CIT2016 – XII Congreso de Ingeniería del Transporte
València, Universitat Politècnica de València, 2016.
DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

Demand prediction model for regional railway services considering spatial effects between stations

Rubén Cordera
Research Fellow, University of Cantabria, Spain
Roberto Sañudo
Research Fellow, University of Cantabria, Spain
Luigi dell'Olio
Professor, University of Cantabria, Spain
Ángel Ibeas
Professor, University of Cantabria, Spain

ABSTRACT

The railways are a priority transport mode for the European Union given their safety record and environmental sustainability. Therefore it is important to have quantitative models available which allow passenger demand for rail travel to be simulated for planning purposes and to evaluate different policies. The aim of this article is to specify and estimate trip distribution models between railway stations by considering the most influential demand variables. Two types of models were estimated: Poisson regression and gravity. The input data were the ticket sales on a regional line in Cantabria (Spain) which were provided by the Spanish railway infrastructure administrator (ADIF – RAM). The models have also considered the possible existence of spatial effects between train stations. The results show that the models have a good fit to the available data, especial the gravity models constrained by origins and destinations. Furthermore, the gravity models which considered the existence of spatial effects between stations had a significantly better fit than the Poisson models and the gravity models that did not consider this phenomenon. The proposed models have therefore been shown to be good support tools for decision making in the field of railway planning.

1. INTRODUCTION

The European Commission transport roadmap (European Commission, 2011) gives priority to the railways because of their proven safety and environmental sustainability compared to road transport. One of the Commission's stated future goals is the creation of a unique European railway space, the introduction of new technological solutions and the construction of new intelligently financed and costed infrastructure.

In order to reach these goals, the European Commission has highlighted the need to evaluate transport projects to guarantee their social profitability and the added value they give to the

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

EU. This evaluation needs to be supported by the available evidence and transport demand models which allow user behaviour to be accurately simulated.

Among the group of transport demand models are trip distribution models which allow the interaction between origin and destination points to be simulated. The most well-known and widely used distribution model has traditionally been the gravity model which, based on the analogy with Newtonian physics, has later been theorized from a probabilistic perspective as a maximum entropy model (Wilson and Bennett, 1985). The state of the art provides many calibration techniques for the parameters of both origin and destination as well as for impedance (Ortúzar and Willumsen, 2011). Other researchers have insisted on the need to use Poisson type regression models given the discrete and positive nature of the journeys (Flowerdew and Aitkin, 1982).

This article proposes the estimation of trip distribution models based on the boarding and alighting data of passengers on a regional railway line. The data used has been obtained from ticket sales on the line provided by the Spanish railway infrastructure administrator (ADIF – RAM). The models were estimated based on two methods: a Poisson type nonlinear regression without any kind of constraint and a Wilson type gravity model doubly constrained to origins and destinations. Both types of models are compared by considering their goodness of fit with the data, in order to determine if the greater number of parameters estimated in the gravity models really does provide greater significance. The models have also been estimated with additional variables to consider the existence of spatial effects between stations to determine if these effects are significant and increase the explanatory capability of the models. The results show that gravity models restricted to origins and destinations with additional variables which consider spatial effects like contiguity between stations have a significantly better goodness of fit.

A brief review of the state of the art in the field of trip distribution models and distribution models applied to the railways is presented in the following section. The methodology followed is summarised in Section 3 concentrating on Poisson type regression models and doubly constrained gravity models. Section 4 provides a description of the study area and presents and discusses the results obtained by the models. Finally, the conclusions drawn are summarised in Section 5.

2. STATE OF THE ART ABOUT TRIP DISTRIBUTION MODELS

Spatial interaction models were applied very early on in multiple fields of study for simulating the effects of spatial interaction such as the movement of people between urban areas (Ravenstein, 1885) or commercial flows (Huff, 1959). The first models proposed were based on an analogy with Newtonian gravity theory with the sizes of origins and destinations and the distances between them as explanatory variables. This type of model has a

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

reasonably good fit to the data although they lack theoretical justification. The theoretical base was provided by Wilson (1970) who showed the possibility of deriving a great number of models from the principal of maximum entropy by which the most probable distribution matrix is the one which maximises the microstates of a given macrostate (Fotheringham et al., 2000). Other authors have later insisted on the convenience of using Poisson type non lineal regression models given their greater adaptability to the trip distribution phenomenon (Flowerdew and Aitkin, 1982).

