Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 160 (2014) 430 – 439

Procedia
Social and Behavioral Sciences

XI Congreso de Ingenieria del Transporte (CIT 2014)

Advanced trip generation/attraction models

Alexandre A. Amavi[a], Juan P. Romero[b], Alberto Dominguez[a], Luigi dell'Olio[a]*, Angel Ibeas[a]

[a]*University of Cantabria, Av. de los Castros s/n, Santander 39005, Spain*
[b]*IDOM, Av Zarandoa Etorbidea 23, Bilbao 48015, Spain*

Abstract

In this paper, advanced trip generation/attraction models are proposed. A multiple linear regression (MLR) model has been created from zonal data. These models are compared to each other by analyzing their hypothesis and the required adjustments. Additionally, advanced generation/attraction models considering spatial correlation are proposed, and their improvements with reference to previous models not considering spatial correlation are analyzed.
A global spatial correlation model that conducts a joint review of every unit in the sample to determine whether the spatial units are randomly distributed or in accordance with a certain pattern is specified. To consider cluster situations in a given area, a local spatial correlation model aimed to measure the spatial autocorrelation to place each observation is defined.
The models are applied to the Santander metropolitan area (in Spain) in order to obtain advanced generation/attraction models in that city. For Santander, the models considering spatial dependence among observations have better results than the MLR models.

Keywords: *generation/attraction models; spatial correlation models; spatial lag; spatial error*

1. Introduction

Synthetic models were developed assuming that every trip has a generation and an attraction. This models essentially link generations to attractions. In the case of home-based trips, the generation is always the home. However,

* Corresponding author. Tel.: +34-619-257-353; fax: +34-942-201-734.
E-mail address: delloliol@unican.es

doi:10.1016/j.sbspro.2014.12.155

the home is the origin of the home-based trips only for those trips whose destination is the workplace (or the study place or shopping, etc.); for the return trip the home becomes the destination of the trip (Ortuzar and Willumsen, 2001).

Trip generation/attraction models can consider a higher level of disaggregation, since both home-based trips (home leave and return) and non-home-based trips have different explanatory variables.

From the transport demand perspective, the users' most relevant characteristic is their socioeconomic level. Nevertheless, since it is hard to determine this level for every particular user, a classification of the households is used instead. Each household has a given family income and a given motorization rate; this variables are used to categorize the households and, by extension, those who live in them. To do this, the number of demand categories is decisive to determine the dimensions, information requirements and calibration of the model, as well as the accuracy of the results.

In this study, the models for determining the trip generation and attraction in Santander were devised. First, a multiple linear regression (MLR) model was proposed from zonal data; these models are compared by analyzing their hypothesis and settings. In addition, advanced generation/attraction models considering spatial correlation are proposed (**Spatial Lag** and **Spatial Error**), not to mention the **basic model**. A regression in made using the commercial software **GeoDa** (Anselin and Bera, 1998); afterwards, these new models are compared with the models that does not consider spatial correlation, analyzing their advantages.

These models are applied to the Santander metropolitan area (in Spain) in order to obtain advanced generation/attraction models in that city. For Santander, the models considering spatial dependence among observations have better results than the MLR models.

2. State-of-the-art

2.1. Trip generation/attraction model

The operation of the transport model requires the origin-destination trip vectors for each analysis period, classified by trip purpose and demand category, as input data (MIDEPLAN – FDC, 2010). Those trip vectors are estimated through the *generation/attraction models*.

Ideally, one origin and another destination vector should be estimated for each purpose and demand category; however, in practice the demand category classification is not always possible. Since these categories are defined from the households' levels of income and motorization rates, it is easy to classify the trips whose origin is the home (trip productions), which are most of the trips during the morning rush hour and an important part of the trips during the off-peak hours.

Nevertheless, most trips during these time periods have destinations different from home; therefore, a prospective classification of the destinations (trip attractions) would be arbitrary at best.

Having regard to the above, the proposed model assumes that only origins are classifiable according to demand purpose, while destinations are classifiable only according to trip purpose. Thus, the transport model has the following input data: i) an Origin vector for each trip purpose and each demand category (O_i^{pn}); and ii) a Destination vector for each trip purpose, in which all demand categories are grouped together (D_j^p). In addition, it must be also satisfied:

$$\sum_i \sum_p \sum_n O_i^{pn} = \sum_j \sum_p D_j^p \tag{1}$$

O_i^{pn}	Number of trips generated in zone ***i*** of the category ***n*** with trip purpose ***p***.
D_j^p	Number of trips attracted by zone ***j*** with the purpose ***p***.

