UNIVERSITÀ DEGLI STUDI DI TORINO

ALMA UNIVERSITAS TAURINENSIS

Founded in 1404

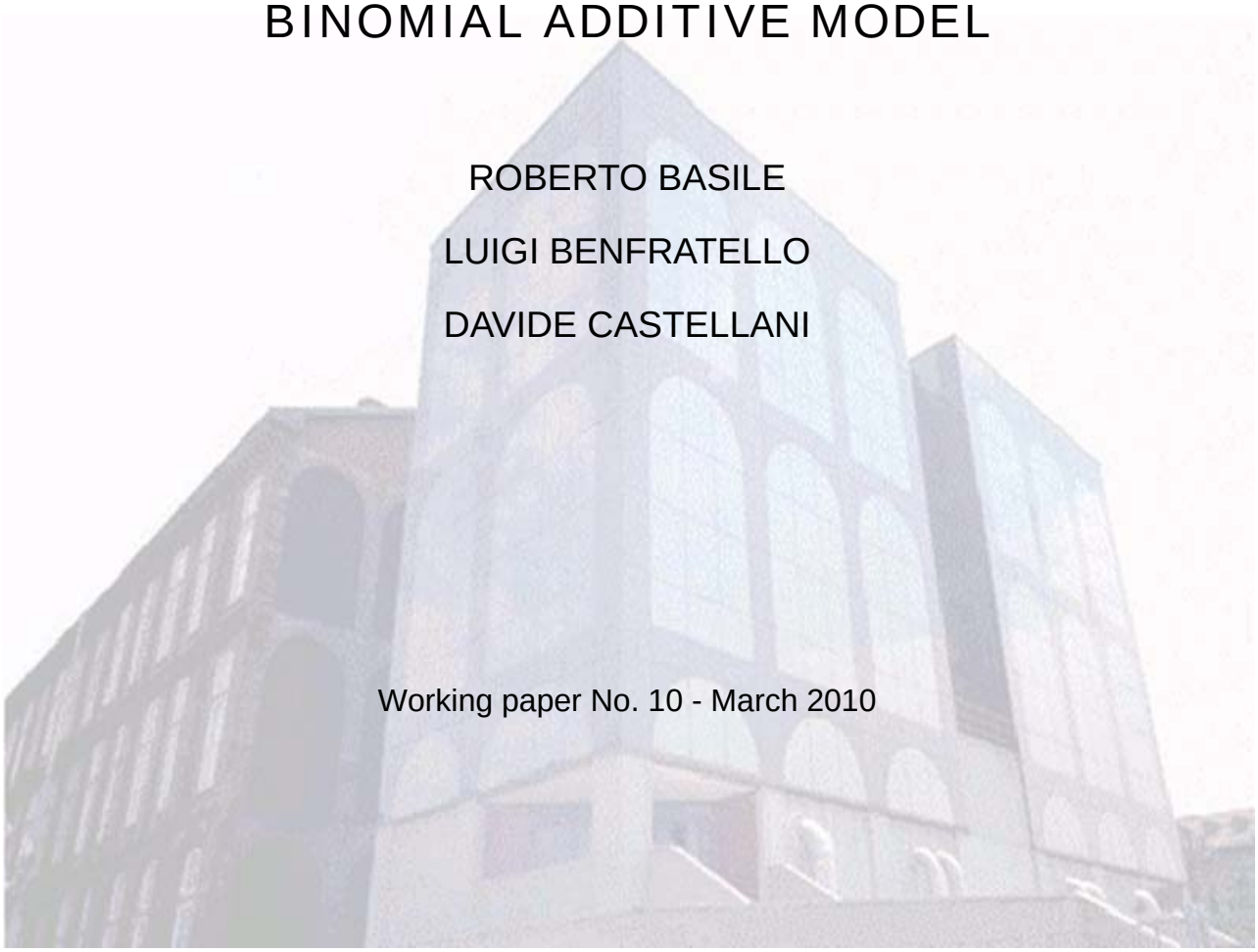
DEPARTMENT OF ECONOMICS AND PUBLIC FINANCE "G. PRATO" WORKING PAPER SERIES

LOCATION DETERMINANTS OF GREENFIELD FOREIGN INVESTMENTS IN THE ENLARGED EUROPE: EVIDENCE FROM A SPATIAL AUTOREGRESSIVE NEGATIVE BINOMIAL ADDITIVE MODEL

ROBERTO BASILE

LUIGI BENFRATELLO

DAVIDE CASTELLANI



Working paper No. 10 - March 2010

Location determinants of greenfield foreign investments in the Enlarged Europe: evidence from a spatial autoregressive negative binomial additive model

Roberto Basile[♣], *Institute for Studies and Economic Analysis and University of Macerata*

Luigi Benfratello[♠], *University of Turin and Ceris-CNR*

Davide Castellani[♦], *University of Perugia and Centro Studi Luca d'Agliano*

March 12, 2010

Abstract

This paper addresses two important methodological issues in the analysis of industrial location: spatial dependence and nonlinearities. To this end, we estimate a semi-parametric spatial autoregressive negative binomial model using data on the number of inward greenfield FDI occurred over the 2003-2007 period in 249 European regions. Results support the view that multinational firms' location choices are very spatially dependent, even controlling for a large number of regional characteristics. A spatial lag model with a non-parametric spatial filter allows us to purge the residuals from spatial dependence and yields sensible changes in the magnitude of some estimated coefficients. We also provide robust evidence of nonlinearities. In particular, we find that the effect of agglomeration economies fades down as the density of economic activities reaches some limit value.

JEL codes: C14 C21 F14 F23

Keywords: Multinational firms, greenfield FDI, count data, spatial econometrics, semiparametric econometrics

♣ Corresponding author: Piazza Indipendenza 4, 00185, Rome. E-mail: r.basile@isae.it
♠ E-mail: benfratello@econ.unito.it
♦ E-mail: davide.castellani@unipg.it

1. **Introduction**

The location of new plants draws considerable attention among economists and policy makers. By attracting new plants regions can foster their economic development, thereby justifying the investments of (local, national and supranational) institutions to attract firms' establishments. This is all the more true when it comes to the location of multinational firms which can bring foreign technology into a local context and eventually generate significant knowledge and pecuniary externalities (Barba Navaretti and Venables, 2005). The alleged positive effect of foreign establishments contributed to motivate a considerable amount of economic research into the determinants of such a location process. A recent survey of empirical works on the topic reports more than 50 econometric studies, mostly from the last decade and in large proportion focussing on foreign-owned firms (Arauzo-Carod et al., 2010). Despite this extensive amount of research, two issues have received very little attention so far: spatial dependence and nonlinearities (spatial parameter instability). The former refers to the correlation of regional attributes over space whereas the latter implies that not all regions obey a common linear specification of the industrial location model. Both issues may lead to biased estimates and unreliable significance tests.

In this paper, we address these two issues in the case of the location of greenfield foreign investments in the NUTS2 regions of the Enlarged Europe over the 2003-2007 period. We submit that both spatial dependence and nonlinearities (spatial parameter instability) may be relevant in this case. On the one hand, it has been widely documented that FDI are not randomly distributed in space, but rather they tend to concentrate in a few clusters of (neighbouring) regions. This clearly violates the assumption of spatial independence. Regional characteristics included as explanatory variables in the model may partially account for the clustering of foreign firms. However, most likely some spatial correlation remains in

the errors, affecting efficiency and consistency of estimates. This may occur because of *(i)* unobserved factors and *(ii)* covariates displaying their effect beyond regional borders.

On the other hand, spatial parameter instability is also very likely to occur in a large sample of very heterogeneous regions, such as the NUTS 2 regions of Europe. For example, theory suggests that the effect of agglomeration economies on firms' location decisions could be different as agglomeration rises. As a matter of fact, as agglomeration reaches some critical value, a congestion effect may eventually kick-in reducing the attractiveness of a given location. Following this reasoning, a dearth of author have postulated an inverted-U shaped relation modelled by using squared terms for agglomeration measures (Arauzo-Carod, 2005; Viladecans-Marsal, 2004). Admittedly, this is only one of several, competing parametric restrictions which may capture a nonlinear relation. Indeed, parameter heterogeneity can be better accommodated in a semi-parametric framework, where nonlinear relations for each regional characteristics can be tested against parametric specifications and the actual shape of the partial effect can be assessed using smooth functions. Within the same framework, one can also specify a spatial lag model which accounts for spatial dependence.

These two methodological issues are not only far from trivial *per se* but, in our case, are complicated by the fact that, in line with the vast literature on plants location, we need to explain the number of new ventures in each region. Therefore, our dependent variable is a count, which takes discrete and non-negative values, so that a Generalised Linear Model (GLM) framework, assuming a negative binomial distribution for the conditional expectation of the number of foreign plants creation in each region, is called for. As will be discussed in greater details in section 4, GLMs lend themselves to an extension into the semi-parametric framework, by adding smooth functions of covariates linearly in the conditional expectation. This class of models is called *Generalized Additive Models* (GAMs).

In sum, we use for the first time a spatial lag specification of a Negative Binomial Generalized Additive Model (SAR NB-GAM) and apply it to estimate the determinants of multinational firms' greenfield investment location in the regions of the Enlarged Europe. To the best of our knowledge only another work uses a semi-parametric approach to estimate the determinants of new plant creation in Spanish provinces (Arauzo-Carod and Liviano, 2007). Our work is different from this one since we thoroughly address spatial dependence alongside with non-linearity, we focus on multinational plants and we take a much broader perspective by studying location in the Enlarged Europe. As for the application of spatial econometric methods to FDI location models, the literature is also scarce (Coughlin and Segev, 2000; Blonigen et al., 2007; Baltagi et al. 2007, 2008) and none of the existing studies considers nonlinearities.

Results support the view that multinational firms' location choices are spatially dependent, even controlling for a large number of regional characteristics, such as employment density, market size, Jacobs externalities, human capital, labour cost, unemployment and density of transport infrastructure and for nonlinear spatial trends. By estimating a spatial lag model and applying a non-parametric spatial filter (by means of a smooth interaction between latitude and longitude), we are able to purge the errors from spatial dependence. We show that this yields a remarkable change in the magnitude of some estimated coefficients, although, in our case, do not alter the statistical significance of the various location determinants. The semi-parametric approach allows us to appreciate that some regional characteristics have indeed a linear effect on FDI counts; but also that some important nonlinearities emerge. In particular, we provide evidence that, in line with theoretical predictions, the effect of agglomeration economies fades down as the density of economic activities reaches some limit value. However, nonlinearity does not seem to be inverted-U shaped: it is always positive but the marginal effect decreases as agglomeration rises and, for a significant portion of our sample,

the relation is flat. Thus, no matter how dense economic activity becomes, our data suggest that congestion (or competition) effects would never overcome positive agglomeration externalities.

The rest of the paper is organized as follows. Section 2 introduces the dataset on foreign greenfield investments in the European regions and reports the results of an exploratory spatial data analysis which provides some important insights for modelling FDI counts. Section 3 presents the theoretical framework which motivates the choice of location determinants included in the empirical model and further justifies the inclusion of a spatial lag term along with nonlinearities. Section 4 introduces the econometric methodology whereas Section 5 reports the econometric results. Section 6 concludes.

