This is a post-peer-review, pre-copyedit version of an article published in Networks and Spatial Economics (2020) 20:785–802. The final authenticated version is available online at https://doi.org/10.1007/s11067-020-09495-5

Residential location econometric choice modeling with irregular zoning: common border spatial correlation metric

José-Benito Pérez-López (http://orcid.org/ 0000-0003-0487-6141)

Universidade da Coruña, Group of Railways and Transportation Engineering, Department of Economics, Facultad de Economía y Empresa, Campus de Elviña, 15071 A Coruña, Spain. Corresponding author: benito.perez@udc.es

Margarita Novales (http://orcid.org/0000-0003-0581-6933)

Universidade da Coruña, Group of Railways and Transportation Engineering, Departament of Civil Engineering, ETS Ingenieros de Caminos, Canales y Puertos, Elviña, 15071 A Coruña, Spain.

Francisco-Alberto Varela-García (http://orcid.org/0000-0001-9705-2154)

Universidade da Coruña, cartoLAB, Grupo de Visualización Avanzada e Cartografía, Departament of Civil Engineering, ETS Ingeniería de Caminos, Canales y Puertos, Campus de Elviña, 15071 A Coruña, Spain.

Alfonso Orro (http://orcid.org/0000-0003-0688-3417)

Universidade da Coruña, Group of Railways and Transportation Engineering, Departament of Civil Engineering, ETS Ingenieros de Caminos, Canales y Puertos, Elviña, 15071 A Coruña, Spain.

Keywords: mixed GEV models; spatial correlation; econometric choice modeling; residential location choice modeling; land use-transport interaction models; behavioral demand modeling.

Abstract

Residential location choice (RLC) predicts where and how people choose their residential location in the framework of land use–transport interaction models (LUTI). This paper seeks an efficient RLC model in the context of irregular zoning of location alternatives. The main current proposals in the field are discrete choice models. In RLC modeling, the alternatives are spatial units, and spatially correlated logit (SCL) is an efficient approach when the analyst cannot pre-define groups of alternatives that efficiently reflect the systematic substitution patterns among the alternatives. The SCL uses the spatial information on the contiguity of the zones to determine spatial correlation among the alternatives. Urban residential location choice usually uses administrative zoning, which is very irregular in many cities (mainly historic cities); however, SCL is not efficient in this context owing to the limitations of the binary contiguity spatial variable employed as a spatial correlation metric (SCM). This paper proposes an extension of the mixed SCL model, with an SCM based on the proportion of common border length in contiguous zones, which is more efficient in the irregular urban zoning context. The proposed model is applied to an urban case study of LUTI RLC

modeling with irregular zoning, based on the administrative divisions of the city of Santander (Spain) and is shown to be empirically more efficient than the previous approaches.

1. Introduction

Public administrations and companies must plan infrastructure supply and services in advance, including their location and capacity, to manage resources efficiently. To that end, they need to predict demand with the highest possible reliability. The demand models of activity-transportation type are based on the geographic distribution of population and activities. This paper deals with the spatial and behavioral aspects of modeling urban land use and transportation demand. The most extensively-used current approach involves mathematical simulation models of the interaction between land uses and transportation (LUTI), with a structure of the zonal space and two main components: a subsystem of land uses and that of transportation (Torrens 2000).

In LUTI models, the choice of the residential location (RLC) is part of the subsystem of land uses (together with the location of activities). It influences the transportation subsystem due to the necessity of people's mobility and goods transport (Wegener 1994) and is influenced by it through the accessibility or ease with which an activity can be accessed from a given location using a transportation system (Geurs and van Wee 2004). Therefore, LUTI models need to develop RLC models, which aim to predict where and how people choose their residence location. The study of these models dates to Alonso's (1960; 1964) work.

For any family, choosing the characteristics of their residence is crucial, as it determines their price of purchase or rent (in most cases it is one of the most important components of the family finances), and the available services (sometimes, even the type of social activity and personal relationships). For example, the array of transportation services available will influence the time available for other activities.

Location is one of the most essential characteristics of a residence and the most commonly-used approach in LUTI RLC modeling is based on discrete choice models (DCM). This approach, consistent with random utility theory (Thurstone 1927), groups the available locations in spatial residential zones. The economically rational decision maker chooses the zone that maximizes his/her perceived utility. This utility depends on the characteristics of the residences themselves or on the services in the zone, such as transportation, quality of social life, and so on. The equation of a DCM with I available alternatives is as follows:

$$U_{ni}=V_{ni}+\varepsilon_{ni},\ i=1,\ldots,I \tag{1}$$

where dependent variables (U_{ni}) are the unobserved utility perceived by each decision maker (n) for each alternative (i). Some characteristics can be observed: V_{ni} is the observed component with a parametric function of the observed explanatory variables. However, others cannot be observed: ε_{ni} is a perturbation term summarizing the contribution of the unobserved variables.

Each econometric DCM assumes a function of the observed utility and a joint distribution of the perturbations and these lead to an expression of the probability that a decision maker n chooses alternative i from I available P_{ni}. The most extensively-used

approaches are linear-in-parameter utilities and extreme value for the joint distribution of unobserved utilities, called logit choice models.

The simplest and the most common logit choice model is multinomial logit (MNL) (McFadden 1974; Domencich and McFadden 1975), which assumes that unobserved utilities are uncorrelated and homoscedastic, according to the extreme value type I (or Gumbel) distribution. The variance-covariance matrix of MNL is a scalar matrix

$$\Omega = \frac{\pi}{\mu\sqrt{6}} \cdot I_I \tag{2}$$

where $\mu \in (0,1]$ is the scale parameter of the Gumbel distribution. The MNL has a closed form; that is, probability is calculated without integration as follows:

$$P_{ni} = \frac{y_{ni}^{\mu}}{\sum_{j=1}^{I} y_{nj}^{\mu}} \tag{3}$$

where $y_{ni} = e^{V_{ni}}$. To identify the coefficients of observed utility, the properties "the scale of utility is arbitrary" and "only differences in utility matter" (Train 2003: 23) allow to innocuously fix the variance of utility (and, hence, μ) to an arbitrary value if the data are not from different sources (i.e., different groups may have different variances). The usual normalization of μ in MNL is equal to one. This paper will perform this normalization in every logit model described.

