American Journal of Theoretical and Applied Statistics
2015; 4(4): 291-299
Published online July 3, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150404.19
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

The Log Normal and the Poisson Gravity Models in the Analysis of Interactions Phenomena

Giuseppe Ricciardo Lamonica

Department of Economics and Social Sciences, Polytechnic University of Marche, Ancona, Italy

Email address:
g.ricciardo@univpm.it

To cite this article:
Giuseppe Ricciardo Lamonica. The Log Normal and the Poisson Gravity Models in the Analysis of Interactions Phenomena. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 4, 2015, pp. 291-299. doi: 10.11648/j.ajtas.20150404.19

Abstract: Three problems often encountered when bilateral interaction data are analyzed by means of the log-normal gravity model: the bias created by the logarithmic transformation, the failure of the homoscedasticity assumption and the treatment of zero valued flows. When the interaction are count data type that takes non-negative integer values, to overcome these problems the literature suggests to use a Poisson gravity model instead of log-normal model. In this paper, using a real interaction phenomenon a comparative analysis of the two models is carried out. The most important results obtained highlights that if the phenomenon is correctly specified, the two specification of the gravity model have a very similar behaviour.

Keywords: Gravity Model, Poisson Model, Log Normal Model, Comparisons, Count Data

1. Introduction

The analysis of interactions (or flows) phenomena of any type is an area of particular interest. Its aim is to describe, explain and predict the interactions that arise between the units of a collective.

So many models have been developed for this purpose in the literature that it is impossible to list them here. However, put briefly, it is possible to cluster them into two classic categories: stochastic models and econometric models.

The former are probabilistic models of Markov type, and they aim to highlight the fundamental constants of the interactions (see for example [1]). The latter are all those models characterized by a number of variables (covariates) considered explanatory, and they try to explain and predict the interactions.

Belonging in the latter group is the gravity model, which is considered one of the most important models for the analysis of interaction phenomena, and to which the literature has devoted close interest especially from an empirical point of view (as in: [2], [3], [4], [5], [6], [7], [8], [9], [10] and[11]).

The idea underlying this model is that the interaction which arises between two units of a collective, in conformity with Newton's gravitational law, is directly proportional to the masses of those units and inversely proportional to the distance between them. In its classic form, the model is set out as follows:

$$f_{ij}=\beta_0 \frac{p_i^{\beta_1} p_j^{\beta_2}}{d_{ij}^{\beta_3}} \varepsilon_{ij} \quad (1)$$

Where f_{ij} is the interaction whose origin is the i-th unit whose destination is the j-th unit; p_i and p_j represent the masses of the two units; d_{ij} is the distance between them; and ε_{ij} is the residual variable. Finally, β_0 is a constant of proportionality which, together with the parameters β_1, β_2 and β_3, is subject to estimation.

Considering the logarithm of (1) a double log-linear model (henceforth 'log-normal model') is easily estimated with the Ordinary Least Squares method:

$$\lg(f_{ij})=\lg(\beta_0)+\beta_1\lg(p_i)+\beta_2\lg(p_j)+\beta_3\lg(d_{ij})+\lg(\varepsilon_{ij}) \quad (2)$$

The literature has highlighted that the log-normal model (2) based on the classic hypothesis has a series of potential drawbacks (see for example [12] and [13]). In particular:

- The use of the logarithmic transformation produces estimates of the logarithm of the covariates, not of the covariates themselves. The antilogarithms are biased estimates (because of Jensen's inequality). A consequence of this is the underprediction of large flows.
- Model (2) assumes the homoscedasticity of the residual variable and, as in [12],the variance of $\lg(f_{ij})$is identical for all ij pairs. Thus, an observed flow of 2in relation to an estimate of 1is as likely as an observed flow of 200

in relation to an estimate of 100. This property, homoscedasticity, is implausible for data sets where there is a wide variation in the size of interaction flows. The first consequence is that the standard errors of the least-squares estimates of the regression parameters are incorrect, and the confidence intervals and tests of hypotheses that use them are invalid. Since standard computer software packages use these formulas, they are inappropriate when heteroscedasticity is present. The second consequence is that the OLS estimators of the parameters of the regression model lose some of their desirable statistical properties. They remain unbiased but no longer have minimum variances even if the correct formulas are used to estimate these variances. This is so because it can be shown that the estimators with minimum variance are the generalized, not the ordinary, least-squares estimators. Generalized least-squares (GLS) is usually applied by an appropriate transformation of the regression model that makes the resulting disturbances homoscedastic ([14]);

- When the flows are zeros, the logarithmic transformation cannot be computed. To avoid this problem, a small positive constant $\alpha \in (0; 1]$ is added to all observations. However, as in [12], when there are many zero flows, the choice of this constant has a considerable impact on the parameters of the model and on its explanatory power.
- If the residual variable of model (2) is normal distributed, also the $\lg(f_{ij})$ is normal and f_{ij} is log-normal distributed. This is unlikely because the flows are nonnegative integer values.

When the dependent variable is of a count data type that takes non-negative integer values – for example the number of people that move from one place to another – to avoid these pitfalls the literature suggests using a Poisson regression model.