1. Spatial distribution of FDI within the Enlarged Europe

The data used in this paper are retrieved from fDI Markets, a recently released database which provides information on almost 60,000 foreign investments projects worldwide. We selected 1,930 greenfield investments in the creation of manufacturing plants carried out by both European and non-European MNFs in the enlarged Europe (including both the "old" Western European Union countries and the new Eastern accession countries) over the 2003-2007 period. For each project detailed information is available on the investor (name and state/country of origin and sector of activity, including both manufacturing and services) and on the destination area (country, state and city). This allowed us to count the number of projects in each NUTS2 region. Five countries (Bulgaria, Latvia, Cyprus, Luxemburg and Malta) and three Spanish regions (Comunitad Autónoma de Ceuta, Comunitad Autónoma de Melilla and Canarias) have been excluded due to data availability problems. Therefore, the overall sample is composed of 249 NUTS2 regions in 22 EU countries.

The distribution of the 1,930 foreign greenfield investments in the manufacturing sector is right skewed (Figure 1, Panel A), with a share of zeros of about 14%, suggesting a high degree of overdispersion in the raw data. It is also characterized by strong positive spatial autocorrelation as suggested by the results of Moran's *I* statistical tests (Figure 1, Panel B).

- INSERT FIGURES 1 AND 2 ABOUT HERE-

For stationary Data Generating Processes (DGP), positive spatial autocorrelation has important interpretations in terms of interregional spillovers and spatial contagion (Anselin, 2004). Indeed, when the DGP is non-stationary, the evidence of spatial dependence may be induced by the presence of spatial trends so that, after removing them, test statistics may reveal the absence of spatial dependence or a random dispersion pattern (Diggle and Ribeiro, 2007). The circle plot in Panel C in Figure 1 suggests that foreign investments are mainly directed towards regions belonging to the New Member States, either to exploit relatively cheap and high skilled labour forces or to serve growing markets. Be that as it may, we checked whether spatial correlation is driven only by DGP non-stationarity in the mean by estimating a model with only a spatial linear trend as a covariate.[1] Panel D in Figure 1 shows the results of Moran's *I* tests applied to the residuals from this model: even after removing the effect of a linear trend, spatial dependence still appears, so that the spatial dependence is not wholly determined by a simple trend.

Spatial trends in the data can be better captured by a nonparametric estimation of the regional FDI counts on the smooth interaction between latitude and longitude[2]. Figure 2 reports the

[1] Precisely, we have estimated a parametric negative binomial model where the dependent variable is the regional FDI count and the only explanatory variables are latitude and longitude. Only the parameter associated with longitude turned out to be positive and significant, confirming the preponderance of response values towards Eastern regions, whereas the parameter associate to the latitude was not significant.

[2] In this case, we used a semiparametric approach, using a thin plate regression spline as smoother (see Wood, 2006a). See Section 4 for further discussion on this method.

estimation results of this model, showing that a complex non-monotonic spatial trend surface characterizes the regional distribution of these data.

In sum, the exploratory spatial data analysis suggests to model FDI counts bearing in mind the issues of overdispersion, spatial dependence and spatial non-stationarity. In particular, the latter must be carefully faced by modelling the mean function though a set of covariates able to capture the spatial trends and by including spatial trends surfaces. In the next section, we discuss some theoretical hypotheses on the location determinants of FDI which suggest a proper set of explanatory variables needed to model the expected mean function.

2. Determinants of inward FDI distribution

The spatial distribution of economic activities, including FDI investments, can be considered as the result of the interaction between centripetal (or agglomeration) and centrifugal (or competition) forces (including higher land prices, higher factor prices–wages *in primis*–or strong competition).[3] Among agglomeration forces, we focus on the role of urbanization externalities, while as for centripetal forces we consider the effect of labour cost along with other labour market characteristics.

2.1. Urbanization economies

Since Hoover (1948), it is common to distinguish between two sources of agglomeration externalities: *a*) economies external to the firm but internal to the sector (the so-called Marshallian externalities) and *b*) economies external both to the firm and to the sector (the so-called urbanization externalities). According to Marshall, industrial firms tend to localize

[3] Agglomeration economies are also known as "Second Nature" advantages in contrast to "First Nature" explanations of spatial concentration (Krugman, 1993), connected to the presence of exogenous factors, such as natural resources, climate and so on. Agglomeration economies are endogenously determined by the actions of firms and workers.

where other firms of the same industry are already established. The well known benefits of this form of externality are three-fold: *i*) access to a more stable labour market, *ii*) availability of intermediate goods, production services and skilled manpower and *iii*) knowledge spillover between adjacent firms. Marshallian externalities are therefore more suitable to explain "small scale" agglomeration phenomena, such as the emergence of Industrial Districts, that is spatial clusters of firms operating in the same (mainly traditional) industry (for example, clothing or footwear). However, in the present context, we do not have detailed information on the specific industry where the new plant operates so that we are only able to count the total number of manufacturing FDI in the regions. Therefore, we are forced to consider only the role of urbanization externalities[4] which we model using four different variables: 1) the size of the regional market, 2) the overall employment density in the manufacturing sector, 3) the degree of sectoral diversification and 4) the level of road infrastructures.

The *size of regional market* (measured by the log of total value added) is intended to capture externalities which have to do with the "home market effect". As first noted by Krugman (1980), under increasing returns to scale, the appeal of a country (or region) as a production site depends crucially on the size of its domestic market. Firms will locate in the country where they can exploit economies of scale to a greater extent and, eventually, export to neighbouring countries[5].

Employment density in manufacturing (measured as total manufacturing employment per square km) represents the scale of urbanization economies. We expect that regions with higher density of economic activity attract more FDI due to agglomeration effects. However,

[4] It is worth mentioning that empirical evidence has provided some support to the view that urbanization economies seem to outweigh industry-specific localisation economies (Guimarães et al. 2000, Arauzo-Carod et al., 2007).

[5] A market size effect may occur both from the consumers' demand and from the demand of other firms, that is input-output linkages between firms, as emphasized by Venables (1996) and Puga (1999).

the occurrence of congestion costs (including higher land prices, higher crime rates, environmental pollution, traffic jams, excess commuting and so on) may compensate the positive effect of agglomeration economies and, thus, determine a threshold effect in the positive impact of employment density. In other words, regions tend to attract FDI if, *ceteris paribus,* agglomeration economies overcome congestion costs. Therefore, a nonlinear effect of employment density on the number of inward FDI is expected. Some empirical studies upfront assume an inverted-U shaped relationship between agglomeration and location, thereby inserting the measure of agglomeration economies squared as additional regressor (Viladecans, 2004; Arauzo-Carod, 2005). This is certainly the easiest way to deal with such a nonlinearity in a parametric framework. Needless to say, this is only one of many possible nonlinear parameterizations. In particular, this specification assumes that at some point congestion costs would be higher than positive agglomeration externalities so that an increase in employment density would discourage new investments.

Sectoral diversification within regions (measured by the median of the sectoral specialization indexes[6]) is meant to capture urbanization externalities deriving from diversity or variety of the regional economy (*Jacobs externalities*). According to Jacobs (1969), a diverse sectoral structure increases the chances of interaction, generation, replication, modification and recombination of ideas and applications across different industries. Moreover, a diverse industrial structure protects a region from volatile demand and offers the possibility to switch between input substitutes.

The extent of *road infrastructures* (kilometres of motorways per squared kilometres) should pick up the component of urbanization economies due to the provision of *public goods*. A

[6] The specialisation index for each sector s and region i is the following employment location quotient: $S_{si} = (E_{si} / \sum_s E_{si}) \Big/ (\sum_i E_{si} / \sum_i \sum_s E_{si})$, where E denotes employment. Alternatively one could use other indicators, such as the inverse Hirshman-Herfindahl, Gini and Theil indicator.

higher level of public goods (in particular infrastructures) is likely to increase firm productivity and to reduce transport costs, lowering the cost of inputs sourced and facilitating the access to markets. The ensuing increase in private returns to investments makes locations with better infrastructure provisions more attractive for both domestic and foreign investments.

2.2. Labour market characteristics

The role of labour market characteristics as a determinant of inward FDI and new plant creation is well established in the literature (Friedman et al., 1992). In this paper we follow previous works by specifying the regional labour market characteristics using three different variables: the average *wages* (measured by the total compensation to labour divided by the number of employees in the region), *labour availability* (approximated by the unemployment rate) and *human capital* endowment (proxied by the share of population aged 24 or more holding a tertiary education degree). The impact of wages and unemployment is not univocal, however. Lower wages may in fact attract firms seeking for lower labour costs (that is firms pursuing *cost reducing strategies*), but high wages may signal highly skilled workers which in turn attract location of higher value added activities. Furthermore, firms may interpret unemployment both as a measure of a large supply of labour, which would attract firms, and as an indicator of a relatively rigid labour market, which would discourage them. In sum, the effects of basic labour market conditions may be characterized by some nonlinearities which should be taken into account when modeling FDI location decisions.

2.3. Spatial dependence

The spatial extent of the many forms of externalities mentioned above is an important issue. Agglomeration and Jacobs externalities can be geographically bounded to the region in which the new technological knowledge is created if these spillovers involve a significant share of

tacit knowledge and their transmission thus depends on distance. This introduces the need for geographical proximity and creates an impetus for firms to concentrate in regions where other firms are located, in order to capture their knowledge spillovers. However, some knowledge might spill over to other regions generating spatial contagion (dependence) effects (that is, inter-regional spillovers or inter-regional externalities). Therefore, the location of new foreign manufacturing investments in a specific region can be influenced not only by the presence of other manufacturing firms (belonging either to the same industry or to a different one) within that region, but also by the presence of other manufacturing firms in adjacent regions.

Spatial dependence occurs also through the effect of market size and infrastructure. In fact, a corollary to Krugman's home market effect is that firms will prefer location which enjoy a large market potential, which would be a function of market size of neighbouring countries and their accessibility. Some recent spatial analyses on FDI (Blonigen et al., 2007, Baltagi et al, 2007, 2008) motivate the presence of spatial effects on the ground of the theories on export platforms and complex FDI proposed in the international trade and international business literature (Dunning, 1993, Ekholm et al., 2007, Neary, 2009, Yeaple, 2003). The same argument can be proposed with regards to the effect of infrastructure. In this perspective, the location in a specific area is affected not only by the quantity and quality of transport infrastructures in that area, but also by the quantity and quality of infrastructures in nearby territories, as well as between regions. Some empirical studies have used measures of market potential, obtained by the weighted sum of the values of the GDP (that is, market size) of all regions in the system, where the weights are either the simple inverse bilateral distance between any two location (e.g. Basile, 2004), or a more articulated weights where actual trading costs are also taken into account (Head and Mayer, 2004). However, as it will be discussed in Section 4, the presence of multiple forms of spatial externalities makes a spatial

lag model–which allows us to take into account the global spatial multiplier effect induced by the various source of interregional externalities–more appropriate.

Finally, spatial dependence may occur (and should be controlled for when specifying a location model) because of unobserved variables. For example, a number of factors related to culture, policy actions (incentives, corporate taxes, and other institutional characteristics), infrastructures (different from road infrastructures) and various forms of amenities can affect the regional FDI attractiveness.[7] Unfortunately, these factors are either unobservable or cannot be properly measured, especially in large samples. In so far as these variables are spatially correlated, the errors will be spatially correlated too. As discussed in LeSage and Pace (2009), a regression model that includes a spatial lag of the dependent variable vector can capture some of these influences and, thus, reduce possible biases. Residual unobserved heterogeneity can be captured through non-parametric spatial trends (Augustin et al. 2009).

