UNIVERSITAT
ROVIRA I VIRGILI

WORKING PAPERS

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

"Agglomeration and Location: a Nonparametric Approach"

Josep-Maria Arauzo-Carod
Daniel Liviano

Document de treball nº -5- 2007

**DEPARTAMENT D'ECONOMIA
Facultat de Ciències Econòmiques i Empresarials**

UNIVERSITAT
ROVIRA I VIRGILI

Edita:

Departament d'Economia
http://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
432004 Reus
Tel. +34 977 759 811
Fax +34 977 300 661

Dirigir comentaris al Departament d'Economia.

Dipòsit Legal: T-1571 - 2007
ISSN 1988 - 0812

DEPARTAMENT D'ECONOMIA
Facultat de Ciències Econòmiques i Empresarials

Agglomeration and Location: a Nonparametric Approach

(Very preliminary version. Please, do not quote)

Daniel Liviano (♣,♦): daniel.liviano@urv.cat
Josep-Maria Arauzo-Carod (♣): josepmaria.arauzo@urv.cat

Abstract:

The aim of this article is to assess the effects of several territorial characteristics, specifically agglomeration economies, on industrial location processes in the Spanish region of Catalonia. Theoretically, the level of agglomeration causes economies which favour the location of new establishments, but an excessive level of agglomeration might cause diseconomies, since congestion effects arise. The empirical evidence on this matter is inconclusive, probably because the models used so far are not suitable enough. We use a more flexible semiparametric specification, which allows us to study the nonlinear relationship between the different types of agglomeration levels and location processes. Our main statistical source is the REIC (Catalan Manufacturing Establishments Register), which has plant-level microdata on location of new industrial establishments.

Keywords: agglomeration economies, industrial location, Generalized Additive Models, nonparametric estimation, count data models.

(♣) Quantitative Urban Regional Economics (QURE). Department of Economics, Universitat Rovira i Virgili. Av. Universitat 1. 43204 – Reus, Catalonia (Spain). Phone: +34 977 759 800.

(♦) Entrepreneurship, Growth and Public Policy Group. Max Planck Institute of Economics. Kahlaische Straße 10. 07745 – Jena, Germany.

This research was partially funded by the CICYT: SEJ2004-05860/ECON, the CICYT: SEJ2004-07824/ECON and the Xarxa de Referència d'R+D+I en Economia i Polítiques Públiques (Catalan Government). We would like to acknowledge the helpful comments of Miguel Manjón and Elisabet Viladecans. Any errors are, of course, our own.

1. Introduction

Industrial location literature is attracting a growing interest from researchers in recent years, in which empirical and theoretical contributions have increased considerably. The reasons of such situation can be found at policy implications that derive from those papers. Nowadays, economic activity is being more mobile than before and this means that traditional sources of competitiveness are being modified. Therefore, it is important to know which the location determinants of those manufacturing firms are, and why some territories seem to be in a better position to receive new firms than others.

In this paper we analyse such location issues by focusing in a specific area that has received a little attention until now, which is how agglomeration economies (which are one of the most important location determinants) are shaped according to other territorial characteristics. Our assumption is that the incidence of agglomeration economies over entry decisions is not constant across space and varies significantly according to the degree in which there is a different industrial structure in each part of the territory. So, the incidence of those agglomeration economies is not the same, for instance, in a specialised area than in a diversified area.

For doing that, we use a different methodological approach than previous empirical contributions, and we use a more flexible semiparametric specification, which allows us to study the nonlinear relationship between the different types of agglomeration levels and location processes.

We have structured the paper as follows. In section 2 we discuss the literature about location determinants and about agglomeration and disagglomeration economies. In section 3 we present the model and the data and we also review main models used by scholars in previous contributions. In section 4 we show the estimates and the results. Finally, in section 5 we summarise our main conclusions.

2. Literature Review

2.1 Determinants of location decisions

Location decisions of manufacturing firms have been analysed by a wide range of scholars departing from different theoretical approaches about what are the main determinants of those decisions. Nevertheless this theoretical diversity, most of those contributions can be grouped into three main approaches (Hayter 1997): a neoclassical approach, a behavioural approach and an institutional approach.

The neoclassical approach is mainly related to classical location theory and focus the analysis on profit maximization and cost-minimising strategies. Scholars that contribute inside this approach consider that location determinants are, mainly, quantitative, and are related with issues like wages, land costs, transportation costs, etc. This is mainly quantifiable variables that can (sometimes) easily be obtained for empirical analysis.

The behavioural approach deals with situations of imperfect information and uncertainty. Unlike the neoclassical approach, location decisions are taken considering (also) non economic issues, like the social and family environment of the entrepreneur or his / her locational preferences. This approach needs detailed entrepreneur’s information that usually is not available to researchers.

Finally, the institutional approach point analyses over institutional issues that also influence firm’s location decisions. Among those issues some of them have been identified: characteristics of suppliers and customers, firm networks, public policies and trade union’s strategies. This later approach also needs a huge amount of information that it is not easy to obtain and that usually is qualitative and need to be transformed into categorical variables.

Given the scarcity of information scholars usually follow the first approach and use those types of quantitative variables in order to analyse location decisions.

Among the empirical papers that focus on this approach there are those of Arauzo (2005), that focus on agglomeration economies; Holl (2004a, 2004b and 2004c), that focus on transport infrastructures; Guimarães et al. (2000), that focus on agglomeration economies and transport infrastructures; and List (2001) and List and McHone (2000), that focus on wages, market size, taxes, agglomeration economies and environmental regulations. Nevertheless, there are some scholars that take into account entrepreneur's point of view (behavioural approach), like Figueiredo et al. (2002).

2.2 Agglomeration and disagglomeration economies

Usually agglomeration economies have been identified as a major source of location of economic activity. Apart from other contributions by Alfred Weber (the impact of transportation costs on location decisions), Johann Heinrich Von Thünen (land use model), Walter Christaller (Central Place Theory) and William Alonso (Central Business District), Alfred Marshall (1890) needs to be underlined because he showed the idea of external economies. That is to say, benefits derived from the concentration of jobs and firms in an area.

Initially, Marshall (1890) introduced the idea of external economies as a source of competitiveness additionally to the well known internal economies. Specifically, Marshall classified those external economies as a specialised labour market, supplier's availability and knowledge spillovers. Later, Hover (1937 and 1936) offered a more detailed classification of those external economies by dividing them into localisation economies and urbanisation economies. The first ones relates to the concentration of similar activities close to each other, while the second one relates to concentration of economic activity as a whole, no matters the industry to which those activities belong. According to Hoover (1937 and 1936) classification, localisation economies are external to the firm but internal to the industry while urbanisation economies are both external to the firm and to the industry.

Hoover's classification has become so popular among scholars, so there is an endless list of contributions that rely on this classification[1]. Most of those contributions analyse which type of external economies is of more importance when explaining location decisions of firms. This is a key issue, because it explains whether firms prefer to be surrounded by firms of the same industry or they just prefer to be in a site with economic activity regardless of its type. An extensively review of several classifications regarding agglomeration economies can be found in Parr (2002).