In the context of LUTI, the samples for estimation are mainly based on cross-sectional data from revealed preferences surveys. This implies that temporal correlations and correlations among the responses of the same individual are absent. Nevertheless, there will be correlation among the unobserved responses of the decision makers due to the dependency among the alternatives. In RLC modeling, the alternatives are spatial units; there is, therefore, spatial dependence at the very least.

Spatial dependence among spatial units could be defined as the existence of a functional relationship between events at a certain point in space and those elsewhere (Moreno and Vayá 2000), that is, the values adopted by a variable in a particular zone and in neighboring zones. The spatial correlation among residential zones in the unobserved utility refers to the presence of systematic patterns of substitution among neighbors in the choice process (Anselin 1988; Hunt, Boots and Kanaroglou 2004).

Train (2003: 46) describes the limitations of MNL in the context of relaxing two assumptions: uncorrelation and homoscedasticity. The unobserved heterogeneity in the preferences of the decision makers, that can also be present in LUTI context, can be introduced in the model with superimposed mixed specifications of random coefficients. Section 2 reviews the logit models that relax both restrictions, with special emphasis on spatial correlation, and compares them in the RLC modeling context.

In Section 3, we introduce the mixed generalized extreme value (MGEV) (Chernew et al. 2003; Bhat and Guo 2004; Hess, Bierlaire, and Polak 2005) family of models, with kernel spatially correlated logit (SCL) (Bhat and Guo 2004) that incorporates spatial correlation based on contiguity. However, this model has limitations in RLC modeling when zonings based on administrative areas are highly irregular. This paper proposes an extension of SCL with a spatial correlation metric (SCM) that is more efficient than the SCL and other approaches and is consistent with the assumptions of SCL. Section 4 empirically applies the proposed specification to a case study of RLC with irregular zoning and compares its

performance with the other approaches reviewed in the paper. We present the conclusions in Section 5.

2. Logit choice models

Mixed models and the MGEV family of models are the main approaches to incorporate unobserved correlation among alternatives and unobserved heterogeneity of the decision makers. In this section, we review the evolution of these models and analyze their adaptation to RLC modeling.

2.1. Mixed models

The advances in statistical simulation techniques have increased the use of the mixed models approach, where the probabilities are calculated as logit probabilities integrated over a density of parameters (Train 2003: 139). This integral does not have a closed-form solution but can be approximated using simulation, so maximum simulated likelihood is used for estimating the parameters. Of these models, the most well known is the mixed MNL (MMNL) or mixed logit, which can approximate any DCM derived from random utility maximization as closely as one pleases (McFadden and Train 2000). These models can be specified from two conceptually different but mathematically equal approaches (see Ben-Akiva and Bierlaire 2003):

- Random coefficients logit (RCL) captures the unobserved heterogeneity of individuals' preferences. The coefficients of the utility function, that reflect the valuations of the observed variables, are considered random variates with a pre-determined distribution and parameters to be estimated (for example, Bhat 2000).
- Error components logit (ECL) allows focusing on the analysis of the substitution patterns between alternatives based on the structure of the non-Gumbel component of the error (for example, Bhat 1998).

Both approaches can be combined to capture unobserved heterogeneity (RCL) and correlation between alternatives (ECL). Nevertheless, this approach for correlation between alternatives is not considered appropriate in the RLC models, where there is a high number of alternatives zones and, therefore, an extremely high number of error parameters to be specified.

2.2. GEV models

Williams (1977), Daly and Zachary (1978), and McFadden (1978) defined, independently, the nested logit (NL), which allows a richer pattern of substitution among alternatives than the MNL maintaining a closed-form probability, though with some additional parameters estimated jointly with the coefficients of observed utility.

The NL model introduced the concept of nests of alternatives, pre-defined by the analyst, to incorporate correlation among the alternatives in a logit model. Each nest is a group of alternatives, where each one has correlated unobserved utility with the alternatives of the same nest, but uncorrelated with the alternatives of other nests. Each alternative belongs to only one nest. All the alternatives of the same nest have the same scale parameter $\mu_k \in (0,1]$, which reflect dissimilarity among the alternatives of nest k, because it is inversely related to the correlation between alternatives. The correlation among pairs of alternatives is:

$$Corr(U_i, U_j) = (1 - \mu_k^2) \cdot \delta_k(i,j) \quad (4)$$

where $\delta_k(i,j) = 1$, if the alternatives *i* and *j* belong to nest k, else 0.

Different nests can have the same dissimilarity parameter but this requires justification. In the case of every dissimilarity parameter being equal to one, NL collapses into an MNL.

McFadden generalized logit models with nests of alternatives in the generalized extreme value (GEV) (McFadden 1978) family of logit models. The GEV models are DCM with unobserved utility of the alternatives (ε_{n1}, …, ε_{nJ}) jointly distributed GEV, according to the generating function $G(y_{n1}, \dots, y_{nI})$, which must fulfill a series of properties. The cumulative distribution function is given by:

$$F_{\varepsilon_{n1},\dots,\varepsilon_{nI}}(y_{n1}, \dots, y_{nI}) = e^{-G(e^{-y_{n1}},\dots,e^{-y_{nI}})} \quad (5)$$

The parameters can be jointly estimated through maximum likelihood. The probability of choice of alternative *i* for an individual *n* in GEV models maintains a closed-form expression:

$$P_{ni} = \frac{y_{ni} \cdot \frac{\partial G(y_{n1},\dots,y_{nI})}{\partial y_{ni}}}{G(y_{n1},\dots,y_{nI})} \quad (6)$$

An MNL model is defined by the following GEV generating function:

$$G(y_{n1}, \dots, y_{nI}) = \sum_{i=1}^{I} y_{ni} \quad (7)$$

An NL model is defined by the following GEV generating function:

$$G(y_{n1}, \dots, y_{nI}) = \sum_{k=1}^{K} \left(\sum_{i \epsilon Nest_k} {y_{ni}}^{1/\mu_k} \right)^{\mu_k} \quad (8)$$

where *I* is the number of alternatives and *K* is the number of nests (decided by the analyst). The NL approach was further made flexible with different specifications of a cross-nested logit (CNL) (Small 1987; Vovsha 1997; Ben-Akiva and Bierlaire 1999) and the generalized NL (GNL) (Wen and Koppelman 2001). Both CNL and GNL are GEV models that allow each alternative to be in more than one nest and, therefore, permit a richer pattern of substitution among alternatives than the NL with a closed-form probability. In this paper, we will use the GNL specification. GNL uses the following GEV generating function:

$$G(y_{n1}, \dots, y_{nI}) = \sum_{k=1}^{K} \left(\sum_{i \epsilon Nest_k} (\alpha_{ik} \cdot y_{ni})^{1/\mu_k} \right)^{\mu_k} \quad (9)$$

where $\alpha_{ik} \geq 0$ are the allocation parameters (or logsum) of each alternative *i* in each nest *k*. GNL adds the restriction $\sum_{k=1}^{K} \alpha_{ik} \ \forall i$ to facilitate the interpretation of α_{ik} as the portion of each alternative *i* assigned to each nest *k*. GNL are hierarchically nested with MNL (like NL) because they collapse into an MNL when every allocation parameter is equal to 1.