3. Modelling regional inward FDI counts

3.1 Overdispersion and zero-inflation

[7] Several cross-regional studies have investigated the role of regional policies in affecting location choices of multinational firms (among others, Wheeler and Mody, 1992; Head et al., 1999; Crozet et al., 2004; Basile, 2004). Other studies have analyzed the effect of national policies and national institutional settings (corporate tax, labour market institutions, bureaucratic efficiency and corruption, legal system and intellectual property right protection, product market regulation and openness to FDI) on regions' performance in attracting foreign investors (Basile et al. 2006; Barrios et al., 2008). Finally, European policy (such as the Structural and Cohesion funds allocated to EU laggard regions) can also be important factors affecting the attractiveness of a location (Basile et al., 2008). Unfortunately, comparable data on institutional variables would not be available for all the regions in our sample, so should we control for these characteristics we would have to reduce our sample size significantly.

Research on foreign firms' location choice usually appeals to discrete-choice models (conditional, nested and mixed logit models) that rely on the Random Utility Maximization (RUM) framework. In this framework, decision probabilities are modelled in a partial equilibrium setting where foreign firms maximize profits subject to uncertainty that derives from unobservable characteristics. In the present study, however, the use of discrete choice models is hindered by the large dimension of the choice set which makes estimation very burdensome. As an alternative, data can be aggregated at the elementary choice level by counting the number of times a given alternative is chosen (in our case, the number of greenfield foreign investments in each region *i*). Therefore, the dependent variable used in the econometric analysis assumes discrete, non-negative integer values (so-called count data). The standard framework for count data is the Poisson regression model. Let y_i, be the dependent variable, X_i a $k\times1$ vector of explanatory variables and β a $k\times1$ vector of regression parameters. The Poisson regression model for these data is defined by the following conditional distribution:

$$P\left(Y_i = y_i \mid X_i'\beta\right) = \frac{\exp\left(-\mu_i\right)\mu_i^{y_i}}{y_i!} \qquad (1)$$

with $\mu_i = E\left(Y_i\right) = Var\left(Y_i\right)$, the so-called equidispersion condition. The Poisson regression is a special case of the Generalized Linear Model (GLM) framework (McCullagh and Nelder, 1989). The canonical link is $\eta_i = g\left(\mu_i\right) = \log\left(\mu_i\right) = X_i'\beta$, resulting in a log-linear relationship between mean and linear predictor.

Fortunately, Guimaraes et al. (2004) have demonstrated that, under mild conditions, the coefficients of a Poisson regression are equivalent to those of a conditional logit model. Therefore, also the Poisson regression model can be thought as derived directly from a RUM process. In practice, however, the classical Poisson regression model for count data is often of

limited use in a regional location analysis since empirical inward FDI counts typically exhibit overdispersion.

A way of dealing with overdispersed count data is to assume a negative binomial distribution for $y_i \mid X_i$ which can arise as a gamma mixture of Poisson distributions.[8] One parameterization of its probability density function is

$$P\left(Y_i = y_i \mid X_i'\beta, \theta\right) = \frac{\Gamma\left(y_i + \theta\right)}{\Gamma\left(\theta\right) \cdot y_i!} \cdot \left(\frac{\mu_i}{\mu_i + \theta}\right)^{y_i} \cdot \left(\frac{\theta}{\mu_i + \theta}\right)^{\theta} \qquad (2)$$

with mean μ and shape parameter θ; $\Gamma(.)$ is the gamma function. The variance function is now $V\left(\mu\right) = \mu + \mu^2\theta^{-1}$. Note that, for large θ, the model approaches the Poisson model.

Recently, a large number of studies on regional inward-FDI counts have used the negative binomial regression model to address the overdispersion issue (Kogut and Chang, 1991; Zhou et al., 2002; Coughlin and Segev, 2000; Barry et al., 2003; De Propis et al., 2005; Arauzo-Carod and Viladecans-Marsal, 2007)[9]. Random effects extensions of negative binomial regression for panel data have also been considered by Blonigen (1997), Basile (2004) and Basile et al. (2006).

Even though negative binomial regression models capture overdispersion quite well, they are not always sufficient for modelling excess zeros. To overcome this problem, Mullahy (1986) and Lambert (1992) have introduced zero-augmented models that incorporate a second model component capturing zero counts. Zero-inflation models (Lambert, 1992) are mixture models that combine a count component and a point mass at zero. Hurdle models (Mullahy, 1986) take a somewhat different approach and combine a left-truncated count component with a right-censored hurdle component. Examples of applications of Zero-inflated Poisson (ZIP)

[8] The issue of overdispersion can be also addressed by estimating a quasi-Poisson model with sandwich covariances.

[9] For a detailed description of these and other works, see Arauzo-Carod et al. (2010).

and Zero-inflated Negative Binomial (ZINB) models to FDI location analyses are in Tadesse and Ryan (2004), Basile (2004), Tomlin (2000) and Iannizzotto and Miller (2002).

3.2 Nonlinearities: Negative Binomial Additive Models

Both the Poisson and the Negative Binomial model used in the recent literature on inward FDI counts assume a log-linear relationship between FDI and its determinants. In our case, this amounts to assume that all regions obey a common linear specification of the location model, disregarding the possibility of nonlinearities reflecting spatial instability in the behaviour of economic agents. In particular, as already mentioned, we cannot disregard possible threshold effects in the impact of agglomeration externalities on regional FDI attractiveness.

Nonlinearities can be addressed in different ways. Firstly, polynomial expansions up to a cubic can be considered within a GLM approach. This would be rather easy to implement, but the risk of introducing multicollinearity is very high. Secondly, Geographically Weighted Regression (GWR) models represent a standard method for properly handle spatial instability problems. However, as far as we know, there are no extensions of GWR models for overdispersed data. A third solution, the one considered in this paper, is the Negative Binomial Additive Model (NB-GAM), recently introduced by Thurston et al. (2000). NB-GAMs are a special extension of Generalized Additive Models (GAM) to handle Negative Binomial responses.

The GAM framework (Hastie and Tibshirani, 1990) extends the GLM by allowing nonlinearity in the relationship between η_i and the covariates:

$$\eta_i = g(\mu_i) = X_i^{'*}\beta^* + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots + \varepsilon_i \qquad (3)$$

where $\mu_i = E(Y_i)$, $Y_i \sim NegativeBinomial$, $f_j(.)$ are unknown smooth functions of the covariates, X_i^* is a vector of strictly parametric components and β^* is corresponding parameter vector.

Each smooth term in (3) can be represented as $f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j)$, where the $b_{jk}(x_j)$ are known basis functions, while the β_{jk} are unknown parameters to be estimated. One or more measures of '*wiggliness*' $\beta_j' \mathbf{S}_j \beta_j$, where $\mathbf{S}_j$ is a matrix of known coefficients, is associated with each smooth function. Typically, the wiggliness measure evaluates something like the univariate spline penalty $\int f_j''(x_j)^2 dx$ or its thin-plate spline generalization, but it may also be more complex, such as tensor product smooth penalty with multiple $\beta_j' \mathbf{S}_j \beta_j$ terms.[10]

Given bases for each smooth term, model (3) can be re-written as a GLM, $g\{E(y_i)\} = X_i' \beta$ where X includes the columns X^* and columns representing the basis functions evaluated at the covariate values, while β contains β^* and all the smooth coefficient vectors β_j. For fixed smoothing terms λ_j, GAMs are estimated by minimizing the penalized deviance (which corresponds to maximise the penalised likelihood):

$$D(\beta) + \sum_j \lambda_j \beta' \mathbf{S}_j \beta \qquad (4)$$

where $D(\beta) = 2\{l_{\max} - l(\beta)\}\phi$, with l the log-likelihood and $l_{\max}$ the maximum possible value for l given the observed data, which is obtained by considering the MLE of a model with one parameter per datum.

[10] The penalization is necessary to avoid overfitting of the model which would increase the variance of the estimate. To provide the intuition of the relationship between $\int f_j''(x_j)^2 dx$ and $\beta_j' \mathbf{S}_j \beta_j$, consider that once the basis are known, the second derivatives of the function *f* will be a known function of the β_j parameter vector. The quadratic term in the integral explains the quadratic form in β_j.

The most popular approach for estimating the β vector in GAMs is the back-fitting algorithm (Hastie and Tibshirani, 1990). This method, however, presents some shortcomings with respect to model selection and inference issues. Therefore, the Penalized Iteratively Reweighted Least Squares (PIRLS) is a better alternative and a grid search provides estimates for λ_j based on Generalised Cross Validation (GCV) (Wood, 2006a).[11] Wood has implemented all these techniques in the R package *mgcv*.

3.3 Spatial autocorrelation: a control function approach

If the errors of a statistical model are spatially autocorrelated, one of the key assumptions of standard statistical analyses, that errors are independent and identically distributed (*iid*), is violated. The violation may bias parameter estimates and can increase type I error rates (falsely rejecting the null hypothesis of no effect). Spatial autocorrelation in the residuals may occur because of the existence of spatial dependence either in un-modelled effects (when excluded variables that are subsumed in the error term jointly follow a spatial random process) and/or in modelled effects (when the *X* terms affect the left hand side of the model through a "*global multiplier effect*", i.e. both x_i as well as a set of x_j throughout the spatial systems affect y_i) (Anselin, 2004). As discussed in Section 3.3, an example would be when inward FDIs are set ~~in~~ as a function not only of the local market, but also of the market size of the neighbours, and their neighbours' neighbours, and so on (the so-called "*market potential effect*"). Similar examples can refer to the effect of transport infrastructure and of agglomeration externalities.

[11] Wood (2008) has recently proposed an alternative method for automatic and integrated smoothing parameters selection for GAMs termed "outer iteration". Another estimation alternative is to resort to Restricted Maximum Likelihood, due to the possibility of rewriting the Penalised GAMs as a Generalised Additive Mixed Model (see Wood 2006b and Ruppert et al 2003). We checked that our results are not affected by estimating the model with these alternative techniques (results are available upon request to the authors).

Dealing with spatial externalities within a nonparametric framework is a challenging task and at the research frontier in spatial econometrics. In a parametric linear setting, such as $y = X\beta + \varepsilon$, global multiplier effects are modelled by replacing X by $\left(I - \rho W\right)^{-1} X$ and ε with $\left(I - \rho W\right)^{-1} \varepsilon$, where I is an identity matrix, ρ is the parameter of spatial externality and W is a spatial weights matrix. In the present context, the inverse spatial transformation of X and ε suggests that the attractiveness of region i is affected not only by its own characteristics and random shocks, but also by the features and random shocks of all other regions. Thus, every location is correlated with every other location in the system. However, given the characteristics of the standardized spatial weights matrix, the strength of spatial dependence between observed regions declines with the distance between them. In other words, neighbouring units exhibit a higher degree of spatial dependence than units located far apart ("spatial diffusion with friction"). The introduction of the spatial multiplier effect in the model yields a reduced form as $y = \left(I - \rho W\right)^{-1} X\beta + \left(I - \rho W\right)^{-1} \varepsilon$ and the structural form becomes the standard Spatial Autoregressive model (SAR) $y = \rho W y + X\beta + \varepsilon$.

