(Optimal) Spatial Aggregation in the Determinants of Industrial Location*

JOSEP-MARIA ARAUZO-CAROD
MIGUEL MANJÓN-ANTOLÍN
QURE and Department of Economics, Rovira i Virgili University
Av. de la Universitat, 1 - 43204 Reus
Spain
josepmaria.arauzo@urv.cat
miguel.manjon@urv.cat

Abstract

Empirical studies on the determinants of industrial location typically use variables measured at the available administrative level (municipalities, counties, etc.). However, this amounts to assuming that the effects these determinants may have on the location process do not extent beyond the geographical limits of the selected site. We address the validity of this assumption by comparing results from standard count data models with those obtained by calculating the geographical scope of the spatially varying explanatory variables using a wide range of distances and alternative spatial autocorrelation measures. Our results reject the usual practice of using administrative records as covariates without making some kind of spatial correction.

Keywords: industrial location, count data models, spatial statistics

JEL classification: C25, C52, R11, R30

1) Introduction

The determinants of industrial location have been widely investigated both theoretically and empirically (HAYTER, 1997; ARAUZO et al., 2010). However, little is known about the *geographical scope* of these determinants (ROSENTHAL and STRANGE 2003; JOFRE-MONSENY, 2009). Most empirical studies use dependent and explanatory variables measured at the available administrative level (provinces, regions, states, etc.) and therefore implicitly assume that the effects the covariates may have on the dependent variable are restricted to span over the geographical area defined by the administrative unit. This assumption, however, is at odds with the theoretical foundations of the New Economic Geography (FUJITA et al., 1999; FUJITA and THISSE, 2002; COMBES et al., 2008).

What, then, is the (possibly optimal) level of spatial aggregation that should be used when investigating the determinants of industrial location? This is a central question in empirical studies because the use of spatial units that differ from those effectively used by agents may bias results to an unknown extent (see, for example, AMRHEIN, 1995).[1] This is also critical for the suitable design and implementation of local public policies aimed at supporting the creation of new firms (LEE, 2008) since a misleading choice of the geographical unit may result in underperformance of a government's investment and, ultimately, in a waste of public funds. Yet this issue has received scarce attention in the literature.

Some studies have acknowledged the importance of accounting for the geographical scope of the covariates by including among them distances (for example, to a major city or infrastructure, as in e.g. GUIMARÃES et al., 2000 and FIGUEIREDO et al., 2002), spatial effects and/or spatially lagged variables (AUTANT-BERNARD, 2006, LAMBERT et al., 2006, ALAÑÓN et al., 2007). The statistically significant coefficients they generally report indeed indicate that the assumption of no geographical scope in the determinants of industrial location does not hold. However, the question of what geographical extension these determinants may have remains unaddressed.

In this paper we seek to (partially) fill this gap in the literature by examining the extent to which the establishment of new concerns in a particular site is driven by the characteristics of that particular site and/or by the (average) characteristics of the surrounding area. Specifically, we estimate standard count data models for the (per period) number of new concerns created in a particular site and compare the results obtained from using explanatory variables constructed from administrative records with those obtained by calculating the geographical scope of the explanatory variables that have spatial variation. In particular, we use different distances (some of which roughly define functional territorial units such as Travel-To-Work-Areas—TTWAs—and administrative territorial units such as counties) and spatial autocorrelation measures calculated at the global and local level to construct the spatially lagged variables. Results using data from Catalonia indicate that the usual practice of using administrative records as covariates without making some kind of spatial correction may provide misleading conclusions.[2]

The rest of this paper is organised as follows. Section 2 reviews previous related studies and discusses the empirical strategy. Section 3 deals with the evidence: we first present the data set, then the spatial explanatory analysis and finally the main econometric results. Section 4 provides our conclusions.

2) What do we know about the geographical scope of the determinants of industrial location and how can we empirically investigate it?

2.1 Related studies

When an individual entrepreneur or an established firm has to choose the location of a new establishment, do they look at the characteristics of just a narrow area (as defined, for example, by municipalities, TTWA or counties) or do they care more about the characteristics of a larger, broader area (as defined, for example, by provinces, regions, or states)? For

example, if the availability of skilled labour is one of the main determinants of location decisions (COUGHLIN and SEGEV, 2000), where do new concerns look for it? Do they require skilled workers who live where they are planning to locate or is it enough for them to have this input scattered over a nearby area (perhaps with good commuting infrastructures)?

These questions are key for economic policymakers because they ultimately determine the level of geographical aggregation that should be taken as reference in the design and implementation of local policies (LEE, 2008). They are also important for researchers interested in the topic because omitting the geographical scope of the determinants may entail a severe specification error (AMRHEIN 1995). However, as ARAUZO et al., 2010 show in their recent review of the literature on empirical industrial location, these questions have only been partially and/or indirectly investigated.[3]

Among the studies that have partially investigated this issue (in the sense that they focus on a single determinant of the industrial location process) is the seminal contribution of ROSENTHAL and STRANGE, 2003: 377 on "the geographic scope of agglomerative externalities". These authors conducted a microgeographic analysis of agglomeration using ZIP codes as a geographical unit, therefore measuring the spatial extent of agglomeration economies in terms of mile-rings rather than administrative units. "The paper's most important finding is that agglomeration economies attenuate with distance".[4]

Among the studies that have indirectly addressed the issue, one strand of the literature has used distance (either in time or in km/miles) to link the sites where firms locate with the principal characteristics of the surrounding areas. We may therefore find extensive evidence on the role of distance to transport infrastructures (e.g. HOLL, 2004a, 2004b), to main cities (e.g. GUIMARÃES et al., 2000 and FIGUEIREDO et al., 2002), to Central Business Districts (e.g. FINNEY, 1994 and WU, 1999), to markets (e.g. KITTIPRAPAS and MCCANN, 1999 and VAN DIJK and PELLENBARG, 2000), to suppliers (e.g. VAN DIJK and PELLENBARG, 2000 and KLIER and MCMILLE, 2008), to universities (e.g. EGELN et al., 2004 and WOODWARD et al., 2006), and to the home country in the case of FDI (e.g. CROZET et al., 2004 and DISDIER and MAYER, 2004). Notice, however, that these studies implicitly assume that all the geographical scope that may exist in the determinants of industrial location is somehow embedded in the measure of distance.

Other related studies have analysed how the results are affected by the selection of a particular territorial level. ARAUZO and MANJÓN, 2004, for example, compare results from different aggregation levels and conclude that firms seem to choose mainly between medium-size administrative units ("*comarques*") rather than between small administrative units (municipalities). ARAUZO, 2008 extended this work by adding functional territorial units (TTWAs) to the analysis but found little differences in the determinants of industrial location across functional and administrative units. Notice, however, that both studies use dependent and explanatory variables measured at the same administrative/functional level, which means that, *de facto*, they cannot analyse the geographical scope of the determinants. In a different but complementary way, MAYER and MUCCHIELLI, 1999 approach such location decisions from different perspectives using several nested structures, namely centre-periphery and country-region.