For methodological reasons, the trip generations (origins) and the trip attractions (destinations) are modeled independently, although their results must be consistent, as appears from equation (1). On the other hand, since the conceptual development of the trip generation models exceeds that of the trip attraction models, analysts usually rely more on the results of the former and, consequently, the trip attractions are usually adjusted to the trip generations.

An easy way to adjust those trips is to estimate a correction factor for **each trip purpose**, as follows:

$$f_p = \frac{\sum_i \sum_n O_i^{pn}}{\sum_j D_j^p} \tag{2}$$

Then, this factor is multiplied by the components of the destination vector of the corresponding trip purpose, thus obtaining the adjusted values:

$$D_j^{p(a)} = f_p \cdot D_j^p \tag{3}$$

Sometimes, for some particular trip purpose, the calibration of the attraction model could deliver results more trustworthy than those from the generation model. In those cases, it is advisable to adjust the trip generation to the trip attraction, by estimating the factor f_p as the relation between the total number of trips attracted and the total number of trips generated (equation 4); then, the origins are adjusted (equation 5):

$$f_p = \frac{\sum_j D_j^p}{\sum_i \sum_n O_i^{pn}} \tag{4}$$

$$O_i^{pn(a)} = f_p \cdot O_i^{pn} \tag{5}$$

Two types of models are used to explain the trip generation: linear regression and cluster analysis. The election between one or another depends on the characteristics of the trips whose origins or destinations are to be explained. Although the cluster analysis models are conceptually more suitable, their scope of application is exclusively centered on home-based trips. On the other hand, though the linear regression models are not especially suitable to explain the trip generation, they are the only methodological tool to study non-home-based trip attractions and generations.

2.1.1. Trip generation models

The objective of trip generation is to estimate the number of trips originated in each zone within the study area, which it is usually correlated to the socioeconomic characteristics of the resident population in that zone (Ortuzar, 2000). Therefore, the trip generation is the sum of the home-based trips (HB) plus the non-home-based trips (NHB).

Since the 95% of the resident trips are home-based, the socioeconomic variables used to adjust the generation models are closely related to the population in the transport zones and the characteristics of the households.

This method relies on the relation between the trip generation, obtained from the origin-destination surveys, and the information from the land-use.

For each zone in the study area it is necessary to determine the area covered by each land-use to estimate the trips to and from that zone.

In other words, trip generations can be classified into: trips leaving from home (home-based trips), trips returning to home (home-based trips) and non-home based trips. The first ones are estimated through Multiple Classification Analysis (MCA) models, whereas the second and third types of trips can be estimated using Multiple Linear Regression (MLR) models. Therefore, the origins in a given zone can be expressed as:

$$O_i^{pn} = O_{i(hbl)}^{pn} + O_{i(hbr)}^{pn} + O_{i(nhb)}^{pn} \tag{6}$$

O_i^{pn} Number of trips generated in zone ***i*** of the category ***n*** with trip purpose ***p***.
$O_{i(hbl)}^{pn}$ Number of home-based trips leaving home in zone ***i*** of the category ***n*** with trip purpose ***p***.
$O_{i(hbr)}^{pn}$ Number of home-based trips returning home in zone ***i*** of the category ***n*** with trip purpose ***p***.
$O_{i(nhb)}^{pn}$ Number of non-home-based trips in zone ***i*** of the category ***n*** with trip purpose ***p***.

This differentiation is methodological important for the following reasons: First, the relative importance of each type of trip depends on the time period modeled; thus, home-based trips leaving home are mainly carried out during the morning peak hour. Secondly, the socioeconomic variables associated with the travelers' household explain the home-based trips leaving home; conversely, the variables associated with activities can explain both home-based trips returning home and non-home-based trips. These latter two types of trips are more important in off-peak periods.

2.1.2. Fundamentals of trip generation

The objective of the trip generation stage is to acquire appropriate identification and quantification of the trips to and from the different zones in which the study area is divided (Aldana, 2006).

The number of trips generated are usually very difficult to determine and estimate directly. Some information characterizing the zones can explain the trip generation more precisely than directly estimating the trips; this information is known as explanatory variables. The information regarding land-use, the socioeconomic characteristics in the zones within the study area and the characteristics of the transportation system are usually regarded as explanatory variables. The trip generation models are made up of functional relationships between the trips generated and the explanatory variables, so that the trip demand in a given future scenario can be accurately estimated by knowing the explanatory variables in that scenario.