These arguments can be extended to the semiparametric GAMs, with the obvious difference that the effect of spatial externalities may not be homogenous over space. So, the NB-GAM framework described above can be extended by including the linear term $Wy = \sum_j w_{ij} y_j$ on the right hand side (SAR-NB-GAM):

$$\begin{aligned} \eta_i = g\left(\mu_i\right) = X_i^{'*}\beta^* + f_1\left(x_{1i}\right) + f_2\left(x_{2i}\right) + f_3\left(x_{3i}, x_{4i}\right) + \\ f_4\left(Lat_i, Long_i\right) + \rho\sum_j w_{ij} y_j + \ldots + \varepsilon_i \end{aligned} \qquad (4)$$

Model (4) includes also a smooth term $f_4(Lat, Long)$, where *Lat* and *Long* are the spatial coordinates of the region's centroid. This term helps controlling for the non-stationarity, i.e.

for the presence of spatial trends in over- or under-predictions of the non-spatial regression model.[12]

Because of the feedbacks between y and its spatial lag term, Wy enters endogenously in equation (4), that is Wy and ε are correlated.[13] For the case of linear spatial autocorrelation regression models, Kelejian and Prucha (1998) have proposed a 2SLS procedure with spatial lags of the strictly exogenous variables as instruments. Wy is first regressed on a set of exogenous and predetermined variables and—in the second stage—the fitted values from the first stage are used in place of the endogenous variable. The motivation for this form of 2SLS is the replacement of the endogenous regressor with that part of Wy (its linear projection on the set of spatial lags of the exogenous variables) that is uncorrelated with the error term. As emphasized by Blundell and Powell (2003), however, this procedure is not suitable for the estimation of nonparametric and semiparametric models. In particular, the replacement of the endogenous term with fitted values of the first stage generally yields inconsistent estimates of Wy. Therefore, Blundell and Powell (2003) have proposed a general solution which is appropriate for the estimation of nonparametric models. This method consists of extending the "control function" method to additive nonparametric models.

The control function approach applied to the linear model $y = X\beta + \varepsilon$ has its antecedent in the interpretation of the 2SLS estimator β_{2SLS} as the coefficients on X in a OLS regression of y

[12] While rarely considered for modelling economic data, spatial and spatio-temporal trends are widely included in biological models using generalized additive models (see, for example, Augustin et al. 2009).

[13] As clearly pointed out by Anselin (2004, p.6), the interpretation of the role of Wy can generate some confusion: "While it may be intuitive to interpret such a variable as relating values for y at i to its neighbors, this is only partially the case, since the neighboring values in turn depend on y_i. More precisely, the particular spatial pattern between locations and their neighbors can be considered to be the equilibrium outcome of a process that follows from global spatial correlation in the X and error terms. Hence, any economic interpretation of y_i depending on y_j actually works through the spatial patterns in the X and u".

on X and the residuals v from a linear regression of X on a set of instruments Z.[14] Application of the control function approach to nonparametric and semiparametric settings is straightforward. It consists of two steps. In the first one, an auxiliary nonparametric regression of the form $Wy= g(Z)+v$ is considered, with Z being a set of appropriate instruments (also including the exogenous subset of the original covariates X) and v a sequence of random variables satisfying $E(v|Z)=0$. Moreover, if Z and ε are independent, then it yields that $E(\varepsilon|v,Z)=E(\varepsilon|v)$. It follows from the last assumption that $E(\varepsilon|Wy)\neq 0$ arises when $E(\varepsilon|v)\neq 0$. The second step consists of estimating an additive model of the form

$$\eta_i = g(\mu_i) = X_i^{*'}\beta^* + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + f_4(Lat_i, Long_i) + \rho\sum_j w_{ij}y_j + f_5(\hat{v}_i) + \ldots + \varepsilon_i \quad (5)$$

4. Evidence from parametric and semi-parametric regressions

4.1 Results of parametric Poisson and negative binomial regression models

The coefficients estimated with parametric Poisson and negative binomial regression models are reported in Table 1, along with standard errors (in parenthesis) and p-values (in square brackets). The dependent variable is the regional number of greenfield investment projects from foreign multinationals directed to each region over the 2003-2007 period. As discussed in Section 3, the explanatory variables are *Mkt* (the market size, approximated by the regional total value added), *Infra* (a measure of transport infrastructure), *Jacobs* (a proxy for Jacobs externalities), *Empdens* (the employment density in manufacturing), *Wage* (the average labour cost), *Ur* (the regional unemployment rate) and *Ter* (the level of tertiary education).[15] All coefficients turn out to be statistically significant and with the expected sign both in the Poisson and in the negative binomial regression models. Since all variables are in logarithms,

[14] The Z vector includes first order spatial lags of all exogenous variables.
[15] See the appendix for a thorough definition of the variables.

coefficients can be interpreted as elasticities. First, the positive coefficient of market size confirms that foreign firms concentrate where demand is highest and possibly serve smaller markets via exporting. Second, the expected number of foreign greenfields increases with the density of transport infrastructures. Third, Jacobs externalities have a strong positive effect, indicating that a more diversified regional economy is conductive of new foreign firms. Fourth, a higher employment density increases the expected number of greenfield investments into the region. Therefore, on average, it seems that congestion costs are more than counteracted by agglomeration externalities. Finally, high wages seem to discourage FDI, while high regional unemployment and tertiary education attract foreign investors.

-INSERT TABLE 1 ABOUT HERE-

-INSERT FIGURES 3 AND 4 ABOUT HERE-

Table 1 also reports a series of diagnostics tests and measures of goodness of fit. The value of the AIC clearly works in favour of the Negative Binomial model. The Negbin model seems to perform rather well both against the Poisson model, in that is able to account for overdispersion, and against a Zero-Inflated Negbin model (ZINB) and a Hurdle Model, as the null hypothesis of no excess of zeros with respect to the prediction of the NegBin distribution, cannot be rejected.[16] Despite this favourable diagnostics, Panel B of Figures 3 and 4 show that substantial spatial correlation remains in the residual of both the Poisson and the Negbin suggesting the need to explicitly model spatial dependence

[16] As the standard ZINB is not-nested in the NB model, a Vuong test is applied. This test calculates the logarithm of the ratio of the conditional probability of the dependent variable, conditional on the independent variables, for two alternative distribution hypotheses. In our case, the Vuong test statistic proves to be non-significant, so that the existence of zero-inflation can be excluded. The Wald test for the NB against an Hurdle model also suggests that the Negative Binomial is–from this point of view–correctly specified.

4.2 Results of semiparametric additive models

Table 2 reports the estimation results of three nested semiparametric Negbin additive models. The first specification is a simple semiparametric model where we do not introduce any control for spatial dependence, Model 2 is the spatial lag NB-GAM whereas Model 3 introduces the smooth interaction between latitude and longitude to capture nonlinear spatial trends. The value of the AIC ranges from 1,309 (Model 1) to 1,189 (Model 3) and is lower than the one estimated for parametric Poisson and Negbin models, clearly suggesting that the semi-parametric models are to be preferred, from the statistical point of view, to the fully parametric models. Among the semi-parametric models, the specification with spatial lag and spatial trend (Model 3) encompasses all the others. Diagnostics on the residuals, reported in Figure 5, reveal the lack of overdispersion in all three semi-parametric models, while the Moran *I* tests on the residuals show that spatial autocorrelation cannot be rejected in Model 1 and 2, whereas introducing the spatial trend surface (Model 3) we are able to purge the residuals from spatial dependence.

-INSERT TABLE 2 ABOUT HERE-

-INSERT FIGURE 5 ABOUT HERE-

As for the specification of the model, since we are rather agnostic about which variables should enter non-linearly the model we first let the data speak. We run a fully nonparametric model and test for which variable a parametric specification could not be rejected[17]. This test is based on the Effective Degrees of Freedom (*edf*, henceforth) estimated for each smooth function. If the *edf* are equal to 1, a linear (i.e. parametric) relationship cannot be rejected. A preliminary analysis suggests that *edf* are equal to 1 for *Mkt*, *Jacobs, Infra, ur* and *ter*. In order to reduce the computational burden of subsequent estimations, we enter them in

[17] Results of this model are available upon request.

parametric form so that Table 2 reports the estimated parameters, standard error and p-values for these variables. It is worth mentioning that all the parametric terms keep their sign and remain statistically significant. However, their size is in some cases quite different from those of the parametric model. As the presence of the spatial lag modifies the interpretation of the parameters, we compare the coefficients of the Negbin parametric model and the first column of Table 2: the elasticity of unemployment rate (human capital) almost decrease (increase) to one half, from 0.52 to 0.27 (from 0.452 to 0.738) and the coefficients of several other variables show a less dramatic but still noteworthy change. For *wage* and *empdens* non-linearity cannot not be rejected. For these variables, Table 2 reports χ^2-tests and p-value for the overall significance of the smooth terms, as well as the number of *edf*. Low values of the χ^2-test entail a high probability that the estimated smooth term[18] is not different from zero, while *edf* is a measure of its nonlinearity. Intuitively, the larger the *edf,* the larger the number of points necessary to fit a smooth function, hence the higher the degree of nonlinearity. Results from all three models support that both *wage* and *empdens* are statistically significant determinants in the location of new foreign-owned plants in European regions, and that their relationship with the number of new investments is non linear.

Figure 6 shows the smoothed partial effect of *wage* (right-panel) and *empdens* (left-panel) on the expected number of FDI. The shaded areas highlight the 95% confidence intervals. The wage-plot suggests that regions with low average labour costs tend to attract more FDIs, after controlling for the other variables. However, the effect of a wage drop appears higher for intermediate wages levels and decreases in regions with either very low or very high wages. This latter results is is consistent with the idea that in high-end regions the wage rate does not

[18] Nonparametric terms are estimated using tensor product smoothing splines and applying the method described in Wood (2006a) that allows integrated smoothing parameter selection.

only captures labour cost, but also proxies for its quality. Thus, an increase in wages may not discourage multinationals after all.

As far as *empdens* is concerned, Figure 6 shows that the expected number of inward FDIs increases with the employment density in manufacturing, up to a point where the relation becomes basically flat and not significantly different from zero. This is consistent with the hypothesis that a larger presence of industrial activity can exert a positive externality which favours the location choice of foreign firms but, when the level of agglomeration becomes too high, congestion costs kick-in and gradually reduce the magnitude of the positive externality, up to the point where an increase in employment density as no further effect on foreign entries. It should be noted that, in our sample, the relationship between employment density and location of foreign plants is non-linear but does not appear to be inverted-U shaped, as most studies using parametric specifications had anticipated. In fact, an inverted-U relation would predict that investments would eventually decline for very high values of *empdens*, whereas our smoothing functions do not show such a declining pattern.

-INSERT FIGURE 6 ABOUT HERE-

Furthermore, it is also important to observe that the positive and significant sign estimated to the parameter associated with the endogenous term Wy[19] entails that the attractiveness of a region to foreign investors is influenced not only by its own characteristics (market size, infrastructure and so on), but also by the characteristics of all other regions through a "spatial

[19] For the computation of the term Wy, the spatial weights matrix, $W = \{w_{ij}\}_{i,j=1,\ldots,N}$, is specified so that w_{ii} are set to zero whereas $w_{ij} = d_{ij}^{-2}$ if $d_{ij} < \overline{d}$ and $w_{ij} = 0$ if $d_{ij} > \overline{d}$, with d_{ij} the great circle distance between the centroids of region i and region j and $\overline{d}$ the cut-off distance (equal to 423 km). To control for endogeneity bias, the spatial lags of the exogenous variables have been used as instruments, i.e. as additional exogenous regressors in the first step of the control function. First step results are available upon request.

multiplier effect" which decreases with distance. As emphasized above, in the case of market size, this has a direct interpretation in terms of market potential.