In the field of trip distribution models relating specifically to railways, these models allow different planning alternatives to be evaluated. The currently available demand prediction models can be classified into two large groups depending on the data used: models based on aggregate data which use ticket sales information and models based on surveys which use disaggregate data on an individual level.

Among the aggregate models based on ticket sales, Wardman (2006) proposed an unrestricted distribution model using time series data for the United Kingdom in the 1990s. The estimated model presented variables corresponding to the characteristics of the origin such as the population, GDP and the rate of motorisation, as well as to the journey such as the overall cost. The author found that GDP was the most important factor in explaining the growth of journeys, even though in a complete four stage model these types of variables are usually introduced into trip generation models. In a similar work applied to railway journeys to and from airports, Lythgoe and Wardman (2002) estimated a demand model based on linear regression which calculated elasticities for different variables like GDP, the fare or journey time.

Where disaggregate data is available, models based on user surveys allow researchers to simulate individual choices considering personal characteristics (age, gender, income, etc.) and transport service characteristics as well as origins and destinations (Ben-Akiva and Lerman, 1985). However, this type of disaggregate model based on random utility theory require greater effort during the data collection phase because they are generally estimated using fewer data than models based on ticket sales.

3. METHODOLOGY

Different authors have highlighted the specification problems involved in using a multiple linear regression model (MLR) to estimate the generation and distribution of journeys in a study area (Flowerdew and Lovett, 1988; Thill and Kim, 2005). The dependent variable in distribution models is of a discrete nature, whereas the MLR model assumes a continuous distribution. Therefore, it is desirable to use a model specified with a qualitative dependent variable such as the Poisson regression model (Gujarati and Porter, 2009). This model takes the form:

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

$$P(Y_i) = \frac{\mu^Y e^{-\mu}}{Y!} \quad (1)$$

The Poisson regression assumes that each dependent variable Y_i is extracted from a Poisson type discrete distribution with the distribution parameter μ_i, which is logarithmically linked to a linear combination of explanatory variables:

$$\ln(u_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_k X_{ki} \quad (2)$$

Where:

β_k are parameters to be estimated

X_{ki} are the independent variables

The Poisson model cannot be made linear, meaning that the parameters cannot be estimated using Ordinary least Squares (OLS). Various alternative estimation methods such as maximum likelihood (Greene, 2003) or reweighted least squares have been proposed producing both equivalent results (Green, 1984).

A particular case of the Poisson model appears when all the independent variables are specified as dummy variables. In this case the Poisson model is equivalent to a log-linear model as both the dependent variable and the independent are qualitative. Log-linear models are more frequently used for modelling contingency tables (Agresti and Kateri, 2011). This type of model can be specified as totally saturated, in other words, with a perfect fit to the data as a parameter is specified for each observation. Willekens (1983) has shown how log-linear models are equivalent to the gravity models if they are conveniently scaled, usually by equalling the equilibrium factors of the first origin and destination to 1.

The fit of a Poisson model can be evaluated through different indicators as the Akaike information criteria (AIC), the log–likelihood or through the deviation of the model estimated with respect to the totally saturated model, in other words, using a likelihood reason test (LR) of the following kind:

$$LR = -2\left[L(\hat{\theta}_0) - L(\hat{\theta}_s) \right] \quad (3)$$

Where:

$L(\hat{\theta}_0)$ is the log – likelihood of the estimated model

$L(\hat{\theta}_s)$ is the log – likelihood of the saturated model

This type of test asymptotically distributes χ^2 with *r* degrees of freedom, where *r* in this case is the difference between the number of observations and the number of parameters

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

estimated in the non-saturated model. The LR test can also be used to compare the fit between general models and their constrained versions with fewer parameters.

A trip distribution model estimated using a Poisson regression is usually specified with three variables: a variable of the trips produced by the origin, a variable of trips attracted by the destination and an impedance variable between both zones, where the variables of the produced and attracted trips are usually extracted from a trip generation model (Hall, 2012). Therefore, this type of model would not present any kind of constraint although it could have problems of spatial autocorrelation in the origins or destinations which would be convenient to address to guarantee the reliability of the estimated parameters (Griffith, 2007). One of the techniques which is available for addressing this spatial autocorrelation in nonlinear models is Spatial Filtering (Tiefelsdorf and Griffith, 2007) where the spatial effects are separated from the rest of the non-spatial effects, thereby eliminating the possible correlation present in a neighbourhood matrix.