Trip generation can be divided into:

- Estimation of the number of trips from each zone (productions).
- Estimation of the number of trips into each zone (attractions).

The total number of productions must be equal to the total number of attractions in the study area, although this is not necessary the case in each particular zone. This is due to the fact that the home-based trips (HB) are always produced by the zone in which the home is and attracted by another zone (or the same), no matter what way the trip is; on the contrary, non-home-based trips (NHB) are always produced by the origin zone and attracted by the destination zone.

Consequently, productions are usually higher than attractions in residential areas, whereas attractions are usually higher than productions in commercial, industrial or educational areas.

This differentiation of trips between production and attraction is relevant when a gravitational model is used in the distribution stage; in other distribution models is not important, since they only consider trip origins and destinations.

2.1.3. Variables explaining trip generation

Land-use

Land-use can be easily determined and forecasted with an acceptable accuracy. Three different attributes that have an effect on the trip generation can be identified in this variable: type, intensity and location.

Different **types** of land-use have different characteristics in relation to trip generation; that is why it is important to distinguish them. Land-use is usually classified into residential, commercial, industrial, educational and recreational. The residential land-use produces more trips than the other land-uses, while these other land-uses usually attract more trips than they produce (Ortuzar and Willumsen, 2001).

The land-use **intensity** shows the level of activity characterizing a given zone; it is usually expressed in terms of quantity of density, such as the total number of dwellings in the zone or the number of jobs per unit of area. The land-use intensity has a marked influence in the number and type of trips generated by a certain zone. In general, the lower residential density the more trips generated per person. This also happens with regard to the distribution per trip mode, that is, the lower residential density the more car trips per household.

The **location** of the activities refers to the spatial distribution of the land-use and activities within the study area. The transport modes used in a high-density neighborhood surrounded by low-density neighborhoods differ significantly from the transport modes used in a high-density neighborhood close to the city center.

Socioeconomic characteristics

The households' socioeconomic characteristics that influence the most in trip generation are family income, household size, car ownership, type of housing and principal occupation of the household members.

The **family income** is one of the most important characteristics to determine both the number of trips per household (or per individual) and the transport mode used. The higher the family income the more trips in a given time period and the more trips by car.

The **household size**, that is, the number of household members has a positive effect on the trip generation; in other words, there are more trips as the household size goes up.

Car ownership is directly related to the family income and with the household size. Generally, the lower car ownership the less trips a family generates.

Trip generation varies according to the **type of housing**. Single-family dwellings generate more trips per household member than detached or semi-detached family dwellings, and the latter generate more trips per household member than family dwellings in apartment buildings. This variable is not typically used in studies at urban level, but it is commonly employed to determine trip generations in certain urban developments, such as large buildings, private neighborhoods, etc.

The **principal occupation of the household members** influences trip generation. The occupation of the head of the household highly influences the trip generation, since it determines the family level of income. The more members working the more trips generated per household.

3. Methodology

Advanced trip generation/attraction models are specified in this paper. First, two multiple linear regression (MLR) models are determined from zonal data. A local spatial correlation model that measure the spatial autocorrelation of the observations is calibrated in order to consider cluster situations in a particular area. These models are applied to the Santander metropolitan area (in Spain) in order to obtain advanced generation/attraction models in that city.

3.1. Spatial autocorrelation or spatial dependence

Spatial autocorrelation implies that the value of a variable is conditioned by the value of that variable in the neighboring zone. As discussed later in this paper, vicinity is not necessary defined as contiguity in the material sense; there are quite a few criteria to define the vicinity from a connection matrix (Buzai and Baxendale, 2009). This is one of the biggest methodological weaknesses that experts have risen against the spatial econometrics and its results.

The autocorrelation spatial indexes allow the dependent relationship between locations and values of variables or attributes of interest. At the same time, they are very suitable to enable observation of the fragmented spatial configuration of our times.

If we try to measure the correlation of a particular variable in different adjacent spatial units from a horizontal perspective, three different possibilities can result:

- Positive spatial autocorrelation: adjacent spatial units have similar values of the variable; this shows a tendency towards spatial unit grouping.
- Negative spatial correlation: adjacent spatial units have dissimilar values of the variable; a tendency towards spatial unit dispersion is shown by this correlation.
- No autocorrelation: adjacent spatial units have neither similar nor dissimilar values of the variable; therefore, spatial units are randomly located.