The correlation between two alternatives in a GNL model is obtained from the joint cumulative distribution function, which cannot be written in a closed form; it is therefore calculated using numerical integration. Papola (2004) proposed the following approximation to value the correlation:

$$\widehat{Corr}(U_i, U_j) = \sum_{k=1}^{K} \alpha_{ik}^{1/2} \alpha_{jk}^{1/2} (1 - \mu_k^2) \quad (10)$$

Hence, the correlation among alternatives not only depends on the dissimilarity parameters of the nests, but also on the allocation parameters.

The allocation parameters can be estimated jointly with the rest of the parameters (observed utilities' coefficients and dissimilarity parameters) or can be pre-defined by the analyst. In the RLC models, there are many alternatives, so the first approach has a limitation in terms of a high number of parameters to be estimated. In a model with *I* alternatives and *K* nests, where each one belongs to all nests, there are *I·K* allocation parameters to estimate. The approach of pre-defined allocation parameters depends on the chosen criterion.

In any case, the CNL model has the same dependence as the NL on the nest structure pre-defined by the analyst. It has to efficiently reflect the patterns of substitution among the alternatives.

The paired combinatory logit model (PCL) (Chu 1989; Koppelman and Wen 2000) does not need a pre-defined structure of nests because it considers them as pairs of alternatives. A paired GNL model (PGNL) (Wen and Koppelman 2001) generalizes the PCL model by incorporating the allocation parameters and specifies it as a GNL model with the following GEV generating function:

$$G(y_{\mathrm{n}1}, \dots, y_{nI}) = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \left(\left(\alpha_{i,ij} y_{ni}\right)^{1/\mu_{ij}} + \left(\alpha_{j,ij} y_{nj}\right)^{1/\mu_{ij}} \right)^{\mu_{ij}} \qquad (11)$$

where every pair of alternatives i≠j are a nest, with $\alpha_{i,ij}$ and $\alpha_{j,ij}$ (not necessarily equal) allocation parameters and dissimilarity parameter μ_{ij}. PGNL is hierarchically nested with MNL (like NL and CNL) because when all allocation parameters are equal to one, it collapses into an MNL.

The limitation of PGNL is the high number of nests when there are many alternatives, like in the case of RLC modeling and consequently, a high number of allocation parameters to be estimated or pre-defined by the analyst.

2.3. Mixed GEV models

GEV models are logit models that allow unobserved correlation among alternatives maintaining a closed-form probability. They can have different patterns of substitution based on a structure of nests of alternatives, defined by the analyst, or nesting each pair of different alternatives. However, this approach is compatible with a mixed superimposed specification of random coefficients, named the MGEV family of models, that allows unobserved heterogeneity in the preferences of decision makers.

The MGEV model assumes that the coefficients of the observed utility vector β follows a multivariate random distribution (usually normal) with vector θ of underlying moment parameters and multivariate density function *f*. Then:

$$P_{ni} = \int_{-\infty}^{\infty} (P_{ni}|\beta) f(\beta|\theta) d\beta \qquad (12)$$

The MGEV model can be estimated using maximum simulated likelihood method. Under rather weak regularity conditions, these estimators are consistent and asymptotically efficient and normal (Hajivassiliou and Ruud 1994; Lee 1992). The simulated log-likelihood function is:

$$SL(\theta) = \sum_n \sum_{i=1}^{I} y_{ni} log \tilde{P}_{ni}(\theta) \qquad (13)$$

where $\tilde{P}_{ni}(\theta)$ is the average of the realizations of the probability applying simulation techniques; that is, an unbiased estimator of $P_{ni}(\theta)$.

3. Spatial Logit models

In the logit choice models, where the alternatives are spatial zones, the correlation among alternatives have two components: spatial correlation and non-spatial correlation (see introduction in this paper). In addition to the information obtained from the sample, spatial variables are considered. These variables are used to include information on spatial correlation in the model before the estimation process.

3.1. Spatially Correlated Model

Bhat and Guo (2004) proposed an MGEV model, MSCL, with a kernel that does not depend on a nest structure pre-defined by the analyst because it is a specification PGNL. The correlation among the alternatives in a PGNL (and in all GNL, see equation 10) is defined from both the allocation and dissimilarity parameters. The MSCL kernel, SCL, separates both the components of correlation among the alternatives. The model uses a metric of spatial variables to establish pre-calculated allocation parameters; it is named spatial correlation metric (SCM) in this paper. The non-spatial correlation is collected through the dissimilarity parameters, which are estimated jointly with the rest of the parameters. This property allows defining and analyzing both correlation components separately.