5. Conclusions

This paper contributes to the extensive literature on the determinants of industrial location by addressing two largely unexplored issues: spatial dependence and nonlinearities. Using data on greenfield projects in the NUTS2 European regions, we have estimated a semi-parametric count data model. Results have shown significant spatial dependence, even controlling for a large number of regional characteristics, such as urbanization externalities and labour market characteristics and for nonlinear spatial trends. By estimating a spatial lag model, we have been able to purge the residuals from spatial dependence and shown that this yield a significant change in the magnitude of the estimated coefficients. The semi-parametric approach allowed to identify some important nonlinearities. In particular, we provided evidence that, in line with theoretical predictions, the effect of agglomeration economies fades down as the density of economic activities reaches some limit value. Furthermore, the elasticities of some variables entering parametrically our semi-parametric spatial model remarkably change their magnitude. Overall, our result do support the use of more flexible and general models than those traditionally employed in the analysis of FDI location.

Appendix: definition of explanatory variables

- ***Market size***: log of total value added in the region (source: Cambridge Econometrics).
- ***Jacobs externalities***: median specialisation index for each sector *i* and region *j*. Each sectoral index is calculated as the following employment location quotient: $S_{ij} = (E_{ij} / \sum_i E_{ij}) \Big/ (\sum_j E_{ij} / \sum_j \sum_i E_{ij})$, where E denotes employment (source: Cambridge Econometrics)

- ***Employment density:*** number of people employed in the manufacturing industry per km^2.
- ***Public infrastructure:*** kilometres of highways and other roads divided by total population in the region.
- ***Tertiary education:*** share of adults (population aged 25-64) with tertiary education (ISCE97 codes 5 and 6) averaged over the 1999-2002 period (source: Eurostat). For the regions DE41 and DE42 data on tertiary education were available only for the years 2004 and 2005.
- ***Labour cost***. Source: Cambridge Econometrics. For German and UK NUTS2 regions, for which data are available only at the NUTS1 level, we have attributed the value of the NUTS1 they belong to.
- ***Unemployment rate***. Source: Cambridge Econometrics.

Since we aim at estimating the effect of all these variables over the 2003-2007 period, we used—if possible—all explanatory variables averaged over the 2000-2002.

References

Anselin L. (2004) Spatial Externalities, Spatial Multipliers and Spatial Econometrics. *International Regional Science Review*, 26, 153-166.

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 Liviano D. (2007) Agglomeration and Location: a Nonparametric Approach. Working Paper 5-2007, Universitat Roviri I Virgili.

Arauzo-Carod J.M., Liviano-Solis D. and Manjon-Antolin M. (2010) Empirical Studies in Industrial Location: An Assessment of their Methods and Results. *Journal of Regional Science*, forthcoming.

Arauzo-Carod J.M., Viladecans-Marsal E. (2007) Industrial location at the intra-metropolitan level: the role of agglomeration economies. *IEB Document de Treball* n. 5.

Augustin N., Musio M., von Wilpert K., Kublin E., Wood S. and Schumacher M. (2009) Modelling spatiotemporal forest health monitoring data. *Journal of the American Statistical Association*, 104 (487), 899-911

Baltagi, B. H., Egger, P. and Pfaffermayr, M. (2007) Estimating models of complex FDI: Are there third-country effects? *Journal of Econometrics*, 140(1), 260-281.

Baltagi, B. H., Egger, P. and Pfaffermayr, M. (2008) Estimating regional trade agreement effects on FDI in an interdependent world. *Journal of Econometrics*, 145(1-2), 194-208.

Barba Navaretti G. and Venables A. (2005) *Multinational Firms in the World Economy*, Princeton University Press.

Barrios S., Huizinga H., Laeven L., and Nicodème N. (2008) Tax and multinational firm location decisions. CEPR DP 7047

Barry F., Görg H. and Strobl E. (2003) Foreign Direct Investment, Agglomerations and Demonstration Effects: An Empirical Investigation. *Review of World Economics / Weltwirtschaftliches Archiv*, 139, 583-600.

Basile R. (2004) Acquisition Versus Greenfield Investment: the Location of Foreign Manufacturers in Italy. *Regional Science and Urban Economics*, 34, 3-25.

Basile R., Benfratello, L. and Castellani D. (2006) Attracting Foreign Direct Investments in Europe: Are Italian Regions Doomed? In Malagarini M. and Piga G. (eds.), *Capital Accumulation, Productivity and Growth. Monitoring Italy 2005*, Palgrave Macmillan, 319-354.

Basile R., D. Castellani, and Zanfei A. (2008) Location Choices of Multinational Firms in Europe: the Role of National Boundaries and EU Policy. *Journal of International Economics*, 74, 328-340.

Blonigen B. (1997) Firm-Specific Assets and the Link between Exchange Rates and Foreign Direct Investment. *The American Economic Review*, 87, 447-465.

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

Blundell R. and Powell J. (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In Dewatripont M., L. Hansen, and Turnsovsky S.J. (eds), *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge University Press.

Coughlin C. and Segev E. (2000) Location determinants of New foreign-owned manufacturing plants. *Journal of Regional Science*, 40, 323-351.

Crozet M., Mayer T. and Mucchielli J. (2004) How do firms agglomerate? A study of FDI in France. *Regional Science and Urban Economics*, 34, 27– 54

De Propris L., Driffield N. and Menghinello S. (2005) Local Industrial Systems and the Location of FDI in Italy. *Int. J. of the Economics of Business*, 12, 105--121.

Diggle P.J. and Ribeiro P.J. (2007) *Model-based Geostatistics*. New York : Springer

Dunning J.H. (1993) *Multinational Enterprises and the Global Economy*, Wokingham: Addison Wesley

Ekholm K., Forslid R. and Markusen J.R. (2007) Export-Platform Foreign Direct Investment. *Journal of the European Economic Association*, 5(4), 776-795.

Friedman, J., Gerlowski, D. and Silberman, J. (1992) What attract foreign multinational corporations? Evidence from branch plant location in the United States. *Journal of Regional Science* 32, 403-418.

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

Guimaraes P., Figueiredo O. and Woodward D. (2004) Industrial location modeling: extending the random utility framework. *Journal of Regional Science*, 44, 1--20.

Hastie T. J. and Tibshirani R. J. (1990) *Generalized Additive Models*, London, Chapman and Hall.

Head K., Ries J. and Swenson D. (1999) Attracting foreign manufacturing: Investment promotion and agglomeration. *Regional Science and Urban Economics*, 29 (1999) 197–218

Head K. and Mayer T. (2004) Market Potential and the Location of Japanese Investment in the European Union. *The Review of Economics and Statistics*, 86, 959-972

Hoover E.M. (1948) *The Location of Economic Activity*. New York: McGraw Hill.

Iannizzotto M. and Miller N.J. (2002) The effect of exchange rate uncertainty on foreign direct investment in the United Kingdom, Working Paper, *International Economic Association World Congress*, Paris.

Jacobs, J. (1969) *The Economy of Cities*, Random House.

Kelejian H.H. and Prucha I.R. (1998) A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *Journal of Real Estate Finance and Economics*, 17, 99-121.

Kogut B. and Chang S.J. (1991) Technological capabilities and Japanese foreign direct investments in the United States. *The Review of Economics and Statistics*, 73, 401-413.

Krugman P. (1980) Scale Economies, Product Differentiation, and Patterns of Trade. *American Economic Review* 70, 950-959**.**

Krugman P. (1993) First nature, second nature and metropolitan location. *Journal of Regional Science*, 33(2), 129-144.

Lambert D. (1992) Zero-inflated Poisson regression, with an application to detects in manufacturing. *Technometrics*, 34, 1-14.

LeSage J. and Pace R.K. (2009) *Introduction to Spatial Econometrics*. Taylor &Francis Group, LLC.

McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models*, 2nd edition, London: Chapman and Hall.

Mullahy J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341-365.

Neary P. (2009) Trade costs and foreign direct investment. *International Review of Economics and Finance*, 18(2), 207-218.

Puga D. (1999) The rise and fall of regional inequalities. *European Economic Review*, 43, 303--334.

Ruppert D., Wand M.P., Carroll R.J. (2003) *Semiparametric Regression*. Cambridge University Press

Tadesse B. and Ryan M. (2004) Host market characteristics, FDI, and the FDI – trade relationship. *The Journal of International Trade & Economic Development*, 13: 2, 199-229.

Thurston S.W., M.P.Wand and Wiencke J.K. (2000) Negative Binomial Additive Models. *Biometrics*, 56, 139-144.

Tomlin K. (2000) The Effects of Model Specification on FDI Models: An Application of Count Data Models. *Southern Economic Journal*, 67(2), 460-468.

Venables A.J. (1996) Equilibrium locations of vertically linked industries. *International Economic Review*, 37, 341--359.

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

Wheeler D. and Mody A. (1992) International investment location decisions. The case of US firms. *Journal of International Economics*, 33, 57-76.

Wood S.N. (2006a) *Generalized Additive Models. An Introduction with R*, Boca Raton, Chapman and Hall.

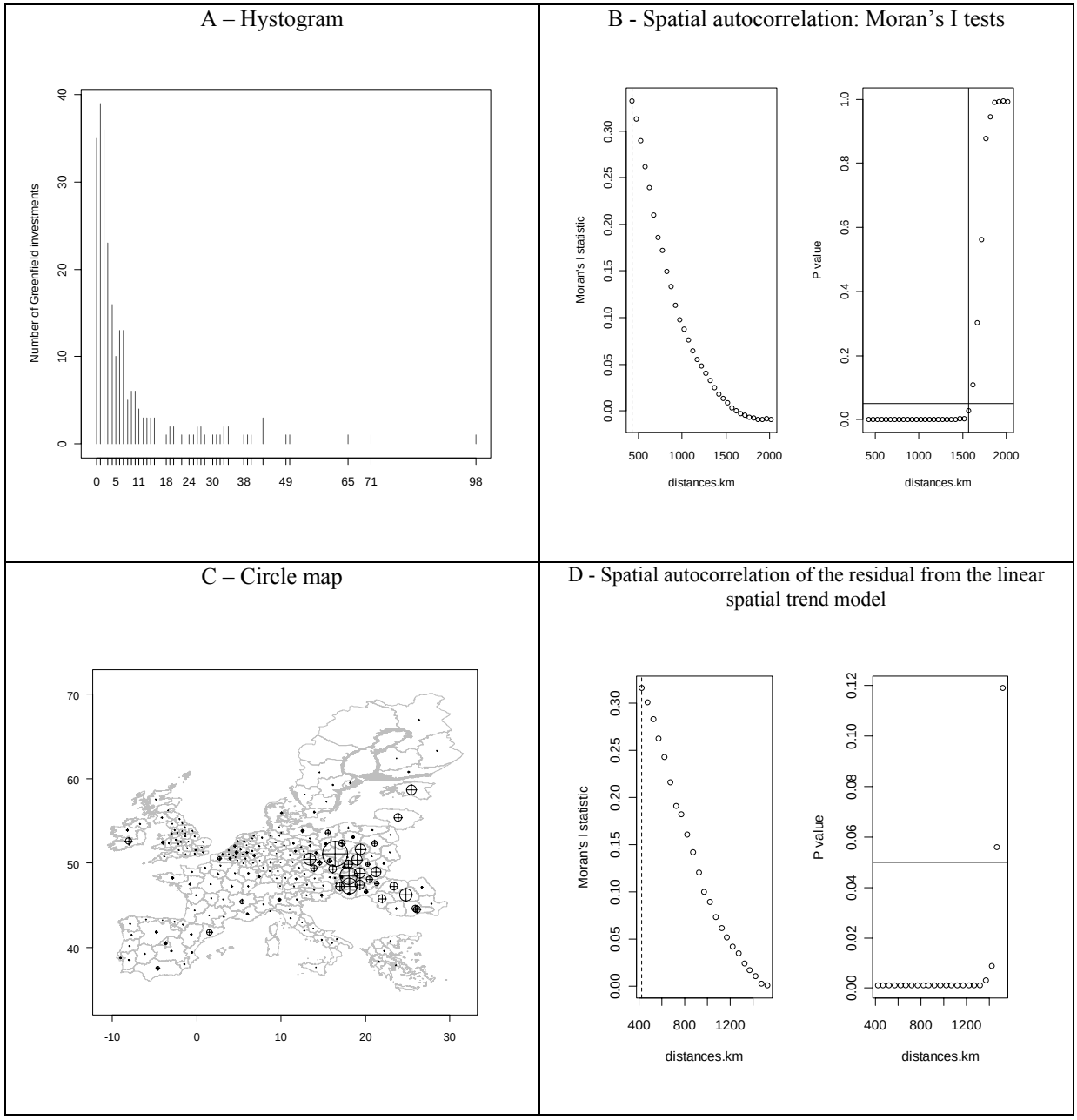
Wood S.N. (2006b) Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, 62, 1025–1036