This model (see [15]) is based on the hypothesis that if the probability of interaction between two generic units is small and constant, then it is possible to assume that f_{ij} is a realization of the variable F_{ij} with Poisson probability distribution and mean λ_{ij}. Thus the probability that $F_{ij}=f_{ij}$ is:

$$\Pr(F_{ij}=f_{ij})=\frac{e^{-\lambda_{ij}}\lambda_{ij}^{f_{ij}}}{f_{ij}!} \quad (3)$$

Moreover, the parameter λ_{ij} is logarithmically linked with the covariates:

$$\lambda_{ij}=e^{\beta_0+\beta_1\lg(p_i)+\beta_2\lg(p_j)-\beta_3\lg(d_{ij})} \quad (4)$$

If on the one hand, the Poisson gravity model does not present the drawbacks previously mentioned, on the other also this model has some pitfalls. The most important of them is that the model is characterized by one parameter which represents the mean and variance distribution.

When real data are used the variance is often greater than the mean (over-dispersion) and the Poisson regression may not be appropriate for count data.

Another problem with Poisson regression is the excess of zeros, i.e. real data have more zeros than a Poisson regression would predict.

Referring for the details to the numerous econometric manuals existing in the literature, the aim of this paper is to analyzethese two models by means of a real phenomenon in order to identify theirshared characteristics and those specific to each of them.

In particular,the paper will focus on the problem of zero values and that of homoscedasticity of the residual variable in the normal gravity model. The main result obtained is that the normal gravity model is still a reference scheme of undoubted interest for describing and interpret the interactions phenomena and, contrary to several claims in the literature, the benefits of using the Poisson regression are minor and only theoretical.

The paper is organized as follows: section 2 describes the data used and the results obtained, section 3 concludes.

All the analysis were performed using the SAS System software ver. 9.3.

2. Data and Results

To compare the log-normal gravity and the Poisson gravity models the analysis reported by this paper considered as interaction phenomena the migratory flows of resident foreigners for the year 1995 among the Italian regions (see Figure 1 of the appendix) corresponding to the second level of the Nomenclature of Territorial Units for Statistics (NUTS 2).

We are aware that the data used are not really recent. However, this does not limit the goodness of the obtained results that are independent from the age of the data.

Table 10 in the Appendix reports the data used for the inquiry. Excluding the movements within the Italian regions (i.e., f_{ii} for i=1,..,20) from the analysis, the following Table 1 shows the frequency distribution of the observed flows:

Table 1. *Frequency distribution of the observed flows.*

Class interval	Number of flows	Frequency (%)
$0 \le f_{ij} \le 20$	253	65.58
$20 < f_{ij} \le 50$	52	13.68
$50 < f_{ij} \le 100$	40	10.53
$f_{ij} > 100$	35	9.21
Total	380	100

As will be seen, in 1995, the mean size of flows of resident foreigners among the Italian regions was 33.9 and the variance was 3673.52. Furthermore, 65.58% of the flows did not exceed 20 movements and 9.21% were greater than 100. The largest flow was recorded from Lazio to Lombardia and involved 398 migrants. By contrast, 11.32% (43/380) of the flows were zeros.

The distinctive features of the data set considered are that it includes a very large number of zero and small flows, so that it is particularly suited to the type of experimentation carried out in this paper.

When real phenomena are analysed, the gravity model is usually extended in order to consider, besides the classic

determinant, other potential factors that may influence the phenomenon under investigation.

Consequently, as reported by a large body of literature (see for example: [16], [17], [18], [19] and [20]), migratory phenomena are influenced not only by masses and distance but also by economic, social and demographic disparities among the territorial units considered. Hence, for the purposes of the analysis, it was decided to consider, for each Italian region, 18 variables (see the Appendix), in that they were deemed able to measure the principal aspects of the characteristics just mentioned.

Preliminary examination of these indexes revealed the presence of correlations such to counsel against their direct use in the gravity regression model. Consequently, the 18 indicators were synthesised by means of factor analysis.

The results of this analysis are set out in Table13 of the appendix. They show that the factor structure identified has a considerable power of synthesis. The first two factors, considered on the basis of the usual criteria for factorial choice, can be immediately interpreted.

The high and positive coefficients of correlation between the first factor and all the variables of an economic nature suggest identification of this factor as a complex index of the economic structure, while the close correlations of the second factor with the remaining indexes suggest its identification as a complex index of the demographic structure.

For the purposes of the analysis, the following log-normal gravity model(5) was considered, where $F1_i$, $F1_j$ are the first factor (economic factor) in the origin and the destination regions of flows, while $F2_i$, $F2_j$ are the second factor (demographic factor) in the origin and the destination regions of flows. Finally, $\lg(\beta_0)$ and β_i(for i=1,…,7) are the parameters of the model.

$$\lg(f_{ij})=\lg(\beta_0)+\beta_1\lg(p_i)+\beta_2\lg(p_j)-\beta_3\lg(d_{ij})+\beta_4F1_i+\beta_5F1_j+\beta_6F2_i+\beta_7F2_j+\lg(\varepsilon_{ij}) \quad (5)$$

In order to estimate the model parameters, the masses (p_i and p_j) were calculated as the geometric average of the population at the beginning and at the end of the year. The distances (d_{ij}) between the regions were instead calculated by considering the Euclidean distance between the demographic barycentres of each region. The pairs of co-ordinates identifying each regional demographic barycentre were determined by calculating the arithmetic average, weighted with the population, of the latitude and longitude of each provincial capital in the same region.

Since some observed flows were zeros, as in[12], the following experimentation was conducted: a constant α taking values from 0.1 to 1 by 0.1 was added to all flows, and model (5) was fitted.

The results of this analysis are shown in Table 2, which, as said above, does not consider the intra-region flows (i.e. the f_{ii} for i=1,..,20).