The SCL model assumes a spatial association scheme based on contiguous zones (when they have a common border) and proposes an SCM based on the spatial information of contiguity among the zones. It is based on the binary contiguity variable among pairs of zones, δ_{ij}, which is equal to 1 if zones *i* and *j* are different and contiguous and 0 otherwise. This is the simplest spatial spillover variable from spatial statistics and is pre-calculated using a map, but the analyst does not need the support of a geographic information system (GIS). The allocation parameters proposed in SCL are the "space weights" to normalize δ_{ij}:

$$\alpha_{i,ij} = \frac{\delta_{ij}}{\sum_{k=1}^{I} \delta_{ik}} \tag{14}$$

The nests with non-zero allocation parameters are pairs of contiguous zones. It is assumed that the allocation parameters are all the same, named μ. We will name this hypothesis the equality assumption of dissimilarity parameters. The GEV generating function of an SCL is:

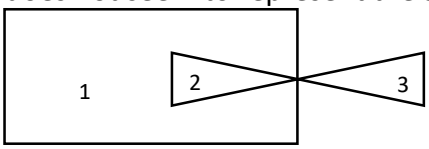
$$G(y_{\mathrm{n}1}, \dots, y_{nI}) = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \left(\left(\alpha_{i,ij} y_{ni}\right)^{1/\mu} + \left(\alpha_{j,ij} y_{nj}\right)^{1/\mu} \right)^{\mu} \tag{15}$$

The simple SCM of an SCL can be efficient in collecting the spatial correlation among alternatives in the context of regular zoning. However, in the LUTI RLC context, the zoning is usually based on administrative zones, which are sometimes highly irregular, particularly in historic cities or cities that have grown without urban planning.

Irregular zoning implies spatial zones with different shapes and sizes, so there could be multiple possibilities of contiguity. In this context, the simple and rigid SCM of an SCL does not seem efficient for collecting spatial correlation. For example, in the irregular

Figure 1. Example of irregular zoning where SCL is not efficient in collecting spatial correlation

zoning showed in Figure 1, the SCM of the SCL values the pairs 2-1 and 2-3 with equal magnitude but that does not seem to represent the spatial correlation.



In this situation, it may be necessary to employ a different approach to collect spatial correlation. One possible approach is to extend the SCL with a different SCM that uses more complex but flexible spatial variables.

3.2. Distance-based SCL

A distance-based SCL model (DSCL) (Sener, Pendyala and Bhat 2011) is an extension of SCL using this approach. It proposes the following SCM to extend an SCL:

$$\alpha_{i,ij} = \frac{d_{ij}^{\phi}}{\sum_{k=1}^{J} d_{ij}^{\phi}} \tag{16}$$

where d_{ij} is the Euclidean distance between the centroids of zones *i* and *j* (requires GIS support) and a scalar ϕ to be estimated. This SCM should be efficient in a regular zoning context but not necessarily with irregular zoning. For example, Figure 2 shows an actual irregular urban zoning where the red points are the centroids of each zone. With this SCM $d_{25,10}^{\phi} = d_{25,13}^{\phi}$, but it does not seem to represent the spatial correlation between these zones.

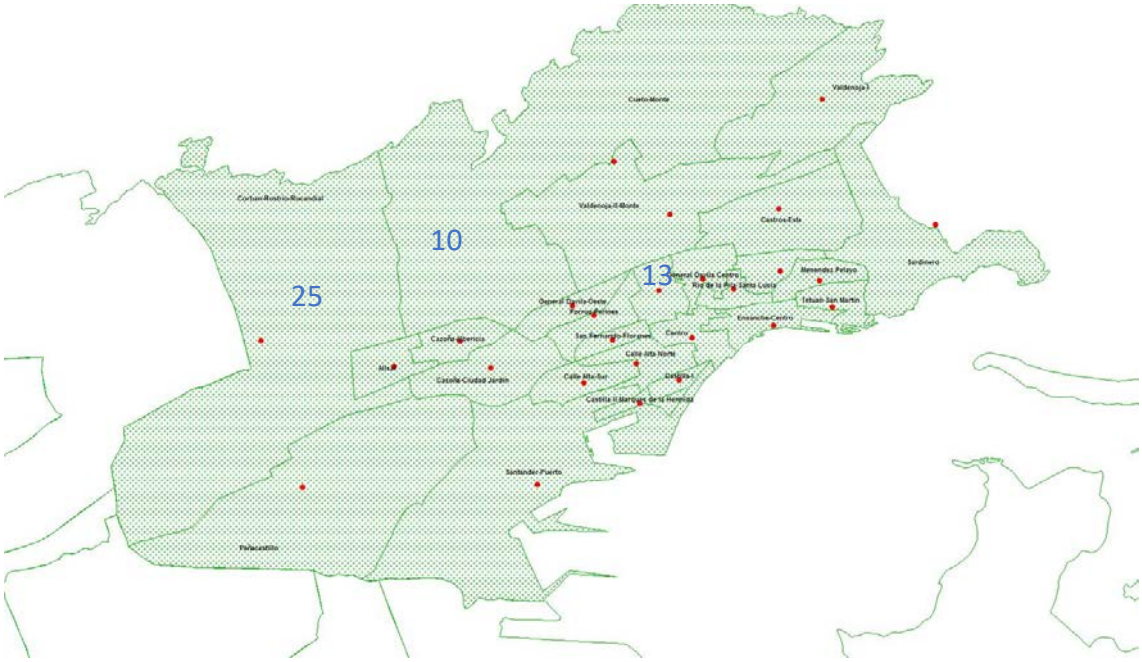


Figure 2. Actual irregular administrative urban zoning of Santander (Spain)

The equality assumption of dissimilarity parameters of SCL should be reviewed for DSCL. While in SCL only the contiguous alternatives have not-null allocation parameters, in DSCL all the pairs of alternatives have it.

This SCM allows allocation parameters to not be completely pre-determined when ϕ is the parameter to be estimated. In this case, the DSCL does not maintain different parameters for spatial and non-spatial correlation, like the SCL. Pérez-López and Orro

(2016) presented a specification of DSCL, with fixed parameter $\phi = -2$ named gravitational DSCL (GDSCL). It is more parsimonious than the DSCL and maintains different parameters for spatial and non-spatial correlation, like the SCL. The allocation parameters are the following:

$$\alpha_{i,ij} = \frac{d_{ij}^{-2}}{\sum_{k=1}^{I} d_{ij}^{-2}} \tag{17}$$

3.3. Common border length-based SCL

This paper seeks an efficient logit choice model in the LUTI RLC modeling context with irregular zoning. The spatial correlation among alternatives could depend on its distance or the length of the common border (Stetzer 1982; Anselin and Rey 1991; Florax and Rey 1995). The use of the common border length between pairs of zones has been explored in the field of spatial autocorrelation in linear regression models (Dacey 1969; Ma et. al 1997). The common border approach maintains the contiguity requirement, so it is consistent with the equality assumption of dissimilarity parameters. Moreover, in case of irregular zones, it seems more efficient than the SCM of SCL and DSCL because it considers the degree of contiguity among zones, regardless of the form.