While previous contributions emphasized the positive role played by agglomeration of economic activity, Townroe (1969) introduced the idea that an excessive concentration of economic activity generates negative effects that can overcome the benefits derived from the urbanization economies. This later assumption implies that there are some limits to concentration and, therefore, growing and increasing economic activity in an area is not ever the best solution. Those too large cities must face some problems like environmental pollution[2], traffic congestions, excess commuting, higher wages or high land prizes. Other scholars, like Quigley (1998), focus on poverty (caused by greater segregation) and crime rates (Glaeser, 1998, p. 152) given that there are "higher returns to crime in cities, perhaps due to scale economies in stolen goods or a greater market of potential victims". Additionally, Henderson (1997) shows that cost-of-living raises, which means that a wage premium is needed[3]. It seems to be some kind of trade-off between agglomeration and disagglomeration economies as Basile (2004, p. 8) points out: "Admittedly, agglomeration economies tend to reach limit values and agglomeration diseconomies eventually emerge. Indeed, firms operating on markets with a relatively large number of firms face stronger competition in product and labor markets. This act as a centrifugal force, which tends to make activities dispersed in space. Once the centrifugal forces surpass the effects of the agglomeration economies in a region, firms will look for locations in contiguous

[1] See, among others, Arauzo (2005), Duranton and Puga (2000), Figueiredo et al. (2002), Guimarães et al. (2000), Parr (2002), Rosenthal and Strange (2004), Viladecans (2004).

[2] Glaeser (1998) demonstrates that while some pollutants are correlated with city size, others are not affixed to the area that creates them.

[3] For a wider discussion about diseconomies see, among others, Keeble and Walker (1994), Krugman (1998), Moomaw (1988) and Zeng (1998).

regions where production costs are lower, while at the same time taking advantage of some degree of external economies, given the short distances involved."

Therefore, one could argue that cities grow (in terms of population, for instance) until the size in which the benefits of agglomeration are overwhelmed by the costs of congestion (Tolley, 1974). Then, there is a decline in urbanisation that has already been checked for some countries like U.S. (Glaeser, 1998) and France (Le Jeannic ,1997), among others[4]. This process has been extensively analysed by Brueckner (2000) and Mieszkowski and Mills (1993), among others.

So, there is some borderline between urbanization economies and disurbanization economies. The identification of this border is not an easy task and, usually, scholars have used some rough measures of disurbanization economies. At the empirical literature, the most common way to measure this negative phenomenon has been by using squared urbanization economies as a regressor, trying to catch up the effects of a very high concentration of economic activity.

From previous theoretical and empirical contributions there are two main conclusions that must be highlighted. The first one is that disurbanization economies seem to be of high importance, given that they contribute in a negative way to urban competitiveness. The second one is that the empirical approaches to these phenomena have been so much vague and, therefore, there is not an accurate way by which those negative effects are measured.

[4] An extensive review on urban sprawl for several countries (UK, France, Switzerland and US) can be found at Richardson and Baie (2004).

3. Model and data

3.1 Data

Our data refers to local units in Catalonia[5], and we have two types of datasets, on the one hand the data about firm entries and, on the other hand, the data about municipal characteristics.

The database about entries is the REIC (Catalan Manufacturing Establishments Register), which has plant-level micro data on the creation and location of new manufacturing establishments. The REIC provides data both about new and relocated establishments, and since they may be attracted to the territory by the same variables, we use both of them without making any distinction[6]. We also selected only those establishments with codes 12 to 36 (NACE-93 classification), and we drop out the incomplete registers. Then, we have the aggregated entries of manufacturing establishments on 907 municipalities (out of 946) over the years 2002-2004.

The database about territorial characteristics comes mainly from the Trullén and Boix (2004) database about Catalan municipalities, from the Catalan Statistical Institute (IDESCAT) and from the Catalan Cartographical Institute. Our data covers almost all the Catalan municipalities[7], and we have considered these variables for the year 2001. The variables about territorial characteristics used as regressors are classified into the following groups:

[5] Catalonia is an autonomous region of Spain with about 7 million inhabitants (15% of the Spanish population) and an area of 31,895 km2. It contributes 19% of Spanish GDP. The capital of Catalonia is the city of Barcelona.

[6] See Manjón and Arauzo (2006) for a detailed analysis of interrelations between locations and relocations.

[7] Due to lack of 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) we have drop them out.

Agglomeration Economies

We consider several variables measuring agglomeration economies, since it is a multidimensional concept that cannot be reduced to a single variable. Following the recent literature, agglomeration economies can be classified into two types: urbanization economies and localization economies (Henderson et al., 1995). As explained in the previous section, urbanization economies are associated with a city's population and employment levels and the diversity of its productive structure, while localization economies are associated with a city's specialization in one specific sector. There is no clear evidence on whether either urbanization or localization economies are more important for the location of new firms, being the empirical evidence mixed and inconclusive in that regard. Combes (2000) provides a discussion on that topic.

We try to proxy urbanization economies with three variables. *EMPD* stands for employment density and *POPD* stands for population density. In both measures the denominator is the surface in km2 of urbanised land. The third measure is intended to reflect the diversity of the productive structure of each municipality. In this respect we consider the Manufacturing Diversification Index (*MDI*), which is based on the correction for differences in sectoral employment shares at the national level of the inverse of the Hirschman-Herfindahl index proposed by Duranton and Puga (2000):

$$MDI_i = 1/\sum_j | s_{ij} - s_j | .$$

Where s_{ij} is the share of manufacturing activity *j* in manufacturing employment in municipality *i*, and s_j is the share of manufacturing activity *j* in total manufacturing employment.

Localization economies of each municipality are approached by the Relative Specialization Index (*RSI*), which is computed as:

$$RSI_i = \frac{1}{2}\sum_j | \frac{s_{ij}}{s_i} - \frac{s_j}{s} | .$$

Where s_{ij} is the share of manufacturing activity *j* in manufacturing employment in municipality *i*, s_j is the share of manufacturing activity *j* in total manufacturing employment, s_i is the share of total manufacturing activity in municipality *i*, and e is total manufacturing employment.

Human Capital

We consider three different variables that measure the level of human capital of Catalan municipalities. *HC1* is the percentage of science and technology workers over the total occupation of a municipality. *HC2* stands for the percentage of workers with a university degree over the total occupation of a municipality, and *HC3* is the average schooling years of the population over 25 years old.

Spatial Effects

The aim of the variables included here is to reflect the possible influence of agglomeration as well as human capital variables of neighbouring municipalities on a municipality's locations[8]. These variables are the spatial lags of agglomeration and human capital variables, and are computed by means of the product of a neighbourhood (or weight) matrix *W* with the independent variables. The resulting spatial lag stands for the averaged value of the values of the regressor in neighbouring municipalities. The weight matrix defines the concept of neighbourhood considered, and has a $n \times n$ dimension, being *n* the number of municipalities of the sample. Each element of this matrix (w_{ij}) stands for the geographical relationship between two municipalities. Here we consider a simple binary matrix, with $w_{ij} = 1$ when two municipalities share a border and $w_{ij} = 0$ otherwise. The final matrix has been standardised, so that each row sums 1[9]. The resulting independent variables are *W-EMPD, W-POPD, W-MDI, W-RSI, W-HC1, W-HC2* and *W-HC3*.