Table 2. *Results of the log-normal gravity model for various values of α.*

α values	Intercept	$\lg(p_i)$	$\lg(p_j)$	$\lg(d_{ij})$	$F1_i$	$F1_j$	$F2_i$	$F2_j$
0.1	-28.83	1.06	1.16	-0.80	-0.05	0.70	-0.02	0.13
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.715)	(<.0001)	(0.439)	(0.055)
0.2	-26.21	0.97	1.07	-0.74	-0.03	0.68	-0.03	0.12
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.574)	(<.0001)	(0.504)	(0.017)
0.3	-24.64	0.91	1.02	-0.71	-0.01	0.67	-0.04	0.11
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.810)	(<.0001)	(0.357)	(0.017)
0.4	-23.49	0.87	0.98	-0.69	0.00	0.65	-0.05	0.10
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.994)	(<.0001)	(0.270)	(0.018)
0.5	-22.58	0.84	0.95	-0.67	0.01	0.64	-0.05	0.10
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.833)	(<.0001)	(0.213)	(0.020)
0.6	-21.830	0.82	0.92	-0.65	0.02	0.63	-0.05	0.09
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.702)	(<.0001)	(0.174)	(0.021)
0.7	-21.18	0.80	0.90	-0.64	0.02	0.63	-0.06	0.09
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.596)	(<.0001)	(0.147)	(0.023)
0.8	-20.61	0.78	0.88	-0.62	0.02	0.62	-0.06	0.09
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.509)	(<.0001)	(0.126)	(0.025)
0.9	-20.10	0.76	0.86	-0.61	0.03	0.61	-0.06	0.08
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.438)	(<.0001)	(0.110)	(0.027)
1	-19.64	0.74	0.85	-0.60	0.03	0.60	-0.06	0.08
	(<.0001)	(<.0001)	(<.0001)	(<.0001)	(0.379)	(<.0001)	(0.097)	(0.029)

Legend: p-values in parenthesis

For various values of α, the estimates of the constant (intercept) and the parameters associated with the population size of the regions ($\lg(p_i)$ and $\lg(p_j)$), as well as the parameter relative to the distance ($\lg(d_{ij})$), were always highly significant. The parameter sign of the latter variable was negative and consistent with expectations.

The estimates of the parameters associated with the economic factor ($F1_i$) in the regions of origin were always not significant. By contrast, in the destination regions of flows ($F1_j$) they were always highly significant.

According to the signs, this factor was a push determinant in the regions of origin and a pull determinant in the

destination regions of flows, while in the absolute values a predominant effect of the pull rather than push determinant was evident.

Consideration of the demographic factor of the places of origin ($F2_i$) and of the places of destination ($F2_j$), found that the estimates of the associated parameters were non-significant.

Moreover (Table 3), the White and the Kolmogorov-Smirnov tests showed that the regression residuals were, respectively, homoscedastic and normally distributed. Finally, the index of determination (corrected R^2) was found to be very high (from 77% to 81%).

Similarly the results in [12], also in this analysis if the constant α increases, the parameter estimates associated with the intercept, the masses of the regions, and the distance decrease.

However, due to the inclusion of the two factors, the parameters estimates of the model, contrary to the results in [12], are much more stable, highlighting a quasi-constant effect of α.

Even if we admit that an α-effect exists on the parameters of the model, this is a problem easily solved because the criteria shown in Table 3 indicated that α=1 should be assigned as the optimal value.

This choice concurs with that of several studies which have recommended the use of the lowest possible non-zero count in this situation (see, for example, [21]).

Table 3. *Criteria for assessing goodness of fit of the log-normal gravity model for various values of α.*

α values	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
R^2	0.77	0.79	0.80	0.80	0.81	0.81	0.81	0.81	0.81	0.81
White test of heteroskedasticity	71.82 (0.00)	62.41 (0.01)	55.11 (0.02)	50.88 (0.04)	47.92 (0.07)	45.74 (0.11)	44.02 (0.14)	42.58 (0.18)	41.32 (0.21)	40.20 (0.25)
Kolmogorov test of normality	0.07 (<0.01)	0.06 (<0.01)	0.06 (<0.01)	0.05 (0.03)	0.05 (0.02)	0.05 (0.06)	0.04 (0.10)	0.04 (0.10)	0.04 (0.08)	0.04 (0.11)
Log likelihood	-531.7	-483.8	-456.2	-437.0	-422.5	-410.9	-401.2	-392.9	-385.8	-379.4
AIC	1081.4	985.6	930.4	982.1	863.0	839.8	820.4	803.9	789.5	776.8
BIC	1116.9	1021.1	965.8	927.5	898.5	875.2	855.87	839.4	824.9	812.3
Pearson Chi-Square	5525.6	5218.3	5146.1	5141.0	5162.6	5196.5	5235.9	5278.1	5321.0	5363.8

Legend: p-values in parenthesis

Table 4 shows a more detailed analysis of the goodness of fit of the log-linear model. In particular, it reports the cross tabulation between the fitted and the observed flows. As will be seen, about 57% (the sum of the frequencies on the main diagonal) of the predicted flows match the observed flows because they are classified in the same classes. By contrast, 17% (frequencies above the main diagonal) are overestimated and 26% are underestimated. This latter situation is stronger for the flows greater than 40.

