UNIVERSITAT
ROVIRA I VIRGILI
DEPARTAMENT D'ECONOMIA

CENTRE DE RECERCA EN ECONOMIA
INDUSTRIAL I ECONOMIA PÚBLICA

WORKING PAPERS

Col·lecció "DOCUMENTS DE TREBALL DEL DEPARTAMENT D'ECONOMIA - CREIP"

Industrial Location and Space: New Insights

Daniel Liviano
Josep-Maria Arauzo-Carod

Document de treball nº -9- 2011

DEPARTAMENT D'ECONOMIA – CREIP
Facultat de Ciències Econòmiques i Empresarials

UNIVERSITAT
ROVIRA I VIRGILI
DEPARTAMENT D'ECONOMIA

creip
CENTRE DE RECERCA EN ECONOMIA
INDUSTRIAL I ECONOMIA PÚBLICA

Edita:

Departament d'Economia
www.fcee.urv.es/departaments/economia/public_html/index.html
Universitat Rovira i Virgili
Facultat de Ciències Econòmiques i Empresarials
Avgda. de la Universitat, 1
43204 Reus
Tel.: +34 977 759 811
Fax: +34 977 300 661
Email: sde@urv.cat

CREIP
www.urv.cat/creip
Universitat Rovira i Virgili
Departament d'Economia
Avgda. de la Universitat, 1
43204 Reus
Tel.: +34 977 558 936
Email: creip@urv.cat

Adreçar comentaris al Departament d'Economia / CREIP

Dipòsit Legal: T -1093- 2011

ISSN 1988 - 0812

DEPARTAMENT D'ECONOMIA – CREIP
Facultat de Ciències Econòmiques i Empresarials

Industrial Location and Space: New Insights

Daniel Liviano [♠] [♣]

Josep-Maria Arauzo-Carod [♠] [♦]

Abstract

This paper tries to resolve some of the main shortcomings in the empirical literature of location decisions for new plants, i.e. spatial effects and overdispersion. Spatial effects are omnipresent, being a source of overdispersion in the data as well as a factor shaping the functional relationship between the variables that explain a firm's location decisions. Using Count Data models, empirical researchers have dealt with overdispersion and excess zeros by developments of the Poisson regression model. This study aims to take this a step further, by adopting Bayesian methods and models in order to tackle the excess of zeros, spatial and non-spatial overdispersion and spatial dependence simultaneously. Data for Catalonia is used and location determinants are analysed to that end. The results show that spatial effects are determinant. Additionally, overdispersion is descomposed into an unstructured *iid* effect and a spatially structured effect.

Keywords: Bayesian Analysis, Spatial Models, Firm Location.

JEL Classification: C11, C21, R30.

This research was partially funded by SEJ2007-64605/ECON, SEJ2007-64686/ECON, the "Xarxa de Referència d'R+D+I en Economia i Polítiques Públiques", the SGR Program (2009-SGR-322) of the Catalan Government and the PGIR program N-2008PGIR/05 of the Rovira i Virgili University. This article has benefited from discussions with M. Manjón and E. Viladecans. We would also like to acknowledge the helpful and supportive comments received from seminar participants at the Industrial Dynamics and Economic Geography workshop (DIME - Utrecht University). Any errors are our own.

(♠) Universitat Rovira i Virgili, Departament d'Economia (QURE-CREIP); Av. Universitat 1; 43204 Reus (Catalonia); Phone: +34 977758902; Fax: +34 977759810.

(♣) Universitat Oberta de Catalunya (UOC); Av. Tibidabo 39-43; 08035 Barcelona (Catalonia); Phone: +34 932542169.

(♦) Institut d'Economia de Barcelona (IEB); Av. Diagonal 690; 08034 Barcelona (Catalonia).

1 Motivation

What is the real importance of space in industrial location processes? This is presumably a key issue, since industrial location takes place in specific geographic areas with characteristics, together with those of the surrounding area, which are surely taken into account by entrepreneurs when deciding their firms' final location. From an academic point of view, many significant breakthroughs were made in the analysis of firm location behaviour in recent decades (McCann and Sheppard 2003). These studies have highlighted the existence of a new group of factors that have changed the patterns of firm location, such as the rise of economic integration processes and free trade areas, as well as the growth in new communication technologies[1]. However, a thorough and in-depth look at the role of spatial factors in this field has yet to be undertaken. Indeed, as Arauzo-Carod et al. (2010) argue, "the scarce use of spatial econometric techniques may be due to the lack of appropriate tools, while future developments in spatial econometrics should shortly be followed by applications to industrial location".

Industrial location data generally present several characteristics which ought to be taken into account in the estimation of empirical models. Overdispersion appears whenever the variance of the estimated model is greater than its mean, and an excess of zero counts consists of an excessive number of zero values in the dependent variable (the number of new located firms/establishments). Furthermore, a key characteristic of industrial location data is that it is georeferenced, i.e. the location of production units is distributed across the geography, and zones where almost no locations take place coexist with highly agglomerated zones. In other words, firm locations tend to take place in urban areas, which also tend to be close to each other. It is also reasonable to assume that when making the decision about where to locate, firms not only consider the characteristics of a single area, but also the characteristics of other nearby areas. A consequence of all this is the high level of spatial heterogeneity in the data as well as crossed effects between the different spatial units[2], which is called spatial dependence[3].

[1] This group of studies can be related to a discipline called New Economic Geography.

[2] See Arauzo-Carod and Manjón-Antolín (2009) for a discussion about the spatial aggregation problem in the context of industrial location literature.

[3] In fact, how to deal with these effects in the context of applied research has led to the development of spatial statistics and econometrics.

In methodological terms, the approach used in this article focuses on the territory, studying the effect of territorial characteristics on the number of new units located in each spatial unit by estimating Count Data (CD) models[4]. In practice, the main concern of empirical researchers has been the existence of both overdispersion and excess of zero counts in the data, since these render the results of the basic Poisson regression inefficient and biased. In fact, the existence of these problems has triggered the use of Negative Binomial (NB) and With Zeros (WZ) models.

However, although the empirical evidence suggests that the location of new units is linked with space, the explicit study of spatial factors has traditionally been neglected in the empirical industrial location literature, most probably due to the fact that spatial methods for Count Data models were not yet available. Fortunately, there have been many methodological developments in the field of spatial statistics and econometrics outside the framework of classical linear models in recent years. Among these, models for generalized linear models (GLM) accounting for spatial effects in different forms seem suitable for the empirical study of industrial location. Nonetheless, although a few contributions trying to incorporate these effects in the context of firm location studies have been proposed[5], the exhaustive consideration of these effects and the use of new methodological quantitative tools allowing such analysis are still a matter pending.

However, why is the inclusion of spatial effects so important in the estimation of CD models? In other words, what are the consequences of not modelling these effects explicitly? On the one hand, it seems quite obvious that firms do not only take into account the characteristics of the single spatial unit where they intend to locate, but also the characteristics of the surrounding area. A straightforward consequence of this is a violation of the assumption of data independence, i.e. individual observations are no longer independent of each other. If such interactions are not properly taken into account, the model is likely to suffer from

[4] The other approach to the study of industrial location is the estimation of discrete choice models. For a discussion and a comparison of both approaches in the context of industrial location literature, see Arauzo-Carod, Liviano and Manjón-Antolín (2010).

[5] These studies are reviewed in Section 2.

misspecification[6]. On the other hand, the (uneven) distribution of data across space is a source of overdispersion, and therefore the use of CD models accounting for overdispersion partially captures such spatial effects. However, depending on the data, model and phenomenon analysed, the omission of the spatial component in the model is likely to render the estimations inefficient and inaccurate. This fact should not be underestimated, since after all, the presence of spatial effects violates the assumption of independently and identically distributed (*iid*) errors of most statistical procedures, and they can even invert the slope of estimated coefficients from non-spatial analysis (Kühn 2007), which may inevitably lead to false and wrong conclusions. Nonetheless, this problem can be tackled using a Bayesian approach in which the overdispersion can be modelled by specifying two random effects: an unstructured *iid* effect and a spatially structured effect. This constitutes an important step forward, because sources of overdispersion can help to explain fims location decisions and consequently provide a better understanding of this location process, which is a necessary condition for any firm entry promotion policy. In addition, it should be noted that the detection of spatial effects in CD models is not easy. According to the standard literature, a common procedure to do so is post-estimation analysis of the estimated model's residuals as a way to determine the existence of omitted spatial effects in the model, i.e. residuals showing spatial autocorrelation are the proof that the model has not been correctly specified, since the spatial structure of the data is missing in the specification. However, as Lin and Zhang (2007) demonstrate, the use of Moran's I test for residuals in Generalised Linear Models (GLM) is speculative, since the test tends to show that there is no spatial autocorrelation in the residuals, even when such correlation in fact exists. All in all, and in line with Jacqmin-Gadda et al. (1998), a great deal of work must be done in order to reach a satisfactory spatial autocorrelation test for the residuals of GLM.