The Poisson regression can also be specified with constraints on the origins or destinations by estimating a different parameter for each zone. The case of a doubly constrained model with an impedance variable leads to the well-known gravity distribution model derived from the principle of maximum entropy (Wilson, 1970):

$$T_{ij} = A_i O_i B_j D_j \exp(-\beta c_{ij}) \tag{4}$$

Where:
T_{ij} are the trips between zones i and j
O_i are the trips produced by zone i
D_j are the trips attracted by zone j
C_{ij} are the costs between zone i and zone j
β is an impedance parameter to be estimated

The impedance parameter β can be estimated using different procedures like the method proposed by Hyman (1969) or using a log linear model (Dennett, 2012). The balancing factors A_i and B_j are codependent, meaning they need to be estimated iteratively:

$$A_i = \frac{1}{\sum_j B_j D_j \exp(-\beta c_{ij})} \qquad B_j = \frac{1}{\sum_i A_i O_i \exp(-\beta c_{ij})}$$

Given the constraints on the origins and destinations of the model, the resulting fits are usually high. However, it is possible to introduce new variables into the model in order to consider other spatial effects. Flowerdew (2010) has proposed inserting dummy variables into the model to consider zonal contiguity, as depending on the type of trip being modelled, the contiguous zones may be a more or less likely destination than the rest of the areas. This type of spatial effect may help in improving the fit of the models by adapting them to the

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

peculiarities of each study area.

4. STUDY AREA AND RESULTS

4.1 Available data and the study area

The trip distribution models have been estimated using data provided by the Spanish railway organisation ADIF – RAM about ticket sales on a narrow gauge regional line in Cantabria (Spain). The ticket sales provide information on both the origins and destinations of the passengers meaning the trip matrix gives an exact representation of travel on the line.

The studied line has a total of 23 stations being the two terminals located at Santander and Cabezón de la Sal (see Figure 1). The data obtained corresponds to the week from 19th to 25th January 2015 and counted 26,371 passengers. The stations with the highest production and attraction trips were the two largest towns in the region, Santander and Torrelavega, which accumulated more than 50% of the passengers given their higher demographic weight.