[8] For a contribution that considers a similar approach, see Viladecans (2004).

[9] For a seminal work on spatial econometric techniques, see Anselin (1988).

Size of the market

With the use of these variables, we try to assess the effect of the size of each local market on the attraction of new firm locations. We follow the study of Arauzo (2007), and consider employment (*EMP*) and population (*POP*) as proxies for the size of the market.

Industrial mix

These variables are aimed to reflect the industrial composition of each municipality, so that *EMP-MAN* is the share of manufacturing employment over total employment and *EMP-SER* is the share of service employment over total employment.

Geographic position

This set of variables is meant to control for the geographic position of each municipality. *COAST* is a dummy variable that takes the value one if the municipality belongs to a shore-line area, *DIS-CC* is the distance in kilometres to the nearest county capital, and *MAB*, *MAG*, *MAL*, *MAT*, and *MAM* take the value one if the municipality belongs to one of the five biggest metropolitan areas of Catalonia, which are the ones surrounding the towns of Barcelona, Girona, Lleida, Tarragona and Manresa, respectively.

3.2 Models used in previous contributions

Most of recent research work on location decisions is based on Count Data Models (CDM). These models allow the study of the location of firms from the point of view of the chosen geographic space, so that the contributions that consider this methodology focus on the effects that specific territorial factors have on the territory's probability of being chosen to locate a productive activity[10]. CDM are based on the Poisson as well as related distributions, and

[10] The other main stream of the literature considers discrete election models, since this methodology focus on the effect that individual characteristics exert on the location decision.

allow the analysis of the determinants of the expected number of new firms or establishments created in a certain location per unit of time. For this reason, they are the natural econometric resource when the location is analysed from the point of view of the territory. The use of CDM departs from the assumption that the location decision of a firm is based on the maximization of expected profits, and that these profits contain an i.i.d. stochastic term. Then, the probability that a firm chooses a certain location can be expressed in terms of a discrete random variable containing the result of this choice. Nevertheless, by adding these individual decisions this probability could easily refer to the number of entries carried out in a certain territory and period of time, and such random variable would follow a certain distribution which could be approximated by CDM[11].

CDM have been widely used in the study of firm location. The most popular distribution has been the Poisson distribution, since it is very suitable with highly disaggregated territorial variables. The reason has to do with the fact that some of the spatial units are likely to receive no establishments at all, and by means of a Poisson regression the covariates also help explaining these cases. In this literature this situation is known as the "zero problem"[12]. Barbosa et al. (2004) for the case of Portugal, Arauzo (2005) and Arauzo and Manjón (2004) for the case of Catalonia have considered a Poisson model in their firm location analysis.

However, Poisson models make two important assumptions. The first one is that the mean and the variance should be equal, but this is usually violated when we deal with industrial location decisions, because of the concentration of entries in some areas (this causes the variance to be greater than the mean, which is known as the "overdispersion problem"). The existence of overdispersion is explained in terms of unobserved heterogeneity in the mean function. The second assumption is about the "zero problem". Poisson models can deal with situations in which there are a high number of observations with

[11] For a discussion about the firm location decision problem, see Guimarães et al. (2000b).

[12] See Cameron and Trivedi (1998) for detailed information about how zero observations contribute to the likelihood function.

value zero, but some problems arise when this number is excessive. That is, the excess of zeros is likely to be due to the fact that zero and non-zero counts may follow a different probability distribution.

To solve such shortcomings, several studies have used alternative models to study firm location. The Negative Binomial has been the most used alternative model, and assumes that there exists unobserved individual heterogeneity among the observations, and therefore the overdispersion is accounted for by means of a more accurate modelling of the variance. Empirical contributions using this model are Coughlin and Segev (2000) and Smith and Florida (1994) for the U.S. case, and Cieslik (2005) for the case of Poland. When the data is longitudinal, fixed and random effects estimation have been applied both to Poisson and Negative Binomial models. Examples of the use of such models are Holl (2004a,b) for both Spain and Portugal. Besides, Conditional Poisson models, which assume that the mean value function is a stochastic process, have been used by List and McHone (2000) and Becker and Henderson (2000) for the U.S. case. In order to solve the "excess of zeros" problem, two models have been considered: the Zero Inflated Poisson model and the Zero Inflated Negative Binomial model. Both models assume different probability models for the zero and nonzero counts. After all, the selection of the model to be used on each case relies on the characteristics of the data set. In that regard, there exists a battery of tests that may be a guide in the selection of the final model[13].

A feature of the existing literature on firm location is that so far only purely parametric models have been considered. This approach is likely to suffer from a misspecification problem, especially when the relationship between the regressors and the dependent variable is not linear. This may be the case of the variables related to agglomeration economies and diseconomies. As stated before, the agglomeration of economic activity may boost the new location of firms on a particular location, because of the positive externalities it supposes. Up to a point, however, these advantages may turn into disadvantages, because a too high level of agglomeration may cause congestion and, hence,

[13] For a complete manual on CDM, see Cameron and Trivedi (1998).

diseconomies. The common way of measuring this nonlinear relationship on the literature has been to include two regressors for agglomeration: one in levels and the other its squared value. By including these covariates, a parabolic relationship between agglomeration and location is being assumed. Among the scholars that have used this rough measure there are Arauzo (2005) and Arauzo and Manjón (2004) analyzing the case of Catalonia. They obtained a positive value for the coefficient of the agglomeration variable in levels and a negative value for the coefficient of the squared value of that variable. Such a result confirms that both economies and diseconomies of agglomeration appear, but even so the true nonlinear relationship among these variables may still remain veiled. Viladecans (2004) have used a similar measure, which is squared population, while Barrios et al. (2006) just uses population to proxy land prizes and Carlino (1978) also uses the same variable. Other scholars use square root of population density (Keeble and Walker, 1994).

Even though the idea of agglomeration economies seems so clear, there are some problems about their empirical estimations. There are some scholars that share this argument, like Carlino (1979), for instance (p. 363): "While agglomeration economies have been well articulated, they have nonetheless been hard to measure. Most measurement techniques have been indirect rather than direct". So, measurement problems are not exclusive of disagglomeration economies but also of agglomeration economies.

3.3 Model

The model proposed in this article belongs to a class of statistical models for a univariate response variable which is called Generalize Additive Models for Location, Scale and Shape (GAMLSS), proposed by Rigby and Stasinopoulos (2005). This class of statistical models is a development of the Generalized Linear Models (GLM) and the Generalized Additive Models (GAM). GLM were introduced by Nelder and Wedderburn (1972) and further developed by McCullagh and Nelder (1989), and its basic feature is that the regression function, i.e. the expectation $\mu = E(Y \mid X)$ of Y is a monotone function of the index $\eta = X\beta$. The *link function* G [14] relates μ and η, so that:

$$E(Y \mid X) = G(X\beta) \;\Leftrightarrow\; \mu = G(\eta).$$

The GLM framework assumes that the distribution of Y is a member if the exponential family, which covers a broad range of distributions[15]. One of the most notorious developments of GLM are the GAM, which were introduced by Hastie and Tibshirani (1986 and 1990). GAM extends GLM by allowing the covariates be related to the dependent variable nonparametrically adopting a semiparametric additive structure:

$$E(Y \mid X) = G\left\{c + \sum_k g_k(x_k)\right\}.$$

Where $g_k(\cdot)$ are nonparametric smooth functions. Another class of models are the Generalized Additive Partial Models (GAPM), which allow the modelling of part of the covariates parametrically and the rest nonparametrically[16].

