value	SE	StC	Value	SE	StC	Value	SE	StC
Estimation	β_{JT}	−0.18	0.0518	−0.708	−0.104	0.0271	−0.409	−0.104	0.0271	−0.409
	β_{FO}	−1.12	0.447	−0.251	−1.00	0.300	−0.224	−0.892	0.284	−0.200
	β_{HO}	2.05	0.417	0.467	1.55	0.305	0.353	1.29	0.283	0.294
	β_{PR}	−2.63	0.596	−0.333	−2.17	0.426	−0.275	−1.99	0.399	−0.252
	$\beta_{PS \bullet H}$	1.58	0.396	0.136	1.22	0.262	0.105	1.02	0.260	0.088
	$\beta_{SC \bullet H}$	0.302	0.0752	0.453	0.210	0.0460	0.315	0.173	0.0469	0.259
	μ^{-1}	1.74	0.0979		1			1		
	μ_A^{-1}				1.25	0.147		3.11	1.42	
	μ_B^{-1}				1.26	0.128		2.27	0.773	
	μ_C^{-1}				1.05	0.089		1.49	0.430	
	No. est. par.	7			9			9		
GoF	*LL*	−1663.183			−1661.270			−1659.038		
	ρ^2	0.0441			0.0452			0.0464		
	$\bar{\rho}_H^2$	0.0420			0.0426			0.0438		
	AIC	0.04003			0.03998			0.0413		
	LRT-MNL	9.570	**		13.396		**	17.860		**
Val.	*PG-CV*	0.0437			0.0438			0.0440		

income level (*H*) with the prestige of the area (*PS*) and with the number of primary and secondary education centers near the centroid of the area (*SC*). The theoretically expected sign of estimated coefficients is negative in the case of *JT, FO* and *PR*, and positive in the case of *HO* and the interactions *PS* • *H* and *SC* • *H*. The mixed GEV specifies the coefficient of the regressor *SC* • *H* like a random variable with a normal distribution.

The zoning used in this section is based on the map of administrative areas of the city and resulted in 26 alternatives with very irregular shapes, as shown in Fig. 1. The spatial metric based on the length of the common border between pairs of alternatives is more efficient than the previous metrics in the context of alternatives with irregular shapes (Pérez-López et al., 2020). We are using this spatial metric in this application on a specification of a SCL-based model (BSCL) and on a specification of the SCNL model proposed in this paper (BSCNL).

The nests structure shown in Fig. 2 consists of three nests (A, B and C), leaving two alternatives in the root nest, that is, uncorrelated with each other or with other alternatives. Nest A consist of two different areas. This nested structure has a strong spatial component in order to capture the spatial correlation patterns between alternatives and those resulting from non-spatial characteristics, for the NL model. Thus, this design will test the ability of the SCNL model to capture spatial correlations between alternatives that have not already been identified by the nested structure, thereby verifying the ability of the SCNL model to complement the NL model.

4.2. Results analysis

All GEV models are estimated in this section by maximum likelihood (maximum simulated likelihood with 1000 iterations in the case of mixed specifications) using the Biogeme program (Bierlaire, 2003), applying the same DONLP2 (Spellucci, 1993) optimization algorithm in all estimates. In the models, the coherence of the signs of the estimated coefficients with those theoretically expected (described in the previous subsection) was verified. The relevance of the regressors was also checked using the asymptotic *t*-test at 5% of significant level of the corresponding estimated coefficients. The relative influence of the regressors was ordered using standardized coefficients, even though they are measured on different scales (in fact, in this case there are continuous, qualitative, and even dichotomous regressors). Different statistics are used to standardize the estimated coefficients (see Menard, 2004; Menard, 2011). In this case, we will use the statistic proposed by Menard (1995) and Agresti (1996) which is obtained by multiplying every estimated coefficient and the sample standard deviation of its regressor. The higher the absolute standardized value of the regressors is, the stronger their relative influence on the decision will be.