Wood S.N. (2008): Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society, Series B* (Statistical Methodology), 70, 495-518.

Yeaple S. (2003) The complex integration strategies of multinationals and cross country dependencies in the structure of foreign direct investment. *Journal of International Economics*, 60, 293–314.

Zhou C., Delios A. and Yang J.Y. (2002) Locational Determinants of Japanese Foreign Direct Investment in China. *Asia Pacific Journal of Management*, 19, 63-86

Figure 1 – FDI distribution



Notes:

- Panels B and D display the values of Moran's I statistics and the corresponding bootstrapped *p*-values obtained from 999 random permutations. Several row-standardized spatial weights matrices (*W*) have been used to compute Moran's *I* statistics in order to check the robustness of the evidence of spatial autocorrelation. Elements w_{ii} on the main diagonal of each matrix are set to zero whereas $w_{ij} = 1$ if $d_{ij} < \overline{d}$ and $w_{ij} = 0$ if $d_{ij} > \overline{d}$, with d_{ij} being the great circle distance between the centroids of region *i* and region *j* and $\overline{d}$ the cut-off distance ranging from 423 km (the minimum distance allowing all regions to have at least one neighbor) up to and including 2,023 km at 50 km intervals. A monotonic relation between $\overline{d}$ and spatial autocorrelation emerges: Moran I statistics reach a maximum when the cut-off distance is equal to 423 km, but they are always positive and significant at 5% up to a 1,573 km cut off.
- Each circle in the plot in panel C, centred at the regional centroids, is proportional to the regional percentage share of FDI on the total number. The spatial coordinates are in millions of feet, hence the East-West extent of the enlarged Europe is approximately 12,000 kilometres, while the South-North extent is about 10,700 km.

Figure 2 – Spatial trend surface model: smooth interaction effect of latitude and longitude

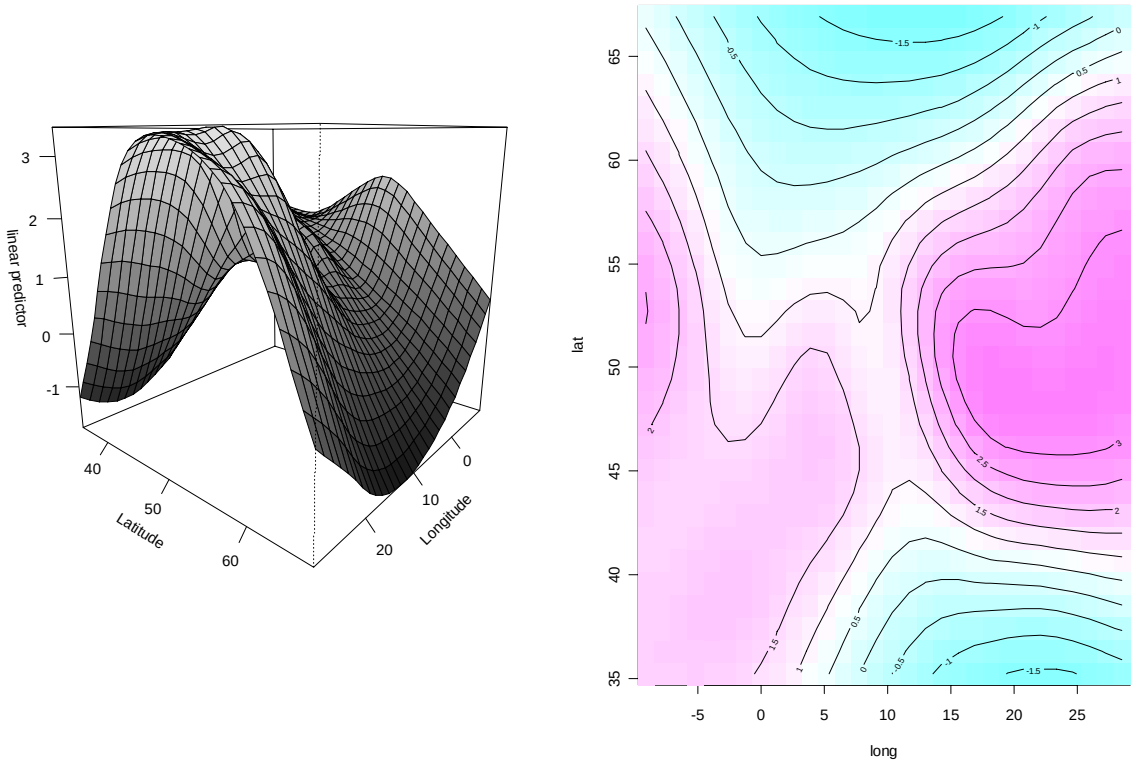


Table 1 - Econometric results: parametric model

	Poisson	**Negative Binomial**
mkt	0.349 (0.037)	0.521 (0.098)
	[0.000]	[0.000]
infra	0.329 (0.033)	0.305 (0.087)
	[0.000]	[0.000]
Jacobs	1.263 (0.151)	1.295 (0.380)
	[0.000]	[0.001]
empdens	0.422 (0.030)	0.351 (0.075)
	[0.000]	[0.000]
wage	-1.232 (0.046)	-1.576 (0.146)
	[0.000]	[0.000]
ur	0.520 (0.042)	0.309 (0.130)
	[0.000]	[0.018]
ter	0.452 (0.067)	0.669 (0.203)
	[0.000]	[0.001]
AIC	2,078	1,335
Overdispersion	10.090[0.001]	0.267[0.605]
$\hat{\theta}$		1.183
Vuong		0.309
		[0.378]
Wald-Hurdle		0.281
		[0.998]

Notes: AIC is the Akaike Information Criterion. The tests of spatial dependence, based on Moran's I statistics, use different distance neighbours weights matrices ranging from 423 km to 1,423 km (shown in the next Figure). The test of overdispersion is based on the estimation of the simple model $|e| = f(y)+u$, where $|e|$ is the absolute value of the residuals of the model and y is the vector of fitted values. Under the null hypothesis of equi-dispersion, the smooth term f(y) must be estimated with one degree of freedom and, according to a F test, it should have an insignificant effect on $|e|$. $\hat{\theta}$ is the estimated Negative Binomial scale parameter. Vuong is a non-nested hypothesis test statistic asymptotically distributed N(0,1) under the null that the models ZINB and NB-GLM are indistinguishable. Wald-Hurdle tests the null hypothesis that no-zero-hurdle is required in hurdle regression models for count data. The same set of regressors is used in the hurdle model for both the count component and the zero hurdle component. Standard errors are in parentheses and p-values are in square brackets.

Figure 3 - Diagnostics based on parametric Poisson regression model residuals

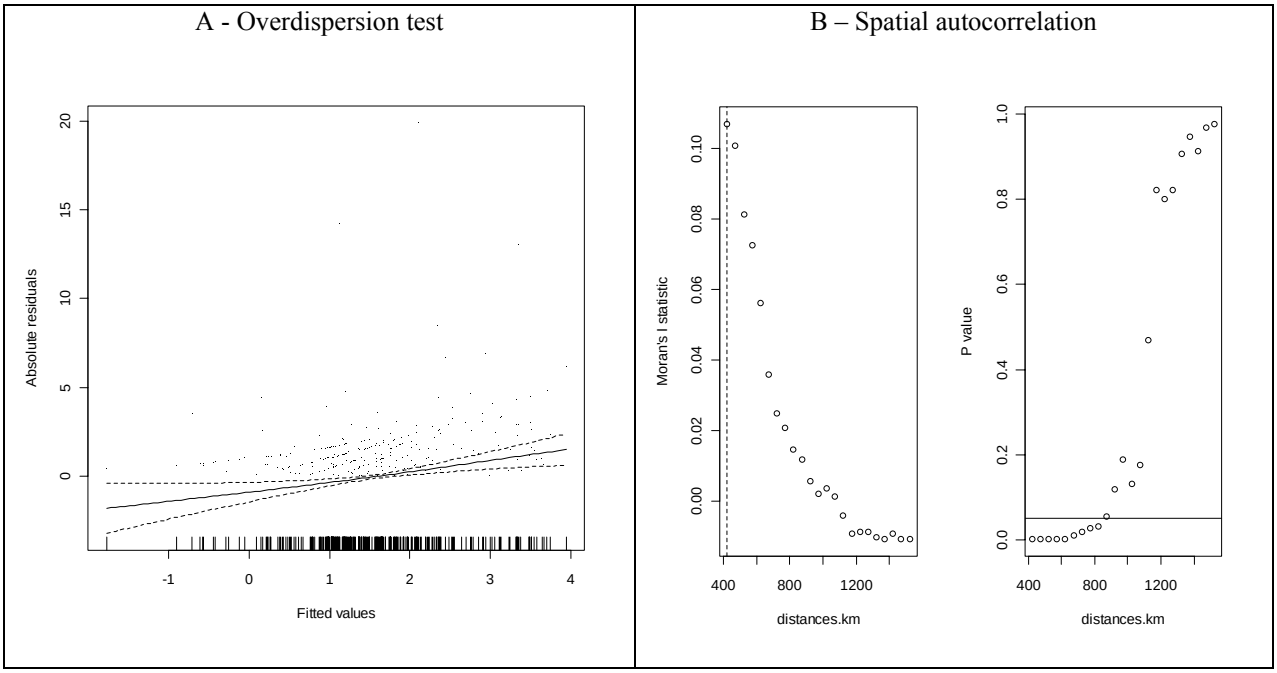


Figure 4 - Diagnostics based on parametric negative binomial regression model residuals