GAMLSS introduce several improvements on GLM and GAM models. In these previous class of models the mean μ of the dependent variable Y is modelled as a function of explanatory variables, depending the variance of Y on the mean (μ) as well as on a constant dispersion parameter (ϕ),.i e. $V(Y) = \phi \upsilon(\mu)$. Other characteristics of the distribution of Y, as the skewness or the kurtosis, are not modelled explicitly in term of the covariates, but implicitly through their dependence on η. This problem is accounted for in GAMLSS. Besides, the exponential family assumption is relaxed and replaced by a very general distribution family.

In order to choose the model presented in this article, several candidate distributions for the dependent variable y have been considered. Bearing in mind that the distribution of y has two important features, i.e. (i) a high right-

[14] It should be noticed that Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) actually denote G^{-1} as the link function.

[15] Some of the distributions that belong to the exponential family are: Bernoulli, Binomial, Poisson, Negative Binomial, Normal, Gamma and Inverse Gaussian.

[16] For a manual on nonparametric and semiparametric modelling and estimation, see Härdle et al. (2004).

skewness and (ii) a very high percentage of zeros[17], the candidate distributions were: Poisson, Negative Binomial Type I and II, Poisson-Inverse Gaussian, Sichel, Zero-Inflated Poisson and Zero-Adjusted Inverse Gaussian. The criteria followed to select the model that better fits the data were the generalized Akaike Information Criterion (GAIC) and the Schwarz Bayesian Information Criterion (SBC). Both indicators reached its minimum value with the Zero Adjusted Inverse Gaussian (ZAIG) model, so it has been finally selected. Figure 2 shows the fitted Probability Distribution Function (PDF) of the ZAIG distribution for some mean values of its distributional parameters. A comparison of the fitted PDF with the histogram of the dependent variable shown in Figure 1 appears to indicate that this distribution fits reasonably the data.

Let y_i be the number of entries on the municipality i during the period 2002-2004, for $i = 1,...,n$, being n the total number of municipalities. The ZAIG model assumes that the distribution of y can be written as a mixed discrete-continuous probability function, so that:

$$f(y_i) = \left\{ \begin{matrix} 1 - p_i & \Leftrightarrow & y_i = 0 \\ p_i \, g(y_i) & \Leftrightarrow & y_i > 0 \end{matrix} \right\}. \tag{1}$$

Where p_i is the probability of having at least one firm located for a certain municipality i, and $g(y_i)$ is the inverse gaussian density of the positive values of y_i:

$$g(y_i) = \frac{1}{\sqrt{2\pi y_i^3}\sigma_i} \exp\left[-\frac{1}{2y_i} \left(\frac{y_i - \mu_i}{y_i \sigma_i} \right)^2 \right]. \tag{2}$$

This mixed distribution allows to model two different phenomena separately by means of different specifications: (i) the fact that a municipality has new firms located in it, and (ii) the amount of new firms located in a municipality. This is due to the fact that the ZAIG model allows the explicit modelling in terms of explanatory variables of its three distribution parameters: the mean μ, the standard deviation σ and the shape parameter ν, which is equal to $1 - p$. Therefore, this model consists of three different equations to be estimated:

[17] Figure 1 shows the histogram of the dependent variable.

$$\log(\mu_i) = m_1(X_1) \quad (3)$$

$$\log(\sigma_i) = m_2(X_2) \quad (4)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = m_3(X_3)\,. \quad (5)$$

Being $m_r(\cdot)$ a flexible function of the subset of covariates r, for $r = 1,2,3$. This function stands for an additive regression structure, where the regressors may be related to the independent variable either parametrically or through a nonparametric smooth function. Also interaction terms among the regressors are allowed. It is worth noting that in equation (5) covariates are incorporated through the logit link function on p_i. The criterion considered to define these subsets of regressors and the estimation method are explained in the next section.

4. Estimation and results

The econometric software used in the estimations carried out in this section has been R. Specifically, the ZAIG regression model is incorporated into the *gamlss* package in R (Stasinopoulos et al., 2006). The estimation method is the maximum penalized likelihood, and the penalized log likelihood functions have been maximized iteratively using the RS and CG algorithms of Rigby and Stasinopoulos (2005). These algorithms use a backfitting algorithm to perform each step of the Fisher scoring procedure[18].

The first estimated model is shown in Table 1. This model consists of three different estimated equations, one for each distributional parameter (μ, σ and ν). The equations with an economic interpretation are the ν equation, since this equation explains the existence or not of new located firms in a territory, and the μ equation, because its regressors explain the amount of new firms in a territory conditional on the fact that at least one new firm has been located in that territory. For this reason, the whole set of regressors has been included in

[18] For a thorough explanation of the Fisher scoring and the backfitting method, see Hastie and Tibshirani (1990) and Härdle et al. (2004).

both equations as additive linear parametric terms. With respect to the σ equation, its sole aim is to model the variance of the distribution, and only a subset of significant regressors has been included[19]. The results of the estimation show that only few variables are significant in both equations. With respect to the ν equation, the estimated coefficient of employment density (*EMPD*) is negative, while the population density (*POPD*) coefficient is positive. This result seems to indicate that, on average, areas with a high employment density may have reached a level of agglomeration that hinders new firms from locating there. With respect to education, only the variables related to the percentage of workers with a university degree are significant. Thus, the coefficient of *HC2* is positive and the coefficient of *W-HC2* is negative. This result may indicate that the existence of workers with university degrees enhances the probability that at least one firm will be located in that municipality, and the negative result for *W-HC2* indicates that the *HC2* averaged value of the neighbours of a municipality reduces the probability of the occurrence of firm locations. This may be a surprising result, and could indicate the presence of a certain negative spatial autocorrelation of the presence of firms and workers with degrees, i.e. firms that require workers with university degrees tend to be clustered in certain isolated municipalities. The negative sign of *EMP-MAN* indicates that municipalities with a high percentage of manufacturing employment may have reached a level that deters new firms from locating in the municipality. This result could suggest that municipalities with a high percentage of manufacturing activity have surpassed the threshold that delimits economies and diseconomies of agglomeration. The estimation of the μ equation shows that a different subset of regressors enters significantly. The *COAST* parameter is negative, which could imply that shore-line municipalities may have reached a level of congestion that keeps the amount of new located firms from being bigger. The distance to the county capital (*DIS-CC*) has a significant coefficient, which appears to be negative but very small. Finally, the coefficients of the variables *MAB* and *MAM*, which indicate that a municipality belongs to the metropolitan area of the towns of Barcelona and

[19] In order to select the subset of covariates, a stepwise model selection algorithm based on the Generalized Akaike Information Criterion (GAIC) has been used. See Stasinopoulos et al. (2006) for details.

Manresa, are positive, which indicate that locating in these areas implies benefiting from positive externalities, while the result for the metropolitan area of Lleida (MAL) is negative. The results obtained in these two equations (μ and ν) must be taken cautiously and as partial results, because neither nonlinearities nor interaction among regressors have been accounted for.