Finally, other studies have used spatial econometric techniques to control for the fact that industrial location data are georeferenced. LAMBERT et al., 2006, for example, assume that the marginal effects of the explanatory variables vary across locations due to unobserved specific factors in a Geographically Weighted Regression and a Poisson Spatial Generalized Linear Model, whereas AUTANT-BERNARD, 2006 and ALAÑÓN et al., 2007 find that the spatially lagged explanatory variables included in their Conditional Logit and (Bayesian) Probit specifications, respectively, help to explain the location decisions of firms. Notice, however, that none of these studies concerns the level of geographical aggregation but the mere existence of spatial effects.

All in all, we can conclude that there exists indirect evidence to show that some space-related measures are required in the study of industrial location. However, no previous study seems to have empirically addressed the question of what is the geographical scope of the determinants of industrial location (beyond the case of agglomeration economies). It seems necessary, therefore, to provide a brief discussion of how we intend to investigate the issue.

2.2 Empirical strategy

Our empirical strategy is similar to that used to study the MAUP (DURANTON and OVERMAN, 2005 and 2008, BRIANT et al., 2008) and the geographical scope of agglomeration economies (ROSENTHAL and STRANGE, 2003, JOFRE-MONSENY, 2009). In essence, our idea is to compare results from a baseline specification that reproduces a widely used model in the literature (using variables measured at the smaller available functional/administrative unit) with those obtained when *A)* we replace the explanatory variables that have spatial variation by their spatially lagged versions, or *B)* we add to the baseline specification spatially lagged versions of the original spatially varying explanatory variables. Individually and jointly significant coefficients associated with these spatially lagged variables would provide the first evidence against the usual implicit assumption of no geographical scope in the determinants of industrial location. In addition, specification tests and model selection criteria should indicate that the extended specifications (i.e. with spatially lagged variables) perform better than the baseline specification.[5]

We implement this idea using count data models.[6] In particular, since our data is characterised by "overdispersion" and an "excess of zeros" (see MULLAHY, 1997 and Table 3 below), we estimate commonly used extensions of the standard Poisson regression model that deal with these characteristics: the Negative Binomial Model (henceforth NBM), the Zero Inflated Poisson Model (ZIPM) and the Zero Inflated Negative Binomial Model (ZINBM). As for the tests and information criteria, we follow CAMERON and TRIVEDI, 1998, 2005 in using the following statistics to determine which specifications perform better for our data: the value of the log-likelihood function (denoted by "Log L" in the tables of results), the Akaike Information Criterion ("AIC"), the Chi-Square Goodness-of-Fit test ("GoF Test"), the Likelihood-Ratio test for the joint significance of the model ("LR Joint Test"), an LR-type test between the ZIPM in the ZINBM based on the null hypothesis of "equidispersion" ("LR Inflated Test"), and a non-nested testing procedure that discriminates between Poisson and Negative Binomial models and their respectively inflated specifications, ZIPM and ZINBM (the so-called "Vuong Test").

Notice, however, that to address the central question of which level of spatial aggregation should be used when investigating the determinants of industrial location, we ultimately need to compare estimates obtained from using spatially lagged covariates constructed for different distances. Ideally, this would mean estimating alternative specifications for each combination of the set of explanatory variables with spatial variation that results from alternative spatially lagged values calculated over the range of possible distances (for example, from zero to the maximum distance between the spatial units of the geographical area we may have considered), but this is clearly unfeasible. For example, a vector of five explanatory variables in a setting where distances measuring the geographical scope are assumed to vary every 10 km in an area where the largest distance is 100 km would result in 10^5 alternative specifications. In fact, the problem becomes even more involved if one takes into account that alternative spatial correlation measures can be used to construct the spatially lagged variables. Consequently, we need to impose several restrictions if we want to compare a relatively small number of specifications.

In this paper, therefore, we do not consider the possibility that different variables have different geographical scopes. This may seem a strong assumption but, given the number of explanatory variables with spatial variation in our data set (18), the number of possible combinations would make comparisons between specifications practically impossible. We have also limited comparisons across spatial correlation measures to a comparison between a measure calculated at the global level and one calculated at the local level. We have therefore used either the Global Spatial Autocorrelation (Moran's I) or the Local Index of Spatial Association (LISA) developed by ANSELIN, 1995 to compute the spatially lagged variables.[7]

Within these two major constraints, we have considered a wide range of distances that are consistent with the characteristics of the region we are analysing (Catalonia). In particular, we have explored 10 km variations ranging from 10 km to 100 km. The lower limit of the range and the 10 km increments were given by the average distance (rounded to the first digit) between municipalities (5.8 km). Another reason for using 10 km variations was the fact that, since the average distance between TTWAs and between "counties" in Catalonia is 20.8 km and 27.9 km, respectively, this allow us to somehow consider functional and administrative units.[8] That is, we can interpret results obtained from 20 km and 30 km as being representative of functional and administrative units, respectively. Finally, the upper limit arises from calculations of Moran's I, which becomes practically negligible for distances over 100 km (see Figure 1 below).

In summary, our empirical strategy entails comparing results from alternative count data models (NBM, ZIPM and ZINBM) that use explanatory variables constructed from administrative records with those obtained by either replacing them with or adding to them spatially lagged variables calculated using different distances (10 km variations ranging from 10 km to 100 km) and alternative spatial correlation measures (Moran's I and LISA). In particular, we estimate the following specifications:

- Specification 1: Baseline model using municipalities data.
- Specification 2.A: We replace the explanatory variables of the baseline model that have spatial variation with spatially lagged variables calculated using Moran's I and a neighbourhood criterion varying from 10 km to 100 km (10 km variation).

- Specification 2.B: We add to the explanatory variables of the baseline model spatially lagged variables calculated using Moran's I and a neighbourhood criterion varying from 10 km to 100 km (10 km variation).
- Specification 3.A: We replace the explanatory variables of the baseline model that have spatial variation with spatially lagged variables calculated using Moran's I and a neighbourhood criterion varying from 10 km to 100 km (10 km variation) if LISA indicates a significant spatial autocorrelation (otherwise the value of the original variable remains unchanged).[9]

3) Data, spatial exploratory analysis and econometric results

3.1 Data

To perform the empirical analysis described in the previous section, we use data on the location of new manufacturing establishments in the municipalities of Catalonia (provided by the Catalan Manufacturing Establishments Register) and data on several characteristics of these municipalities (provided by the Catalan Statistical Institute, the Catalan Cartographical Institute and TRULLÉN and BOIX, 2005). Table 1 lists the definitions and descriptive statistics for both dependent and explanatory variables. In particular, the dependent variable for the count data models used in this study is the number of new manufacturing establishments (codes 12 to 36 of NACE classification) created in each Catalan municipality in 2002. Data for the explanatory variables refer to 2001 (except for residential population change, which is defined over the period 1991 to 2001) and cover most of the factors that have been investigated in the literature (see Arauzo et al. 2010). Below we list these factors and the variables used to proxy them:

- Agglomeration economies. Residential population change between 1991 and 2001 (RES_VAR), urbanisation economies (URB), disurbanisation economies (DISURB), jobs (JOB) and population density (DENS).
- Industrial mix. Manufacturing concentration index (CONC), percentage of manufacturing jobs (JOB_IND) and percentage of jobs in services (JOB_SER).
- Education. Percentage of population older than 10 years of age with technical secondary school (TEC_SEC), percentage of population older than 10 years of age with secondary school (SEC), percentage of population older than 10 years of age with a 3-year degree (DEG), and percentage of population older than 10 years of age with a 4-year degree or a PhD (DEG_PHD).
- Transport infrastructures. Travel time to the capital of the province (TT_CP), dummy for rail station (RAIL), travel time to the closest airport (TT_AIR) and travel time to the closest port (TT_PORT).
- Knowledge. Jobs in high-tech industries (JOB_HT) and in high-tech manufacturing industries (JOB_HT_MA).
- Commuting. Population working and living at municipality "j" over jobs at "j" (POP_JOB) and population working and living at "j" over population living at "j" and working at "j" or elsewhere (POP_JOB_E).
- Population. Population aged between 20 and 44 (POP_20_44).[10]
- Location. Dummies for the municipalities of each province (GIRONA, LLEIDA and TARRAGONA, with Barcelona's municipalities as the residual category), a dummy

for the capitals of "comarques" (CAP_COM), a dummy for shore-line areas (COAST), distance (km) to the nearest city with at least 100,000 inhabitants (DIST_100), and distance (km) to the capital of Catalonia (DIST_CAT).

- Firms. Percentage of small firms (FIRM_SMALL).

[INSERT TABLE 1 HERE]

3.2 Spatial Exploratory Analysis

We calculate the spatially lagged variables as $W\_X = WX$, where X is a matrix containing the spatially varying explanatory variable and W is an appropriate (row standardised) spatial neighbour matrix. This spatial neighbour matrix is a symmetric matrix with 1/0 values (divided by its row sum) depending on whether every two sites are considered as neighbours (here neighbourhood is defined in terms of a predefined distance). In particular, we use 10 neighbour matrices, where the neighbourhood criterion ranges from 10 km to 100 km with 10 km variations. This means that, for example, two municipalities located within an area of 20 km are not considered neighbours according to the first criterion (i.e. they have a value of 0 in the corresponding 10 km spatial neighbour matrix) but are considered neighbours according to the second criterion (i.e. they have a value of 1 in the corresponding 20 km spatial neighbour matrix). These matrices are also used to calculate the Global Spatial Autocorrelation (Moran's I) and the Local Index of Spatial Association (LISA) of each spatially varying explanatory variable (except for those variables that are based on distances, since it would be meaningless to do so).

[INSERT FIGURE 1 HERE]

In Figure 1 we graphically report Moran's I for each of these spatially varying explanatory variables. Spatial autocorrelations are non-negligible for variables related to population, education and agglomeration economies, which means that, at least for these variables, using values measured at the municipality level may lead to biased estimates due to the omission of their geographical scope (AMRHEIN, 1995). Also, Moran's I estimates are generally small and diminish with distance. In fact, estimates for distances of over 100 km are practically zero.

[INSERT FIGURE 2 HERE]

However, as the Global Spatial Autocorrelation is the result of simultaneous measurements for many locations, it is often useful to compare its values with those obtained using local measures of spatial autocorrelation. In this way we can assess the extent to which the results are driven by spatial autocorrelation phenomena occurring in specific areas of the analysed territory. To illustrate the problem, let us consider the urbanization economies. As this variable has the highest Moran's I (0.4399) when the 10 km neighbour matrix is used, one could conclude that this variable is spatially correlated across all the municipalities in Catalonia.

[INSERT FIGURE 3 HERE]

This conclusion would be misleading, however. As Figure 2 shows, urbanization economies are in fact not spatially correlated for a great deal of municipalities. Also, Figure 3 shows that a wider definition of neighbourhood (in this case, municipalities within a 40 km range) implies lower values of the local spatial autocorrelation and a higher number of positively autocorrelated municipalities. This example illustrates the importance of comparing results across different neighbourhood criteria.

3.3 Model Selection and Estimates

The large number of specifications (99) and explanatory variables (47 in specification 2.B, and 29 in the baseline model and specifications 2.A and 3.A) we consider makes it unfeasible to report the whole set of econometric results. We will therefore compare and select some of the specifications in terms of their fit and report detailed results for these only. In Table 2 we therefore report the values of the log-likelihood function (Log L), the Akaike Information Criterion (AIC) and the Chi-Square Goodness-of-Fit test ("GoF Test") for the various specifications (baseline, 2.A, 2.B and 3.A), models (NB, ZIP and ZINB) and distances (10km to 100km).

[INSERT TABLE 2 HERE]

Several trends can be observed for the figures in Table 2.[11] First, the inflated versions of the count regression (ZIPM and ZINBM) perform better than the NBM in terms of log-likelihood and AIC. Second, the baseline model provides a good fit but is generally beaten by the specifications that account for the geographical scope of the determinants (particularly by specifications 2.B and 3.A). Third, the best fit seems to be obtained in the 40 km–80 km range. Fourth, except for specification 2.A with a neighbourhood criterion of 60 km (and less clearly for a neighbourhood criterion of 70 km), the GoF Test indicates that most specifications are likely to be misspecified.

However, this exercise of model comparison and selection may be subject to a pre-testing bias. For our purposes this means that we cannot categorically conclude that the geographical scope of the determinants of industrial location spans a range of exactly 60 km. However, we can confidently reject the usual practice of using covariates measured at the available administrative level. This conclusion is further supported by the individual and joint statistical significance of the spatially lagged variables.[12]

Bearing in mind these caveats, we then concentrate on analysing the baseline specification and the models that provide a good fit while appearing to be well specified. In Table 3, therefore, we only report tests and estimates of the marginal effects obtained from the baseline model and specification 2.A, in which we replaced the explanatory variables of the baseline model that have spatial variation with spatially lagged variables calculated using Moran's I and a neighbourhood criterion of 60 km. We focus on Negative Binomial and Zero Inflated Negative Binomial models because, although the latter shows signs of misspecification

according to the GoF Test, results from the Vuong Test (reported at the bottom of Table 3) indicate that it fits the data better.

[INSERT TABLE 3 HERE]

The estimated marginal effects of the baseline model were as expected (see ARAUZO et al., 2010). They show the importance of agglomeration economies, the industrial mix, knowledge, population and institutional and geographical characteristics as determinants of industrial location. In contrast, education, commuting and the presence of small businesses were not statistically significant. Interestingly, most of these results hold when the geographical scope of these determinants is considered. However, there are also several important differences. First, agglomeration economies are no longer relevant. Second, the effect of the industrial mix variables is the opposite. Third, proxies for commuting and, to a lesser extent, education now become significant. Finally, marginal effects differ, sometimes considerably. All in all, these estimates show that using the baseline specification to make inferences about the determinants of industrial location may be misleading.