This paper proposes an extension of SCL with a SCM based on the proportion of common border length between contiguous zones:

$$\alpha_{i,ij} = \frac{\beta_{ij}}{\sum_{k=1}^{I} \beta_{ik}} \tag{18}$$

where β_{ij} is the common border length between contiguous zones i≠j, and 0, otherwise. This value needs GIS support for its calculation. This extension of SCL will be named BSCL in this paper and its probability is of the following closed form:

$$P_{ni} = \sum_{j \neq i} P_{ni|j} \cdot P_{nij} =$$

$$\sum_{j \neq i} \frac{\alpha_{i,ij} {y_{ni}}^{1/\mu}}{\alpha_{i,ij} {y_{ni}}^{1/\mu} + \alpha_{j,ij} {y_{nj}}^{\mu}} \cdot \frac{(\alpha_{i,ij} {y_{ni}}^{1/\mu} + \alpha_{j,ij} {y_{nj}}^{1/\mu})^{\mu}}{\sum_{k=1}^{I-1} \sum_{l=i+1}^{I} (\alpha_{k,kl} {y_{ni}}^{1/\mu} + \alpha_{l,kl} {y_{nj}}^{1/\mu})^{\mu}} \tag{19}$$

where $P_{ni|j}$ is the probability that an individual *n* chooses alternative *i* having chosen the pair of alternatives *ij* and P_{nij} is the probability that an individual *n* chooses the pair of alternatives *ij.* The superimposed mixed specification with random coefficients of BSCL and MBSCL allows unobserved heterogeneity in the preferences of the decision makers.

This approach can be generalized with a new SCM, for example, combining the proportion of the common border length between zones with their Euclidean distances (see Cliff and Ord 1973 in the field of spatial autocorrelation in linear regression models) or considering the smallest distances between the borders. These extensions have the advantage of considering the spatial correlation that could exist between zones that are near but not adjacent. In any case, it would be necessary to evaluate its consistency with the equality assumption of dissimilarity parameters.

4. Empirical analysis

This section empirically applies the proposed approach, BSCL and its mixed specification, MBSCL, to the real-life case study of RLC modeling in Santander, a historic city in the north of Spain (Europe) with highly irregular administrative areas (see Figure 1). We also

compare it with the other spatial logit models without pre-defined nests reviewed in the paper: the SCL, GDSCL, and its mixed specifications.

4.1. Methodology

To compare the models, each one was applied to the same conditions: data, observed utility function, and software tools.

The estimations were performed using BisonBiogeme (Bierlaire 2003), setting 1000 draws in all mixed specifications. The spatial variables of the SCM were calculated using public mapping of the administrative zoning and in the case of the GDSCL and BSCL models, using QGIS (2018).

The utility function specification was developed in two steps: perform the kernel's observed utility (i.e., with fixed coefficients) and perform the mixed specification with random coefficients.

The first step is a backward process using MNL, starting from the observed utility specification with all variables of the sample (see next sub-section). In each iteration, we validate the sign of the estimated coefficients with the theoretically correct, individual relevance using t-test and the improvement of the global goodness-of-fit (GoF) with likelihood-ratio tests and indexes. The maximum level of significance used for the tests is 5%. The second step uses the same approach to select the coefficients of the MNL specification that are considered random, supposedly with a normal distribution, starting with all of them being random.

The likelihood ratio test (LRT) is used to compare the global GoF of nested models[1]. The null hypothesis of LRT is that the likelihood of both models are equal at significance level α and the alternative hypothesis is that the likelihood of the nested model is significantly higher[2]. The likelihood ratio of the nested models is not a known distribution, but its transformation, named Wilks's statistic[3], is asymptotic $\chi^2_{(p_2-p_1)}$, where $p_2 - p_1$ is the difference in the number of estimated parameters between them (Wilks 1938). Chi-square distribution is not appropriate when some variance component is zero or insignificant with respect to others. This can be violated with the mixed models, so LRT is not used in this paper to compare the nested models when some of them are mixed. LRT is significant when the Wilks's statistic is higher than the correspondent $\chi^2_{(p_1-p_2),\alpha}$quantile.

To compare the global GoF of non-nested models, we can use the LR with a model nested with each one (at least all the models are nested with the null model) or a likelihood ratio index (LRI). McFadden's LRI[4] is a pseudo-R^2 between 0 and 1 that compares the log-likelihood of each model with that of the null model and allows the comparison of non-nested models, or nested models when some of them are mixed

[1] A model is nested with other when it contains the same parameters and at least one more.

[2] The likelihood of a nested model with other is at least equal to that of the other.

[3] The Wilks's statistic of a model with likelihood L_1, nested with other with likelihood L_2, is $-2log\frac{L_2}{L_1} = -2(LL_2 - LL_1)$, where $LL = logL$, named log-likelihood.

[4] McFadden's LRI=$1 - \frac{LL(model)}{LL(null\_model)}$

models, with an equal number of estimated parameters. The closer the value is to one, the better the GoF. The Akaike-adjusted LRI[5] (aLRI) considers the number of estimated parameters, and so allows the comparison of models with a different number of estimated parameters.

The data collection was performed in the TRANSPACE LUTI project and previous studies (Ibeas, Cordera, Dell'Olio and Coppola 2013; Dell'olio, Cordera, and Ibeas 2016) from three sources, with spatial and non-spatial information: zoning of the available residential locations from public administrative areas, socio-economic and transportation information of each alternative zone from public information, and the microdata of an ad-hoc preferences survey.

Santander city was divided in 26 zones, based on administrative areas, that fulfill the Anas condition (1981), with a mean area of 0.33 km^2. The color-coding on the map of Figure 3 represents the tertiles of the number of individuals in the sample who have chosen that area per square kilometer, to show patterns of decisions less dependent on zoning.

The irregular zoning can be appreciated, especially in the city center. For example, in the SCL approach, the pair of zones 14-13 and 14-15 have the same SCM value, but not in BSCL; it is much higher in the second relationship because it considers the common border length. In addition, the pair of zones 25-10 and 25-16 have a similar SCM value when using GDSCL, but the SCM value is much higher for 25-10 with BSCL. The GDSCL assumptions are not in accordance with the expected spatial correlation in this case.