Spatial autocorrelation is an extension of the temporal correlation to a two-dimension space. This is a subject of increasing concern arising from GIS-aided analysis of the natural resources and the environment. More specifically, spatial autocorrelation occurs when the dependent variable or the error in a particular point is correlated to the dependent variable or the error in other points. In order to be statistically significant, the regression must correct the spatial autocorrelation, so that the estimators obtained allow making accurate economic calculations.

3.2. Sources of spatial autocorrelation

The main sources of spatial autocorrelation can be measurement errors and spatial interaction of the units. In financial terms, spillover effects can generate spatial autocorrelation. This process has been strengthen by the economic integration processes.

3.3. Spatial weight matrix

When analyzing time series, the influence of past observations on the sequence dynamics and the current value are usually obtained through a delay term considering the runtime. This is also important because of the possibility of making predictions. Therefore, the influence of the time delay is unidirectional: past observations affect current and future observations, according to a dynamic structure. However, multidirectional relationships are established in the spatial analysis framework; therefore, it is necessary to build a matrix to properly include them in the analysis.

The spatial weight matrix (also known as connection matrix or spatial proximity matrix), represented by W, is a non-stochastic N × N square matrix (N being the number of spatial units) whose elements (w) show the intensity of the interdependence between each pair of regions i and j (Moreno and Vaya, 2000).

3.4. Spatial lag

A ***substantive*** autocorrelation occurs when the value of the dependent variable in each geographic unit is actually determined by its value in the neighboring units. Therefore, the autocorrelation is still present after other explanatory variables. If we overlook this type of spatial dependence, the coefficients estimated by the least-squares regression will be biased. The alternative to model this type of autocorrelation in the spatial lag model, which takes into account the spatial dependence by introducing a lag spatial variable. According to Baller et al. (2001), a spatial lag model represents the interactive relationship between the independent variables and the dependent variable in the neighboring units.

Spatial lag models are similar to delayed dependent variable models in time series analysis; however, the correlation coefficient cannot be easily estimated (Anselin, 2003). The problem is that a spatial weight matrix is required to estimate the coefficient, but it is not clear what the matrix should be similar to, that is, what the real spatial relationship is.

3.5. Spatial error

The error term autocorrelation can be handled by a spatial error model (estimated with maximum likelihood). This model assumes that the spatial dependence found in the dependent variable is the result of the geographic distribution of the explanatory variables and the error term autocorrelation; the latter suggests that our model is not well specified (Anselin, 1992a; Baller et al., 2001).

Spatial autocorrelation arises when the error terms are correlated through observations, that is, when the observation error affects the errors in the neighboring units. This is similar to the serial correlation in time series analysis with impartial ordinary least squares coefficients, which make them inefficient (Anselin, Syabri and Kho, 2004). Since this is really an annoying problem, spatial errors are also called "annoying dependency error".

Several cases arise from spatial error. For example, on the lines of time series, a strong correlation can be derived from non-measured variables through the area or from the aggregation of spatially correlated variables and systematic measurement errors.

Then, what should be done if there is good reason to believe that there is no spatial error? Moran's I is allegedly the most famous test to detect the problem graphically; it is based on the residuals of the regression and it is also related to the Moran's dispersion of the residuals. Other statistics can be used to solve the problem in different ways, such as the Lagrange multiplier and risk coefficient tests. If there is good reason to believe that spatial error is a problem, the way forward is through modeling the error directly or using autoregressive methods.

4. Application with GeoDa

4.1. General characteristics

GeoDa design consists of an interactive environment combining maps with statistical graphs through dynamically linked windows (Anselin and Bera, 1998). In general, its functions can be classified into:

- Spatial data utilities and handling: data input, data output, data conversion.
- Data transformation: variable creation, variable transformation.
- Mapping: creation of thematic maps, cartograms and animated maps.
- Exploratory Data Analysis (EDA): creation of different types of statistical graphs, such as histograms, box-plots and scatter plots.
- Spatial autocorrelation: global and local spatial autocorrelation statistics with inference and visualization.
- Spatial regression: diagnosis and estimation on spatial regression models parameters.