The GoF statistics calculated in every estimated model are the following likelihood ratio indexes: McFadden (ρ^2,1974), Horowitz ($\bar{\rho}_H^2$, 1983), and Akaike Information Criterion (AIC; Ben-Akiva and Swait, 1986). The last two penalize the number of parameters that have been estimated; thus, they are useful for comparing models that estimate different numbers of parameters. The AIC penalizes more the incorporation of parameters, which favors more parsimonious models. To compare the GoF of two of the estimated models, we will use the following procedure. If the two models estimate the same parameters, they will be compared using ρ^2. If the two models estimate different parameters, different criteria are used, depending on the situation. If the two models being compared are nested (where one model can be determined through linear constraints of the parameters of the other model) we use the likelihood ratio test (LRT). If the two models being compared are not nested we use the following criteria described in Horowitz (1983). First, the LRT of each model is performed with respect to a model with which both are nested (in this case, the null model or the MNL model with the same utility function). If one of the LRTs is significant and the other is not, the model with the significant LRT will be chosen. If both are significant, GoF statistics will be used to penalize the incorporation of additional parameters. Horowitz proposed $\bar{\rho}_H^2$, in this research, we will also analyze the AIC. If different conclusions are reached with each parameter, we will consider the result inconclusive.

A cross-validation process with $K = 10$ groups has been conducted, randomly partitioning the sample. The accuracy of the