4) Conclusions

Consistent with the main tenets of the New Economic Geography, this paper shows that there is some geographical scope in the determinants of industrial location. Whereas previous empirical studies have typically resorted to explanatory variables measured at the available administrative level, in this study we explored the use of spatially lagged variables. Specifically, we estimated count data models on the number of new establishments created in each municipality of Catalonia (Spain) using covariates calculated for different distances and alternative spatial autocorrelation measures and compared these estimates with those obtained with variables measured at the municipality level.

Our results show that the best fit was achieved when we used spatially lagged variables defined by a neighbourhood criterion of 60 km. Admittedly, this figure may be subject to a certain pre-testing bias. Also, it may be different in geographical areas that have different characteristics (institutional, physical, legal, etc.) from the one investigated here. We will leave for future research the question of whether our conclusions hold for alternative settings (for example, if they are derived from discrete choice models and/or different correlation measures). In any case, our estimates soundly reject the usual practice of using administrative records as covariates without making some kind of spatial correction. This calls into question some of the conclusions from previous studies while supporting those based on microgreographic data.

References

[1] HAYTER, R.: *The dynamics of industrial location. The factory, the firm and the production system*, Wiley. (1997)

[2] ARAUZO, J.M.; LIVIANO, D. and MANJÓN, M.: Empirical Studies in Industrial Location: An Assessment of their Methods and Results. In: *Journal of Regional Science*, forthcoming. (2010)

[3] ROSENTHAL, S.S. and STRANGE, W.C.: Geography, industrial organization and agglomeration. In: *The Review of Economics and Statistics* Vol. 85 (2) (2003)**,** pp. 377-393.

[4] JOFRE-MONSENY, J.: The scope of agglomeration economies: Evidence from Catalonia. In: *Papers in Regional Science*, forthcoming. (2009)

[5] FUJITA, M.; KRUGMAN, P. and VENABLES, A.: *The Spatial Economy. Cities, Regions and International Trade*. Cambridge (Mass.), MIT Press. (1999)

[6] FUJITA M. and THISSE, J. -F.: *Economics of agglomeration. Cities, industrial location, and regional growth*. Cambridge (UK), Cambridge University Press. (2002)

[7] COMBES, P.-P.; MAYER, T. and THISSE, J.-F.: *Economic Geography*, Princeton University Press. (2008)

[8] AMRHEIN, C.: Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations. In: *Environment and Planning A* Vol. 27 (1995), pp. 105-119.

[9] OPENSHAW, S. and TAYLOR, P.J.: A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem. In: WRIGLEY N. (Eds), *Statistical Applications in the Spatial Sciences* (1979), pp.127-144. London, Pion.

[10] LEE, Y.: Geographic redistribution of US manufacturing and the role of state development policy". In: *Journal of Urban Economics* Vol. 64 (2008), pp. 436-450.

[11] GUIMARÃES, P.; FIGUEIREDO, O. and WOODWARD, D.: Agglomeration and the Location of Foreign Direct Investment in Portugal. In: *Journal of Urban Economics* Vol. 47 (2000), pp. 115-135.

[12] FIGUEIREDO, O.; GUIMARÃES, P. and WOODWARD, D.: Home-field advantage: location decisions of Portuguese entrepreneurs. In: *Journal of Urban Economics* Vol. 52 (2002), pp. 341-361.

[13] AUTANT-BERNARD, C.: Where Do Firms Choose to Locate their R&D? A Spatial Conditional Logit Analysis on French Data. In: *European Planning Studies* Vol. 14 (2006), pp. 1187-1208.

[14] LAMBERT, D.M.; MCNAMARA, K.T. and GARRETT, M.I.: An Application of Spatial Poisson Models to Manufacturing Investment Location Analysis. In: *Journal of Agricultural and Applied Economics* Vol. 38 (2006), pp. 105-121.

[15] ALAÑÓN, Á.; ARAUZO, J.M. and MYRO, R.: Accessibility, agglomeration and location, in: ARAUZO J.M. and MANJÓN M. (Eds.), *Entrepreneurship, Industrial Location and Economic Growth* (2007), pp. 247-267, Chentelham: Edward Elgar.

[16] COUGHLIN, C.C. and SEGEV, E.: Location determinants of new foreign-owned manufacturing plants. In: *Journal of Regional Science* 40 (2000), pp. 323-351.

[17] DURANTON, G. and OVERMAN, H.G.: Testing for Localization Using Microgeographic Data. In: *Review of Economic Studies* Vol. 72 (2005), pp. 1077-1106.

[18] BRIANT, A.; COMBES, P.-P. and LAFOURCADE, M.: Dots to Boxes: Do the Size and Shape of Spatial Units Jeoparsize Economic Geography Estimators?, Working Paper No. 6928 (2008), Centre for Economic Policy Research (CEPR).

[19] HOLL, A.: Transport Infrastructure, Agglomeration Economies, and Firm Birth. Empirical Evidence from Portugal. In: *Journal of Regional Science* Vol. 44 (2004a), pp. 693-712.

[20] HOLL, A.: Manufacturing Location and Impacts of Road Transport Infrastructure: Empirical Evidence from Spain. In: *Regional Science and Urban Economics* Vol. 34 (2004b), pp. 341-363.

[21] FINNEY, M.: Property tax effects on intrametropolitan firm location: further evidence. In: *Applied Economic Letters* Vol. 1 (1994), pp. 29-31.

[22] WU, F.: Intrametropolitan FDI firm location in Guangzhou, China: A Poisson and negative binomial analysis. In: *Annals of Regional Science* Vol. 33 (1999), pp. 535-555.

[23] KITTIPRAPAS, S. and MCCANN, P.: Industrial location behaviour and regional restructuring within the Fifth 'Tiger' Economy: evidence from the Thai electronics industry. In: *Applied Economics* Vol. 31 (1999), pp. 37-51.

[24] VAN DIJK, J. and PELLENBARG, P.H.: Firm Relocation Decisions in the Netherlands: An Ordered Logit Approach. In: *Papers in Regional Science* Vol. 79 (2000), pp. 191-219.

[25] KLIER, T. and MCMILLEN, D.P.: Evolving Agglomeration in the U.S. Auto Supplier Industry". In: *Journal of Regional Science* Vol. 48 (2008), pp. 245–267.

[26] EGELN, J.; GOTTSCHALK, S. and RAMMER, C.: Location Decisions of Spin-offs from Public Research Institutions. In: *Industry and Innovation* Vol. 11 (2004), pp. 207-223.

[27] WOODWARD, D.; FIGUEIREDO, O. and GUIMARÃES, P.: Beyond the Silicon Valley: University R&D and high-technology location. In: *Journal of Urban Economics* Vol. 60 (2006), pp.15–32.

[28] CROZET, M.; MAYER, T. and MUCCHIELLI, J.-L.: How do firms agglomerate? A study of FDI in France. In: *Regional Science and Urban Economics* Vol. 34 (2004), pp. 27-54.

[29] DISDIER, A.-C. and MAYER, T.: How Different is Eastern Europe? Structure and Determinants of Locational Choices by French Firms in Eastern and Western Europe. In: *Journal of Comparative Economics* Vol. 32 (2004), pp. 280-296.