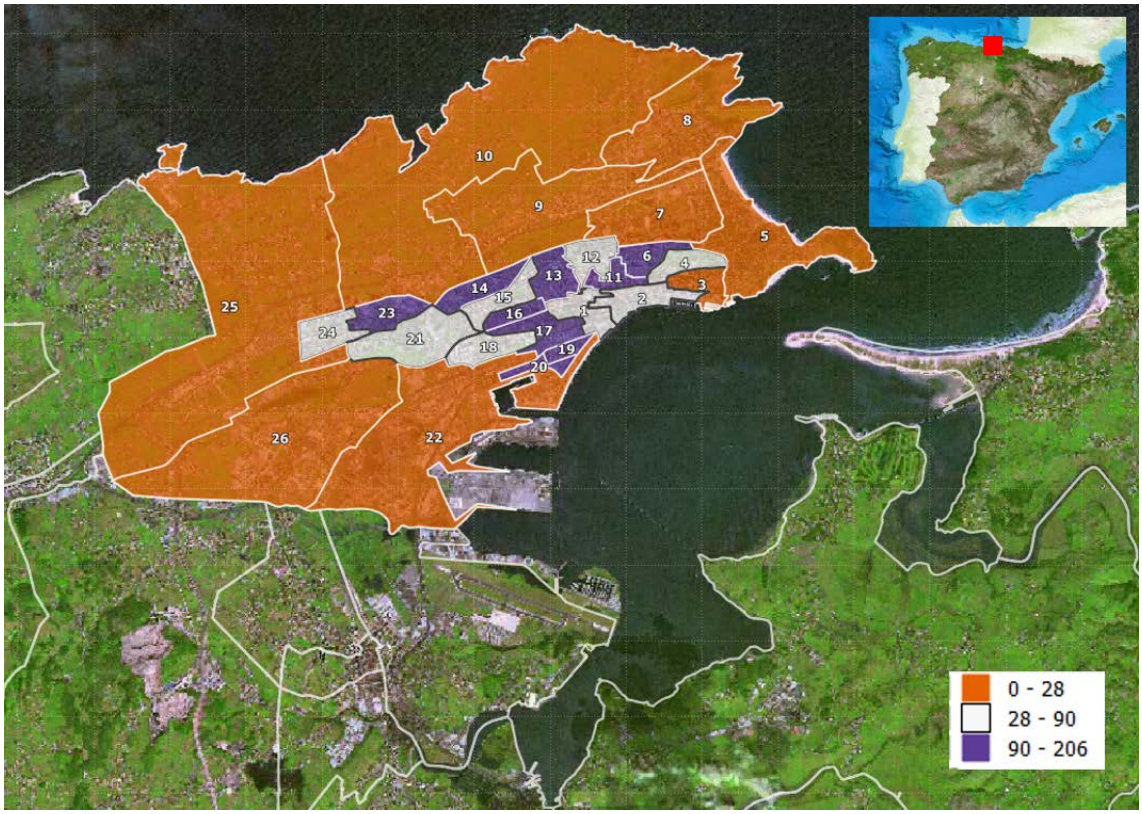


Figure 3. Map of the scheme of the alternative residential zones in Santander and the number of selections per km^2

[5] Akaike-adjusted LRI=$1 - \frac{LL(model)-p}{LL(null\_model)}$

The ad-hoc survey was carried out on 534 individuals, with displacements for work or study with origin and destination in Santander city. The microdata includes the spatial information on the location of family residence and the place of work or study, transportation information of the displacements between the two, and individual and family characteristics.

The microdata of the survey was enriched with socioeconomic and transportation information of the residential zones. The sample includes nine variables shown in Table 1, where TRA type refers to transportation level of service variables, ENV type refers to socio-economic environment variables, and IND type refers to individual attributes of the family unit.

Type	Variable	Name	Description
TRA	Accessibility	AC	Indicator of active accessibility of the zone of Coppola and Nuzzolo (2011)
TRA	Journey time	JT	Journey time in minutes between the centroid of the residential land use zone and the centroid of the work place land use zone.
TRA	Waiting time	WT	Average public transportation waiting time in minutes at the stops in the zone
TRA	Interzonal	IN	Dummy variable equal to 1 where the residential and work place zones coincide.
ENV	Foreigners	FO	Number of non-EU foreigners in the zone (in thousands)
ENV	Housing density	HO	Natural logarithm of the number of dwellings in the zone
ENV	Prestige	PG	Dummy variable equal to 1 where the zone has special prestige (subjective).
ENV	Price of housing	PR	Average price of housing in the zone (in millions of €).
ENV	Schools	SC	Number of schools in the zone
IND	High incomes	H	Dummy variable equal to 1 when family monthly incomes > 2,500€ Net

Table 1: Description of the sample variables (based on Ibeas et al. 2013)

4.2. Utility specification

In the following tables, the color-coding represents the significance results of the relevance and GoF tests: blue (***)—when at least 99%; green (**)—at least 95% but less than 99%; orange (*)—at least 90% but less than 95%; and red (·)—when less than 90%.

The utility function is a linear specification without a specific constant, performed using the methodology explained in sub-section 4.1 (see Perez-Lopez and Orro 2016, for a more detailed explanation).

The first step starts with the saturated model with seven sample variables and their eight interactions (except INs) with the IND type variable HI. Table 2 shows the initial estimation results of the first step, MNL-full, and the final MNL.

Coefficient		MNL-full			MNL		
Var.	Inter.	Est. value	S.E.	Sig.	Est. value	S.E.	Sig.

AC	-	0.00579	0.00994	·			
	H	-0.0166	0.0188	·			
JT	-	-0.102	0.0398	**	-0.114	0.0291	***
	H	-0.00702	0.0712	·			
FO	-	-0.864	0.428	**	-0.870	0.307	***
	H	-1.57	1.00	·			
IN	-	0.235	0.221	·			
HO	-	1.39	0.343	***	1.49	0.265	***
	H	1.49	0.822	*			
PG	-	-0.897	0.3	***			
	H	1.90	0.553	***	0.224	0.272	***
PR	-	-1.30	0.639	**	-2.16	0.429	***
	H	-0.629	1.10	·			
SC	-	-0.0878	0.0451	*			
	H	0.242	0.0847	***	0.224	0.0516	***
WT	-	-0.132	0.0811	·			
	H	0.0135	0.187	·			

Statistic	Value	Sig.	Value	Sig.
p	17		6	
LL	-1,658.581		-1,667.968	
McFadden LRI	0.04670		0.04130	
Akaike aLRI	0.03692		0.03785	
LR-Null	162.486	***	143.712	***
LR-MNL ($\chi^2_{(11)}$)	18.774	*		

Table 2. First step of the observed utility equation performed: initial and final specification results respectively.