The second estimated model departs from the first estimated one, but allows some variables to be related to the dependent variable nonparametrically through a smooth function. There are several smoothing methods to estimate these functions, and the method chosen has been the cubic smoothing splines[20]. The variables that have been found significant in the first model have been included parametrically, and for the other variables cubic smoothing splines have been computed. For the equation ν no significant nonlinearities have been found, while for the equation μ the main nonlinearities found are summarized in Figure 3. These graphs show the partial marginal contribution of the regressors *MDI*, *RSI*, *EMP-MAN* and *EMP-SER* to the dependent variable, once the effect of the other variables has been accounted for. The *MDI* variable does not show a positive significant effect on μ, while for the *RSI* variable the confidence bands indicate that there is a positive and significant effect of *RSI* on μ only up to the value *0.4*. The most significant relationship has been found for the variable *EMP-MAN*, where an inverted U-shape relationship has been found. According to this result, a percentage of manufacturing employment smaller than 30% fosters the location of new firms, and once this threshold is surpassed the effect becomes negative. A similar shape is obtained for the percentage of services employment (*EMP-SER*), although the inverted U-shape is not so clear.

[20] Cubic smoothing splines computation is the solution to an optimization problem, since it minimizes the penalized residual sum of squares:

$$\sum_i \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{ f''(t)\}^2 \, dt .$$

Being λ a fixed constant. For a thorough explanation of this method as well as other smoothers, see Hastie and Tibshirani (1990) and Härdle et al. (2004).

The third estimated model intends to shed light on the relationship between the industrial composition of a municipality, proxied by the variables *MDI* and *RSI*, and the threshold that divides economies and diseconomies of agglomeration, measured here as employment density (*EMPD*). We depart from the hypothesis that a different level of industrial specialisation or diversity may determine where diseconomies of agglomeration appear. To achieve this aim, the methodology considered has been the estimation of varying coefficient terms, introduced by Hastie and Tibshirani (1993). A thorough examination of the nonlinear relationship among these variables for both ν and μ equations has yielded significant results just for the latter. Then, the equations to be estimated are:

$$\begin{aligned}\log(\mu_i) = \alpha_1 + g_1(EMPD_i) + g_2(MDI_i) + g_3(RSI_i) + \\ \beta_1(MDI_i)\,EMPD_i + \beta_2(RSI_i)\,EMPD_i + \gamma_1 X_{1i}^{'}\end{aligned} \tag{6}$$

$$\log(\sigma_i) = \alpha_2 + \gamma_2 X_{2i}^{'} \tag{7}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_3 + \gamma_3 X_{3i}^{'} \tag{8}$$

In the first equation, the additive smooth functions $g(\cdot)$ are complemented by a varying coefficient $\beta(\cdot)$. Varying coefficients are functions that relate how the variables *MDI* and *RSI* change the coefficient of *EMPD* for their whole range. In the three equations, the matrix *X* is related to the dependent variable linearly and parametrically, and it includes the variables the estimated parameter of which was found to be significant in the first model. The results of the third model show similar values for these linear parameters with respect to the first model estimation, so results are not shown for being redundant. With respect to the varying coefficient parameters of the equation (6), their estimates are shown in Figure 4. As it can be seen on the left figure, for small values of *MDI*, an increase of the manufacturing diversification index pushes up the limit between economies and diseconomies of agglomeration, so that areas with a high density of employment and a relatively high level of industrial diversity (no more that 2) will still attract new firms. Nonetheless, once the diversity surpasses a threshold, this effect becomes negative and is eventually not significant, yet the graph become more scattered. With respect to the *RSI* varying coefficient, the function is close to being monotonically negative. This fact indicates that, the more specialised a region is, the sooner a certain value of employment density will cause diseconomies of agglomeration. If we analyse the two graphs together, they seem to indicate that economies of agglomeration (considering as agglomeration the employment density) are stronger in moderately diversified environments, and hence being greater the number of firms located in a municipality that determines the threshold between economies and diseconomies of agglomeration.

5. Conclusions

We have analysed the impact of several territorial variables over the industrial location process in Catalonia. A main group of variables has allowed us analysing how agglomeration economies and diseconomies impact over industrial location. We have analysed as well how the industrial structure of municipalities may determine the existence of economies as well as diseconomies of agglomeration. Our contribution allows understanding in a more detailed way how agglomeration economies contribute to new entries, which is of extreme importance if entry promoting policies want to be applied.

Methodologically, we have proposed the consideration of different classes of models such as the Generalized Linear Models (GLM), Generalized Additive Models (GAM) and Generalized Additive Models for Location, Scale and Shape (GAMLSS). These classes of models appear to be suitable for industrial location analyses, both with respect to the modelling of the distribution of the dependent variable (which is a capital issue, considering the overdispersion and excess of zeros that location data often show) as well as regarding the modelling of the functional form of the regressors. Specifically, in our contribution we have specified the regressors as additive terms that enter the equation parametrically, nonparametrically in the form of smooth functions, and also as nonparametric interaction terms among regressors. We have also highlighted the importance of considering several measures of agglomeration, since it is a very complex phenomenon that has multiples dimensions as well as ways of being measured. Besides, it is relevant to include variables accounting for possible spatial effects between municipalities. In that regard, we have included as regressors the spatial lags of agglomeration as well as human capital variables[21].

Our main results show that the variables regarding the industrial structure of municipalities, both those related to the distinction between manufacturing and service activities and those indicating the extent to which the economic structure of a municipality is specialised or diversified, are relevant for the location of new firms. Not only do they directly influence the location decisions on municipalities, but also determine the apparition of economies as well as diseconomies of agglomeration. Specifically, the percentage of manufacturing employment seems to foster new locations up to a point upon which the effect becomes negative. Municipality levels of productive diversity and specialization are also relevant for location. In that regard, the evidence that we have obtained seems to indicate that in moderately diversified economic environments economies of agglomeration work better and keep congestion effects from arising.

These conclusions notwithstanding, more work needs to be done in this area. Our future research in this field should focus in issues like alternative definitions of agglomeration economies and a spatial econometric analysis of how the influence of those agglomeration economies varies across space. It is also important to take into account some specific industry effects that could shape the way in which agglomerations economies behave.

[21] Anselin et al. (2004) review the main advances in spatial econometrics in the measurement of spatial externalities as well as urban growth and agglomeration economies.

References

Anselin, L. (1988): *Spatial Econometrics: Methods and Models,* Dordretcht: Kluewer Academic Publishers.

Anselin, L., Florax, R.J.G.M. and Rey, S.J. (2004): Advances in *Spatial Econometrics,* Springer-Verlag Berlin Heidelberg.

Arauzo, J.M. (2005): "Determinants of Industrial Location. An Application for Catalan Municipalities", *Papers in Regional Science* **84 (1)**: 105-120.

Arauzo, J.M. (2007): "Industrial Location at a Local Level: Some Comments about the Territorial Level of the Analysis", Working Paper # 4. *Documents de treball del departament d'economia, Universitat Rovira i Virgili.*

Arauzo, J.M. and Manjón, M. (2004): "Firm size and geographical aggregation: An empirical appraisal in industrial location", *Small Business Economics* **22**: 299-312.