As a result of all the above, the research presented in this article is intended to shed some light upon this issue by proposing an alternative type of model and estimation method. The primary aim of this approach is to propose an estimation framework which is flexible enough to deal with the issues described above simultaneously and to provide tools to select the final model for consideration.

[6] See Arauzo-Carod and Manjón-Antolín (2009) for an empirical exercise about the misleading results obtained when such spatial interactions are not considered.

To date, this analysis would be hard to carry out using purely frequentist estimation methods such as maximum likelihood. Indeed, issues like overdispersion, spatial effects and excess zeros are likely to occur simultaneously but with a different intensity and different effects upon the estimation results. As a result, alternative modelisations of these issues are likely to yield different results, and as such a coherent framework of models estimation and selection appears necessary.

The development of Bayesian estimation techniques, combined with the increasing availability of more powerful computers and specific software, has made Bayesian methods a valuable alternative to classical frequentist methods, especially when the models to be estimated are particularly complex. In specific terms, a type of model called Integrated Nested Laplace Approximation (INLA, Rue and Martino 2006) is proposed, which is a recent approach to statistical inference for latent Gaussian Markov Random Fields (GMRF) and Bayesian Hierarchical models. The empirical analysis in this article has been carried out by using R[7].

The article is structured as follows. Section 2 reviews the main contributions to the field of empirical industrial location, focusing specifically on the methods and models used in order to tackle problems arising from the data. The data set (Catalan municipalities) and the variables used in the empirical application are presented in Section 3. Section 4 presents a previous univariate exploratory analysis of the main variables, the aims of which are to study the spatial distribution of these variables, and to detect the existence of spatial autocorrelation[8]. Section 5 covers the regression analysis carried out in this paper. Section 5.1 presents the empirical models and the Bayesian methods used to estimate them, and Section 5.2 is devoted to presenting the empirical results and comparing them with the results from standard, classical non-spatial estimations of the model. Finally, Section 6 concludes the paper, summarising the main results of the analysis and giving several hints for future research.

[7] This is open-source software which is available under *http://www.r-project.org*.

[8] This analysis includes the computation of Moran's I plots and Geary's c correlograms and variograms. See Anselin (1988) for details.

2 Literature Review

As stated above, one current in the empirical literature approaches industrial location from the viewpoint of the territory, studying the effect of territorial characteristics on the number of new firms located in each spatial unit by estimating CD models. In specific terms, changes in location characteristics can affect, ceteris paribus, the conditional expectation of the number of firms created in the geographical location i over a certain period of time. In more formal terms, the empirical framework can be defined as

$$E(y_i|x_i) = f(x_i, \beta), \;\; i = 1, ..., N. \tag{1}$$

where y_i is the dependent variable, x_i is a set of regressors, $E(y_i|x_i)$ is the conditional expectation (or conditional mean) of the dependent variable (the number of entries), $f(\cdot)$ is a certain function governing the relationship between the regressors and the conditional mean, and β is a parameter vector. At this point, it is crucial to choose the model and the estimation method used to analyse the effects of the regressors on the dependent variable correctly[9]. In this case, the natural candidate is the Poisson regression model, which assumes that y_i, conditional on x_i (henceforth conditional distribution), follows a Poisson distribution, i.e. $y_i|x_i \sim Po(\mu_i)$. The Poisson model was thus the starting point for an important current of empirical research on industrial location[10]. However, the Poisson regression model is restrictive in practice and has several drawbacks, including (a) overdispersion, which occurs when the data is overdispersed, and the variance in the data therefore exceeds the variance assumed by the model; and (b) excess zeros, which occurs when the number of zero counts exceeds the number of zero counts expected by the model (see Cameron and Trivedi, 1998).

From a methodological point of view, several solutions to tackle these two problems have been implemented. One of the most popular alternatives has been the mixture Poisson model as a way to extend the basic Poisson regression model,

[9] Because of the count nature of the dependent variable ($y_i = 0, 1, 2, \ldots$), a linear model would not be appropriate, since the dependent variable (and therefore the error term) would not follow a normal distribution.

[10] Among these contributions are Smith and Florida (1994), Wu (1999), List (2001), Arauzo-Carod and Manjón-Antolín (2004), Barbosa et al. (2004), Gabe and Bell (2004), Arauzo-Carod (2005, 2008), Autant-Bernard et al. (2006), Alañón et al. (2007) and Arauzo-Carod and Viladecans (2009).

which consists of adding some distributions to the underlying distribution, thus adding flexibility by improving the fit of the resulting distribution to the observed data. These models can be grouped into continuous and finite mixture models. Continuous mixed models control for the unobserved heterogeneity by including an unobserved heterogeneity term for each observation, extending the specified distribution from $(y_i|x_i)$ to $(y_i|x_i, \nu_i)$, where ν_i is an idd term that denotes the unobserved heterogeneity and is independent of the covariates, which is regarded as a location-specific random effect in the firm location literature. A very popular and frequently used mixture model is the Negative Binomial (NB) model, which arises as a mixture of a Poisson distribution for the dependent variable and a Gamma distribution for the unobserved heterogeneity term. In the firm location literature, the NB model has been used by Arauzo-Carod and Viladecans (2009), Wu (1999), Cieślik (2005) and Manjón-Antolín and Arauzo-Carod (2010), among others.

The other alternative to the overdispersion and excess zeros problems are the finite mixture models, which are especially suited to handling the excess of zeros. These models outperform continuous mixture models in some cases, because continuous mixtures, despite controlling for heterogeneity, may not properly account for the excess of zero counts[11]. In the context of firm location literature, a class of finite mixture models called *with zeros* (WZ) models[12] has been used, which assumes a discrete representation of the unobserved locational heterogeneity by modifying the probability of the zero outcome by a mixing parameter, which is parametrised using Logit or Probit models. Two specific WZ models have been used in the context of the firm location literature: Zero Inflated Poisson (ZIP) models and Zero Inflated Negative Binomial (ZINB) models. ZIP models have been used by List (2001), Gabe (2003), Basile (2004) and Manjón-Antolín and Arauzo-Carod (2010), whereas ZINB models have been used by Manjón-Antolín and Arauzo-Carod

[11] The reason lies in the fact that since the excess zeros may stem from two sources, i.e. unobserved heterogeneity and an underlying selectivity process, continuous mixture models such as NB are only likely to account for the excess of zeros stemming from the unobserved heterogeneity. Furthemore, data sometimes displays heterogeneity by an excess of zero counts, the number of which exceeds the number of zeros expected by the continuous mixture model.

[12] The excess zeros can be dealt with using two approaches in the context of finite mixture models: hurdle (or conditional) models, which are interpreted as two-part models, and with zeros (or zero inflated models), which are models in which the heterogeneity is introduced in a binary form.

(2010) and Arauzo-Carod (2008).

In recent years, several contributions have included spatial effects in their location studies in different ways. Lambert et al. (2006) study manufacturing investment location using Spatial Poisson Models. Specifically, they estimate a geographically weighted regression (GWR) as well as a spatial generalized linear model (SGLM) to study spatial correlations between observations, thus obtaining evidence of spatial dependency between countries and the manufacturing investment decisions of firms. For the case of Spain, Alañón et al. (2007) study the relationship between accessibility and industrial location by estimating spatial Probit models with spatially lagged dependant variables, spatially lagged explanatory variables and spatially autocorrelated error terms. Blonigen et al. (2007) study spatial autoregressive relationships in empirical foreign direct investment (FDI) models using data on US outbound FDI activity. They estimate a gravity model and find that both the traditional determinants of FDI and the estimated spatial interdependence are quite sensitive to the sample of countries examined. Basile et al. (2010) estimate a semi-parametric spatial autoregressive negative binomial model using data on the number of inward greenfield FDI in European regions. Their results show that multinational firms' location choices are spatially dependent, even controlling for a large number of regional characteristics, and also show that controlling for spatial dependence in the error term yields significant changes in the magnitude of some estimated coefficients.