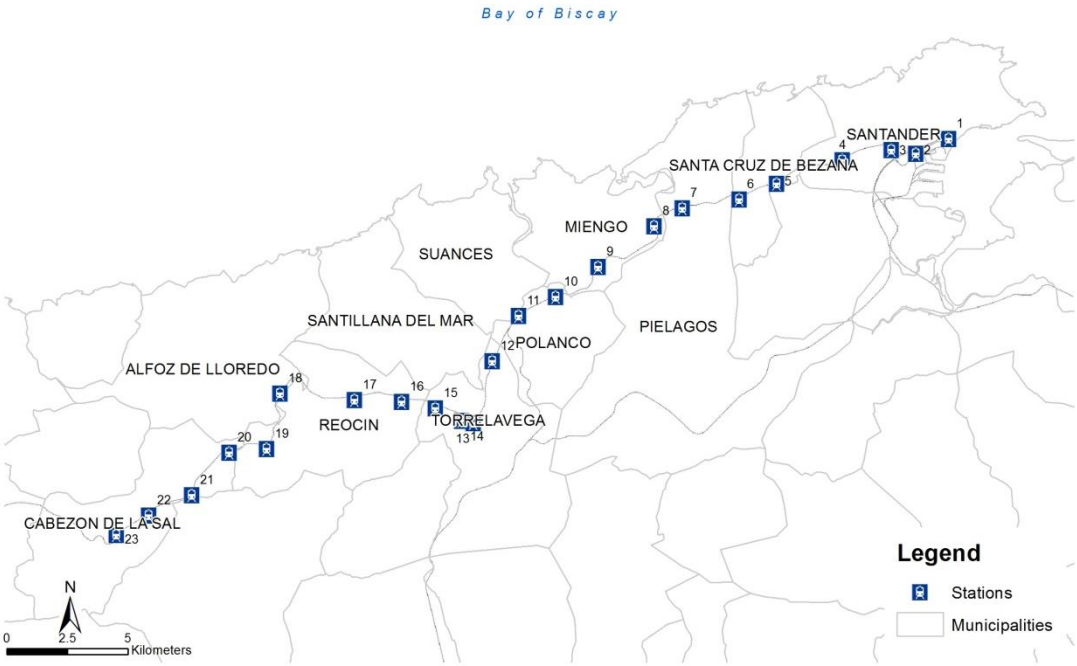

Figure 1 – Stations on the narrow gauge line Santander – Cabezón de la Sal

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

Variable	Description	Units	Average	Standard Deviation	Minimum	Maximum
V_{ij}	Trips between origin i and destination j	No. Trips	52.47	255,40	1	2900
O_i	Trips produced by origin i	No. Trips	1146.57	2083.57	43	8842
D_j	Trips attracted by destination j	No. Trips	1146.57	2066.93	43	8913
C_{ij}	Generalised cost between i and j	Euros	8.59	4.80	1.90	21.05
Cont	Dummy variable if the stations are contiguous	1/0	0.09	0.28	0	1
SantTorre	Dummy variable if the O-D pair ij corresponds to Santander and Torrelavega	1/0	0	0.06	0	1

Table 1 – Descriptive statistics of the variables contained in the database

The variables contained in the database can be seen in Table 1. Between all the O-D pairs there is an average of 52.5 trips with a maximum of 2,900 trips corresponding to the Santander – Torrelavega pair. Impedance between the pairs has been specified through a generalised cost (C_{ij}) measured in euros which combines the journey time between the stations (in minutes) with the fare variable between the stations. The value of time was provided by a previous study based on surveys asked to regional train users with a final weight of 0.25 € per minute of journey time (Grupo de Investigación de Sistemas de Transporte, 2008).

Two dummy variables were also included in the database to consider the possible presence of spatial effects. A variable of contiguity between stations taking a value of 1 if the stations are adjacent, and a variable which takes a value of 1 in the Santander – Torrelavega and Torrelavega – Santander pairs. This latter variable could be important because, as can be seen in Figure 2, the number of trips in the cost interval of 10 – 15 euros increases with respect to the interval 5 – 10 euros due largely to the journeys produced between the two towns.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

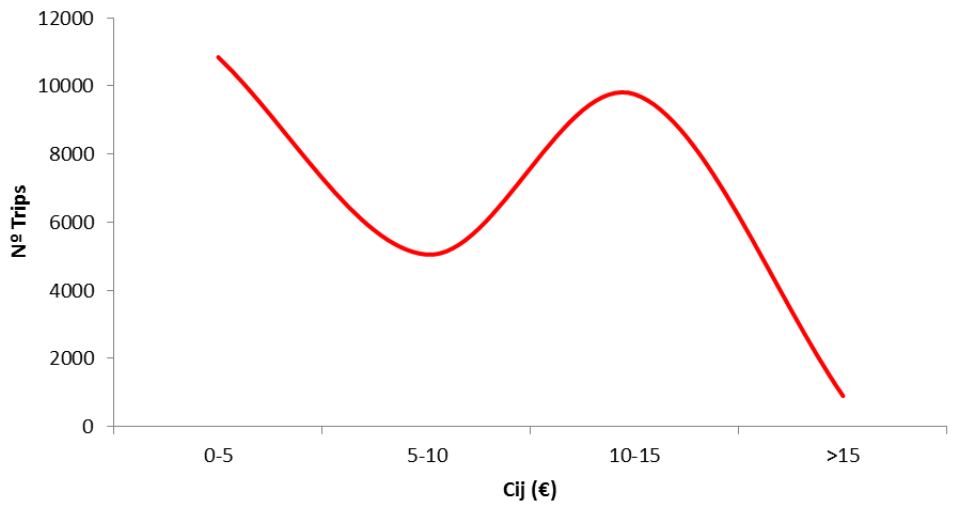

Figure 2 – Histogram of journeys according to generalised cost

4.2 Results and discussion of the models

The parameters estimated for the seven models are summarised in Table 2. The first four (P-1 a P-4) correspond to Poisson type regression models, while the three latter are Wilson type gravity models.