Table 4. *Cross-tabulation between the fitted Vs observed flows of the log-linear model. % frequencies.*

	Fitted flows								
Observed flows	**up to10**	**(10-20]**	**(20-30]**	**(30-40]**	**(40-50]**	**(50-100]**	**(100-200]**	**(200-300]**	**Total**
up to10	40	4.74	1.05	0	0	0	0	0	45.79
(10-20]	6.05	7.89	4.47	0.79	0.53	0	0	0	19.74
(20-30]	1.05	2.37	1.32	2.37	0.53	0	0	0	7.63
(30-40]	0	0.79	1.32	0.53	0	0.53	0	0	3.16
(40-50]	0	0.26	0.53	1.05	0.53	1.05	0	0	3.42
(50-100]	0	1.05	1.58	0.79	2.89	3.95	0.79	0	11.05
(100-200]	0	0	0	0.26	1.05	2.63	2.37	0	6.32
(200-300]	0	0	0	0	0	0.53	0.53	0.53	1.58
over 300	0	0	0	0	0	0	0.53	0.79	1.32
Total	47.11	17.11	10.26	5.79	5.53	8.68	4.21	1.32	100

For comparative purposes, the following Poisson gravity model (6) was estimated and Table 5 shows the results obtained:

$$\lg(\lambda_{ij})=\beta_0+\beta_1\lg(p_i)+\beta_2\lg(p_j)-\beta_3\lg(d_{ij})+\beta_4 F1_i+\beta_5 F1_j+\beta_6 F2_i+\beta_7 F2_j \quad (6)$$

Also in this case, the estimates of the intercept, the parameters associated with the population size of the regions ($\lg(p_i)$ and $\lg(p_j)$), and that relative to the distance ($\lg(d_{ij})$), are significant.

The estimate of the parameter associated with the economic factor ($F1_i$) in the regions of origin is not significant; by contrast, in the destination regions of flows ($F1_j$) it is highly significant.

According to the signs, this factor is a push determinant in the regions of origin and a pull determinant in the destination regions of flows, while in the absolute values a predominant effect of the pull rather than push determinant is evident.

When consideration was made of the demographic factor of the places of origin ($F2_i$) and of the places of destination

($F2_j$), the estimates of the associated parameters were found to be non-significant.

Table 5. *Results of the Poisson gravity model.*

Parameter	Estimates	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chi-Square
Intercept	-22.32	-24.19	-20.45	549.21	<.0001
$lg(p_i)$	0.86	0.77	0.94	391.73	<.0001
$lg(p_j)$	0.91	0.82	1.00	418.65	<.0001
$lg(d_{ij})$	-0.54	-0.64	-0.43	103.52	<.0001
$F1_i$	0.00	-0.07	0.06	0.00	0.9546
$F1_j$	0.72	0.64	0.80	302.89	<.0001
$F2_i$	-0.09	-0.15	-0.02	6.75	0.0094
$F2_j$	0.06	0.00	0.12	4.32	0.0376
Scale	2.98				
Criteria for assessing goodness of fit					
Criterion			Value		
Pearson Chi-square			3791.66		
Pseudo R^2			0.87		
Log likelihood			-2416.96		
AIC			4849.31		
BIC			4881.44		

Also in this case, the parameters associated with the population size of the regions, the distance, and the economic factor in the destination region of the flows are highly significant. Therefore, from this point of view, the two models are equivalent.

The Chi-square and pseudo R^2 indices show that also the Poisson gravity model has high explanatory power. This result is confirmed in Table 6, where, similarly to the previous situation, the cross tabulation between the fitted and the observed flows is reported. In synthesis, 56% of the fitted flows match the observed flows;17% are underestimated; and 27% are overestimated.

Table 6. *Cross-tabulation between the fitted Vs observed flows of the Poisson model. % frequencies.*

	Fitted flows									
Observed flows	up to10	(10-20]	(20-30]	(30-40]	(40-50]	(50-100]	(100-200]	(200-300]	over 300	Total
up to10	38.16	8.95	1.84	0.26	0	0	0	0	0	49.21
(10-20]	2.89	6.84	3.95	2.37	1.05	0.26	0	0	0	17.37
(20-30]	0.79	1.84	1.05	1.84	1.05	0.26	0	0	0	6.84
(30-40]	0.26	0.53	0.79	0.53	0.26	0.53	0	0	0	2.89
(40-50]	0	0	0.53	1.05	0.26	1.58	0.53	0	0	3.95
(50-100]	0	0.79	1.05	1.05	1.32	5.26	1.32	0	0	10.79
(100-200]	0	0	0	0	0.26	2.63	2.37	0.79	0	6.05
(200-300]	0	0	0	0	0	0	0.79	0.26	0.53	1.58
Over 300	0	0	0	0	0	0	0	0.53	0.79	1.32
Total	42.11	18.95	9.21	7.11	4.21	10.53	5	1.58	1.32	100

On comparing the estimates of the Poisson model with the corresponding estimates of the log-normal model (Table 2), in general, no substantial differences are apparent. In particular, using the Euclidean distance between the parameters of the two models as the similarity index, Table 7 shows that for α=0.5 and for α=0.6 the Poisson and log-normal models are extraordinarily coincident. But, if the constant of proportionality is excluded, the similarity between the two models is more marked, with values which decrease with those of the constant. The maximum similarity is reached when the constant is equal to 1.

Table 7. *Similarity (ED) between the Poisson model and the log-normal model for various values of α.*

With the intercept										
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ED	6.5	3.9	2.3	1.2	0.3	0.5	1.1	1.7	2.2	2.7
Without the intercept										
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ED	1.9	1.8	1.7	1.6	1.6	1.5	1.5	1.5	1.5	1.4

Another important characteristic is that the particular similarity between the two models tends to weaken if α approaches 0.1 or 1.

Since the criteria shown in Table 3 and the results set out in Table 7indicated that α=1, a performance analysis was conducted on the predictive power of the Poisson and log-normal (for α=1) models.