These contributions notwithstanding, study of the spatial dimension is not yet generalised in the empirical industrial location literature. However, many methodological contributions proposing methods accounting for spatial effects in models with non-normally distributed response variables have recently been proposed and adopted, mostly in the natural sciences[13]. All of them depart from the notion of Generalized Linear Models (GLM), which enable a flexible generalisation of classical linear models, allowing the distribution of the dependent variable to belong to a broad collection of distributions called the exponential family, including the Poisson and count data related distributions. In this sense, GLM stand for a unified

[13] An up-to-date review of these models and a comparison between them is provided by Dormann et al. (2007).

framework in which several models can be fitted[14] and detailed in McCullagh and Nelder (1989). They assume that the simple regression model $y_i = E(y_i|x_i) + \varepsilon_i$, where $E(y_i|x_i) = x_i^{'}\beta$ can be extended so that $E(y_i|x_i) = h(x_i^{'}\beta)$, with $h(\cdot)$ being a function allowing for a broad collection (or family) of distributions.. In the context of GLM count data models, many efforts have been made to explicitly incorporate the spatial structure of the data in the model's specification[15]. Besag (1974) studied the consequences of the dismissal of the data's spatial structure in the model, thus leading to the model's misspecification and autocorrelated residuals. In recent decades, many contributions have therefore proposed and estimated a variety of methods, mainly in the ecological, biological and medicine fields (Keitt et al., 2002; Dormann 2007 and Miller et al., 2007 are reviews of these methodological contributions). These contributions notwithstanding, there has always been a problem regarding the specification of these models: how to specify the complex spatial structure of the data analytically, and even more importantly, how to estimate it. Classical estimation methods such as maximum likelihood (ML) or quasi-ML have been shown to have certain limitations regarding the computation of such spatial structures. One of the solutions that has been proposed and that has been gaining popularity (mostly due to the increasing availability of faster computers as well as the development of statistical packages such as R and OpenBugs) is the Bayesian methodology[16]. Specifically, Bayesian Hierarchical Models are a framework appropriate for the development of spatially structured models, with the model specified in different layers, each of which accounts for different sources of variation. In the context of georeferenced count data, the key is to decompose the variability of the model into two components: (a) a spatially correlated variable accounting for the neighbourhood relationship between the geographical areas, and (b) a classic area-independent effect[17].

[14] GLM were first described by Nelder and Weddebrurn (1972)

[15] For a comprehensive manual of GLM developments accounting for spatial effects, see Schabenberger and Gotway (2005).

[16] See Gelman (2004) for a gentle introduction to Bayesian models and its application to CD models.

[17] Besag et al. (1991) stands for the seminal article that introduced this decomposition, and Gschlössl and Czado (2008) present regression models for count data allowing for both overdispersion and spatial effects in a Bayesian framework.

3 Data and Variables

The data used in this article refer to local units (municipalities) in Catalonia, and consist of two datasets: data on firm entries and data on municipal characteristics. The database on entries is the REIC (Catalan Manufacturing Establishments Register), which is a compulsory register with plant-level micro data on the creation and location of new manufacturing establishments. The REIC provides data on both new and relocated establishments, and since these may be attracted to the territory by the same variables, here they are used together without distinction[18]. Furthermore, only selected establishments with codes 12 to 36 (NACE-93 classification) are considered. All in all, the analysis includes the aggregated entries of manufacturing establishments in 941 municipalities between 2002 and 2004. The database on territorial characteristics comes from the Trullén and Boix (2004) database on Catalan municipalities, the Catalan Statistical Institute (IDESCAT) and the Catalan Cartographic Institute (ICC). The data cover almost all the Catalan municipalities, and refer to the year 2001[19].

The dependent variable of the analysis is the aggregated number of entries over the period 2002-2004, which has a count nature ($y_i = 0, 1, 2, \ldots$). The regressors are territorial characteristics classified into the following categories and shown in Table 1[20].

A) Agglomeration economies

Agglomeration economies are one of the main determinants of firm location, in the sense that firms consider the presence of population, other firms and economic activity in general when deciding where to locate. Because agglomeration economies is a multidimensional concept that cannot be reduced to a single variable, several variables to proxy it have been considered in the literature to date. Specifically, agglomeration economies have been divided into two types: urbanisation economies and localisation economies. The former involves a city's

[18] See Manjón-Antolín and Arauzo-Carod (2010) for details.

[19] Data for five new municipalities (Gimenells i el Pla de la Font, Riu de Cerdanya, Sant Julià de Cerdanyola, Badia del Vallès and La Palma de Cervelló) have been left out due to lack of data.

[20] See Arauzo-Carod et al. (2010) for a detailed review on firm location determinants and the type of variables used in this literature.

population and employment levels and the diversity of its productive structure, whereas localisation economies involve a city's specialisation in a certain sector. There is no clear evidence as to whether urbanisation or localisation economies are more important for the location of new firms, as the empirical evidence is mixed and inconclusive[21].

Agglomeration economies are proxied using three variables. EMP stands for the log of employment density, where the area of urbanised land in square kilometres is the denominator, and EMP2 is its squared value, which aims to show the existence of diseconomies of agglomeration due to congestion effects. HHI is the Herfindahl-Hirschman Index, which is intended to reflect the diversity of the productive structure of each municipality, and can therefore be regarded as a manufacturing diversification index.Urbanisation economies have been regarded as a location determinant in many studies in the literature. Guimarães et al. (2004), Arauzo-Carod (2005) and Cieślik (2005b) included specifically urbanisation economies in their studies. Furthermore, Arauzo-Carod and Manjón-Antolín (2004), Holl (2004a,b), Arauzo-Carod (2005) and Manjón-Antolín and Arauzo-Carod (2010) include a measure proxying for industrial/sectoral diversity; Bade and Nerlinger (2000), List (2001), Papke (2001), Egeln et el. (2004) and Arauzo-Carod and Viladecans (2009) consider the population density as a covariate; Manjón-Antolín and Arauzo-Carod (2010) control for the density of the economic activity, and Kogut and Chang (1991) include firm density as a regressor.

B) Industrial Mix

These variables (SIZE, ACT, SME and SSE) are intended to reflect the industrial composition of each municipality. SIZE stands for the average establishment size, and is computed as total employment divided by the number of establishments. ACT is intended to capture the activity rate, and is computed as total employment divided by population. SME is the share of manufacturing employment over total employment and SSE is the share of services employment over total employment. These two latter variables have been considered in the empirical literature by Smith and Florida (1994), Blonigen (1997) and Arauzo-Carod (2005).

[21] Combes (2000) provides a discussion on this topic.

C) Spatial Effects

The variables W-EMP, W-HHI, and W-SME reflect the possible influence of agglomeration economies and industrial mix in neighbouring municipalities on the number of locations in a municipality [22]. These variables are spatial lags computed by means of the product of a neighbourhood (or weights) matrix W with certain regressors, representing the average value of regressor values in neighbouring municipalities. In this analysis a distance matrix is considered such that $w_{ij} = (1/d_{ij})$, where d_{ij} is the distance between municipalities i and j, reflecting the idea that the closer these two municipalities are, the stronger the relationship between them. The final choice of spatial lag comes from a multicollinearity analysis, whereby several spatial lags have been discarded.

D) Human Capital

Human capital is proxied by EDU, which stands for the average years of schooling of the population over 25 years old. Human capital measures have been considered in the literature by Coughlin and Segev (2000), Arauzo-Carod and Manjón-Antolín (2004), Egeln et el. (2004), Holl (2004a,b), Cieślik (2005a), Alañón et al. (2007) and Arauzo-Carod and Viladecans (2009).

E) Geographical Position

This set of variables controls for the geographical position of each municipality. ALT is the average municipality's altitude with respect to sea level, which controls for accessibility. TMC is the average transport time to the largest cities[23], CC is a dummy variable with a value of one if the municipality is a county capital, CL is a dummy variable with a value of one if the municipality is coastal, and MAB, MAG, MAT, MAL, and MAM take a value of one if the municipality is within one of the five biggest metropolitan areas in Catalonia (Barcelona, Girona, Lleida, Tarragona and Manresa).

[22] For a similar approach, see Viladecans (2004).

[23] The criterion for a city to be considered large is having at least 100.000 inhabitants.

[TABLE 1]

At this point, it is worth analysing the characteristics of the dependent variable, which ought to be taken into account in the subsequent inference process. Table 2 shows the basic statistics of ENT. As can be seen, the sum of zero counts outnumbers the counts where at least one entry takes place in each year. The proportion of zero counts is somewhat lessened when entries are aggregated over the period 2002 - 2004, since there are some municipalities where entries do not take place every year. Furthermore, the percentile information shows that the distribution of entries is heavily skewed, i.e. there is a small group of municipalities that account for the largest number of entries, while more than a half receive no entries at all. These distributional features can be seen in Figure 1, in which both the histogram and a kernel density estimation of the dependent variable are plotted.

[TABLE 2]

[FIGURE 1]

So far, these characteristics justify the use of models accounting for overdispersion and an excess of zeros. But what about the spatial distribution of entries? Figure 2 shows a map of Catalonia displaying the number of entries in intervals (above) and the corresponding histogram (below). Graphics like these depict how entries tend to be concentrated in the metropolitan area of Barcelona and to a lesser extent, in coastal or urban regions. Furthermore, entries are very scarce in some areas of Catalonia. This result is to be expected, since municipalities are very dissimilar to each other, ranging from small isolated villages in rural areas to huge and densely populated cities. Of course, this piece of evidence points towards the existence of spatial effects, which is addressed in the next section.