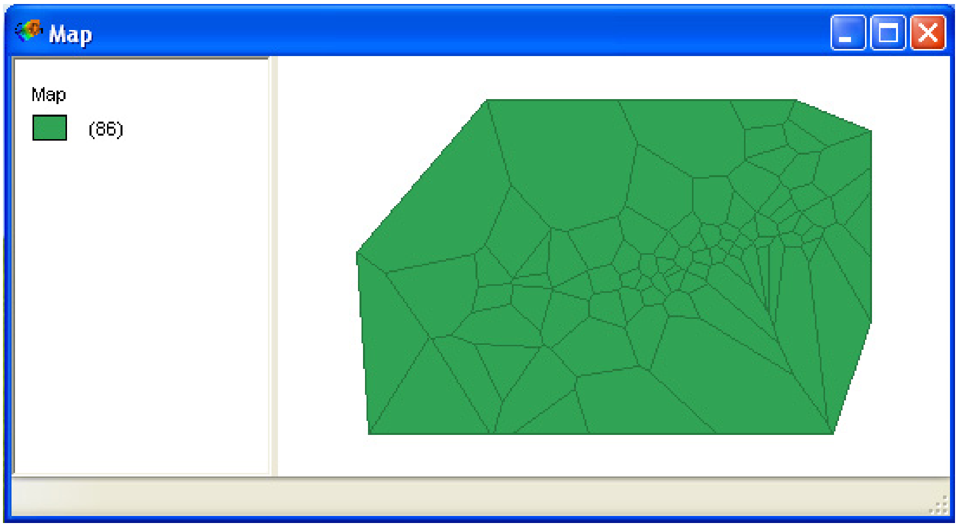

Fig. 1. Construction of Thiessen polygons with the help of GeoDa.

4.2. Weight results using GeoDa

These results are obtained from the Thiessen polygon previously created by GeoDa. It is necessary to open the drop-down menu **Tool** and click the menu **Weights** to create and obtain the following:

- Queen contiguity results.
- Rook contiguity results.
- Threshold distance results.

4.3. Using GeoDa – regression models

Different models were calibrated for different trip purposes during the morning peak hour (**TG_7-9**). Both the dependent and the independent variables were defined at local level (Santander zones); however, several spatial autocorrelation analysis justified the utilization of the variables of the model at a more significant level. The use of **spatial lag** and **spatial error** models improved significantly the statistical adjustment of every trip generation model; it is likely to be the same in the case of trip attraction.

The calibration results were good. In general, the classical regression model, the spatial lag model and the spatial error model calibrated to represent the non-home-based trips (NHB) and the home-based trips leaving home (HBL) are robust, with appropriate adjusted spatial correlation coefficients.

4.4. Description of the variables

The most significant independent variables used in the models are:

- RES_DENS: Residential density.
- EXTR_IND_OC: Extractive industries occupancy.
- AGR_COOP_OC: Agricultural cooperatives occupancy.
- CIV_SERV: Directors and managers of non-agricultural premises, senior officials in the public administration, autonomous communities and local authorities.

The main dependent variables in the models are:

- TG_7-9: Trips generated from 7am-9am.
- TA_7-9: Trips attracted from 7am-9am.

Table 1. Summary of the best results for each model.

Variables	S_E_Basic Model	S_E_Queen		S_E_Rook		S_E_Threshold Distance	
		Spatial Lag Q4	Spatial Error Q2	Spatial Lag R1	Spatial Error R2	Spatial Lag TD1	Spatial Error TD4
CONSTANT	0.041	0.620	0.073	-0.071	0.060	0.265	0.040
	0.447	2.508	0.934	-0.490	0.740	1.221	0.499
RES_DENS	0.185	0.178	0.165	0.192	0.163	0.182	0.178
	4.673	4.768	4.529	4.994	4.349	4.777	4.633
EXTR_IND_OC	0.107	0.097	0.120	0.103	0.128	0.110	0.115
	2.300	2.223	2.661	2.291	2.822	2.444	2.516
AGR_COOP_OC	0.975	0.910	1.013	0.898	1.058	0.864	1.033
	2.162	2.155	2.352	2.045	2.437	1.957	2.364
CIV_SERV	0.015	0.014	0.015	0.015	0.016	0.015	0.016
	4.991	4.898	5.552	4.972	5.651	5.137	5.262
TG_7-9		-0.521		0.118		-0.225	
TA_7-9		-2.524		0.987		-1.104	
LAMBDA			-0.455		-0.450		-0.704
			-1.768		-1.233		-0.921
R squared	0.659	0.683	0.675	0.663	0.668	0.666	0.664
Log likelihood		-46.646	-47.833	-48.754	-48.536	-48.515	-48.787
S.E. of reg.	0.442	0.413	0.418	0.426	0.423	0.424	0.425