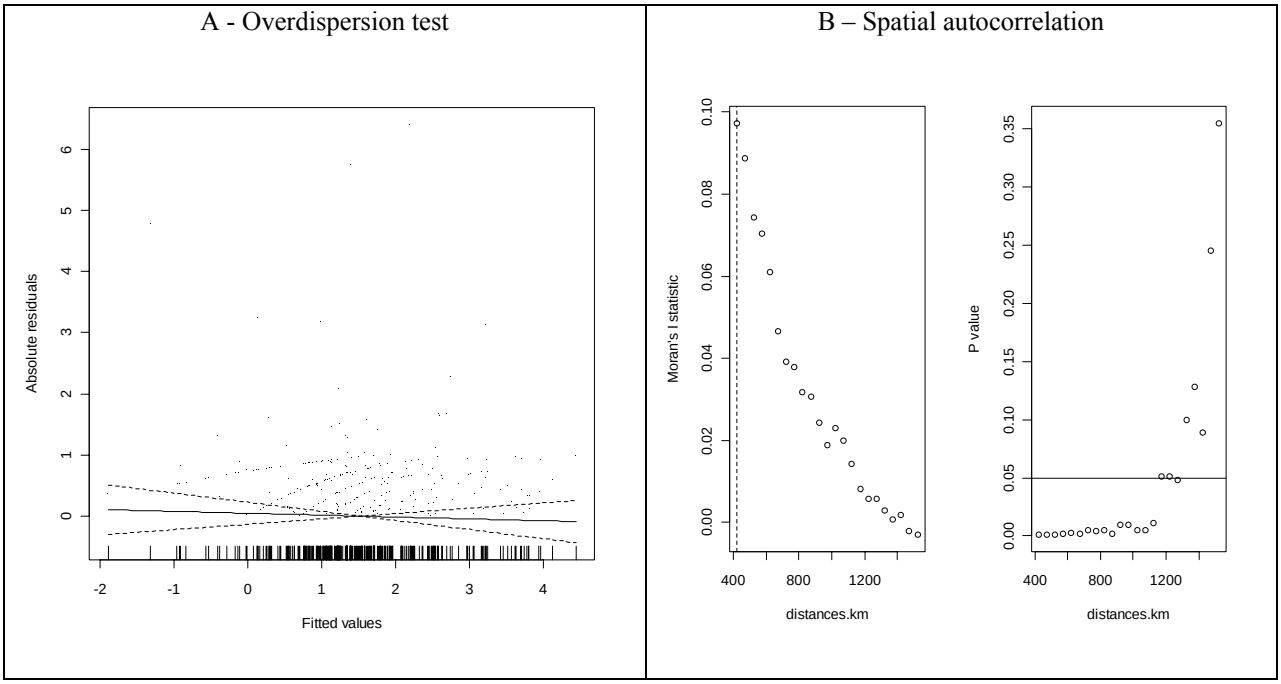


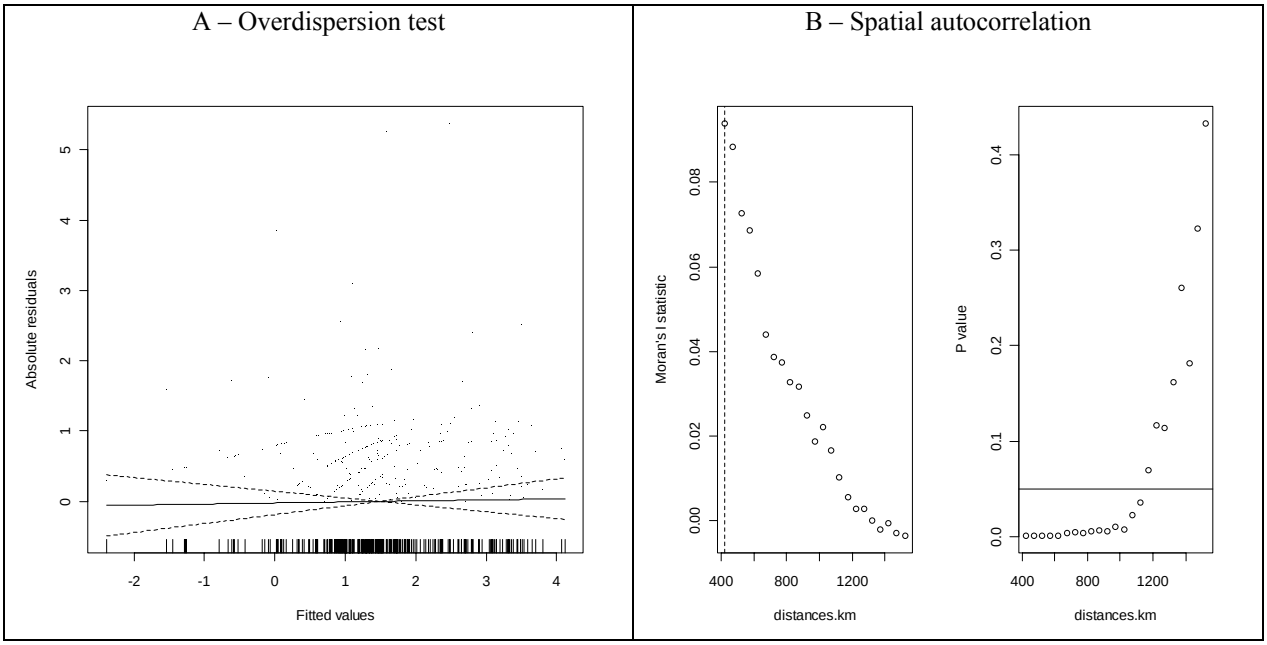
Table 2 - Econometric results of the NB-GAM

		Model 1	**Model 2 (Spatial lag model)**	**Model 3 (Spatial lag model with spatial trend surface)**
Parametric Terms		*Coefficients*	*Coefficients*	*Coefficients*
mkt		0.593 (0.094) [0.000]	0.600 (0.083) [0.000]	0.761 (0.082) [0.000]
infra		0.260 (0.083) [0.001]	0.282 (0.073) [0.000]	0.241 (0.095) [0.011]
Jacobs		0.956 (0.354) [0.007]	0.739 (0.312) [0.018]	0.590 (0.271) [0.030]
ur		0.270 (0.123) [0.028]	0.258 (0.107) [0.015]	0.559 (0.121) [0.000]
ter		0.738 (0.188) [0.000]	0.579 (0.166) [0.000]	0.267 (0.180) [0.139]
Wy			0.064 (0.012) [0.000]	0.065 (0.037) [0.078]
$\hat{v}$			-0.339 (0.055) [0.000]	-0.306 (0.052) [0.000]
Nonparametric terms		χ^2 *test*	χ^2 *test*	χ^2 *test*
$f(empdens)$		24.850 [0.000]	14.120 [0.000]	5.490 [0.068]
	Edf	2.662	2.096	2.060
$f(wage)$		139.110 [0.000]	50.550 [0.000]	99.160 [0.000]
	Edf	3.172	3.056	3.694
$f(lat, long)$				2.977 [0.000]
	Edf			23.495
AIC		1,308	1,262	1,188
Overdispersion		0.069[0.793]	0.001[0.981]	0.773[0.380]
$\hat{\theta}$		1.672	2.513	6.571

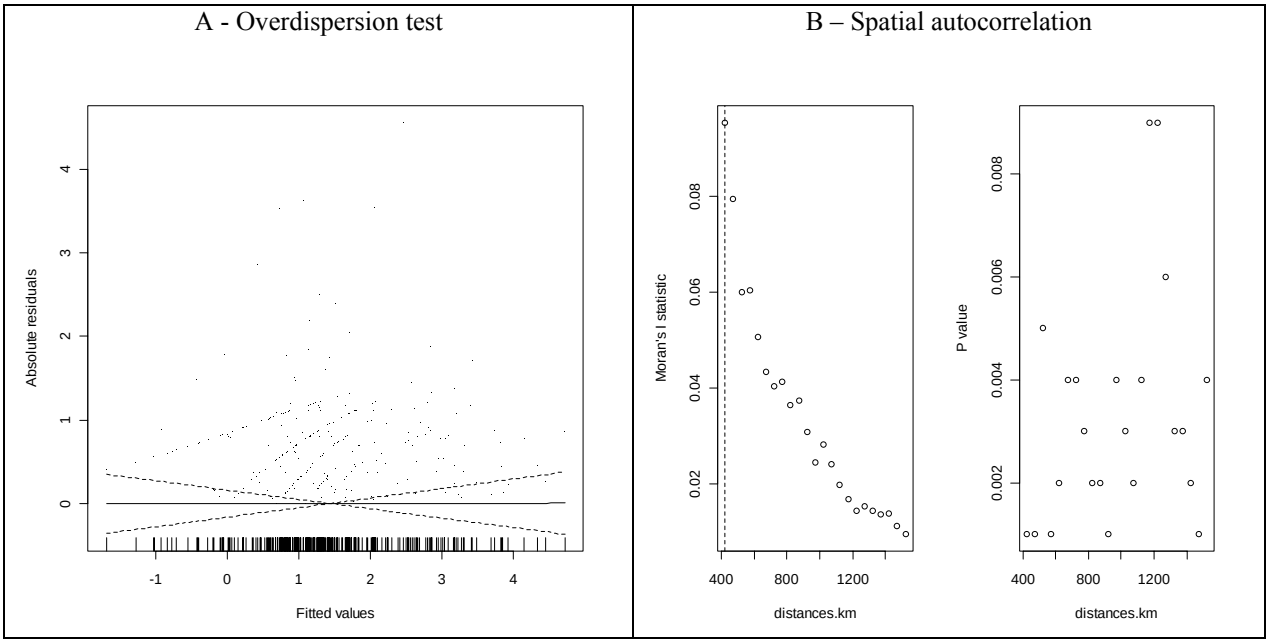
Notes: F tests are used to investigate the overall ("approximate") significance of smooth terms. *e.d.f.* (effective degrees of freedom) reflect the flexibility of the model. An *e.d.f.* equals to 1 suggests that the smooth term can be approximated by a linear term. In such cases, parametric terms have been used. Standard errors are in round parentheses and p-values are in square brackets.