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

Barrios, S.; Görg, H. and Strobl, E. (2006): "Multinationals' location choice, agglomeration economies, and public incentives", *International Regional Science Review* **29 (1)**: 81-107.

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

Brueckner, J.K. (2000): ''Urban Sprawl: Diagnosis and Remedies'', *International Regional Science Review* **23**: 160–171.

Cameron, A.C. and Trivedi, P.K (1998): *Regression analysis of count data*, Cambridge University Press.

Carlino, G. (1979): “Increasing returns to scale in metropolitan manufacturing”, *Journal of Regional Science* **19 (3)**: 363-373.

Carlino, G. (1978): “Economics of Scale in Manufacturing Location: Theory and Measurement”, *Studies in Applied Regional Science* **12**, The Netherlands: Martinus Nijhoff Social Science Division.

Cieślik, A. (2005): “Location of foreign firms and national border effects: the case of Poland”, *Tijdschrift voor Economische en Sociale Geografie* **96(3):** 287-297.

Combes, P.P. (2000): “Economic structure and local growth: France, 1984-1993”, *Journal of Urban Economics* **47:** 329-355.

Coughlin, C.C. y Segev, E. (2000): “Location determinants of new foreign-owned manufacturing plants”, *Journal of Regional Science* **40:** 323-351.

Duranton, G. and Puga, D. (2004): “Micro-foundations of urban agglomeration economies”. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. IV. North-Holland.

Duranton, G. and Puga, D. (2000): “Diversity and Specialisation in Cities: Why, Where and When Does it Matter?”, *Urban Studies* **37 (3)**: 533-555.

Figueiredo, O., Guimarães, P. and Woodward, D. (2002): “Home-field advantage: location decisions of Portuguese entrepreneurs”, *Journal of Urban Economics* **52**: 341-361.

Glaeser, E.L. (1998): “Are cities dying?”, *Journal of Economic Perspectives* **12**: 139-160.

Guimarães, P., Figueiredo, O. and Woodward, D. (2000): "Agglomeration and the Location of Foreign Direct Investment in Portugal", *Journal of Urban Economics* **47**: 115-135.

Guimarães, P.; Figueiredo, O. y Woodward, D. (2000b): "A Tractable Approach to the Firm Location Decision Problem", Working Paper NIMA 2/2000, Universidade do Minho.

Hastie, T.J. and Tibshirani, R.J. (1986): "Generalized additive models (with discussion)", *Statistical Science* **1(2):** 297-318.

Hastie, T.J. and Tibshirani, R.J. (1990): *Generalized Additive Models,* Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hastie, T.J. and Tibshirani, R.J. (1993): "Varying coefficient models (with discussion)", *Journal of the Royal Statistical Society, B.* **60:** 757-796.

Hayter, R. (1997): *The dynamics of industrial location: The factory, the firm and the production system*, Wiley: New York.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004): *Nonparametric and Semiparametric Models,* Springer-Verlag Berlin Heidelberg.

Henderson, V. (1997): "Medium size cities", *Regional Science and Urban Economics* **27**: 583–612.

Henderson, V.; Kuncoro, A. y Turner, M. (1995): "Industrial development in cities", *Journal of Political Economy* **103 (5):** 1067-1090.

Holl, A. (2004a): "Start-ups and Relocations: Manufacturing Plant Location in Portugal", *Papers in Regional Science* **83 (4)**: 649-668.

Holl, A. (2004b): “Transport Infrastructure, Agglomeration Economies, and Firm Birth. Empirical Evidence from Portugal”, *Journal of Regional Science* **44 (4)**: 693-712.

Holl, A. (2004c): “Manufacturing Location and Impacts of Road Transport Infrastructure: Empirical Evidence from Spain”, *Regional Science and Urban Economics* **34 (3)**: 341-363.

Hoover (1936): “The measurement of industrial location”, *The Review of Economics and Statistics* **18**: 162-171.

Keeble, D. and Walker, S. (1994): “New Firms, Small Firms and Dead Firms: Spatial Patterns and Determinants in the United Kingdom“, *Regional Studies* **28(4)**: 411-427.

Krugman, P. (1998): “What’s New About The New Economic Geography?”, *Oxford Review of Economic Policy* **14(2)**: 7-17.

Le Jeannic, T. (1997): “Trente ans de périurbanisation: extension et dilution des villes”, *Économie et Statistique* **307** : 21-41.

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**: 953-973.

List, J.A. and McHone, W.W. (2000): “Measuring the effects of air quality regulations on “dirty” firm births: Evidence from the neo and mature-regulatory periods”, *Papers in Regional Science* **79**: 177-190.

Marshall, A. (1890): *Principles of Economics*, MacMillan: New York.

McCullagh, P. and Nelder, J.A. (1989): *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability,* 2nd edition, Chapman and Hall, London.

Mieszkowski, P. and Mills, E.S. (1993): “The Causes of Metropolitan Suburbanization”, *Journal of Economic Perspectives* **7 (3)**: 135-147.

Moomaw, R.L. (1988): “Agglomeration Economies: Localization or Urbanization?”, *Urban Studies* **25**: 150-161.

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

Parr, J.B. (2002): “Missing Elements in the Analysis of Agglomeration Economies”, *International Regional Science Review* **25 (2)**: 151-168.

Richardson, H.W. and Baie, C.-H.C. (2004): *Urban Sprawl in Western Europe and the United States*, Ashgate Publishing, Ltd.

Rigby, R.A. and Stasinopoulos, D.M. (2005): “Generalized additive models for location, scale and shape”, *Applied Statistics* **54(3):** 507-554.

Rosenthal, S.S. and Strange, W.C. (2004) Evidence, nature and sources of agglomeration economies in Henderson, J.V. and Thisse, J.F. (eds) *Handbook of Urban and Regional Economics*. North Holland.

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.

Stasinopoulos, D.M., Rigby, R.A. and Akantziliotou, C. (2006): “gamlss: a collection of functions to fit Generalized Additive Models for Location, Scale and Shape”, R package version 1.1-0, url=http://www.londonmet.ac.uk/gamlss/.

Tolley, G. (1974): “The Welfare Economics of City Bigness”, Journal of Urban Economics **1 (3):** 325-345.

Townroe, P.M. (1969): “Industrial structure and regional economic growth. A comment”, *Scottish Journal of Political Economy* **16**: 95-98.

Viladecans, E. (2004): “Agglomeration economies and industrial location: city-level evidence”, *Journal of Economic Geography* **4/5**: 565-582.

Zeng, X.P. (1998): “Measuring Optimal Population Distribution by Agglomeration Economies and Diseconomies: A Case Study of Tokyo”, *Urban Studies* **35 (1)**: 95-112.

Tables

Table 1. Estimation of the μ and ν equations with linear parametric terms. Estimation method: penalized likelihood.