[30] ARAUZO, J.M. and MANJÓN, M.: Firm Size and Geographical Aggregation: An Empirical Appraisal in Industrial Location. In: *Small Business Economics* Vol. 22 (2004), pp. 299-312.

[31] ARAUZO, J. M.: Industrial Location at a Local Level: Comments on the Territorial Level of the Analysis. *In: Tijdschrift voor Economische en Sociale Geografie- Journal of Economic & Social Geography* Vol. 99 (2008), pp. 193-208.

[32] MAYER, T. and MUCCHIELLI, J.L. : La localisation à l'étranger des entreprises multinationales. In : *Économie et Statistique* Vol. 326-327 (1999), pp. 159-176

[33] DURANTON, G. and OVERMAN, H.G.: Exploring the Detailed Location Patterns of U.K. Manufacturing Industries using Microgeographic Data. In: *Journal of Regional Science* Vol. 48 (2008), pp. 213-243.

[34] ARAUZO, J.M. and VILADECANS, E.: Industrial Location at the Intra-metropolitan Level: The Role of Agglomeration Economies. In: *Regional Studies* Vol. 43 (2009), pp. 545-558.

[35] MULLAHY, J.: Heterogeneity, Excess Zeros, and the Structure of Count Data Models. In: *Journal of Applied Econometrics* 12 (1997): pp. 337-350.

[36] CAMERON, A.C. and TRIVEDI, P.K: *Regression Analysis of Count Data*, Cambridge: Cambridge University Press. (1998)

[37] CAMERON, A.C. and TRIVEDI, P.KV: *Microeconometrics,* Cambridge University Press. (2005)

[38] ANSELIN, L.: Local Indicators of Spatial Association – LISA. In: *Geographic Analysis* Vol. 27 (1995), pp. 93-115.

[39] BOIX, R. and GALLETTO, V.: Sistemas Locales de Trabajo y Distritos Industriales Marshallianos en España. In: *Economía Industrial* Vol. 359 (2006), pp. 165-184.

[40] TRULLÉN, J. and BOIX, R.: Indicadors 2005, Diputació de Barcelona and Universitat Autònoma de Barcelona. (2005)

[41] MANJÓN, M.: *Chi-Square Tests for Count Data Models*, mimeo. (2009)

* This research was partially funded by SEJ2007-64605/ECON, SEJ2007-65086/ECON, the "Xarxa de Referència d'R+D+I en Economia i Polítiques Públiques" of the Catalan Government and the PGIR program N-2008PGIR/05 of the Rovira i Virgili University (funded by both the Catalan and Spanish Governments). This paper has benefited from discussions with Á. Alañón, D. Liviano, F. Pablo and E. Viladecans. We would also like to acknowledge the helpful and supportive comments of seminar participants at the EEFS 2009 Conference (University of Warsaw) and at the Workshop on "Entrepreneurial Activity and Regional Competitiveness" (Max Planck Institute of Economics & ORKESTRA-Basque Institute of Competitiveness). Any errors are, of course, our own.

[1] This can be seen as a particular case of the so-called "Modifiable Area Unit Problem" (henceforth MAUP) originally described by OPENSHAW and TAYLOR (1979).

[2] Catalonia is an autonomous region of Spain that has about 7 million inhabitants (15% of the Spanish population), covers an area of 31,895 km^2 and contributes 19% of Spanish GDP. The capital of Catalonia is the city of Barcelona. Counties in Catalonia are known as "comarques".

[3] In addition to the different strands of empirical industrial location literature, some related studies have investigated the MAUP (OPENSHAW and TAYLOR 1979). However, these studies were generally not concerned with the determinants of industrial location but with issues such as the spatial distribution of new concerns (DURANTON and OVERMAN 2005 and 2008) and the estimation of wage and gravity equations (BRIANT et al. 2008).

[4] See also JOFRE-MONSENY (2009) for a recent application to the same Spanish region that is investigated here.

[5] As is common in the industrial location literature, our empirical strategy implicitly assumes that the administrative unit to which variables refer is indeed the spatial unit that agents effectively use when taking location decisions. Since we are using municipalities data, we believe that this is a plausible assumption (see, however, ARAUZO and MANJÓN 2004 and ARAUZO 2008). One may still argue that this assumption may not hold for large municipalities and metropolitan areas, so we performed some robustness tests that essentially meant dropping from our data set municipalities with more than 250,000 people (in our case, the city of Barcelona) and those that are part of a metropolitan area (around the cities of Barcelona, Girona, Lleida, Manresa and Tarragona). Though results barely changed in the first case, we found that dropping the metropolitan areas from our sample provided different results from those reported below in terms of preferred specification and neighbourhood criterion (though not much in terms of value and significance of the marginal effects). This may be interpreted as evidence that the location processes in metropolitan and non-metropolitan areas are different (ARAUZO and VILADECANS 2009). However, for the sake of simplicity we do not explore this possibility here but leave it for future research.

[6] One reason for using count data models is that they are probably the most popular specifications in recent empirical studies of industrial location (ARAUZO et al. 2010). Another reason is that, at least in our empirical strategy, they have a comparative advantage over discrete choice models (the other specification used in this literature) when detecting misspecifications arising from a misleading definition of the geographical scope of the determinants. This advantage arises from the fact that as discrete choice models distinguish between location determinants related to the agent taking the decision and those related to the set of spatial units from which the choice is made, detecting such misspecifications would require correctly specifying the vector of entrepreneur/firm characteristics (otherwise the conclusions may be misleading). This is obviously not necessary in count data models because they only consider the characteristics of the spatial units.

[7] When using Moran's I, we replace the value of the original variable with the value of the spatially lagged variable**.** When using LISA, we replace the value of the original variable with the value of the spatially lagged variable if there is a significant spatial autocorrelation in that particular municipality (otherwise the value of the original variable remains unchanged).

[8] Counties, known as "comarques" in Catalonia, are territorial units formed by adjacent municipalities. The average area of the 41 "comarques" in Catalonia is 781 km^2. As for the TTWAs, according to BOIX and GALLETTO (2006) there are 74, with an average area of 433 km^2.

[9] We did not consider Specification 3.B, i.e. one in which we would add (rather than replace the original variables by) the spatially lagged variables calculated as in Specification 3.B because the high correlation between the original variables and these spatially lagged variables (around 0.95 in 6 of the 18 variables) resulted in severe multicollinearity.

[10] We use residential population as the only explanatory variable in the inflated part of the ZIPM and ZINBM. The coefficient associated with this variable was negative and statistically significant in all our specifications.

[11] Note that although we have experimented with alternative sets of explanatory variables (e.g. we have dropped some of the variables related to the agglomeration economies, knowledge and commuting) and computed the GoF tests using different numbers of cells (see MANJON 2009 for details on the computation of this test), these general trends are largely unaffected.

[12] Although some variables were not statistically significant individually, the Wald Test for their joint significance was generally well above standard critical values (results available on request). See Table 3 for an illustrative example of this general trend.

Table 1. Variables: definition, sources and descriptive statistics.