The P-1 model was specified with the totals produced and attracted by the origin and destination stations, using the generalised cost between them as independent variables. The production and attraction parameters were identical and had a positive sign, whereas the impedance parameter was, as expected, negative. Furthermore, all the parameters were clearly significant. The parameters show, using the transformation $100*(e^{\beta}-1)$, that one unit change in production and attraction generates, ceteris paribus, 0.05% more trips. However, an increase of one euro in the generalised cost implies about a 9% reduction in the number of trips being made. According to the AIC index the model had a fit of 20,279 and an R^2 of 0.85 for the estimated journeys compared with the observed journeys. The P-2 model adds to the variables of the P-1 model, the dummy variable of contiguity between stations, which showed a negative sign. This sign provides evidence that, if a greater number of journeys are made between points with low generalised costs (see Figure 2), these are not normally made between adjacent stations given that the parameter implies a reduction of around 72% in the number of journeys. The P-2 model had a slightly better fit than P-1 according to the AIC index as well as a superior R^2 comparing the estimated with the observed journeys. The Poisson P-3 model included an additional dummy variable corresponding to whether the O-D pair was Santander – Torrelavega or Torrelavega – Santander. The sign of the parameter was negative with a reduction of 10% in the number of expected trips which is almost certainly due to the fact that the O and D factors already

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

captured the potential for interaction between the two locations. This model had a slightly better fit than P-2 with all the estimated parameters being clearly significantly different from 0. The specification of P-3 is therefore:

$$\ln(u_{ij}) = \beta_1 + \beta_2 O_i + \beta_3 D_j + \beta_4 C_{ij} + \beta_5 Cont_{ij} + \beta_6 SantTorre_{ij} + \varepsilon_{ij} \quad (5)$$

Variable	P-1	P-2	P-3	P-4	W-1	W-2	W-3
(Intercept)	1.9490	2.2190	2.1460	1.7180	-	-	-
	(.000)	(.000)	(.000)	(.000)			
O / A_i	0.0005	0.0005	0.0005	0.0005	0.0001	0.0002	0.0002
	(.000)	(.000)	(.000)	(.000)			
D / B_j	0.0005	0.0005	0.0005	0.0006	0.8846	0.9690	0.9307
	(.000)	(.000)	(.000)	(.000)			
C_{ij}	-0.0969	-0.1187	-0.1156	-0.1092	-0.1102	-0.1690	-0.1652
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Cont	-	-1.2890	-1.2800	-1.2460	-	-2.3657	-2.4135
		(.000)	(.000)	(.000)		(.000)	(.000)
SantTorre	-	-	-0.1097	-0.5648	-	-	-0.2789
			(.000)	(.000)			(.000)
EvO	-	-	-	-8.9470	-	-	-
				(.000)			
EvD	-	-	-	4.4020	-	-	-
				(.000)			
AIC	20,279	18,614	18,591	16,880	10,295	6,998	6,910
R^2	0.85	0.88	0.89	0.91	0.94	0.98	0.99
Residual Deviation	18,531	16,864	16,839	15,124	8,463	5,164	5,074

Table 2 – Estimated Distribution Models (in brackets the p – value with the statistical significance of the parameters)

Finally, the P-4 model was estimated using the Spatial Filtering technique to eliminate the possible presence of spatial correlation in the origins and destinations (Griffith, 2007). All the pairs with identical origins or identical destinations were considered to have neighbourhood relationships. The spatial filtering selected two eigenvectors, one at origins (EvO) and another at destinations (EvD), which were introduced into the Poisson regression. The fit of the model increased and reduced the AIC to 16,880.

The Wilson gravity type models are summarised in columns W-1 to W-3 in Table 2. Rows A_i and B_j show the average of the 23 balancing factors estimated for origins and destinations, respectively. The rest of the parameters are the same as those specified in the Poisson type models, having been estimated using a log-linear model which also allows their statistical significance to be estimated. The fit of the constrained gravity models was better than that

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

of the Poisson regression models with R^2 superior to 0.9 in all cases, up to a fit of 0.99 for the observed data in model W-3 considering the contiguity of the stations and the specific interaction between Santander and Torrelavega. The W-3 model was specified as:

$$V_{ij} = A_i O_i B_j D_j \exp(\beta_4 C_{ij} + \beta_5 Cont_{ij} + \beta_6 SantTorre_{ij}) \quad (6)$$

If an LR test is conducted between the gravity and Poisson regression models, the former show a test value which is clearly superior to the critical value even considering the greater number of parameters used by the constraints on the origins and destinations. This is the case, for example, with the W-3 model compared with P-4, where the test presented a value greater than 10,000 for a critical value of 95% of the confidence level of 55.8.

If the residual deviation between the estimated models and the completely saturated model is considered, the test value was always superior to the critical value of the distribution, although the Wilson type models clearly got closer to the maximum fit provided by the saturated model.