	μ - parameter equation		ν - parameter equation	
Link function	*log*		*logit*	
	Coefficient	Std. Dev.	Coefficient	Std. Dev.
Intercept	-2.2290**	(1.0470)	10.8200***	(2.4600)
EMPD	-0.0002	(0.0003)	-0.0017**	(0.0007)
POPD	0.0000	(0.0001)	0.0007**	(0.0003)
MDI	0.0431	(0.0791)	-0.2985	(0.4018)
RSI	0.9155	(0.5706)	0.1710	(1.6610)
HC1	1.0380	(1.3880)	-3.6910	(2.8730)
HC2	-0.3628	(1.3950)	9.1230***	(3.0130)
HC3	-0.0248	(0.0659)	-0.0550	(0.1283)
W-EMPD	0.0002	(0.0001)	-0.0005	(0.0005)
W-POPD	-0.0001	(8.4e-05)	0.0002	(0.0002)
W-MDI	0.0740	(0.0609)	-0.2892*	(0.1668)
W-RSI	0.1998	(0.3307)	-1.1310	(0.8598)
W-HC1	1.5800	(0.9947)	1.8250	(2.6240)
W-HC2	-1.4360	(1.0090)	-7.1290**	(2.7760)
W-HC3	-0.0384	(0.0344)	0.1051	(0.0862)
EMP	2.4580	(1.7110)	1.1970	(3.1450)
POP	-1.1530	(1.7240)	-4.1630	(3.1970)
EMP-MAN	0.0210	(0.4813)	-3.1110***	(1.1380)
EMP-SER	-0.1524	(0.2215)	0.1693	(0.4405)
COAST	-0.3549**	(0.1517)	0.4363	(0.4881)
DIS-CC	-9.4e-06**	(3.7e-06)	-7.1370	(1.1e-05)
MAB	0.3043***	(0.1124)	0.2689	(0.3144)
MAG	0.0540	(0.1506)	0.4278	(0.3828)
MAL	-0.2945**	(0.1438)	-0.2075	(0.3664)
MAT	-0.1905	(0.1554)	0.2762	(0.3995)
MAM	0.4162**	(0.1746)	0.0374	(0.6347)

Observations	907
Negrees of freedom for the fit	57
Residual degrees of freedom	850
Global Deviance (GD)	2966
Akaike Info. Crit. (AIC)	3080
Schwarz Bay. Info. Crit. (SBC)	3354

Notes:

(a) The equation for the parameter σ has been nod displayed for the sake of clarity. In such equation only significant terms are estimated, which are: EMP, COAST, MAM, MAB.

(b) ***, ** and * indicate significance of the parameter at the 99%, 95% and 90% levels, respectively.

Figures

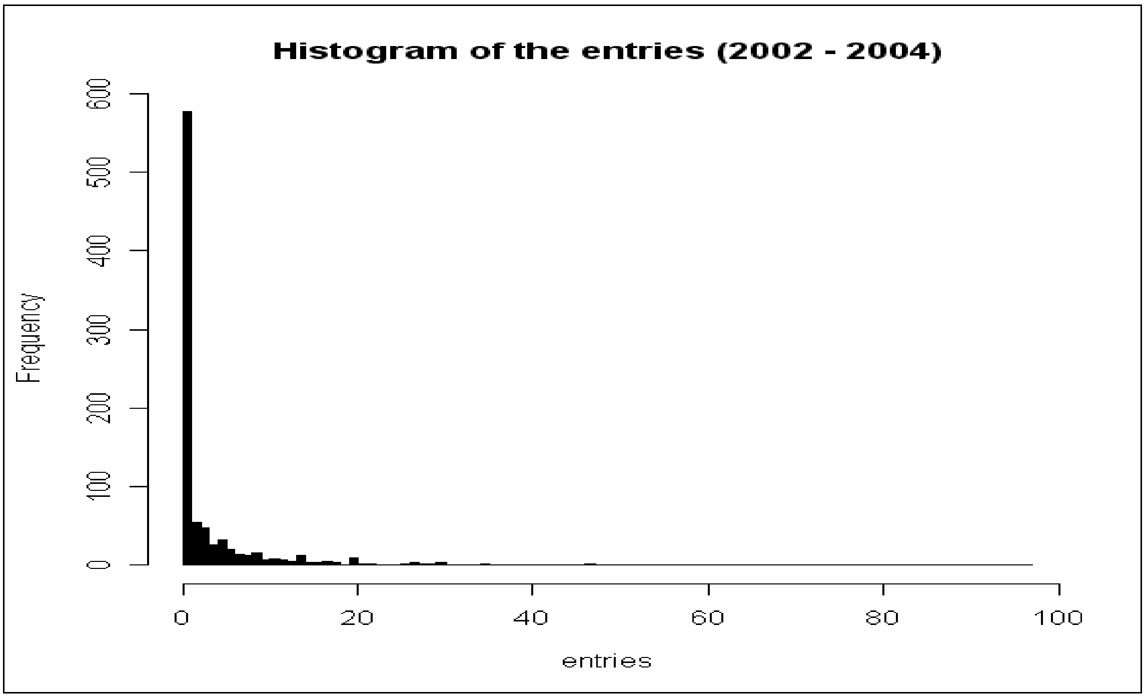

Figure 1. Histogram of the aggregated entries between the years 2002 and 2004. (Note: the values above 100 have been dropped because their contribution to the histogram is marginal.)

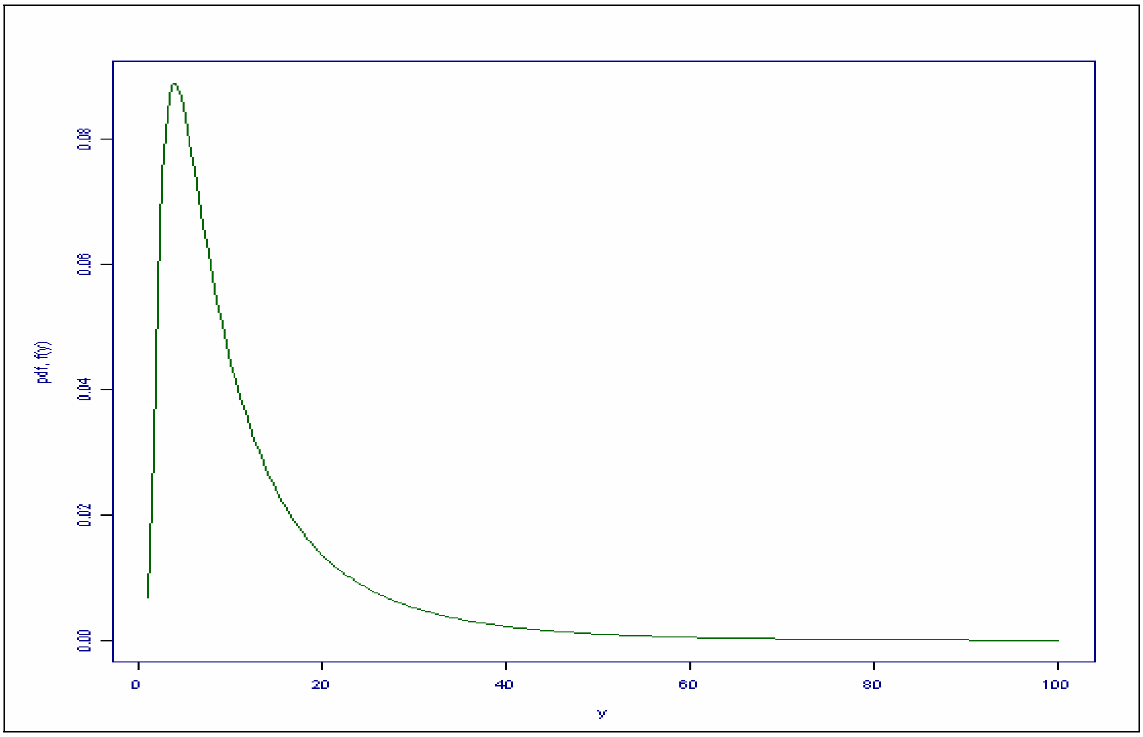

Figure 2. Fitted Probability Distribution Function (PDF) of the ZAIG distribution for the values of the distributional parameters μ =11.71, σ =0.28 and ν =0.003053.