Variable	Typology	Definition	Source (1)	Mean	Std dev	Min	Max
ENTRY	Dependent	New manufacturing establishments (2001-2003)	REI and OC	4.093	14.338	0	258
RES_VAR	Agglomeration eco.	Residential population change between 1991 and 2001	TB2005	0.154	0.344	-0.863	3.043
URB	Agglomeration eco.	Jobs per km^2	TB2005, IDESCAT and OC	263.737	4055.124	0.137	124000
DISURB	Agglomeration eco.	URB^2	OC	1.65×10^7	5.00×10^8	0.019	1.54×10^{10}
JOB	Agglomeration eco.	Jobs	IDESCAT	2977.260	22267.150	12	645682
DENS	Agglomeration eco.	Residential population per km^2	TB2005 and OC	380.107	1520.855	0.765	21020
CONC	Industrial mix	Manufacturing concentration index	TB2005	1.196	1.002	0	3.896
JOB_IND	Industrial mix	Percentage of manufacturing jobs	IDESCAT	0.222	0.116	0	0.609
JOB_SER	Industrial mix	Percentage of jobs in services	IDESCAT	0.473	0.259	0	1
TEC_SEC	Education	% of population older than 10 with technical secondary school	TB2005	10.043	3.225	0.585	23.585
SEC	Education	% of population older than 10 with secondary school	TB2005	9.593	3.482	1.695	28.226
DEG	Education	% of population older than 10 with 3 years degree	TB2005	5.397	2.196	0	24
DEG_PHD	Education	% of population older than 10 with 4 years degree and PhD	TB2005	4.657	2.534	0	21.062
TT_CP	Transport infrast.	Travel time to capital of the province	TB2005	87.010	23.943	0	190
RAIL	Transport infrast.	Dummy for rail station	TB2005	0.107	0.309	0	1
TT_AIR	Transport infrast.	Travel time to the closest airport	TB2005	48.872	33.086	0	190
TT_PORT	Transport infrast.	Travel time to the closest port	TB2005	62.182	33.187	0	197
JOB_HT	Knowledge	Jobs in high-tech industries	TB2005	824.479	12214.780	0	371269
JOB_HT_MA	Knowledge	Manufacturing jobs in high-tech industries	TB2005	16.652	159.379	0	4303
POP_JOB	Commuting	Population working and living at *j* / Jobs at *j*	TB2005	43.444	14.681	0	89.401
POP_JOB_E	Commuting	Population working and living at *j* / Population living at *j* and working at *j* or elsewhere	TB2005	178.869	1730.163	0	52107.410
POP_20_44	Population	Population aged 20-44	OC	29.983	4.596	0	43.050
RES	Population	Residential population (only used in inflated models)	TB2005	6705.190	51711.210	26	1503884
GIRONA	Location	Province of Girona	IDESCAT	0.234	0.424	0	1
LLEIDA	Location	Province of Lleida	IDESCAT	0.243	0.429	0	1
TARRA	Location	Province of Tarragona	IDESCAT	0.194	0.396	0	1
CAP_CO	Location	Dummy for the capitals of the ”comarques”	IDESCAT	0.043	0.204	0	1
COAST	Location	Dummy for shore-line areas	OC	0.074	0.262	0	1
DIST_100	Location	Distance (km) to the nearest city with at least 100,000 inhabitants	CCI	47.073	29.829	0	13.588
DIST_CAT	Location	Distance (km) to the capital of Catalonia (Barcelona)	CCI	86.965	39.671	0	199.590
FIRM_SMALL	Firms	Percentage of small firms (less than 50 workers)	TB2005	83.701	23.671	0	100

Note (1): REI stands for “Register of Industrial Establishments”, OC for “Own Calculations”, TB2005 for Trullén and Boix (2005), IDESCAT for “Catalan Statistical Institute” and CCI for “Catalan Cartographical Institute”.

Table 2: Log L, AIC and Gof Test.

	10 Km	20 Km (TTWAs)	30 Km (Counties)	40 Km	50 Km	60 Km	70 Km	80 Km	90 Km	100 Km	Baseline
Log L											
2.A (Moran's I)	-1051.52 -1089.98 -971.41	-1052.41 -1112.43 -974.78	-1056.99 -1110.54 -978.07	-1048.56 -1112.00 -971.10	-1045.38 -1082.59 -963.55	-1047.05 -1116.40 -971.59	-1050.72 -1115.92 -973.39	-1053.30 -1099.11 -969.03	-1065.68 -1119.40 -978.68	-1056.63 -1117.42 -974.26	-1015.65 (NBM) -1003.58 (ZIPM) -935.79 (ZINBM)
2.B (Moran's I)	-986.96 -945.26 -913.25	-985.74 -970.97 -921.15	-988.40 -974.22 -919.42	-987.12 -978.89 -918.57	-987.54 -968.03 -918.50	-988.38 -969.99 -918.62	-987.45 -980.33 -921.79	-987.07 -977.08 -917.48	-1000.46 -977.89 -926.01	-996.86 -982.31 -923.49	
3. (LISA)	-1028.45 -1077.88 -964.63	-1008.58 -1020.70 -942.47	-1012.81 -1057.93 -950.28	-1012.28 -1060.04 -950.43	-1014.39 -1064.74 -953.38	-1036.59 -1081.06 -964.73	-1005.62 -1051.76 -950.99	-1002.60 -1040.65 -971.72	-950.19 -1047.40 -1092.15	-1002.61 -1053.82 -954.63	
AIC											
2.A (Moran's I)	2072.04 2147.95 1909.83	2073.83 2192.86 1916.56	2082.97 2189.07 1923.14	2066.11 2192.00 1909.21	2059.75 2133.18 1894.10	2063.10 2200.79 1910.17	2070.44 2199.83 1913.78	2075.61 2166.22 1905.06	2100.35 2206.80 1924.36	2082.26 2202.84 1915.53	2000.29 (NBM) 1975.15 (ZIPM) 1838.59 (ZINBM)
2.B (Moran's I)	1924.92 1840.52 1775.50	1922.49 1891.95 1791.30	1927.80 1898.43 1787.85	1925.23 1907.78 1786.14	1926.08 1886.06 1786.00	1927.76 1889.97 1786.24	1925.90 1910.67 1792.58	1925.15 1904.15 1783.96	1951.93 1905.79 1801.02	1944.72 1914.63 1795.98	
3. (LISA)	2025.89 2123.76 1896.26	1986.17 2009.40 1851.94	1994.62 2083.86 1867.56	1993.57 2088.09 1867.87	1997.77 2097.48 1873.76	2042.19 2130.12 1896.47	1980.25 2071.51 1868.98	1974.20 2049.29 1867.38	2063.80 2152.30 1910.44	1974.23 2075.64 1876.25	
GoF Test											
2.A (Moran's I)	94.76^{***} 106.72^{***} 299.50^{***}	48.93^{***} 88.36^{***} 264.80^{***}	52.61^{***} 105.81^{***} 250.06^{***}	57.96^{***} 93.01^{***} 219.90^{***}	43.87^{***} 83.66^{***} 211.02^{***}	12.39 86.63^{***} 220.61^{***}	19.52^{**} 104.67^{***} 227.02^{***}	20.95^{***} 97.75^{***} 256.56^{***}	20.92^{***} 99.06^{***} 266.82^{***}	20.53^{***} 97.65^{***} 222.45^{***}	122.76^{***} (NBM) 138.15^{***} (ZIPM) 295.93^{***} (ZINBM)
2.B (Moran's I)	150.57^{***} 153.01^{***} 272.22^{***}	128.66^{***} 140.06^{***} 263.97^{***}	143.70^{***} 156.78^{***} 262.17^{***}	146.04^{***} 148.57^{***} 245.41^{***}	133.21^{***} 140.62^{***} 197.44^{***}	128.94^{***} 142.97^{***} 234.48^{***}	123.51^{***} 145.73^{***} 262.60^{***}	142.02^{***} 148.72^{***} 296.25^{***}	140.87^{***} 148.12^{***} 272.68^{***}	132.17^{***} 149.56^{***} 260.76^{***}	
3. (LISA)	56.09^{***} 88.54^{***} 252.22^{***}	50.17^{***} 65.43^{***} 180.99^{***}	52.80^{***} 79.45^{***} 215.84^{***}	62.31^{***} 70.41^{***} 240.16^{***}	62.65^{***} 67.74^{***} 229.11^{***}	82.88^{***} 94.43^{***} 305.56^{***}	68.97^{***} 59.95^{***} 111.56^{***}	71.50^{***} 67.85^{***} 202.62^{***}	87.92^{***} 111.48^{***} 315.12^{***}	90.28^{***} 70.53^{***} 235.63^{***}	