An examination of the fit of the models with respect to the observed data by cost ranges (see Figure 3 and Figure 4) shows how the Poisson models have a worse fit for the intermediate cost ranges (5-10 and 10-15 euros). On the other hand, the gravity models and especially the W-2 and W-3 models with dummy variables considering spatial effects showed a better fit over all the cost ranges. The fit provided by these models was significantly better than that of the W–1 model using the LR test with one (W–2) or two degrees of liberty (W–3).

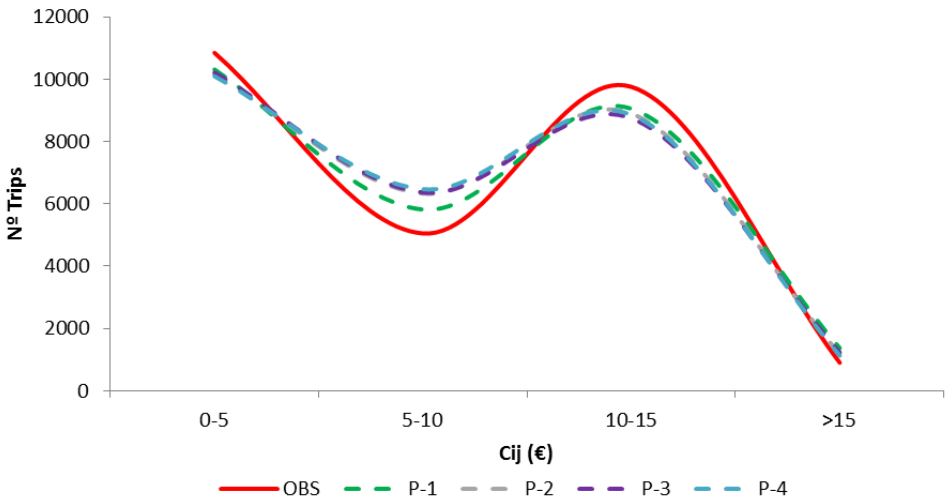

Figure 3 – Histogram of the observed trips compared to estimated trips for the Poisson regression models

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

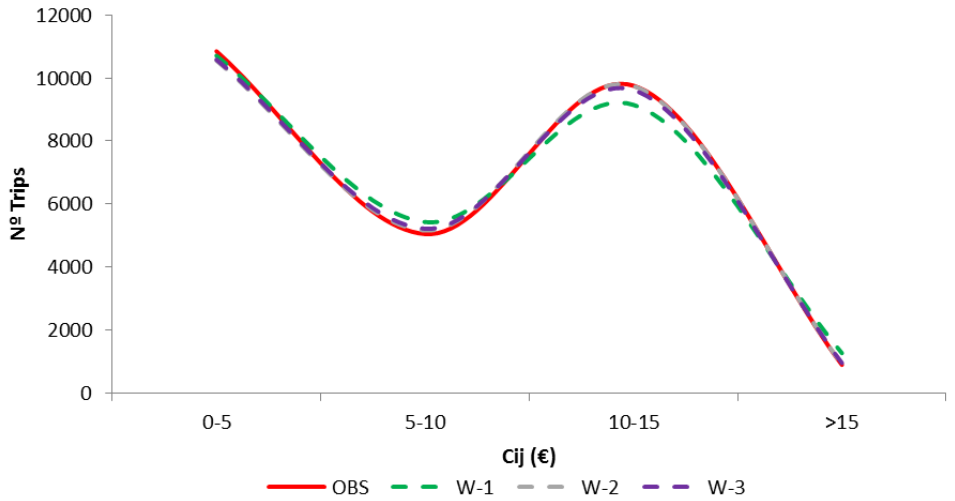

Figure 4 - Histogram of the observed trips compared with the estimated trips for the gravity models

5. CONCLUSIONS

This article has presented the estimation of trip distribution models using two methods: nonlinear Poisson regression and gravity models with constraints on origins and destinations. The goal was to assess whether or not the gravity models fit to the data significantly better considering they require a greater number of parameters. Additional variables have also been introduced to account for the spatial effect of contiguity between stations controlled by the effect of spatial correlation which may be present in the trip distribution data. The estimated models could be useful tools for simulating changes that passengers make in their choice of destination as a result of new policies such as the opening and closing of stations or changes in the service conditions.