[FIGURE 2]

4 Spatial Exploratory Analysis

The first part of the empirical analysis is a close examination of the spatial structure of the main variables under study. Whether the variables are evenly distributed across the space or on the contrary show specific spatial patterns ought to be taken into account, since it may be a sign of either spatial dependence or spatial heterogeneity. The spatial correlation is estimated by using Moran's I statistic (Moran 1948,1950), which has been computed under randomisation of the analysed variable, since it is not distributed normally. Figure 3 plots the value of Moran's I statistic, under the $k-$nearest neighbours criterion, which considers as neighbours the $k-$nearest spatial units, for the range of values $k \in [2, 499]$[24]. Due to the number of variables under study and for sake of clarity, the analysis is presented in two figures. The upper one plots the statistic's value for the activity rate (ACT), establishment size (SIZE), share of manufacturing employment (SME) and the share of services employment (SSE). A first interesting result from it is the comparison of SME and SSE, which leads to the conclusion that manufacturing activity shows a greater spatial concentration than services activity, the spatial concentration is highly reduced and fades quickly as k increases. Besides, ACT shows a value for the statistic of slightly above 0.4 with $k = 2$ and this value fades slowly and smoothly as k rises, which shows that the activity rate shows a moderately high spatial concentration in a broad area. Lastly, SIZE shows only a small spatial concentration (over 0.1). On the other hand, the lower figure plots the statistic's value for the number of entries (ENT), employment density (EMP) and the Herfindahl-Hirschman Index (HHI). First, ENT shows a positive spatial correlation (over 0.2) with a low number of neighbours, and such correlation fades gradually away as k increases. Second, EMP and HHI show a very similar spatial structure, in that spatial correlation is rather high (0.3) in relatively small areas (that is, low k), and then it decreases quickly as k grows. This evidence indicates that agglomeration economies take place in spatially bounded areas, which is a common result found in the literature.

[FIGURE 3]

[24] This analysis has been also carried out considering the neighbours distance criterion, yielding similar results.

5 Regression Analysis

The aim of this section is to estimate different models and to discuss the details of the estimates. In the literature dealing with CD including overdispersion and excess of zeros, four models are considered: Poisson (PO), Negative Binomial (NB), Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB). Each of these models is specified accounting for either (a) no random effects, (b) just an unstructured random *iid* effect and (c) both an unstructured and a spatial effect. This multiple specification will show to what extent the overdispersion in the model is random, or on the contrary, has a spatial structure. Moreover, the effect of the inclusion of a spatial component upon the different estimation results is to be investigated. A Bayesian methodology called Integrated Nested Lagrange Approximation (INLA) is implemented in the estimation of these models[25]. Using Bayesian terminology, the aim of the analysis is to study the distributional characteristics of a set of parameters θ given the observed data y. In other words, the aim is to calculate and interpret the posterior density $p(\theta|y)$ conditional on the distribution of the parameters under study $p(\theta)$ as well as on the empirical distribution of the observed data conditional on the set of parameters $p(y|\theta)$[26]. In Bayesian methodology, the models presented in this section are called Hierarchical Bayesian (HB) models. This category of models assumes that in addition to the distributional model $f(y|\theta)$ for the observed data y given the vector of parameters θ, the latter vector θ is a ransom quantity sampled from a prior distribution $\pi(\theta|\lambda)$, where λ is a vector of hyperparameters. Since these are often unknown, apart from specifying priors (information regarding the prior distribution of θ), hyperpriors (information regarding the prior distribution of λ) will also be needed. In order to give a coherent nomenclature and not to create confusion, the vector of parameters of the model θ is hereinafter denoted as β, and the remaining hyperparameters (overdispersion, zero-inflation parameter, random and spatial effects) will be introduced as required. The four Bayesian regression models (PO, NB, ZIP, ZINB), the random effects (spatial and non-spatial) and the main results of the ten estimated specifications are introduced and commented on below.

[25] For technical details, see Rue et al. (2009).

[26] The Methodological Appendix introduces the basics of Bayesian regression.

5.1 Count Data Models

The general notation of GLM is adopted in the definition of the regression models. The conditional mean is thus denoted as $\mu_i = E(y_i|x_i) = h(x_i^{'\beta)}$, where $\eta_i = x_i^{'\beta}$ is the index or linear predictor, $g(\mu_i) = \eta_i$ the link function and $h(\cdot)$ the response function.

5.1.1 Poisson

The Poisson regression model takes the form $y_i|\eta_i \sim Po(\mu_i)$, assuming that the dependent variable y follows a Poisson distribution with mean and variance μ, with this distribution being

$$p(y|\beta) = \prod_{i=1}^{n} \frac{1}{y_i!} e^{-exp(\eta_i)} (exp(\eta_i))^{y_i}, \tag{2}$$

where $\mu_i = exp(\eta_i)$, with $\eta_i = x_i^{'\beta}$ being the linear predictor and $log(\mu_i) = \eta_i$ the link function. As mentioned above, a characteristic of the Poisson distribution is that its mean is equal to its variance, which is problematic when overdispersion and excess of zeros are present. In the Bayesian framework, the overdispersion can be taken into account within the Poisson regression model by specifying a hierarchical Poisson, which is done by introducing a random effect. Three different specifications are presented in the analysis presented in this section: (a) Poisson without random effects, (b) Poisson with random *iid* effects and (c) Poisson with both *iid* and spatial effects.

5.1.2 Negative Binomial

The Negative Binomial regression model is based on a mixture of Gamma and Poisson distributions, and is intended to capture the overdispersion of the model by including a Gamma-distributed random variable in the model. This regression model takes the form $y_i|\eta_i \sim NB(r, \mu_i)$, and the likelihood of the model is defined as

$$Prob(y) = \frac{\Gamma(r+y)}{y!\Gamma(r)} (1-p)^r p^y, \tag{3}$$

where $\mu = r(1-p)p = exp(\eta)$ and $\sigma^2 = \mu + \mu^2/r$. In this model, r is intended to capture overdispersion. Regarding the Bayesian estiamtion, overdispersion is captured by the hyperparameter $log(r)$, for which a flat prior is set.

5.1.3 Zero Inflated Models

Zero Inflated (ZI) regression models are a type of mixture models in which the heterogeneity is introduced in a binary form, distinguishing zero from non-zero counts. Specifically, assuming that y is distributed either as a Poisson or a Negative Binomial, and that $y = 0$ with probability π, the ZI mixture model consists of a mixture of a degenerate distribution with mass at zero and a distribution function:

$$y \sim \pi I_{\pi} + (1 - \pi) f(y| \ldots), \tag{4}$$

where I_{π} is the degenerate distribution taking the value zero with a probability of one and $f(y| \ldots)$ is either a Poisson or a Negative Binomial probability mass function associated with y. In this expression, π is a hyperparameter where

$$\pi = \frac{exp(\lambda_{\pi})}{(1 + \lambda_{\pi})}, \tag{5}$$

where λ_{π} is the internal representation of π and the value of the initial prior, which is a flat prior in the estimated models. The resulting models are the Zero Inflated Poisson (ZIP) and the Zero Inflated Negative Binomial (ZINB).

5.1.4 Random Effects

As mentioned above, these three regression models are specified by including overdispersion in two ways. The first is the inclusion of a random term u, which is a vector of unstructured random effects with *iid* Gaussian priors with precision λ_u, so that $u|\lambda_u \sim (0, \lambda_u I)$. This effect is to be included in the PO and ZIP models, but not in the NB and ZINB. This is because NB models already account for the unobserved (and unstructured) dispersion, and if an extra random effect was included, this would be a redundancy as well as an overparametrisation of the model. The precision hyperparameter is given a Gamma prior with the values $\lambda_u \sim \Gamma(1, 0.01)$.

The second way to include overdispersion is by considering and specifying the spatial structure of the data. Each observed data y_i is linked to a spatial region s, so that s_i indicates the region the ith observation belongs to. A neighbourhood criterion must be specified in order to introduce a spatially correlated effect. In this article, it is assumed that the sites s_i and s_j are neighbours if they share a common border. In this case, the site j belongs to the neighbourhood of s_i, which is expressed analytically as $j \in \upsilon_i$. Moreover, b_i is assumed to be the number of neighbours of s_i. Given this terminology, $f_s(s_i)$ is the spatial effect for the ith observation. The prior model for f_s is a Conditional Autoregressive (CAR) model[27], so that the spatial process intensity $f_s(s_i)$ follows a conditional Gaussian distribution, defined as

$$f_s(s_i)|f_s(s_j)_{j\in\upsilon_i} \sim \mathcal{N}\left(\frac{1}{b_i}\sum_{j\in\upsilon_i} f_s(s_j), \frac{1}{b_i\lambda_s}\right), \tag{6}$$

where λ_s is an unknown precision parameter. In order to ensure identifiability of μ, a sum-to-zero restriction on $\sum_{i=1}^{N} f_s(s_i) = 0$ is imposed. The precision hyperparameter is given a Gamma prior with the values $\lambda_s \sim \Gamma(1, 0.01)$. The effect $f_s(s_i)$ is to be included in the four regression models, i.e. PO, NB and ZIP and ZINB.