Figure 5 - Diagnostics for the residuals of the semiparametric Negbin model

Model 1



Model 2



Model 3

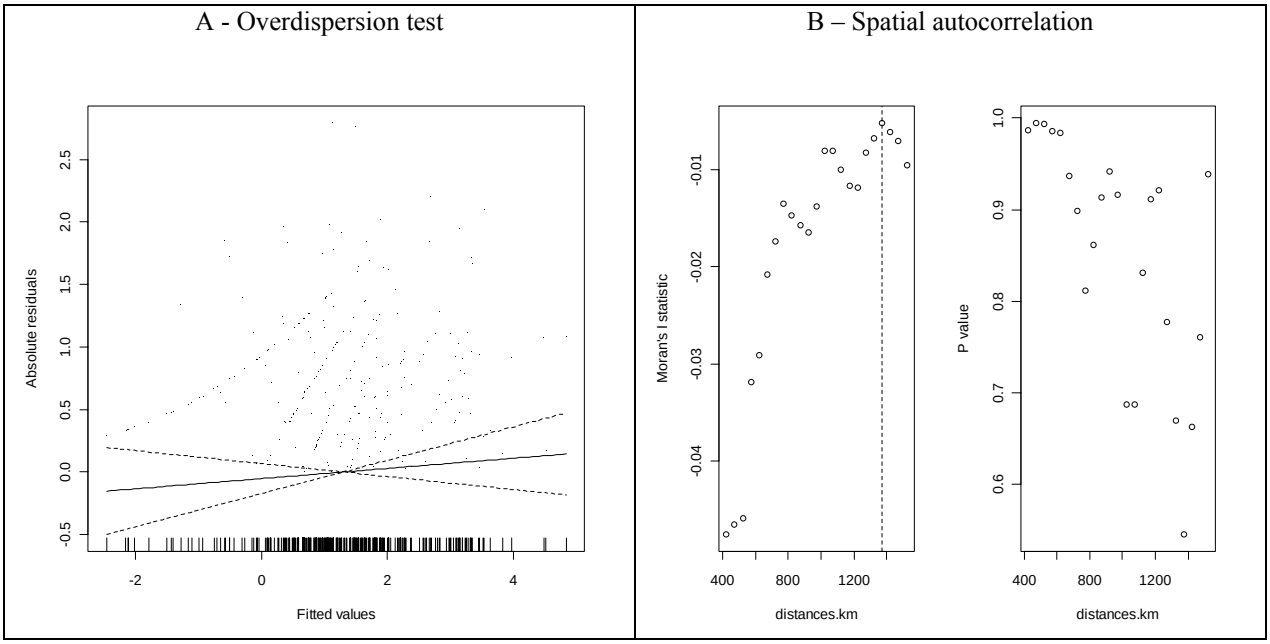
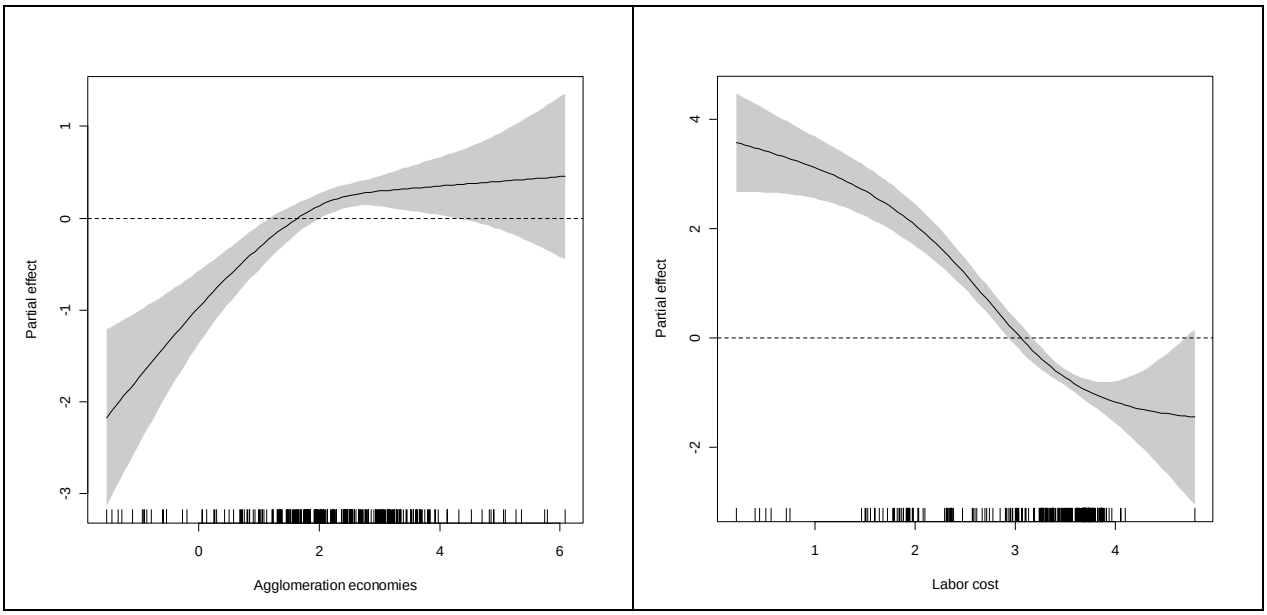
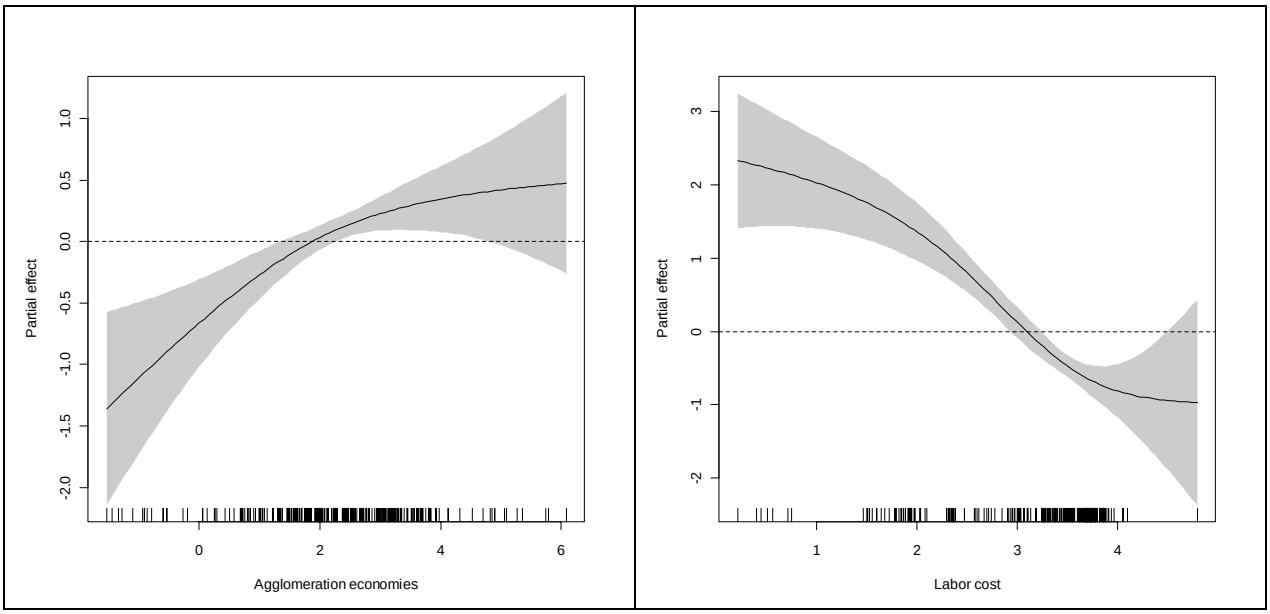

A - Overdispersion test
Absolute residuals
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
-2
0
2
4
Fitted values
B – Spatial autocorrelation
Moran's I statistic
-0.04
-0.03
-0.02
-0.01
400
800
1200
distances.km
P value
0.6
0.7
0.8
0.9
1.0
400
800
1200
distances.km

Figure 6 – Smooth effects of the semiparametric negative binomial regression model

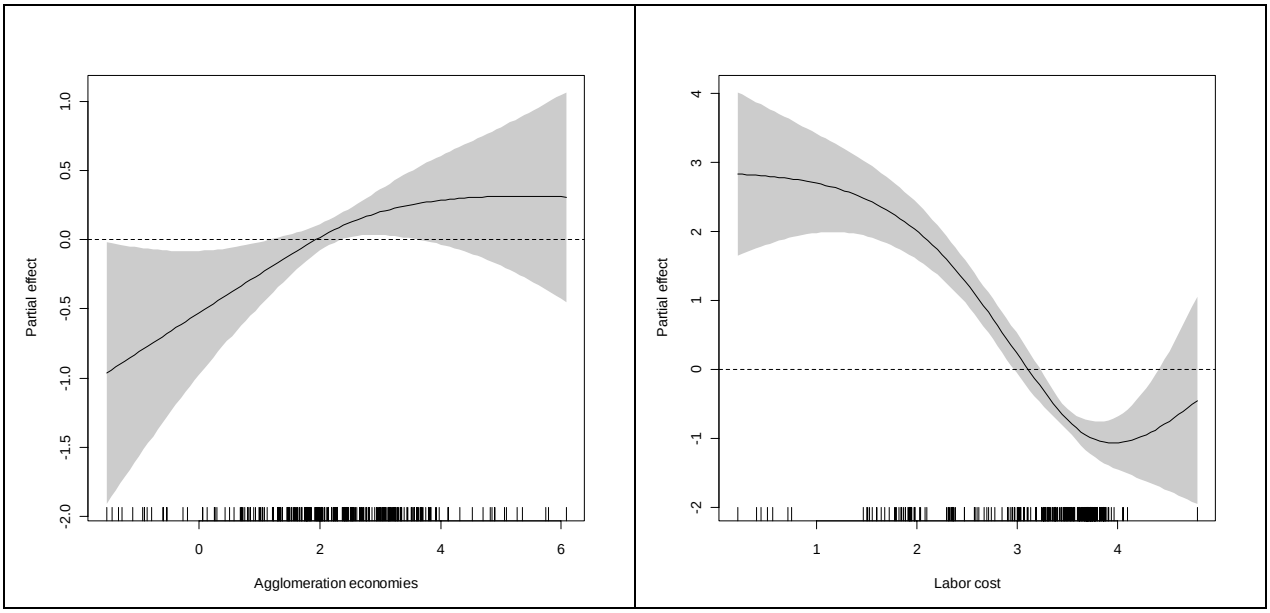
Model 1

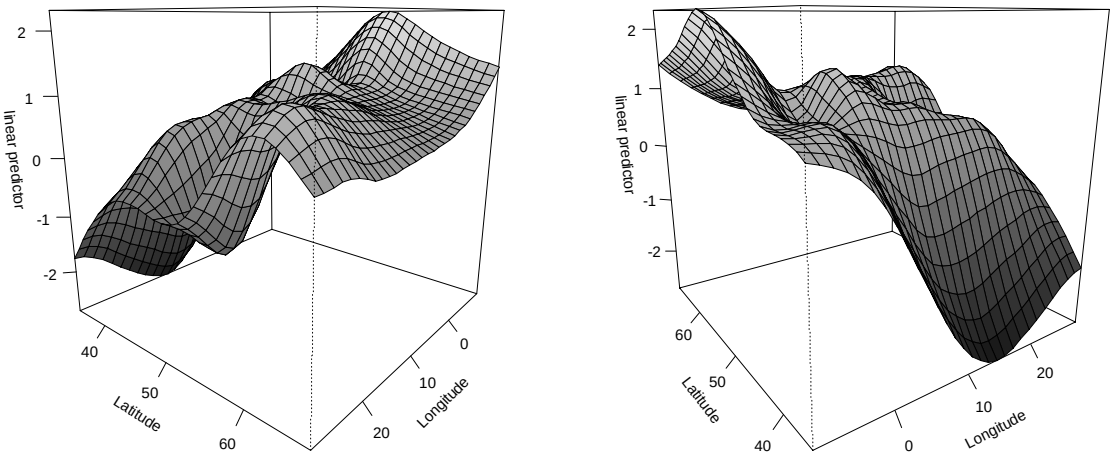


Model 2



Model 3


Partial effect
Agglomeration economies
Partial effect
Labor cost


linear predictor
Latitude
Longitude
linear predictor
Latitude
Longitude

DEPARTMENT OF ECONOMICS AND PUBLIC FINANCE “G. PRATO”
UNIVERSITY OF TORINO
Corso Unione Sovietica 218 bis - 10134 Torino (ITALY)
Phone: +39 011 6706128 - Fax: +39 011 6706062
Web page: http://eco83.econ.unito.it/prato/