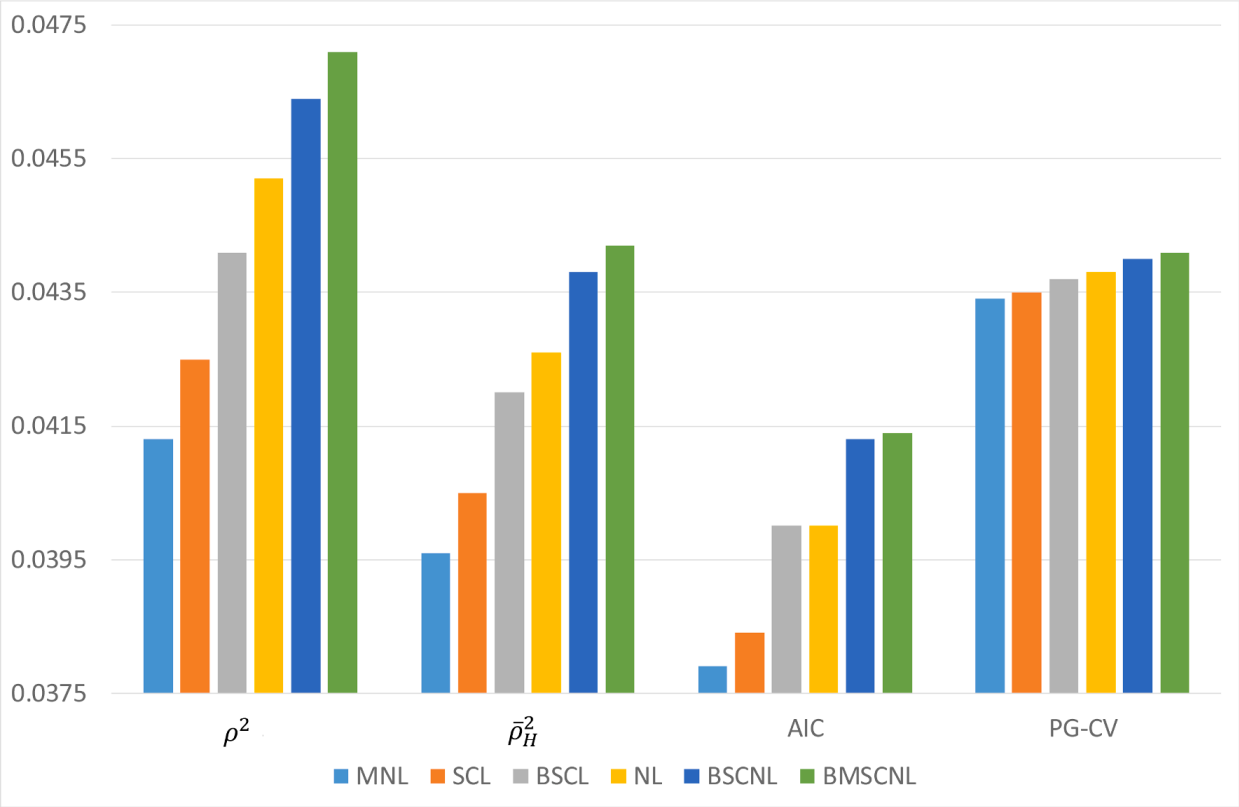

Fig. 3. GoF and validation statistics of estimated models.

predictions has been measured using *PG-CV* statistic, which is the geometric mean value of the ten values of the Predicting Geometric statistics. The Predicting Geometric values are obtained in each iteration of the cross-validation process with the geometric mean value of the correct probabilities (probability according to the model of the alternative chosen by the respondent) in the test sample, albeit using the model estimated with the training sample (Başar and Bhat, 2004; de Luca and Cantarella, 2009; Martínez-Pardo et al., 2020). In this research, we use the geometric mean, instead of the arithmetic mean that is commonly used. The geometric mean has better properties when using probability data.

Table 4 shows the estimation, GoF and validation results to compare NL, BSCL and BSCNL models. The estimation results of each model include, for each estimated parameter, the estimated value (Value), its standard error (SE) and its standardized coefficient (StC). The results table of each estimated model will also show the number of estimated parameters (No. est. par.) and the GoF and Validation results. In the three models, all the estimated parameters are significant, with signs coherent with the theoretically expected and with the same order from standardized coefficients. The most influential explanatory variable in the decision is *JT*, followed by *HO, SC • H, PR, FO* and finally *PS • H*. These results are similar to those obtained with the MNL and SCL models (Pérez-López et al., 2020).

The three models improve the GoF and the validation results of the MNL and SCL models. The NL model improves the validation results of the BSCL model, and some of the GoF results, but is not totally conclusive. They are not nested between them and have a different number of estimated parameters, therefore to compare their GoF, we use $\bar{\rho}_H^2$ and *AIC*. The NL model has a higher $\bar{\rho}_H^2$ value than BSCL but a lower *AIC* value. However, the proposed BSCNL model significantly improves the GoF and the validation results of both of them.

The mixed specification BMSCNL has one more parameter to estimate than its kernel BSCNL. BMSCNL improves the GoF and validation statistics results of BSCNL (but the LRT is not significant). The results are inconclusive, and for this reason it is not included in the table. Fig. 3 compares the values of GoF and validation measures assessed in the different models estimated, including MNL, SCL and BMSCNL that are not shown in table 4. The results of the BSCNL and BMSCNL specifications of the SCNL model improve those of the rest of the models in all the concepts considered.

5. Conclusions

The spatially correlated nested logit (SCNL) model is proposed in this research for spatial location choice modeling. The SCNL model makes it possible to combine the approach of nested alternatives of the nested logit model (NL) with that of extensions based on spatially correlated logit (SCL) model using metrics of spatial correlation between alternatives. Thanks to this combination, the SCNL model improves the explanatory and predictive power of the models that use the previous approaches.

The SCNL model can improve the explanatory and predictive power of the NL, even when the nested structure of the NL model has a strong spatial component, with the same unknown parameters (if the spatial metric does not require additional parameters, as the one

employed in the application). This improvement occurs when the spatial metric selected in the SCNL model is able to capture the spatial correlation between alternatives that the nested structure designed by the analyst for the NL model was unable to capture. The SCNL model achieves this improvement over the NL model by modeling the correlation between pairs of alternatives belonging to the same nest using spatial metrics.

The SCNL model can also improve the explanatory and predictive power of SCL-based models with the same metric of spatial correlation between alternatives. This improvement occurs when non-spatial correlation between alternatives is captured by the nested structure or when this nested structure is capable of detecting spatial correlation between alternatives in addition to that captured by the spatial metric. The SCNL model achieves this improvement over these models thanks to the flexibility of the dissimilarity parameter between nests, which makes it possible to model the correlation between pairs of alternatives with greater flexibility than SCL-based models. Furthermore, unlike SCL-based models, SCNL models do so considering not only spatial but also other correlation factors.