5.2 Estimation and Results

Summing up the models and effects described above, there are ten resulting specifications to be estimated. In all of them, the vector of unknown parameters β is given a zero-mean Gaussian prior with a known and fixed precision parameter, so that $\beta|\lambda_\beta \sim (0, (1/\lambda_\beta)I)$ with $\lambda_\beta = 0.01$. Table 4 shows the diagnostics of the estimation of all specifications, including hyperparameter posteriors. In addition, two measures of fit are computed and reported. On the one side, Spiegelhalter et al. (2002) suggest the use of the Deviance Information Criterion (DIC) for comparison of Bayesian hierarchical models. Assuming a probability model $p(y|\beta)$, this indicator is defined as $DIC = E[D(\beta|y) + p_D$, where $E[D(\beta|y)$ is the mean of the Bayesian deviance and p_D is the effective number of parameters, which is proportional to the deviance variance and is regarded as a measure of model complexity. According to DIC, the smallest value is to be preferred. On the other hand, the marginal likelihood is useful because it can be used to rank, select and average different models[28]. Analytically, the marginalized likelihood is the probability of the data given the model type, not assuming any particular model

[27] This model was first introduced and developed by Besag (1974) and Besag et al. (1991). In addition, Banerjee et al. (2004) analyse the CAR model in the context of Hierarchical Bayesian modelling for spatial data, and Rue and Held (2005) do the same in the context of Gaussian Markov Random Fields (GMRF) and call this model an intrinsic GMRF model.

[28] Han and Carlin (2001) give a very comprehensive review of the computation of the marginal likelihood, and Nandram and Kim (2002) use this indicator to investigate the improvement in the goodness of fit of hierarchical Poisson regression models.

parameters. For a specific model M, the marginal likelihood therefore takes the form $p(x|M) = \int p(x|\beta, M)\, p(\beta|M)\, d\beta$. As in the case of DIC, a value closer to zero is to be preferred when comparing models.

[TABLE 3]

Several conclusions can be drawn when comparing the different estimations shown in Table 4.. First, all general models (PO, NB, ZIP and ZINB) improve the fit when random and spatial effects are taken into account. Second, of all these models, surprisingly the best fit is obtained with a simple PO model. The most likely explanation is that once random *iid* and spatial effects are properly specified, then overdispersion and excess of zeros are properly taken into account and thus neither the overdisperion parameter nor the zero-inflated parameters included in NB and zero-inflated models seem necessary. In other words, NB and ZINB models with random *iid* and spatial effects suffer from overparametrization, since the slight improvement in the Marginal Likelihood is not offset by the penalization due to the increasing number of parameters, and this is shown in a higher DIC value. For this reason, only the results of the three PO model estimations are shown in Table 5, i.e. estimations considering i) no effects, ii) random *iid* effects and iii) random *iid* and spatial effects. Although these three possibilities are presented, the results from DIC show that model improves greatly from i) to ii) and, to a lesser extent, from ii) to iii).Random *iid* and spatial effects should therefore be taken into account and, the analysis consequently only focuses on model iii). In any case, the results from model iii) of Zero Inflated Poisson are quite similar and could also be considered for the analysis, but for the sake of simplicity we have decided to concentrate our discussion on only one estimation (i.e., the one with the best fit).

[TABLE 4]

As usual in empirical location literature agglomeration economies are a strong locational determinant[29], although this effect has been measured in many different ways (see Arauzo-Carod et al. 2010 for a review on several variables that proxy this phenomena). In this paper, agglomeration economies are proxied using

[29] See Arauzo-Carod (2005), Coughlin and Segev (2000) and Guimarães et al. (2000), among others.

employment density. Specifically, local (municipal) employment density (EMP) has a positive effect on the location of new manufacturing plants but if this density is so high (EMP2) the effect becomes negative due to congestion costs associated to densely populated areas. This result fits perfectly with an inverted U-shape profile related to the effect of concentration of economic activity over site attractiveness: initially it is positive but after a certain threshold is crossed it becomes negative. Industrial diversity (HHI) has a negative effect on the location of new plants as shown in previous research for Catalonia (Arauzo-Carod 2005) but these results are in contrast to those obtained for Portugal (Holl 2004a, 2004b). The structure of firms at a local level also influences location decisions. In this regard, new plants are favoured by areas where average establishment size is lower (SIZE). This results confirms previous empirical evidence (Bade and Nerlinger, 2000) regarding the positive effects of networks of SMEs on the location of new plants. New entrants are also positively influenced by activity rate (ACT) at a local level, which means that a positive local economic climate helps to attract additional activity. This latter result is confirmed by the positive effect of the share of manufacturing employment (SME)[30], but the existence of service activities (SSE) do not seem to have any effect over entries.

The geographical position of sites also matters since it implies important differences in terms of infrastructures, accessibility, site availability and markets. Although there are several ways to proxy these differences, one of them refers to altitude above sea level (ALT). In specific terms, the higher the altitude, the less flat the potential sites, and the lower the level of accessibility, which implies fewer entries. Sites located at higher altitudes are therefore less convenient for new plants and are chosen less often than other areas at sea level. Another way to proxy geographical position is to consider the time distance to main cities (TMC). Since firms prefer to be located close to economic and administrative centres, the greater the distance to these areas, the lower the levels of entry of new plants, as other scholars point out (see, for instance, Arauzo-Carod, 2008; Cieślik, 2005b; Holl, 2004b and List, 2001). Good accessibility is therefore an important issue to be taken into account when choosing a site, as demonstrated by most of the empirical literature for different types of areas. As well as good accessibility, new plants also prefer to

[30] The results of Arauzo-Carod (2005) and Smith and Florida (1994) point in the same direction.

be located in county capitals (CC) where they can access main facilities and markets and also in areas with higher quality of life standards, such as coastal areas (CL).

Although this is a surprising result, the negative effect of education level of individuals (EDU) on the entry of new firms fits the characteristics of new plants (most of which have medium-low technology levels and consequently, a low dependence on skilled labour) and with previous empirical evidence regarding Catalonia (Arauzo-Carod, 2005). In any case, the effect of human capital on firm entries is still unclear, since scholars have provided contradictory results: positive (Coughlin and Segev, 2000; Smith and Florida, 1994 and Woodward, 1992) negative (Cieślik, 2005a and Bartik, 1985) and mixed (Arauzo-Carod and Viladecans, 2009).

Apart from previous comments regarding the effect of explanatory variables over the location of new plants, we have descomposed the extant overdispersion into an unstructured *iid* effect and a spatially structured effect. The values of the hyperparameters therefore show that overdispersion is mainly explained in terms of a random process (roughly 66% of the effect) than by spatial issues (roughly 33% of the effect), although standard deviation is higher for the random process. We have thus "identified" about 33% of the overdispersion as being caused by the spatial position of the municipalities and their interaction with neighbours, while there is still room for further research about the origins of the remaining 66% of the effect. In any case, the spatial parameter has been estimated according to a very simple spatial weight matrix (contiguous neighbours) due to computational constraints (i.e., the number of neighbours is lower than if another distance-based criterion is used), so it is possible that the spatial effect would be higher with more "real" neighbours.

6 Summary and Future Research

In this article, several Count Data models have been specified and estimated in order to study industrial location determinants in Catalonia. In specific terms, Poisson, Negative Binomial and Zero Inflated models have been specified and estimated by means of Bayesian estimation techniques. Furthermore, additional *iid* and spatial random effects are included in the models in order to account for unobserved spatial heterogeneity in the data, thus allowing for structured and unstructured overdispersion as well as spatial correlations between observations. The results obtained here show a close relationship between the industrial location and space, and they are summarised as follows:

a) A common characteristic for all the variables is the huge range between the minimum and maximum values, which is due to the enormous heterogeneity present in the data. This result is to be expected, since the municipalities are very dissimilar to each other, ranging from small isolated villages in the mountains to densely populated cities. This factor suggests that when industrial location phenomena are analysed, spatial aggregation level of the analysis should be carefully taken into account. Municipality-level data are used here, but whether other spatial units could be more appropriate (e.g. counties or travel-to-work areas) is an issue that has been postponed for future research.

b) Regarding the characteristics of firm locations, the sum of zero counts outnumbers the number of positive counts each year. Furthermore, the percentile information shows that the distribution of entries is heavily skewed, i.e. there is a small group of municipalities that accounts for the largest number of entries, and more than half have no entries at all, which supports the previous comments regarding data heterogeneity. In addition, the entries tend to be grouped, i.e. they are mostly along the Catalan coast and there are territories in the rural part of the region with few entries, as entries tend to be concentrated in the Barcelona metropolitan area and to a lesser extent, in regions close to the coast and belonging to urban areas.

c) Following a spatial exploratory analysis consisting of estimation of the Moran's I for the dependent variable as well as the main covariates, it can be concluded that the spatial distribution of such variables is not random at all. Considering different neighbourhood criteria, i.e. k-nearest neighbours criterion, distance neighbours criterion (including those territorial units within a circle with a previously specified radius) and the contiguous neighbours criterion (considering those units with which the unit shares a border as neighbours), a positive and significant Moran's I estimation is found for all of them.

d) The results of the various econometric models estimated show that the best fit is achieved by a simple Poisson model allowing for two random effects, i.e. an *iid* and a spatial random effect. As a result, according to the Bayesian indicators Deviance Information Criterion (DIC) and Marginal Likelihood

(ML)indicators, this way of modelling overdispersion is to preferable to Negative Binomial and Zero Inflated models.

e) The results show that spatial effects are determinant and must consequently be taken into account. Likewise, overdispersion is descomposed into an unstructured *iid* effect and a spatially structured effect, while the former is stronger than the later, although there is still room for further research into this issue (e.g. using alternative settings of neighbour criteria).