Note: Log L is the value of the log-likelihood function, AIC is the value of the Akaike information criterion and GoF Test is the value of the Chi-Square Goodness-of-Fit test computed using 10 cells (see Manjón 2009 for details on the computation of this test). As indicated in the Baseline column, each cell reports values obtained for the NBM (first line), the ZIPM (second line) and the ZINBM (third line).

Table 3. Econometric results.

	BASELINE SPECIFICATION			EXTENDED SPECIFICATION	
	NEGBIN	**ZINB**		**NEGBIN**	**ZINB**
RES_VAR	0.0973	-0.1278	W_RES_VAR	-2.2858	-2.1450
	(0.0806)	(0.1574)		(1.7514)	(3.2912)
URB	0.0004	0.0007	W_URB	-0.0071	-0.0148
	(0.0001)***	(0.0002)***		(0.0098)	(0.0187)
DISURB	-0.0000	-0.0000	W_DISURB	0.0000	0.0000
	(0.0000)***	(0.0000)***		(0.0000)	(0.0000)
JOB	0.0000	0.0000	W_JOB	0.0003	-0.0003
	(0.0000)***	(0.0000)***		(0.0005)	(0.0009)
DENS	-0.0001	-0.0003	W_DENS	0.0010	0.0040
	(0.0000)***	(0.0000)***		(0.0027)	(0.0052)
CONC	0.2270	0.6511	W_CONC	-5.0481	-6.0459
	(0.0547)***	(0.1595)***		(1.2531)***	(2.2933)***
JOB_IND	1.0801	1.0445	W_JOB_IND	-0.3051	-9.6818
	(0.3510)***	(0.7779)		(7.2852)	(13.986)
JOB_SER	0.5187	2.0673	W_JOB_SER	-10.8250	-17.1200
	(0.2074)**	(0.6165)***		(2.9130)***	(5.3560)***
TEC_SEC	-0.0002	-0.0000	W_TEC_SEC	-0.0827	-0.3255
	(0.0107)	(0.0241)		(0.1873)	(0.3490)
SEC	0.0131	0.0511	W_SEC	0.6707	1.6596
	(0.0106)	(0.0252)**		(0.3118)**	(0.6191)***
DEG	-0.0211	-0.0335	W_DEG	-0.3594	-0.2558
	(0.0185)	(0.0440)		(0.4597)	(0.9161)
DEG_PHD	-0.0132	-0.0377	W_DEG_PHD	-0.7816	-2.4021
	(0.0150)	(0.0343)		(0.5538)	(1.0735)**
TT_CP	0.0072	0.0130		-0.0130	-0.0015
	(0.0047)	(0.0094)		(0.0060)**	(0.0125)
RAIL	0.1215	0.0155		0.4341	0.4661
	(0.0870)	(0.1231)		(0.1334)***	(0.1751)***
TT_AIR	-0.0015	-0.0058		0.0051	0.0081
	(0.0031)	(0.0061)		(0.0046)	(0.0088)
TT_PORT	-0.0109	-0.0160		-0.0016	-0.0100
	(0.0047)**	(0.0093)*		(0.0063)	(0.0125)
JOB_HT	-0.0000	-0.0001	W_JOB_HT	-0.0004	0.0004
	(0.0000)***	(0.0000)***		(0.0008)	(0.0015)
JOB_HT_MA	0.0002	0.0003	W_JOB_HT_MA	0.0152	0.0308
	(0.0003)	(0.0004)		(0.0265)	(0.0477)
POP_JOB	0.0019	0.0052	W_POP_JOB	-0.0813	-0.0556
	(0.0019)	(0.0037)		(0.0635)	(0.1223)
POP_JOB_E	-0.0000	-0.0000	W_POP_JOB_E	0.0007	0.0016
	(0.0000)	(0.0000)		(0.0006)	(0.0010)
POP20_44	0.0158	0.0315	W_POP20_44	0.4544	0.6780
	(0.0065)**	(0.0130)		(0.1514)***	(0.2800)**

Table 3 (Cont.). Econometric results.

GIRONA	-0.1326	-0.3365		-0.1316	-0.4798
	(0.1310)	(0.2389)		(0.1715)	(0.2804)*
LLEIDA	-0.1969	0.0010		-0.2360	-0.3248
	(0.1230)	(0.3487)		(0.1473)	(0.3506)
TARRAG	-0.3142	-0.1677		-0.0889	-0.1170
	(0.0803)***	(0.2505)		(0.1659)	(0.3454)
CAP_CO	1.6672	1.6281		2.6987	3.2037
	(0.4493)***	(0.4189)***		(0.6654)***	(0.6805)***
COAST	0.2398	-0.0309		0.0019	-0.0568
	(0.1323)*	(0.1532)		(0.1029)	(0.1636)
DIST_100	-0.0054	-0.0067		-0.0161	-0.0232
	(0.0024)**	(0.0051)		(0.0036)***	(0.0075)***
DIST_CAT	0.0034	0.0021		-0.0071	-0.0219
	(0.0020)*	(0.0041)		(0.0059)	(0.1174)
FIRM_SMALL	-0.0021	0.0012	W_FIRM_SMALL	0.0585	0.1098
	(0.0013)	(0.0029)		(0.0450)	(0.0831)
LR Joint Test	588.17***	403.40***		525.37***	331.82***
LR InflatedTest		297.96***			475.91***
Vuong Test		6.27***			6.30***
Wald Joint Test on the Spatially Lagged Variables (W_)				80.03***	47.08***

Note: 946 observations. Standard errors in brackets. Details on the covariates can be found in Table 1 (those starting with "W_" are spatially lagged variables constructed as explained in Section 3.2 using Moran's I and 60 km as neighbourhood criterion).