Tables8 and 9report respectively the predicted (or fitted) flows ($\widehat{f_{ij}}$) of the log-normal gravity and the Poisson gravity model. On analysing these two tables, once again a uniform behaviour of the two models is apparent.

In particular, Table 8 reports a classification of the predicted flows of the two models according to whether they are greater (overestimated) or smaller (underestimated) than the observed flows.

The Poisson model overestimates 63.68% and underestimates 36.32% of the observed flows. Consequently this model is inclined to overestimate the flows. By contrast, the log-linear model exhibits more uniform behaviour.

Moreover, 84.48% of the predicted flows of the two

models are concordant i.e. 36.32% are underestimated and 48.16% are overestimated.

Table 8. *Cross tabulation between the predicted flows of the two models.*

	Poisson model		
Log-normal model (α=1)	Underestimate ($\hat{f}_{ij}<f_{ij}$)	Overestimate ($\hat{f}_{ij}>f_{ij}$)	Total
Underestimate ($\hat{f}_{ij}<f_{ij}$)	138 (36.32%)	59 (15.52%)	197 (51.84%)
Overestimate ($\hat{f}_{ij}>f_{ij}$)	0 (0.00%)	183 (48.16%)	183 (48.16%)
Total	138 (36.32%)	242 (63.68%)	380 (100.0%)

A more detailed analysis is set out Table 9,which shows the cross tabulation between the fitted and observed flows of the two models.

As will be seen, about 73% of the fitted values of the two models are concordant (that is, classified in the same classes) while 27% are discordant (that is, classified in different classes).

Moreover, considering the flows with a discordant classification, it is very clear that the Poisson fitted flows are in general less than the log-normal fitted flows.

Table 9. *Cross tabulation between the fitted flows of Poisson model Vs fitted flows of log-normal (α=1) % frequencies.*

	Fitted flows of the log-normal model								
Fitted flows of the Poisson model	**up to10**	**(10-20]**	**(20-30]**	**(30-40]**	**(40-50]**	**(50-100]**	**(100-200]**	**(200-300]**	**Total**
up to10	41.58	0.53	0	0	0	0	0	0	42.11
(10-20]	5.53	13.42	0	0	0	0	0	0	18.95
(20-30]	0	3.16	6.05	0	0	0	0	0	9.21
(30-40]	0	0	4.21	2.89	0	0	0	0	7.11
(40-50]	0	0	0	2.11	2.11	0	0	0	4.21
(50-100]	0	0	0	0.79	3.42	6.32	0	0	10.53
(100-200]	0	0	0	0	0	2.37	2.63	0	5
(200-300]	0	0	0	0	0	0	1.58	0	1.58
Over 300	0	0	0	0	0	0	0	1.32	1.32
Total	47.11	17.11	10.26	5.79	5.53	8.68	4.21	1.32	100

Put briefly, from the experimentation conducted it clearly emerges that in regard to the phenomenon analyzed:

- The problem of zeros flows may be easily solved, and the solution is in line with those in the literature: a constant greater or equal to 0.5 is a good choice but the optimal choice is a constant equal to 1
- All the classical hypothesis on the residual variable of the log-normal gravity model are verified.
- The estimates of the parameters of the two types of regression considered are very similar.
- The log-normal model tends slightly to underpredict the flows, whereas the Poisson model tends to overpredict the flows.

In conclusion, if the gravity model, as usually happens in real analysis, is extended in order to consider, besides the classic determinant, other potential factors that may influence the phenomenon under investigation,the log-normal model and the Poisson model have the same behaviours and, contrary to claims in the literature, there are no reasons to prefer one model to the other, especially when the analysis is of explanatory type: that is, determining the covariates that influence the interactions.

3. Conclusion

In the analysis of interaction phenomena of count data type, the literature (see e.g. [12]) suggests using the Poisson regression instead of the log-normal regression because the former model does not have certain drawbacks and seems to perform better in real analysis.

Starting from the hypothesis that the results in the literature are not completely convincing owing to the use of a model that suffers from omitted variables, this paper has compared the two regression models by means of a real interaction phenomenon.

In particular, the comparison was carried out usingthe migratory flows of foreign residents among the Italian regions. Following the literature, in addition to the classic covariates of the gravity model, also the economic, social and demographic disparities among the territorial units were considered.

The most important result obtained is that the two models show, in general, very similar behaviours in terms of both parameter estimates and goodness of fit. The only differences are that the Poisson model tends to overestimate small flows, while the log-linear model tends to underestimate the largest flows.

However, in contrast with the literature, the residual variable of the log-normal gravity model satisfies all the classic hypotheses, and the presence of the zero flows is an easily resolvable problem which does not restrict the model's operability.

In conclusion, if the empirical analysis is of explanatory type, i.e. the goal is only to identify the covariates influencing the interaction phenomena, then both models are equally valid for use in practice. However, since the log-normal gravity model is richer with statistical properties and easier to interpret, it may be preferred to the Poisson model.

By contrast, if the analysis is of predictive type, because the Poisson model guarantees non-negative prevision, it may

be preferable if the data do not show over or under-dispersion, and taking into account that the model overestimate the small flows.