A Methodological Appendix: Bayesian Inference

Bayesian inference tools represent an alternative to classical frequentist methods, and it is specifically suitable for the setting and study of complex models for which classical quantitative tools are not appropriate. In fact, it is not just a question of how the model is estimated from a numerical point of view: Bayesian analysis proposes a completely different way of approaching the phenomenon under study[31]. The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis. Let θ be the vector of parameters of interest which is the object of the study (for instance, the coefficients of an econometric specification), and let y denote the observed data for the dependent variable and x a set of explanatory variables. The goal of the analysis is to make an inference, i.e. to draw conclusions about θ. This analysis is made in terms of probability statements conditional on the observed value of y, i.e. the posterior distribution $p(\theta|y)$. However, how can this be done? The starting point is the definition of a full probability model, which consists of the modelling of the joint probability distribution for θ and y, i.e. $p(\theta, y)$, which is the main framework of the analysis because it bounds together all observable and unobservable quantities. The results of the inference largely depend on the model specified, and as a result it should be consistent with knowledge of the phenomenon being studied as well as the nature and characteristics of the data to be used. The density function of this model can be written as a product of two components:

♠ **Prior distribution** $p(\theta)$: refers to the distribution of the parameters under study.

[31] See Gelman et al. (2004) for a complete textbook on Bayesian data analysis.

- ♠ **Sampling distribution** $p(y|\theta)$: the empirical distribution of the observed data conditional on the set of parameters θ.

The joint probability distribution thus has the expression $p(\theta, y) = p(\theta)p(y|\theta)$. In order to make an inference from this expression, we use Bayes' rule. The posterior density is therefore expressed as the joint probability distribution conditioned on the known value of y, i.e.:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{7}$$

Since $p(y)$ does not depend on θ, it can be considered a constant and equation (7) can be expressed as

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{8}$$

which is called unnormalised posterior density, and constitutes the technical core of Bayesian inference: first a full model is specified, and then depending on this model as well as on the observed data, the appropriate posterior distribution is calculated, interpreted and its fit evaluated. In fact, an important step to be taken after the inference is to ask the following questions: does the model fit the data? Are the results sensitive to the assumptions made while setting the full model? Are the conclusions reasonable?

The empirical estimation of Bayesian models relies on simulation processes. The basic idea is to generate samples from a probability distribution and study their histogram. With a large enough sample, the histogram provides accurate information about many features of the distribution of : moments, percentiles and other summary statistics. The generation of samples from the probability distribution is achieved by using (pseudo)random number sequences. In fact, this tricky process has been troublesome to implement for many years, since it is based on complex numerical calculations. Fortunately, in recent years both new estimation algorithms and new specific software have contributed to the increasing availability of Bayesian tools for empirical researchers[32]. These algorithms can be easily implemented by

[32] The most used simulation method in the empirical Bayesian literature is the *Markov Chain Monte Carlo (MCMC)* simulation, which can be implemented using algorithms such as the Gibbs sampler and the Metropolis-Hastings algorithm (see Gelman, 2004 for technical details).

using the Bayesian package BUGS, , for instance, operating within the general statistical package *R*. Likewise, the INLA estimation package considered in this article can also be operated within the R environment, and the University of Munich has created the BayesX package. All these packages enable estimation of a wide array of models, and are thus very flexible.

B Figures and Tables

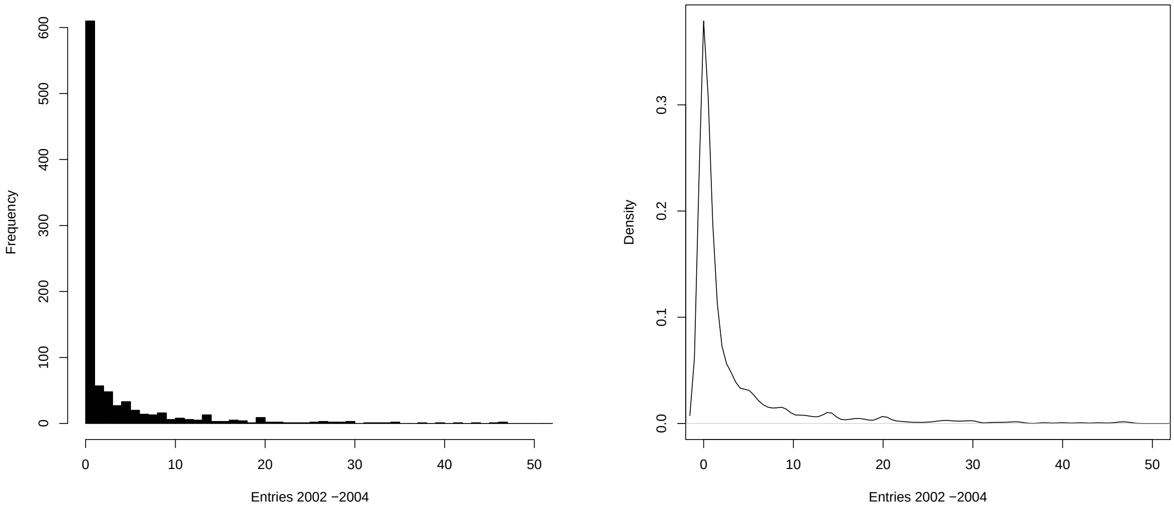

Fig. 1: Establishment entries 2002 - 2004: histogram (left) and kernel density estimation (right).

Tab. 1: Description of variables

Variable	Description	Source
Dependent Variable		
ENT	Number of entries (2002 - 2004)	REIC
A) Agglomeration Economies		
EMP	Employment Density	IDESCAT
EMP2	Squared of Employment Density	IDESCAT
HHI	Hirschman-Herfindahl Index	IDESCAT
B) Industrial Mix		
SIZE	Average Establishment Size	Own
ACT	Activity Rate	Own
SME	Share of Manufacturing Employment	IDESCAT
SSE	Share of Services Employment	IDESCAT
C) Spatial Effects		
W-EMP	Spatial Lag of EMP	own
W-HHI	Spatial Lag of HHI	own
W-SME	Spatial Lag of SME	own
D) Human Capital		
EDU	Average Years of Schooling	IDESCAT
E) Geographical Position		
ALT	Altitude	ICC
TMC	Transport time to Main Cities	ICC
CC	County Capital	ICC
CL	Coast Location	ICC
MAB	Metropolitan Area of Barcelona	T & B
MAG	Metropolitan Area of Girona	T & B
MAT	Metropolitan Area of Tarragona	T & B
MAL	Metropolitan Area of Lleida	T & B
MAM	Metropolitan Area of Manresa	T & B

Sources: Catalan Manufacturing Establishments register (REIC), Catalan Statistical Institute (IDESCAT), Catalan Cartographical Institute (ICC) and Trullén and Boix (2004), (T & B).