Figure 1. Global Spatial Autocorrelation (Moran's I) with neighbour matrices of 10 km to 100 km.

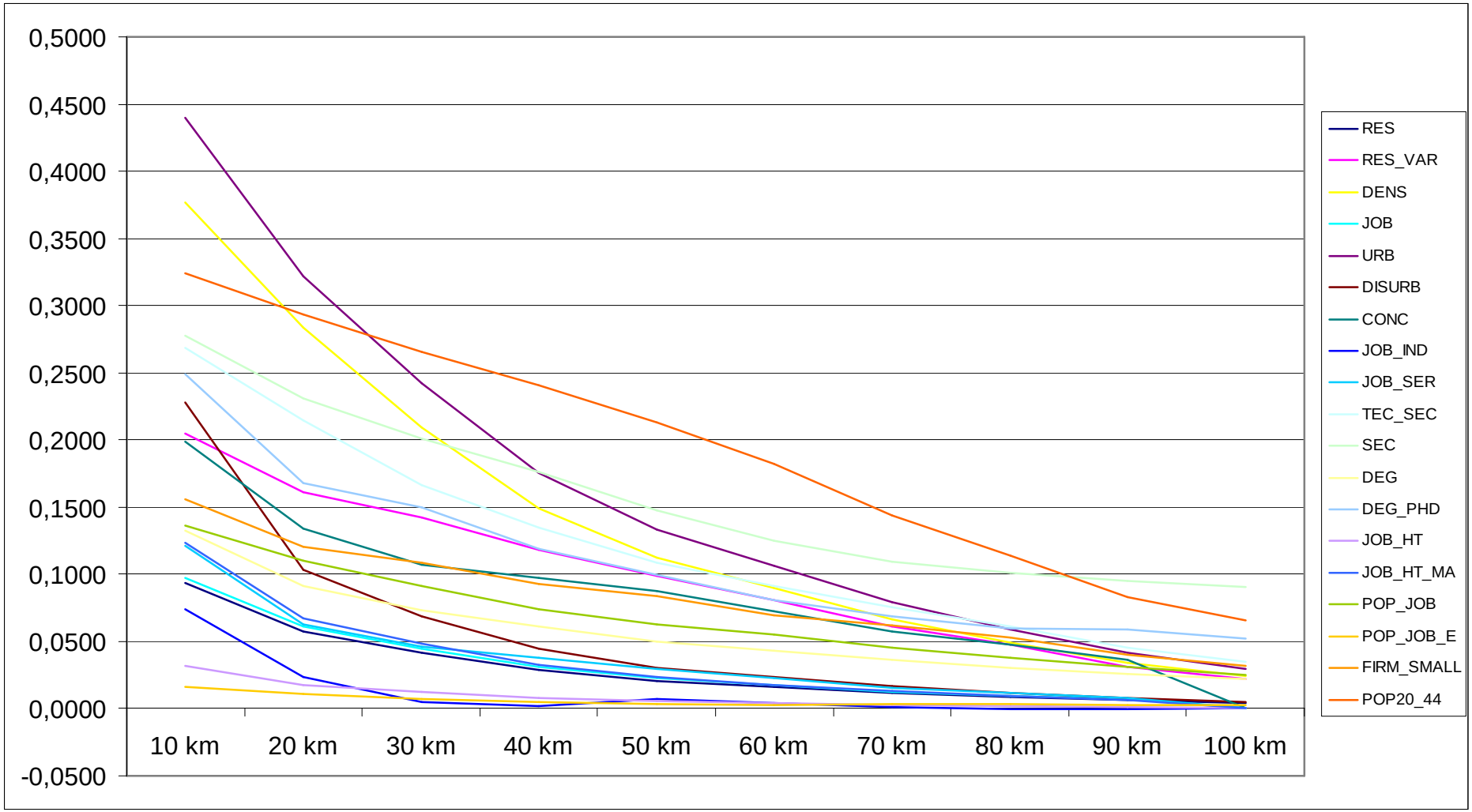

Figure 2. Local Index of Spatial Association (LISA) of the Urbanization Economies with neighbour matrix of 10 km.

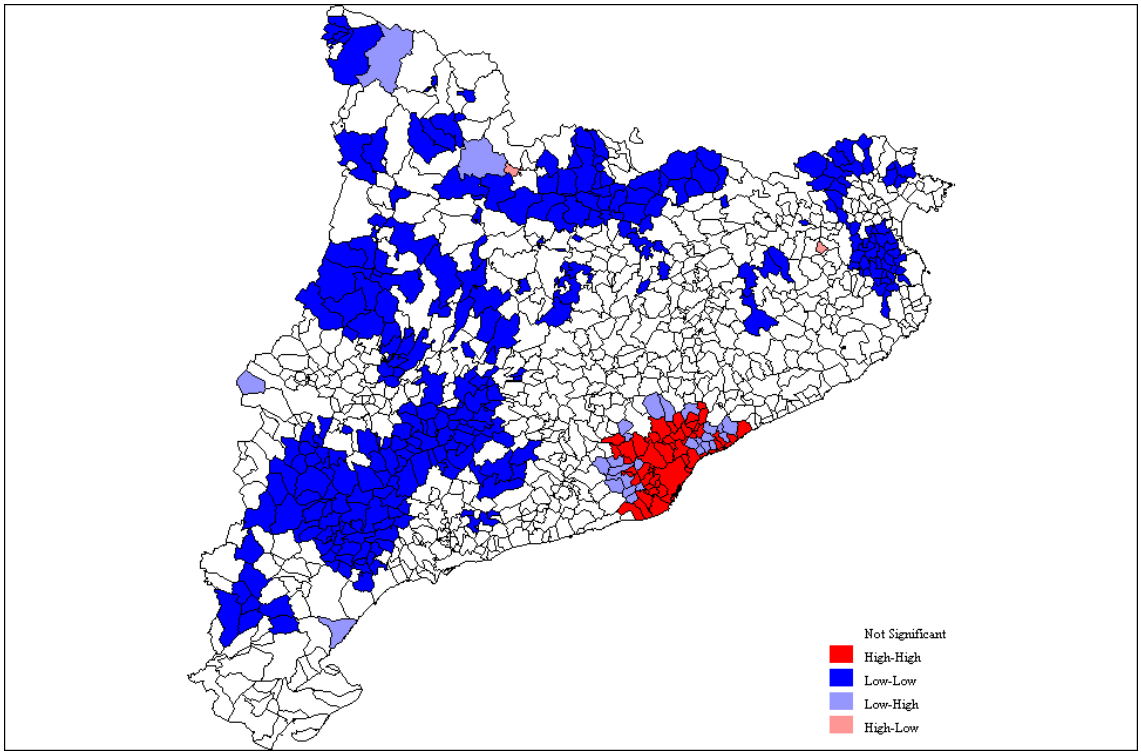

Figure 3. Local Index of Spatial Association (LISA) of the Urbanization Economies with neighbour matrix of 40 km.