Appendix

Figure 1. *Political map of Italy by regions (NUTS-2).*

Table 10. *Migratory flows of foreign residents between the Italian regions Year 1995.*

	Destination region																			
Origin region	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**	**R13**	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**	**R20**
Piemonte (R1)	2439	16	382	17	114	18	85	97	79	20	37	74	7	1	5	12	1	8	31	14
Valle d'Aosta (R2)	13	77	4	-	2	-	-	5	5	-	2	-	-	-	-	-	-	-	-	-
Lombardia (R3)	290	4	8699	59	304	69	109	382	207	14	73	160	19	3	34	33	2	19	57	21
Trentino-Alto A. (R4)	12	-	68	919	75	28	4	23	10	4	7	4	6	-	5	4	-	-	5	2
Veneto (R5)	48	3	281	63	3955	99	12	157	45	10	18	88	6	-	14	11	3	6	13	24
Friuli-V. G. (R6)	9	-	69	12	165	754	10	28	21	-	3	15	-	1	2	5	-	1	2	2
Liguria (R7)	82	3	155	4	26	6	488	36	100	2	18	12	1	-	10	1	3	2	6	4
Emilia-Romagna (R8)	67	-	325	41	135	26	20	3000	90	13	54	46	10	1	11	13	2	6	28	2
Toscana (R9)	46	1	190	14	107	16	49	161	2254	54	50	115	11	4	18	11	-	22	13	10
Umbria (R10)	16	-	58	5	18	7	3	46	65	467	59	75	7	1	4	3	1	2	1	2
Marche (R11)	26	-	52	4	26	10	6	70	21	20	969	16	36	-	8	7	-	-	11	1
Lazio (R12)	69	2	398	65	224	24	48	189	155	88	113	1740	64	6	61	17	2	25	44	9
Abruzzo (R13)	24	-	59	17	38	-	3	42	29	11	73	45	431	10	7	8	3	-	4	2
Molise (R14)	6	1	16	2	1	2	-	4	6	1	11	16	12	28	1	7	-	2	-	-
Campania (R15)	33	-	191	11	107	10	11	112	77	15	20	75	9	1	582	14	4	6	18	5
Puglia (R16)	49	-	272	19	131	17	10	137	89	9	45	50	21	10	14	371	10	12	5	1
Basilicata (R17)	7	-	34	3	13	4	6	12	6	-	6	5	3	-	15	22	35	8	5	1
Calabria (R18)	26	4	127	15	37	7	5	33	15	6	13	26	1	1	7	19	3	229	22	1
Sicilia (R19)	61	5	283	15	131	10	15	173	73	9	66	53	7	6	15	14	4	26	796	14
Sardegna (R20)	25	-	130	8	51	7	8	27	30	6	6	18	1	3	6	4	1	1	14	305

Table 11. *Flows predicted by the log-normal gravity model for α=1.*

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
R1	-	9.9	278.2	21.8	69.3	20.8	77.9	103.7	69.6	11.9	17.9	54.6	9.9	2.3	16.5	10.6	2.6	5.1	14.6	9.5
R2	11.7	-	16.2	1.4	4.5	1.4	3.7	6.4	4.3	0.8	1.2	3.6	0.7	0.2	1.1	0.7	0.2	0.4	1.0	0.6
R3	245.3	12.1	-	60.5	179.2	47.9	113.5	274.4	155.2	25.4	39.1	112.8	20.8	4.7	33.4	21.5	5.3	10.2	28.3	17.0
R4	28.7	1.6	90.4	-	53.7	12.0	13.8	50.0	25.7	4.9	8.0	21.2	4.1	0.9	6.4	4.2	1.0	1.9	5.2	2.8
R5	82.5	4.6	241.8	48.6	-	44.3	41.5	183.8	90.4	17.8	29.9	74.3	14.6	3.2	21.8	13.9	3.4	6.3	17.0	9.0
R6	23.6	1.4	61.8	10.3	42.3	-	11.7	40.1	24.3	5.8	11.1	25.4	5.5	1.2	8.1	5.4	1.3	2.4	6.1	2.9
R7	74.5	3.0	123.1	10.0	33.4	9.8	-	54.0	38.9	6.1	8.9	27.7	4.9	1.1	8.1	5.1	1.3	2.5	7.0	4.8
R8	81.0	4.3	243.1	29.6	120.7	27.6	44.1	-	116.5	16.4	24.6	67.6	12.3	2.7	18.6	11.5	2.8	5.4	14.8	8.5
R9	71.1	3.7	179.8	20.0	77.6	21.9	41.5	152.4	-	17.3	23.1	71.2	12.0	2.6	18.1	10.8	2.7	5.1	14.4	8.9
R10	17.6	1.0	42.7	5.5	22.1	7.6	9.4	31.1	25.1	-	13.1	43.5	6.5	1.2	8.0	4.3	1.1	2.0	5.4	2.7
R11	25.2	1.4	62.3	8.5	35.3	13.7	13.1	44.3	31.7	12.4	-	47.2	11.6	1.9	12.4	6.9	1.8	3.1	8.0	3.7
R12	77.1	4.3	180.7	22.7	88.2	31.5	40.9	122.3	98.4	41.4	47.5	-	33.1	6.4	44.8	21.4	5.7	10.3	28.4	13.4
R13	20.9	1.2	49.8	6.6	25.8	10.1	10.8	33.3	24.7	9.2	17.3	49.3	-	2.6	15.1	7.2	1.9	3.2	8.1	3.4
R14	6.6	0.4	15.4	2.0	7.6	3.0	3.4	9.8	7.2	2.3	3.9	13.0	3.5	-	9.4	3.1	0.9	1.3	3.2	1.1
R15	61.9	3.6	142.5	18.3	68.7	26.7	31.7	89.5	66.7	20.3	33.1	119.1	27.0	12.4	-	29.7	9.0	14.1	33.7	10.9
R16	40.4	2.4	92.6	12.1	44.4	18.1	20.2	55.9	40.0	10.9	18.7	57.6	12.9	4.1	30.0	-	9.8	11.9	22.3	6.7
R17	9.9	0.6	22.6	2.9	10.8	4.4	5.0	13.8	10.0	2.8	4.7	15.1	3.4	1.2	9.1	9.7	-	3.1	5.7	1.7
R18	23.6	1.4	53.0	6.7	24.4	9.5	11.9	31.6	23.1	6.2	10.0	33.5	6.9	2.1	17.2	14.4	3.8	-	19.6	4.2
R19	47.7	2.8	105.0	12.9	46.8	17.6	24.1	62.1	46.3	12.0	18.6	65.9	12.6	3.6	29.4	19.2	5.0	14.0	-	9.5
R20	32.3	1.8	65.4	7.2	25.6	8.7	17.2	37.1	29.5	6.2	8.9	32.1	5.4	1.3	9.9	6.0	1.5	3.1	9.8	-