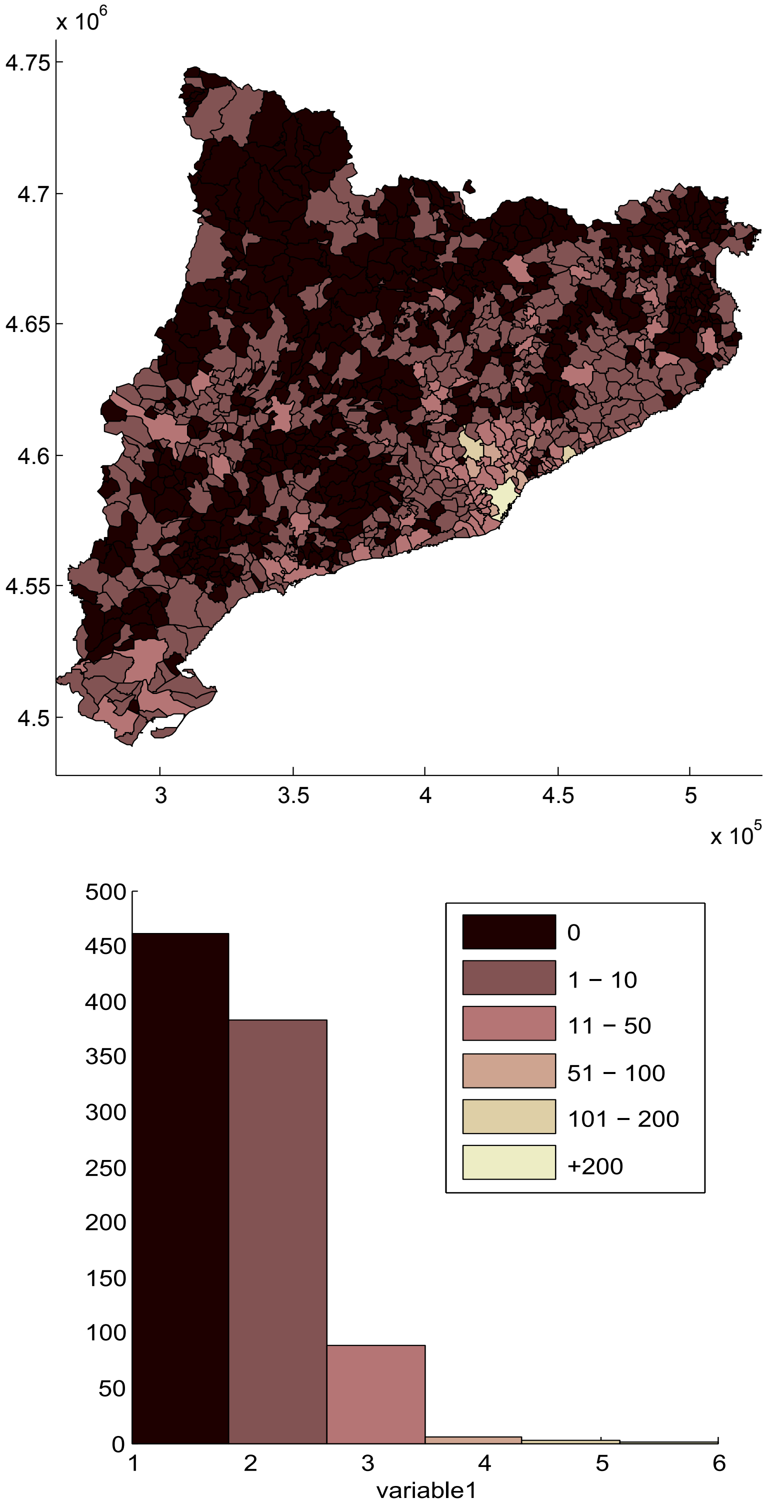

Fig. 2: Establishment entries 2002 - 2004: map representation with entries in intervals.

Tab. 2: Descriptive statistics of establishment entries 2002 - 2004

	Years			
	2002	2003	2004	$2002 - 2004$
Mean	1.420	1.484	1.453	4.359
Std. Dev.	4.889	5.222	4.800	14.590
Zero counts	614	609	612	461
Positive counts	327	332	329	480
Sum.	1337	1397	1368	4102
Min.	0	0	0	0
Max.	87	91	80	258
5th percentile	0	0	0	0
10th percentile	0	0	0	0
25th percentile	0	0	0	0
50th percentile	0	0	0	1
75th percentile	1	1	1	3
90th percentile	4	4	4	11
95th percentile	7	7	7	20

Source: own elaboration.

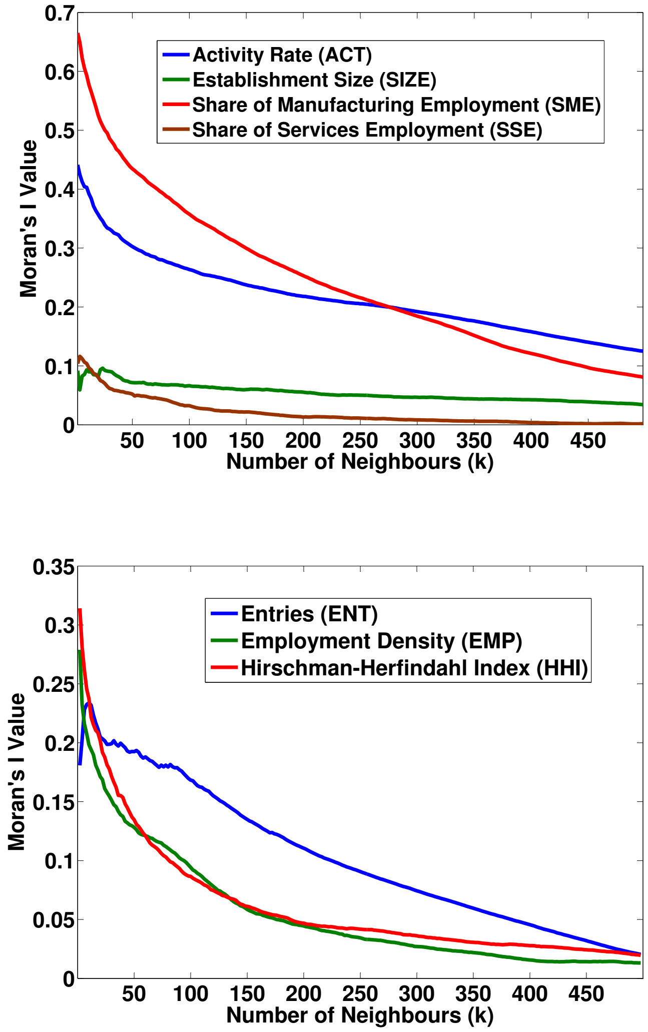

Fig. 3: Moran's I statistic considering an increasing number of neighbours

Tab. 3: Main Estimation Results

	Poisson (PO)			**Zero Inflated Poisson (ZIP)**		
A) No effects						
Measures of fit						
DIC	4659.12			4385.73		
Marginal Likelihood	0.00			−2219.24		
Hyperparameters						
				Mean	Std. Dev.	(2.5%, 97.5%) quant
Zero probability $\hat{\lambda}_\pi$				0.270	0.024	(0.223, 0.319)
B) Random *iid* effect						
Measures of fit						
DIC	2491.33			2497.56		
Marginal Likelihood	−1472.60			−1478.17		
Hyperparameters						
	Mean	Std. Dev.	(2.5%, 97.5%) quant	Mean	Std. Dev.	(2.5%, 97.5%) quant
iid parameter $log(\hat{\lambda}_u)$	1.005	0.100	(0.816, 1.223)	1.019	0.106	(0.829, 1.246)
Zero probability $\hat{\lambda}_\pi$				0.006	0.005	(0.002, 0.020)
C) Random *iid* and spatial effects						
Measures of fit						
DIC	2453.17			2455.09		
Marginal Likelihood	−2140.20			−2145.86		
Hyperparameters						
	Mean	Std. Dev.	(2.5%, 97.5%) quant	Mean	Std. Dev.	(2.5%, 97.5%) quant
iid parameter $log(\hat{\lambda}_u)$	1.840	0.346	(1.293, 2.654)	1.846	0.347	(1.299, 2.667)
Spatial parameter $log(\hat{\lambda}_s)$	1.010	0.317	(0.555, 1.800)	1.010	0.324	(0.552, 1.817)
Zero probability $\hat{\lambda}_\pi$				0.001	0.003	(0.000, 0.010)
	Negative Binomial (NB)			**Zero Inflated Negative Binomial (ZINB)**		
A) No effects						
Measures of fit						
DIC	2913.27			2913.40		
Marginal Likelihood	−1463.68			−1463.96		
Hyperparameters						
	Mean	Std. Dev.	(2.5%, 97.5%) quant	Mean	Std. Dev.	(2.5%, 97.5%) quant
Overdispersion $log(\hat{r})$	0.931	0.086	(0.773, 1.112)	0.932	0.086	(0.773, 1.114)
Zero probability $\hat{\lambda}_\pi$				0.0006	0.002	(0.000, 0.006)
B) Spatial effect						
Measures of fit						
DIC	2912.52			2913.59		
Marginal Likelihood	−2143.90			−2138.16		
Hyperparameters						
	Mean	Std. Dev.	(2.5%, 97.5%) quant	Mean	Std. Dev.	(2.5%, 97.5%) quant
Spatial parameter $log(\hat{\lambda}_s)$	3431.95	1833.45	(796.66, 7865.94)	69832	36943	(17014, 159160)
Overdispersion $log(\hat{r})$	0.928	0.085	(0.771, 1.108)	0.931	0.087	(0.773, 1.114)
Zero probability $\hat{\lambda}_\pi$				0.002	0.005	(0.000, 0.017)