In addition, the SCNL model proposed in this research is compatible with mixed specifications of random coefficients to incorporate heterogeneity into decision-makers' preferences. The mixed SCNL model thus built may have better properties than the kernel SCNL in the presence of heterogeneous preferences.

The application of the different models analyzed empirically confirmed that the proposed SCNL model has good properties. The BSCNL model (using the common spatial border correlation metric) provided better empirical results than the SCL-based model using the same spatial metric and NL models, both in goodness-of-fit and in validation.

CRediT authorship contribution statement

Jose-Benito Perez-Lopez: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft. **Margarita Novales:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Alfonso Orro:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declarations of competing interest

None.

Acknowledgments

The authors acknowledge the financial support provided by the Government of Spain under the projects TRA2012–37659 and RTI2018–097924-B-I00, funded by MCIN/AEI/10.13039/501100011033 and by "ERDFA way of making Europe". Funding for open access charge: Universidade da Coruña/CISUG.

References

Abbe, E., Bierlaire, M., Toledo, T., 2007. Normalization and correlation of cross-nested logit models. Transp. Res. B 41, 795–808.
Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley. P. 129, New York.
Bahamonde-Birke, F.J., 2021. A brief discussion on the treatment of spatial correlation in multinomial discrete models. J Transp. Land Use 14 (1), 521–535.
Başar, G., Bhat, C., 2004. A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. Transp. Res. Part B 38, 889–904.
Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short-term travel decisions. In Hall R (Eds.). Handbook Transp. Sci. 5–34.
Ben-Akiva, M., Francois, B., 1983. Mu-homogenous Generalized Extreme Value Model. Working Paper.
Ben-Akiva, M., Swait, J., 1986. The Akaike likelihood ratio index. Trans. Sci. 20 (2), 133–136.
Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transp. Res. B Methodol. 38 (2), 147–168.
Bierlaire, M., 2003. BIOGEME: a free package for the estimation of discrete choice models. Ascona, Switzerland. In: 3rd Swiss transportation research conference.
Bolduc, D., 1992. Generalized autoregressive errors: the multinomial probit model. Transp. Res. B 26, 155–170.
Chu, C., 1981. *Structural Issues and Sources of Bias in Residential Location and Travel Mode Choice Models*. ***Unpublished Ph.D. Dissertation***. Department of Civil Engineering, Northwestern University, USA.
Chu, C., 1989. A paired combinatorial logit model for travel demand analysis. In: Proceedings of the Fifth World Conference on Transportation Research. Ventura, CA, pp. 295–309 vol. 4.
Daganzo, C., 1979. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press, New York.
Daly, A.J., Zachary, S., 1978. Improved multiple choice models. In: Hensher, DA, Dalvi, MQ (Eds.), Determinants of Travel Choice. Westmead, Saxon House, pp. 335–357.
de Luca, S., Cantarella, G.E., 2009. Validation and comparison of choice models. In: Saleh, W., Sammer, G. (Eds.), Travel Demand Management and Road User Pricing: Success, Failure and Feasibility. Ashgate publications, pp. 37–58.
Dell'Olio, L., Cordera, R., Ibeas, A., (Eds) Alonso, A., Alonso, B., Barreda, R., Comi, A., Coppola, R., González, E., Monzón, A., Moura, J., Nogués, S., Nuzzolo, A., Orro, A., Papa, E., Pérez-López, J.-B., Reques, P., Sañudo, R., Wang, Y., 2016. *Land Use - transport interaction models. The TRANSPACE Model*. 1st edn. Santander: GIST.
Domencich, T.A., McFadden, D., 1975. Urban Travel demand: a Behavioural analysis. American Elsevier, New York.
Garrido, R.A., Mahmassani, H.S., 2000. Forecasting freight transportation demand with the space-time multinomial probit model. Trans. Res. B 34, 403–418.
Guerrero, T.E., Guevara, C.A., Cherchi, E., Ortúzar, J.D.D., 2021a. Forecasting with strategic transport models corrected for endogeneity. Transportmetrica A 1–28. https://doi.org/10.1080/23249935.2021.1891154.
Guerrero, T.E., Guevara, C.A., Cherchi, E., Ortúzar, J.D.D., 2021b. Addressing endogeneity in strategic urban mode choice models. Transportation, 48, 2081–2102.
Guevara, C.A., Ben-Akiva, M., 2006. Endogeneity in residential location choice models. Transp. Res. Rec. (1) 197760-66.
Guevara, C.A., Ben-Akiva, M., 2012. Change of scale and forecasting with the control-function method in logit models. Transp. Sci. 46 (3), 425–437.
Guevara, C.A., 2010. Endogeneity and Sampling of Alternatives in Spatial Choice Models (Doctoral dissertation. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA. Ph. D. Thesis.

Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transp. Res. Part A 82, 240–254.
Hess, S., Bierlaire, M., Polak, J., 2005. Capturing taste heterogeneity and correlation structure with Mixed GEV models. In: Scarpa, R, Alberini, A (Eds.), Applications of Simulation Methods in Environmental and Resource Economics. Springer Publisher, Dordrecht, The Netherlands, pp. 55–76 chapter 4.
Hilbe, J.M., 2009. *Logistic Regression Models*. CRC press. ISBN 978-1-138-10671-0.
Horowitz, J., 1983. Statistical comparison of non-nested probabilistic discrete choice models. Trans. Sci. 17, 319–350.
Ibeas, A., Cordera, R., Dell'Olio, L., Coppola, P., 2013. Modeling the spatial interactions between workplace and residential location. Transp Res A 49, 110–122.
Johnson, N., Kotz, S., 1970. Distributions in Statistics: Constinuous Univariate Distributions. John Wiley, New York. Chapter 21.
Koppelman, F.S., Wen, C.H., 2000. The paired combinatorial logit model: properties, estimation and application. Transp. Res. Part B 34 (2), 75–89.
Martínez-Pardo, A., Orro, A., Garcia-Alonso, L., 2020. Analysis of port choice: a methodological proposal adjusted with public data. Transp. Res. Part A 136, 178–193.
McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P (Ed.), Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics 105–142.
McFadden, D., 1978. Modelling the choice of residential location. In: Karlqvist, A, Jundqvist, L, Snickars, F, Weibull, J (Eds.), Spatial Interaction Theory and Planning Models. North Holland, Amsterdam, pp. 75–96.
Menard, S., 1995. Applied Logistic Regression Analysis. Thousand Oaks, CA: Sage, p. 48. P.
Menard, S., 2004. Six Approaches to Calculating Standardized Logistic Regression Coefficients. Am Stat 58 (3), 218–223.
Menard, S., 2011. Standards for standardized logistic regression coefficients. Soc. Forces 89 (4), 1409–1428.
Pagliara, F., Wilson, A., 2010. The State-of-the-Art in Building Residential Location Models. Residential Location Choice. Springer.
Papola, A., 2004. Some developments on the cross-nested logit model. Transp. Res. B Methodol. 38, 833–851.
Parady, G., Ory, D., Walker, J., 2021. The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: a review of validation practices in the transportation academic literature. J. Choice Modell. 38, 100257.
Pérez-López, J.-.B., Novales, M., Varela-Garcia, F.-.A., Orro, A., 2020. Residential location econometric choice modeling with irregular zoning: common border spatial correlation metric. Netw. Spatial Econ. 20, 785–802.
Sener, I.N., Pendyala, R.M., Bhat, C.R., 2011. Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. J. Transp. Geogr. 19, 294–303.
Small, K.A., 1987. A discrete choice model for ordered alternatives. Econometrica 55 (2), 409–424.
Spellucci, P., 1993. *DONLP2 Users Guide*, Dept. of Mathematics, Technical University at Darmstadt, 64289 Darmstadt, Germany.
Takahashi, K., 2019. Local relaxation of constraints on dissimilarity parameters in the generalized nested logit model. Int. J. Japan Assoc. Manag. Syst. 11 (1), 73–80. December 2019.
Torrens, P.M., 2000. How Land-Use Transportation Models Work. Centre for Advanced Spatial Analysis, London.
Train, K.E., 2009. Discrete Choice Methods With Simulation. Cambridge University Press, New York, New York, USA.
Thurstone, L.L., 1927. A law of comparative judgment. Psychol Rev 34, 273–286.
Vovsha, P., 1997. The cross-nested logit model: application to ode choice in the Tel-Aviv metropolitan area. Transp. Res. Rec. 1607, 6–15.
Weiss, A., Hasnine, S., Habib, K.N., 2019. A comparative study of alternative methods for capturing spatial correlations in discrete choice models through an empirical application on school choice location modelling. In: Paper presented at the 98th Annual Meeting of the Transportation Research Board. Washington, DC, January 13–17 (No. 19-05270).
Wen, C.H., Koppelman, F.S., 2001. The generalized nested logit model. Transp. Res. Part B 35 (7), 627–641.
Williams, H.C.W.L., 1977. On the formation of travel demand models and economic evaluation measures of user benefit. Environ. Plann. 9A, 285–344.