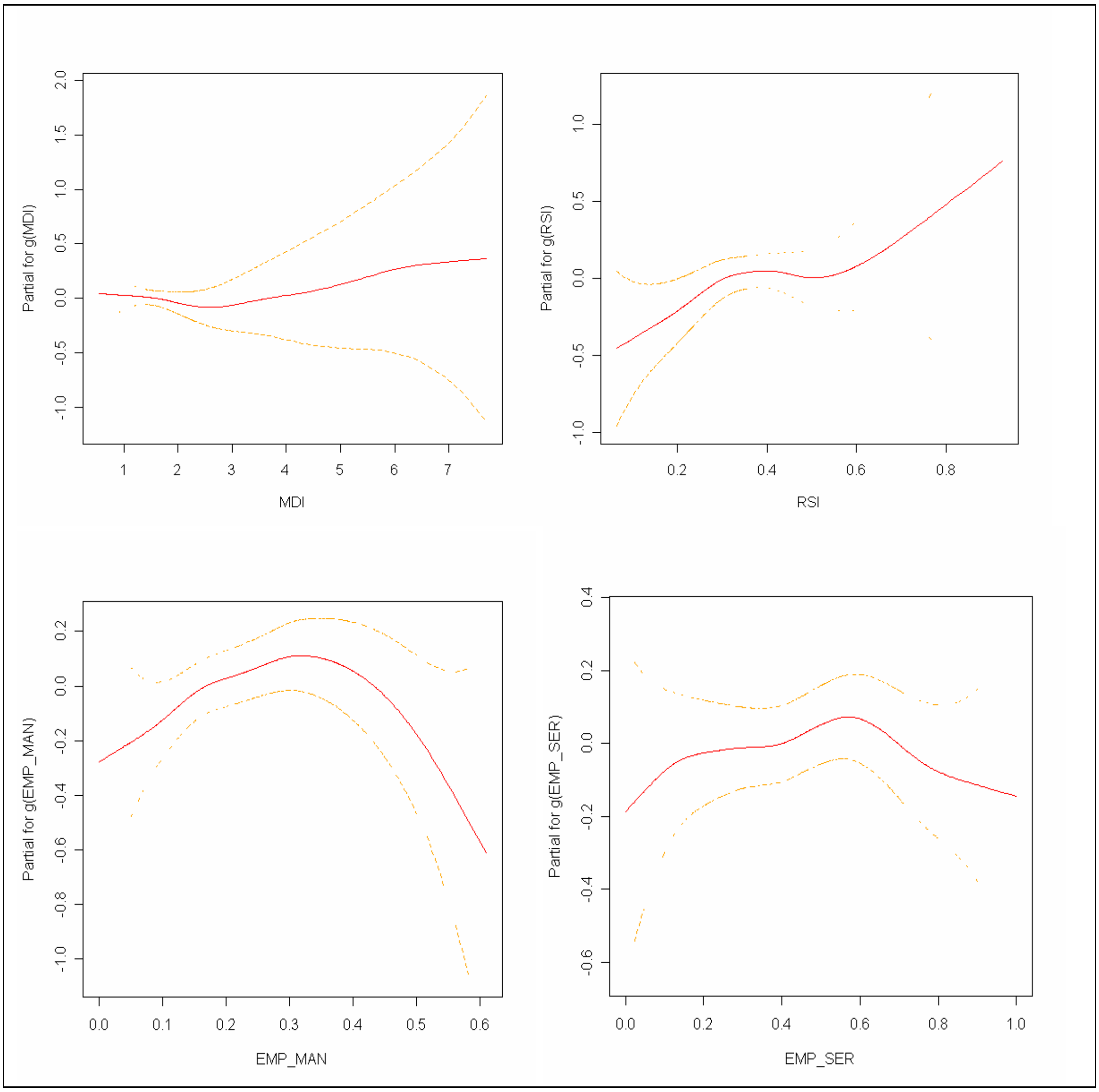

Figure 3. Cubic smoothing spline functions $g(MDI)$, $g(RSI)$, $g(EMP-MAN)$ and $g(EMP-SER)$.

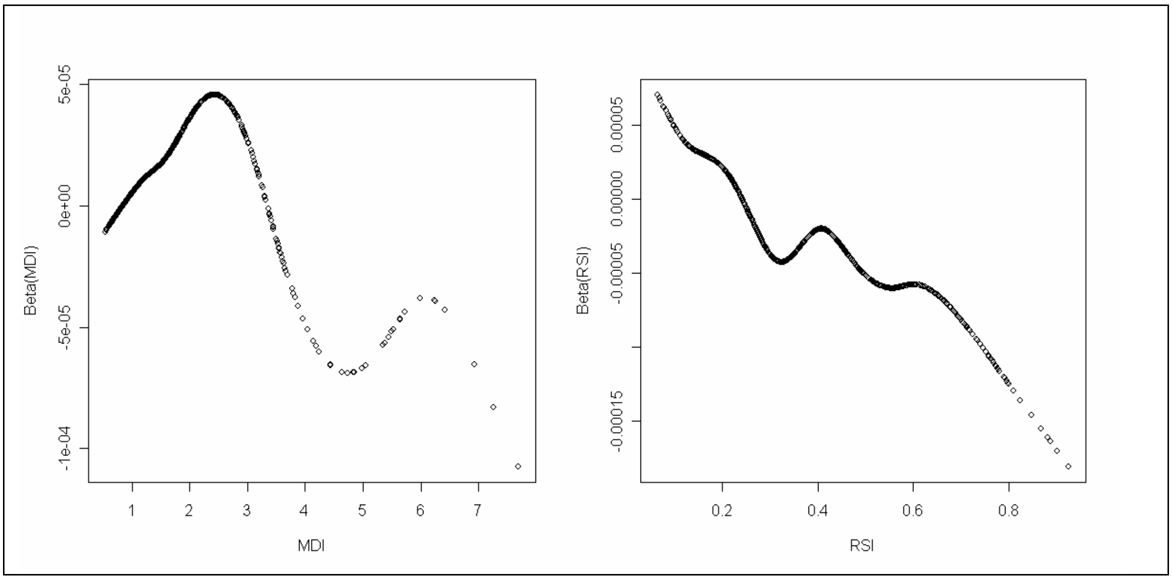

Figure 4. Varying coefficient parameters $\beta(MDI)$ and $\beta(RSI)$ of the variable *EMPD*.