The results confirm that the gravity model with constraints on origins and destinations had a significantly better fit to the data, according to the LR test, than the Poisson regression models without constraints. This fact was true even considering that the gravity models were estimated with 40 to 42 more parameters and that in a Poisson model the presence of spatial correlation was controlled. The models that considered contiguity between stations and the specific effects of interaction also showed a significantly better fit with only one or two more parameters than the models that did not consider these effects. It would therefore seem recommendable to estimate gravity models constrained by production and attraction data obtained from a trip generation model when creating a trip distribution model. Even more so when to estimate a gravity model using a log-linear model does not imply any additional costs other than those involved in the iterations needed to obtain the balancing factors. The possibility of specifying addition spatial variables also gives the model an extra capacity of adaptation to the study area.

A future line of research would be the estimation of gravity models which consider the

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

presence of spatial autocorrelation at origins, destinations and at points of interaction between O-D pairs. The estimation of this type of model currently requires considerable computing power which makes necessary additional research (Griffith, 2009).

ACKNOWLEDGEMENTS

This research was made possible thanks to financing from the Ministry of Economy and Competitiveness of the Government of Spain through the PARK-INFO project (TRA2013-48116-R). The authors would also like to thank ADIF – RAM for having provided the data used.

REFERENCES

Agresti, A., Kateri, M. (2011) Categorical data analysis. Springer.

Ben-Akiva, M.E., Lerman, S.R. (1985) Discrete choice analysis : theory and application to travel demand. MIT Press, Cambridge, Mass.

Dennett, A. (2012) Estimating flows between geographical locations:'get me started in'spatial interaction modelling. Citeseer.

European Commission, (2011) Roadmap to a single European transport area — Towards a competitive and resource-efficient transport system.

Flowerdew, R. (2010) Modelling migration with Poisson regression. Technologies for migration and commuting analysis, 261-279.

Flowerdew, R., Aitkin, M. (1982) A method of fitting the gravity model based on the Poisson distribution. Journal of regional science 22, 191-202.

Flowerdew, R., Lovett, A. (1988) Fitting constrained Poisson regression models to interurban migration flows. Geographical Analysis 20, 297-307.

Fotheringham, A.S., Brunsdon, C., Charlton, M. (2000) Quantitative geography : perspectives on spatial data analysis. Sage Publications, London ; Thousand Oaks, Calif.

Green, P.J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society. Series B (Methodological), 149-192.

Greene, W.H. (2003) Econometric analysis. Pearson Education India.

Griffith, D.A. (2007) Spatial structure and spatial interaction: 25 years later. The Review of Regional Studies 37, 28.

Griffith, D.A. (2009) Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 Germany journey-to-work flows. J Geograph Syst 11, 117-140.

Grupo de Investigación de Sistemas de Transporte, (2008) Optimización de la Coordinación Intermodal mediante la Modelización del Comportamiento del Usuario - INTERCOR. Ministerio de Fomento. Gobierno de España.

Gujarati, D.N., Porter, D.C. (2009) Basic econometrics, 5th ed. McGraw-Hill Irwin, Boston.

Hall, R. (2012) Handbook of transportation science. Springer Science & Business Media.

Huff, D.L. (1959) Geographical aspects of consumer behavior. University of Washington Business Review 18, 27-37.

Hyman, G. (1969) The calibration of trip distribution models. Environment and Planning 1, 105-112.

Lythgoe, W., Wardman, M. (2002) Demand for rail travel to and from airports. Transportation 29, 125-143.

Ortúzar, J.d.D., Willumsen, L.G. (2011) Modelling transport. John Wiley & Sons.

DOI: http://dx.doi.org/10.4995/CIT2016.2016.4053

Ravenstein, E.G. (1885) The Laws of Migration. Journal of the Royal Statistical Society, 68.
Thill, J.-C., Kim, M. (2005) Trip making, induced travel demand, and accessibility. J Geograph Syst 7, 229-248.
Tiefelsdorf, M., Griffith, D.A. (2007) Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environment and Planning A 39, 1193-1221.
Wardman, M. (2006) Demand for rail travel and the effects of external factors. Transportation Research Part E: Logistics and Transportation Review 42, 129-148.
Wilson, A.G. (1970) Entropy in urban and regional modelling. Pion, London.
Wilson, A.G., Bennett, R.J. (1985) Mathematical methods in human geography and planning. Wiley, Chichester [West Sussex] ; New York.
Willekens, F. (1983) Log - linear modelling of spatial interaction. Papers in Regional Science 52, 187-205.