Table 12. *Flows predicted by the Poisson gravity model.*

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
R1	-	9.9	378.3	26.1	88.6	24.8	82.4	130.6	82.8	12.5	19.8	72.4	10.3	2.1	17.9	11.4	2.5	5.1	16.1	9.8
R2	9.1	-	15.0	1.2	3.9	1.1	2.7	5.5	3.4	0.5	0.9	3.2	0.5	0.1	0.8	0.5	0.1	0.2	0.7	0.4
R3	326.8	14.1	-	76.9	245.4	62.2	137.0	369.3	200.9	29.3	47.2	164.3	23.6	4.6	40.0	25.4	5.5	11.1	34.4	19.4
R4	31.3	1.5	107.0	-	54.4	11.7	13.6	52.5	26.3	4.4	7.5	24.0	3.6	0.7	5.9	3.8	0.8	1.6	4.9	2.5
R5	107.8	5.2	345.5	55.1	-	50.5	48.7	225.0	108.1	18.5	32.4	98.7	15.0	2.8	23.8	15.0	3.2	6.4	19.1	9.6
R6	26.6	1.3	76.9	10.4	44.4	-	11.8	43.5	25.2	5.2	10.0	28.5	4.7	0.9	7.4	4.9	1.0	2.0	5.8	2.6
R7	75.8	2.8	145.9	10.4	36.8	10.2	-	58.2	39.2	5.5	8.5	31.6	4.4	0.9	7.6	4.7	1.0	2.1	6.7	4.3
R8	97.5	4.4	318.9	32.6	138.1	30.4	47.2	-	124.6	15.8	25.0	83.4	11.9	2.2	19.0	11.6	2.5	5.0	15.5	8.4
R9	86.0	3.9	241.6	22.7	92.3	24.5	44.3	173.4	-	16.5	23.4	86.5	11.4	2.1	18.4	10.9	2.4	4.8	15.0	8.7
R10	18.8	0.9	50.8	5.4	22.9	7.3	8.9	31.8	23.8	-	10.7	42.2	5.0	0.8	6.7	3.6	0.8	1.6	4.8	2.3
R11	28.9	1.4	79.6	9.0	38.8	13.7	13.4	48.9	32.8	10.4	-	50.8	9.4	1.4	11.1	6.2	1.4	2.6	7.5	3.3
R12	107.9	5.2	283.2	29.7	120.8	39.6	51.1	166.3	124.0	41.9	52.0	-	33.0	5.6	48.0	23.5	5.4	10.4	32.0	14.5
R13	24.5	1.2	65.3	7.2	29.4	10.5	11.3	37.9	26.2	8.0	15.4	52.9	-	1.8	13.3	6.4	1.5	2.6	7.6	3.1
R14	6.7	0.3	17.5	1.9	7.5	2.7	3.1	9.7	6.7	1.8	3.1	12.3	2.5	-	6.6	2.3	0.6	0.9	2.5	0.9
R15	93.8	4.7	242.3	25.9	102.3	36.2	43.1	133.2	92.6	23.5	39.8	168.8	29.1	10.7	-	33.1	8.6	14.6	39.5	12.8
R16	59.6	3.0	153.3	16.6	64.4	23.7	26.8	81.3	54.6	12.6	22.2	82.1	14.1	3.7	33.0	-	8.6	11.7	25.3	7.7
R17	11.4	0.6	29.2	3.2	12.3	4.5	5.1	15.6	10.6	2.5	4.4	16.7	2.9	0.8	7.6	7.6	-	2.4	5.1	1.5
R18	32.4	1.6	81.8	8.6	33.2	11.7	14.6	42.9	29.4	6.6	11.2	44.4	7.0	1.8	17.7	14.2	3.2	-	19.9	4.5
R19	71.2	3.6	176.7	18.2	69.5	23.9	32.3	92.0	64.0	14.1	22.9	95.4	14.1	3.4	33.4	21.5	4.9	13.9	-	10.8
R20	40.8	2.0	94.1	8.8	33.0	10.3	19.4	47.2	34.9	6.4	9.6	40.8	5.4	1.1	10.2	6.2	1.4	3.0	10.2	-

Table 13. *Results of the factor analysis.*

Factor	I	II
Variance explained	60.7	17.6
Correlations between the variables and the first two factors		
X1	0.96	0.02
X2	0.97	-0.12
X3	0.93	-0.26
X4	0.97	-0.10
X5	0.93	-0.19
X6	0.57	0.38
X7	-0.85	0.01
X8	-0.03	-0.46
X9	0.86	-0.18
X10	0.86	-0.41
X11	0.94	-0.01
X12	-0.28	0.70
X13	-0.57	0.64
X14	0.78	0.53
X15	0.73	0.56
X16	0.49	0.81
X17	0.77	-0.25
X18	0.80	0.51

Socio-economic and demographic variables

X1) Employment rate = Employed resident population / Total population resident in the region.