Both models improve the global GoF of the null model significantly. MNL-full has 11 coefficients more than MNL does, but does not significantly improve its global GoF. Further, all the variables of the final estimation MNL show at least 99% relevance. The MNL-observed utility specification for alternative *i* performed in the first step is:

$$V_{n,i} = \beta_{JT} JT_{\mathrm{i}} + \beta_{FO} FO_{\mathrm{i}} + \beta_{HO} HO_i + \beta_{PG*H} PGi_i * H_n + \beta_{PR} PR_{\mathrm{i}} + \beta_{SC*H} SC_i * H_n \quad (22)$$

In the second step, after the process, all the random coefficients were considered fixed, except the interaction of SC and H. Table 3 shows the results of the final estimation of the second step, named MMNL, where SLL is the simulated LL.

	MMNL		
Parameter	**Est. value**	**S.E.**	**Sig.**
β_{JT}	-0.118	0.02934	***

β_{FO}	-0.914	0.309	***
β_{HO}	1.49	0.266	***
β_{PG*H}	1.12	0.287	***
β_{PR}	-2.12	0.433	***
β_{SC*H}	0.160	0.0766	**
$\sigma(\beta_{SC*H})$	0.289	0.127	**
Statistic	**Value**		
p	7		
SLL	-1,666.966		
McFadden LRI	0.04188		
Akaike aLRI	0.03785		

Table 3. Second step of the observed utility equation performed: last results with MMNL

All the parameters estimated in the final specification of the second step are significant. The final mixed specification of the observed utility function of each individual decision maker *n,* is the same as that of the first step, save for the fact that $\boldsymbol{\beta_{SC*H}}$ is considered as a random variable with distribution N(0.160, 0.289). This specification shows the same global GoF as MNL, with one more estimated parameter, using Akaike aLRI.

4.3. Model estimation and validation

This sub-section analyzes the estimation results obtained with the BSCL and MBSCL specifications for residential location choice modeling with irregular zoning.

Table 4 shows the estimation results, obtained in the same circumstances as MNL and MMNL, and the global GoF statistic, which are compared with those obtained in the previous sub-section.

	BSCL			**MBSCL**		
Parameter	**Est. value**	**S.E.**	**Sig.**	**Est. value**	**S.E.**	**Sig.**
β_{JT}	-0.104	0.0264	***	-0.107	0.0264	***
β_{FO}	-0.642	0.263	***	-0.648	0.263	***
β_{HO}	1.18	0.241	***	1.15	0.242	***
β_{PG*H}	0.908	0.243	***	0.814	0.247	***
β_{PR}	-1.51	0.381	***	-1.43	0.380	***
β_{SC*H}	0.173	0.0430	***	0.126	0.0601	**
$\sigma(\beta_{SC*H})$				0.267	0.107	***
μ	1.74	0.296	***	1.80	0.316	***
Statistic	**Value**		**Sig.**	**Value**		**Sig.**
p	7			8		
LL / SLL	-1,663.2			-1,661.8		
McFadden LRI	0.04405			0.04484		
Akaike aLRI	0.04003			0.04025		

LR-MNL ($\chi^2_{(1)}$)	9.570	***	

Table 4. Results of the estimation of the BSCL and MBSCL models

In both models, all the estimated parameters of observed utility are significant. The signs obtained are also coherent with *a priori* notions or theory. A longer commute, greater number of foreigners, or higher prices reduce the utility of the zone perceived by the decision maker, and therefore the choice probability. The presence of more residences increases utility. A zone considered prestigious increases the perceived utility for high-income decision makers, which is as expected. On average, the presence of a greater number of schools increases the utility for high-income decision makers, but with significant variation across the population, which could be due to different family situations.

The dissimilarity parameter is significantly different from 1 with at least 99% confidence level in both specifications (using a one-side test). Therefore, it shows the presence of non-spatial correlation between contiguous alternatives due to non-observed variables.

Both specifications improve the global GoF of MNL and MMNL. The LRT of BSCL with respect to MNL is significant with at least 99% confidence level and its Akaike aLRI is higher than that of MMNL. MBSCL Akaike aLRI is higher than that of MMNL.

4.4. Comparison of spatial correlation approach models

This sub-section compares the MNL models: the three spatially correlated logit models reviewed in the paper and their mixed specifications, all of them estimated in the same circumstances (see in Pérez-López and Orro 2016, the estimation results of GDSCL and SCL kernels and their mixed specifications).

Kernel	MNL	GDSCL	SCL	BSCL
p	6	7	7	7
LL	-1,667.968	-1,667.968	-1,665.940	-1,663.183
McFadden LRI	0.04130	0.04130	0.04247	0.04405
Akaike aLRI	0.03785	0.03728	0.03844	0.04003
LR – MNL ($\chi^2_{(1)}$)		0 ·	4.056 **	9.570 ***
Mixed	**MMNL**	**MGDSCL**	**MSCL**	**MBSCL**
p	7	8	8	8
SLL	-1,666.966	-1,666.905	-1,664.796	-1,661.802
McFadden LRI	0.04188	0.04191	0.04312	0.04484
Akaike aLRI	0.03785	0.03731	0.03853	0.04025

Table 5: Model comparison

Table 5 shows the global GoF statistics. GDSCL does not significantly improve the global GoF of MNL. SCL and BSCL significantly improve the GoF of MNL (and also of the GDSCL), but only BSCL with at least 99% confidence level. BSCL shows the highest McFadden's LRI value of the three spatially correlated kernels and the highest Akaike aLRI value of all the models, except for its mixed specification, MBSCL.

Every mixed specification improves the Akaike aLRI of its kernel, except MMNL, which is the same. MBSCL has the highest McFadden's LRI value of the mixed spatially correlated specifications and the highest Akaike aLRI value of all models.