Conclusions The regression model integrated in GeoDa has been used to calibrate the models analyzed (the results of the best models obtained are shown in table 1):

- S_E_Basic Model: This is the basic model.
- S_E_Queen: This includes Spatial Lag – Q4 (with a fourth-order contiguity matrix) and Spatial Error – Q2 (with a second-order contiguity matrix).
- S_E_Rook: This includes Spatial Lag – R1 (with a first-order contiguity matrix) and Spatial Error – R2 (with a second-order contiguity matrix).
- S_E_Threshold Distance: This includes Spatial Lag – TD1 (with a first-order contiguity matrix) and Spatial Error – TD4 (with a fourth-order contiguity matrix).

The most significant model is S_E_Queen – Spatial Lag – Q4. In order to obtain this result, a minimum of six (6) different points per model have been calibrated; then, the best one has been selected.

N.B.: each variable has its corresponding coefficient and test-t value.

5. Conclusions

In this paper, several applied studies using advanced trip generation/attraction models have been analyzed; these studies describe two main types of methodology: i) to compile background and baseline information; and ii) to use this information in the best possible way through two multiple linear regression (MLR) models from zonal data.

The most significant model obtained from the calibration of spatial data using the regression method integrated in the software GeoDa is S_E_Queen – Spatial Lag – Q4. This model has the best outcomes, that is, the best model fitting according to the test-t, providing the best results for the trips generated.

The results of spatial autocorrelation models for trip attraction will be included in further research with the same database and the same software (GeoDa).

6. Acknowledgements

This work was made possible thanks to a grant from the Spanish Ministerio de Economía y Competitividad with additional financing from FEDER funds to the project TRA2012-37659.

7. References

Aldana, C.M. (2006). *Modelación de la generación y atracción de viajes en el valle de Aburrá*. Medellín: Facultad Nacional de Minas, Universidad Nacional de Colombia Sede Medellín.

Anselin, L. (1992). Spatial dependence and spatial heterogeneity: model specifications issues in the spatial expansion paradigm. In E. Casetti, & J.P. Jones III (Eds.) *Applications of the expansion method* (pp. 334-354). London: Rudledge.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H. Scholten, & D. Unwin (Eds.), *Spatial analytical perspectives on GIS* (pp. 111-125). London: Taylor & Francis.

Anselin, L. (1998). Exploratory spatial data analysis in a geocomputational environment, Proceedings of Conference on Geocomputation 1998. Bristol, United Kingdom.

Anselin, L. (2003). *Geoda 0.9. User's guide*. Urbana-Champaign, IL: Department of Agricultural and Consumer Economics - University Of Illinois.

Anselin, L., & Bera, A. (1988). Spatial dependence in linear regression models with an introduction to spatial econometrics. In: A. Ullah, & D.E.A. Giles (Eds.), *Handbook of applied economic statistics* (pp. 237-289). New York: Marcel Dekker.

Anselin, L. Syabri, I., & Kho, Y. (2004). *GeoDa: An introduction to spatial data analysis - Spatial analysis laboratory*. Urbana-Champaign, IL: Department of Agricultural and Consumer Economics - University Of Illinois.

Baller, R.D., Anselin, L., Messner, S.F., Deane, G., & Hawkins, D.F. (2001). Structural covariates of U.S. county homicide rates: incorporation spatial effects. *Criminology 39(3)*, 561-590.

Buzai, G.D., & Baxendale, C.A. (2009). Análisis exploratorio de datos espaciales. geografía y sistemas de información geográfica. *Geografía y Sistemas de Información Geográfica (GEOSIG), Año 1, n° 1, Sección III*, 1-11.

Coro, C.Y. (2006). *Análisis estadístico de datos geográficos en geo marketing: El programa GeoDa*. Madrid: Departamento de Economía Aplicada, Universidad Autónoma de Madrid.

MIDEPLAN – FDC (2010). *Análisis y formulación de nuevos modelos de generación y atracción de viajes*. Santiago, Chile: Ministerio de Planificación y Cooperación.

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

Ortuzar, J.D. (2000). *Modelos de demanda de transporte*. Santiago, Chile: Ediciones Alfaomega.

Ortuzar, J.D., & Willumsen, L.G. (2001). *Modelling transport*. Chichester, United Kingdom: John Wiley & Sons.