X2) Added value per capita = Regional added value / Total population resident in the region.

X3) Added value per person employed = Regional added value / Employed resident population.

X4) GDP per capita = Regional GDP/ Population resident in the region.

X5) GDP per person employed = Regional GDP/Employed resident population.

X6) % of employed in industry = Share of population resident in the region employed in industry.

X7) % of employed in agriculture = Share of population resident in the region employed in agriculture.

X8) % of employed in other activities = Share of population resident in the region employed in activities other than industry and agriculture.

X9) Consumption per capita = Resident population consumption / Total population resident in the region.

X10) Income per capita = Resident population income / Total population resident in the region.

X11) Units of labour per inhabitant = Number of regional labour units / Total population resident in the region.

X12) Size of unit of labour = Number of employed in the region / Number of regional labour units.

X13) Age dependency ratio = Regional resident population aged 65+ / Regional resident population aged 15-64.

X14) Index of turnover in the active population = Regional resident population aged 15-19 / Regional resident population aged 60-64.

X15) Portion of persons aged 65 and over = Regional resident population aged 65+/ Regional resident population.

X16) Old-age dependency ratio = Regional resident population aged 65+ / Regional resident population aged 15-64.

X17) % of resident foreigners to total population = Number of foreigners resident in the region / Total population resident in the region.

X18) Index of active population structure = Regional resident population aged 40-64 / Regional resident population aged 15-39.

References

[1] R.J. Cook, J.D. Kalbfleisch and G.Yi, "A generalized mover-stayer model for panel data," Biostatistics, 3(3), pp. 407-420, 2002.

[2] E.J.Anderson,"A Theoretical Foundation for the Gravity Equation,"The American Economic Review, 63, pp.106-116, 1979.

[3] Y. Chun and D.A. Griffith, "Modeling network autocorrelation in space-timemigration flow data: An eigenvector spatial filtering approach,"Annals of the American Geographer,101, 3, pp. 523-536, 2011.

[4] S.J. Evenett, and W. Keller, "On the theories explaining the success of the gravity equation," Journal of Political Economy, 110, pp. 281-316, 2002.

[5] W. Isard, "Gravity and spatial interaction models" In: W. Isard, I. J. Azis, M. P. Drennan, R. E. Miller, S. Saltzman and E. Thorbecke, "Methods of Interregional and Regional Analysis", Ashgate, U.K 1998.

[6] D. Karemera, O. V. Iwuagwu, and B. Davis, "A gravity model analysis of international migration to North America," Applied Economics, 32, 2000, pp. 1745-1755.

[7] J.P. LeSage, "Spatial regression models". In: Altman, M., Gill, J., and McDonald, M. (eds). "Numerical Issues in Statistical Computing for the Social Scientist" John Wiley & Sons, New York, 2004, pp. 199-218, 2004.

[8] L. Mathyas, "The Gravity Model: Some Econometric Consideration," The World Economy, 21, pp. 397-401, 1998.

[9] J.P. LeSage, and R. Pace, "Spatial econometric modeling of the origin destination flows," Journal of Regional Science 48, pp. 941-967, 2007.

[10] A. Porojan, (2001) "Trade flows and spatial effects: the gravity model revisited", Open Economies Review, 12, 3, pp. 265-280, 2001.

[11] A. Sen and T. E. Smith "Gravity model of spatial Interaction behaviour" Springer Verlag, Berlin, 1995.

[12] R. Flowerdew and M. Aitkin, "A method of fitting the gravity model based on the Poisson distribution," Journal of Regional Science, 22, pp. 191-202, 1982.

[13] M.J. Burger, F.G. Van Oort and G.J.M. Linders, "On the specification of gravity model of trade: zeros, excess zeros and zero-inflated estimation," Spatial Economic Analysis, 2, pp. 167-190, 2009.

[14] W. B. Stronge and R.R. Schultz, "Heteroscedasticity and the gravity model," Geographical Analysis, vol. X, 3, pp. 279-286, 1978.

[15] A. Lovett and R. Flowerdew, "Analysis of count data using Poisson regression," Professional Geographer, 41, pp. 190-198, 1989.

[16] A. Anjomani,"Regional growth and interstate migration,"Socio Economic Planning Sciences, 36, pp. 239-65, 2002.

[17] G.F. De Jong, D.R. Graefeand S.T. Pierre, "Welfare reform and interstate migration of poor families," Demography, 423, pp. 469-496, 2005.

[18] R. Forrest andA. Murie, "Moving strategies among home owners," In Labour Migration: the internal geographical mobility of labour in the developed world. J.H. Johnson and J. Salt (eds), London, David Fulton Publisher, pp. 191-209, 1990.

[19] M. J. Greenwood, "Human Migration: Theory, models and empirical studies," Journal of Regional Studies, 25, 4, pp. 521-544, 1985.

[20] A. Rogers andL.J. Castro,"Migration," In: Migration and settlement: a multiregional comparative study. A. Rogers and F. Willekens (eds), Dordrecht, Netherland, Kluvert Academic Publishers, 1986.

[21] P.A.G. Van Bergeijk, and B. Steven, "The gravity model in international trade: Advanced and application" Cambridge, Cambridge University Press, 2010.