Transportation Research Part B xxx (xxxx) xxx

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Erratum

Erratum to 'Spatially correlated nested logit model for spatial location choice' [Transportation Research Part B: Methodological Volume 161, July 2022, Pages 1-12]

Jose-Benito Perez-Lopez [a,*], Margarita Novales [b], Alfonso Orro [b]

[a] Universidade da Coruña, Group of Railways and Transportation Engineering, Department of Economics, Facultad de Economía y Empresa, Campus de Elviña, 15071 A, Coruña, Spain

[b] Universidade da Coruña, Group of Railways and Transportation Engineering, Department of Civil Engineering, ETS Ingenieros de Caminos, Canales y Puertos, Elviña, 15071 A, Coruña, Spain

The publisher regrets Table 1 of the above article was published incomplete, with two equations missing in the SCNL model row. The complete table is given below:

The publisher would like to apologise for any inconvenience caused.

Table 1
Direct elasticities of each alternative i∈{1,…,A}.

Model	Direct elasticity
SCNL	$\frac{\sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i\|ij}P_{ij}[(1-P_i)+(\mu_{ij}^{-1}-1)(1-P_{i\|ij})]}{P_i}\beta_m X_{im}$ • If i is in root nest $(1-P_i)\beta_m X_{im}$ • If i is in N_k nest, $k\in\{1,\ \ldots,\ M\}$ $\frac{\sum_{j\notin N_k} P_{i\|ij}P_{ij}(1-P_i)+\sum_{\substack{j\in N_k \\ j\neq i}} P_{i\|ij}P_{ij}[(1-P_i)+(\mu_k^{-1}-1)(1-P_{i\|ij})]}{P_i}\beta_m X_{im}$
SCL-based	$\frac{\sum_{\substack{j=1 \\ j\neq i}}^{A} P_{i\|ij}P_{ij}[(1-P_i)+(\mu^{-1}-1)(1-P_{i\|ij})]}{P_i}\beta_m X_{im}$
NL	• If i is in root nest $(1-P_i)\beta_m X_{im}$ • If i is in N_k nest, $k\in\{1,\ \ldots,\ M\}$ $[(1-P_i)+(\mu_k^{-1}-1)(1-P_{i\|k})]\beta_m X_{im},\ \forall i\in\{1,\ \ldots,\ A\}$
MNL	$(1-P_i)\beta_m X_{im}$

DOI of original article: https://doi.org/10.1016/j.trb.2022.05.007.

* Corresponding author.
E-mail address: benito.perez@udc.es (J.-B. Perez-Lopez).

https://doi.org/10.1016/j.trb.2022.06.008