Tab. 4: Poisson Models Estimations

	No effects				Random effects				Random and spatial effects			
	Mean	St. Dev.	2.5% quant	97.5% quant	Mean	St. Dev.	2.5% quant	97.5% quant	Mean	St. Dev.	2.5% quant	97.5% quant
Parameters												
CONST	−15.559	1.468	−18.462	−12.702	−16.970	3.747	−24.416	−9.619	−15.404	3.863	−22.833	−7.713
HHI	−2.172	0.306	−2.779	−1.577	−2.798	0.656	−4.104	−1.530	−2.388	0.638	−3.656	−1.152
EMP	2.908	0.282	2.364	3.470	2.628	0.609	1.452	3.846	2.253	0.592	1.113	3.438
EMP2	−0.155	0.018	−0.191	−0.121	−0.153	0.040	−0.233	−0.075	−0.133	0.039	−0.211	−0.058
SIZE	−0.132	0.006	−0.143	−0.121	−0.105	0.012	−0.129	−0.081	−0.096	0.012	−0.120	−0.073
ACT	1.326	0.696	−0.040	2.691	6.666	1.598	3.547	9.822	5.414	1.606	2.271	8.578
SME	3.892	0.323	3.259	4.526	4.173	0.742	2.724	5.637	3.904	0.754	2.429	5.389
SSE	0.595	0.138	0.323	0.865	−0.233	0.299	−0.823	0.349	−0.434	0.295	−1.016	0.140
W-HHI	−1.226	1.093	−3.386	0.899	−3.730	2.861	−9.374	1.855	−0.805	3.292	−7.292	5.625
W-EMP	0.721	0.124	0.479	0.966	0.839	0.414	0.029	1.656	0.481	0.410	−0.316	1.295
W-SME	8.503	1.307	5.917	11.045	9.533	3.429	2.810	16.277	−0.267	4.416	−9.048	8.373
ALT	−0.001	0.000	−0.001	0.000	−0.001	0.000	−0.002	−0.001	−0.002	0.000	−0.003	−0.001
EDU	−0.171	0.027	−0.224	−0.117	−0.168	0.067	−0.299	−0.038	−0.179	0.066	−0.309	−0.050
TMC	−0.023	0.002	−0.027	−0.019	−0.021	0.004	−0.030	−0.012	−0.027	0.008	−0.042	−0.011
CC	0.959	0.052	0.857	1.062	1.685	0.202	1.290	2.085	1.645	0.188	1.277	2.016
CL	0.402	0.056	0.291	0.512	0.527	0.190	0.155	0.901	0.591	0.202	0.195	0.989
MAB	0.897	0.066	0.768	1.026	0.663	0.167	0.334	0.990	0.239	0.209	−0.174	0.645
MAG	−0.309	0.113	−0.533	−0.090	−0.528	0.222	−0.966	−0.097	−0.449	0.304	−1.051	0.145
MAT	0.049	0.116	−0.181	0.273	−0.264	0.256	−0.771	0.234	−0.037	0.380	−0.789	0.702
MAL	0.892	0.132	0.629	1.148	0.854	0.274	0.313	1.390	0.274	0.384	−0.486	1.019
MAM	0.521	0.096	0.332	0.706	0.621	0.253	0.122	1.116	0.230	0.327	−0.415	0.868
Hyperparameters												
iid parameter $log(\hat{\lambda}_u)$					1.005	0.100	0.816	1.223	1.840	0.345	1.294	2.654
Spatial parameter $log(\hat{\lambda}_s)$									1.010	0.317	0.555	1.800
DIC	4659.12				2491.33				2453.17			
Marg. Like.	0.00				-1472.60				-2142.20			

References

Alañón, Á., Arauzo-Carod, J.M. and Myro, R. (2007). Accessibility, agglomeration and location. *Entrepreneurship, Industrial Location and Economic Growth, Edward Elgar: Chentelham*.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer.

Arauzo-Carod, J.M. (2005). Determinants of industrial location: An application for Catalan municipalities. *Papers in Regional Science*, 84(1):105–120.

Arauzo-Carod, J.M. and Manjón-Antolín, M. (2004). Firm Size and Geographical Aggregation: An Empirical Appraisal in Industrial Location. *Small Business Economics*, 22(3):299–312.

Arauzo-Carod, J.M. and Manjón-Antolín, M. (2009). (Optimal) Spatial Aggregation in the Determinants of Industrial Location. *Working Papers 10-2009. Rovira i Virgili University. Department of Economics*.

Arauzo-Carod, J.M. and Viladecans, E. (2009). Industrial Location at the Intra-metropolitan Level: The Role of Agglomeration Economies. *Regional Studies*, 43(4):545–558.

Arauzo-Carod, J.M., Liviano, D. and Manjón-Antolín, M. (2010). Empirical Studies in Industrial Location: An Assessment of their Methods and Results. *Journal of Regional Science*, 50(3):685–711.

Arauzo-Carod, J.M., Liviano, D. and Martín, M. (2008). New business formation and employment growth: some evidence for the Spanish manufacturing industry. *Small Business Economics*, 30(1):73–84.

Autant-Bernard, C., Mangematin, V. and Massard, N. (2006). Creation of Biotech SMEs in France. *Small Business Economics*, 26(2):173–187.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall.

Barbosa, N., Guimarães, P. and Woodward, D. (2004). Foreign firm entry in an open economy: the case of Portugal. *Applied Economics*, 36(5):465–472.

Basile, R. (2004). Acquisition versus greenfield investment: the location of foreign manufacturers in Italy. *Regional Science and Urban Economics*, 34(1):3–25.

Basile, R., Benfratello, L. and Castellani, D. (2010). Location Determinants of Greenfield Foreign Investments in the Enlarged Europe: Evidence from a Spatial Autoregressive Negative Binomial Additive Model. *Working Papers. University of Torino, Department of Economics and Public Finance*.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Bivand, R.S. and Portnov, B.A. (2004). Exploring spatial data analysis techniques using R: The case of observations with no neighbors. *Advances in Spatial Econometrics: Methodology, Tools and Applications*, pages 121–142.

Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer.

Blonigen, B.A., Davies, R.B., Waddell, G.R. and Naughton, H.T. (2007). FDI in space: Spatial autoregressive relationships in foreign direct investment. *European Economic Review*, 51(5):1303–1325.

Brett, C. and Pinkse, J. (1997). Those taxes are all over the map! A test for spatial independence of municipal tax rates in British Columbia. *International Regional Science Review*, 20(1-2):131.

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Cieślik, A. (2005). Regional characteristics and the location of foreign firms within Poland. *Applied Economics*, 37(8):863–874.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley and Sons, New York.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Danielkissling, W. and others (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Gabe, T. (2003). Local Industry Agglomeration and New Business Activity. *Growth and Change*, 34(1):17–39.

Gabe, T. and Bell, K. (2004). Tradeoffs between Local Taxes and Government Spending as Determinants of Business Location. *Journal of Regional Science*, 44(1):21–41.

Gelman, A. (2004). *Bayesian data analysis*. CRC press.

Griffith, D.A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. *Practical handbook of spatial statistics*, pages 65–82.

Gschlössl, S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical papers*, 49(3):531–552.

Han, C. and Carlin, B.P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Jacqmin-Gadda, H., Commenges, D., Nejjari, C. and Dartigues, J.F. (1998). Tests of geographical correlation with adjustment for explanatory variables: An application to dyspnoea in the elderly. *Statistics in medicine*, 16(11):1283–1297.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. and Citron-Pousty, S. (2002). Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, pages 616–625.

Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk. *Journal of the American Statistical Association*, 97(459):692–701.

Kühn, I. (2007). Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions*, 13(1):66–69.

Lambert, D.M., McNamara, K.T. and Garrett, M.I. (2006). An Application of Spatial Poisson Models to Manufacturing Investment Location Analysis. *Journal of Agricultural and Applied Economics*, 38(01).

Lin, G. and Zhang, T. (2007). Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky breast cancer data. *Geographical Analysis*, 39(3):293.

List, J.A. (2001). US county-level determinants of inbound FDI: evidence from a two-step modified count data model. *International Journal of Industrial Organization*, 19(6):953–973.

Manjón-Antolín, M. and Arauzo-Carod, J.M. (2010). Locations and relocations: Modelling, determinants, and interrelations. *Annals of Regional Science, forthcoming*.

McCann, P. and Sheppard, S. (2003). The Rise, Fall and Rise Again of Industrial Location Theory. *Regional Studies*, 37(6):649–663.

McCullagh, P. and Nelder, J.A. (1989). Generalized linear models. *Monographs on Staistics and Applied Probability, 2nd edition, Chapman and Hall, London*, 37.

Miller, J., Franklin, J. and Aspinall, R. (2007). Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, 202(3-4):225–242.

Mollie, A. (1996). Bayesian mapping of disease. *Markov chain Monte Carlo in practice*, pages 359–379.

Moran, P.A.P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Moran, P.A.P. (1950). A test for the serial independence of residuals. *Biometrika*, 37(1):178–181.

Nandram, B. and Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72(4):319–340.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman & Hall.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, 71(2):319–392.

Schabenberger, O. and Gotway, C.A. (2005). *Statistical methods for spatial data analysis*. CRC Press.

Smith, D.F. and Florida, R. (1994). Agglomeration and Industrial Location: An Econometric Analysis of Japanese-Affiliated Manufacturing Establishments in Automotive-Related Industries. *Journal of Urban Economics*, 36(1):23–41.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 64(4):583–639.

Trullén, J. and Boix, R. (2004). Indicadors 2005.

Viladecans, E. (2004). Agglomeration economies and industrial location: city-level evidence. *Journal of Economic Geography*, 4(5):565.

Wu, F. (1999). Intrametropolitan FDI firm location in Guangzhou, China. *The Annals of Regional Science*, 33(4):535–555.