5. Conclusions

This work proposes a spatial correlation metric for logit models with spatial correlation, based on the common border of spatial alternatives. It aims to improve efficiency when the zoning of alternatives is irregular, which is a common occurrence in urban areas. The metrics used in previous specifications, such as SCL or DSCL, have limitations in this case.

This paper specifies a generalization of the SCL model using the proposed metric, named BSCL, and its mixed specification MBSCL. Both models were applied to an urban case study of LUTI RCL modeling with irregular zoning, based on administrative divisions in the city of Santander (Spain).

The GoF of the proposed specifications were empirically compared with MNL, MMNL, and the reviewed logit models with spatial correlation: SCL, GDSCL, and their mixed specifications. BSCL significantly improves the GoF of MNL, SCL, and GDSCL with at least 99% confidence level. BSCL also outperforms the mixed specifications of the previous models. MBSCL shows the best GoF values.

Acknowledgements

The authors acknowledge the financial support provided by the Government of Spain under the projects TRA2012-37659 and RTI2018-097924-B-I00 MCIU/AEI/FEDER, UE.

References

Alonso W (1960) A theory of the urban land market. Pap. Reg. Sci. Assoc 6, pp. 149-157.

Alonso W (1964) Location and land use. Harvard University Press, Cambridge, MA.

Anas A (1981) The estimation of multinomial logit models of joint location and travel mode choice from aggregated data. Journal of Regional Science 21 (2), 223–242.

Anselin L (1988) Spatial Econometrics: Methods and Models. Kluwer Academic, Boston, Mass.

Anselin L and Rey S (1991) Properties of tests for spatial dependence in linear regression models. Geogr Anal 23(2):113-131

Ben-Akiva M and Bierlaire M (1999) Discrete choice methods and their applications to short-term travel decisions. In Hall R (ed) Handbook of Transportation Science, pp.5-34.

Ben-Akiva M and Bierlaire M (2003) Discrete Choice Models with Applications to Departure Time and Route Choice. In Hall R (ed), Handbook of Transportation Science, 2nd edn., Kluwer, pp. 7–37.

Bhat CR (1998) Accommodating flexible substitution patterns in multidimensional choice modeling: formulation and application to travel mode and departure time choice, Transportation Research 32B (7), 425–440.

Bhat CR (2000) Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. Transportation Science 34, pp. 228-238.

Bhat CR and 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.

Bierlaire M (2003) BIOGEME: A free package for the estimation of discrete choice models. Ascona, Switzerland, 3rd Swiss Transportation Research Conference.

Chasco C (2002) Econometría Espacial aplicada a la predicción-extrapolación de datos microterritoriales (Doctoral dissertation, Universidad Autónoma de Madrid)

Chernew M, Gowrisankaran G and Scanlon D (2003) Learning and the Value of Information: The Case of Health Plan Report Cards, Submitted to International Economic Review

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, Ord J (1973) Spatial autocorrelation. Pion, London

Coppola P and Nuzolo A (2011) Changing accessibility, dwelling price and the spatial distribution of socio-economic activities. Res. Transp. Econ. 31, pp. 63-71.

Dacey MF (1969) Similarities in the areal distributions of houses in Japan and Puerto Rico. Area, 3; pp. 35-37.

Daly AJ and Zachary S (1978) Improved multiple choice models. In: Hensher DA and Dalvi MQ (eds), Determinants of Travel Choice (Westmead: Saxon House), pp. 335-357.

Dell'Olio L, Cordera R and 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 and Wang Y (2016) Land Use - Transport Interaction Models. The TRANSPACE model. 1st edn. Santander: GIST.

Domencich TA and McFadden D (1975) Urban Travel Demand: A Behavioural Analysis (New York: American Elsevier).

Florax RJGM, Rey S (1995) The impacts of misspecified spatial interaction in linear regression models. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics, advances in spatial science. Springer, Berlin, Heidelberg

Geurs KT and Van Wee B (2004) Accessibility evaluation of land-use and transport strategies: review and research directions. J. Transp. Geogr. 12, 127–140.

Gumbel EJ (1958) Statistics of extremes. Columbia University Press (Facsimile by UMI, Michigan, 1997).

Hajivassiliou VA, Ruud PA (1994) Classical estimation methods for LDV models using simulations. In: Engle R, McFadden D (eds) Handbook of econometrics IV. Elsevier, New York, pp 2383–2441

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

Hunt LM, Boots B and Kanaroglou PS (2004) Spatial choice modelling: new opportunities to incorporate space into substitution patterns. Progress in Human Geography 28-6, pp. 746-766.

Ibeas A, Cordera R, Dell'Olio L and Coppola P (2013) Modeling the spatial interactions between workplace and residential location. *Transportation Reseach A 49*, pp. 110-122.

Koppelman FS and Wen CH (2000) The paired combinatorial logit model: properties, estimation and application. Transportation Research 34B, 75-89.

Lee LF (1992) On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models. Economet Theor 8(4):518–552

Ma J, Haining R, Wise S (1997) SAGE user's guide. Sheffield Center for Geographic Information and Spatial Analysis, University of Sheffield

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

McFadden D (1978) Modelling the choice of residential location. Karlqvist A, Jundqvist L, Snickars F and Weibull J (eds). Spatial Interaction Theory and Planning Models, North Holland: Amsterdam, 75-96.

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

Moreno R, Vayá E (2000) Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial. Edicions Universitat de Barcelona, Barcelona

Papola A (2004) Some developments on the cross-nested logit model. Transp Res 38B(9):833–851

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

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

Sener IN, Pendyala RM and Bhat CR (2011) Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. Journal of Transport Geography 19, pp. 294-303.

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

Stetzer F (1982) Specifying weights in spatial forecasting models: the results of some experiments. Environ Plan 14A(5):571–584

Train KE (2003) Discrete Choice Methods with Simulation. Cambridge University Press, 2003.

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

Thurstone L (1927) A law of comparative judgment. Psychological Review 34, 273-86.

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

Wegener M (1994) Operational Urban Models: State of Art. Journal of the American Planning Association, 60 (1), pp. 17-29.

Wen CH and Koppelman FS (2001) The generalized nested logit model. Transportation Research Part B: Methodological 35 (7), 627–641.

Wilks SS (1938) The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics, Vol. 9, No. 1 (Mar., 1938), pp. 60-62.

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