Basile, Roberto; Benfratello, Luigi; Castellani, Davide

Conference Paper

Geoadditive models for regional count data: an application to industrial location

52nd Congress of the European Regional Science Association: "Regions in Motion - Breaking the Path", 21-25 August 2012, Bratislava, Slovakia

Provided in Cooperation with:
European Regional Science Association (ERSA)

Suggested Citation: Basile, Roberto; Benfratello, Luigi; Castellani, Davide (2012) : Geoadditive models for regional count data: an application to industrial location, 52nd Congress of the European Regional Science Association: "Regions in Motion - Breaking the Path", 21-25 August 2012, Bratislava, Slovakia, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
https://hdl.handle.net/10419/120470

Geoadditive models for regional count data: an application to industrial location

Roberto BASILE[♣]
Second University of Naples

Luigi BENFRATELLO[♠]
University of Turin and Ceris-CNR

Davide CASTELLANI[♦]
University of Perugia and Centro Studi Luca d'Agliano

Abstract

We propose a geoadditive negative binomial model (Geo-NB-GAM) *for regional count data which allows us to simultaneously address some important methodological issues, such as spatial clustering, nonlinearities and overdispersion. We apply this model to study location determinants of inward greenfield investments occurred over the 2003-2007 period in 249 European regions. The inclusion of a geoadditive component (a smooth spatial trend surface) permits us to control for spatial unobserved heterogeneity which induces spatial clustering. Allowing for nonlinearities reveals, in line with theoretical predictions, that the positive effect of agglomeration economies fades as the density of economic activities reaches some limit value. However, no matter how dense the economic activity becomes, our results suggest that congestion costs would never overcome positive agglomeration externalities.*

Keywords: Geoadditive models, Negative binomial models, Industrial location, European Union
JEL classification: C14, C21, F14, F23

We wish to thank three anonymous referees and the editor, Daniel Griffith, for very valuable suggestions on a previous draft. Usual disclaimers apply. Financial support from the Italian Ministry of University and Research (MIUR), PRIN 2009 project on "Production, R&D and knowledge offshoring: economic analyses and implications for Italy", is gratefully acknowledged.

[♣] Corresponding author: Corso Gran Priorato di Malta, 1, 81043, Capua (CE). E-mail: roberto.basile@unina2.it
[♠] E-mail: benfratello@econ.unito.it
[♦] E-mail: davide.castellani@unipg.it

1. Introduction

In modelling regional count data, scholars often face some technical issues related to the non-random spatial distribution of the data, as some degree of spatial clustering is likely to occur, and to overdispersion, since the conditional variance is likely to exceed the conditional mean. Modelling this kind of data is even more complicated as the relationship between the response variable and the covariates is likely to vary over space and/or over the range of the regressors so that the linearity assumption proves to be too restrictive. In this paper we present a method which allows one to tackle all the aforementioned issues and we apply it to the case of the location of greenfield foreign investments in the NUTS-2 regions of the Enlarged Europe over the 2003-2007 period.[1] Needless to say, this method proves to be useful well beyond the specific case of industrial location and can be applied to other analyses both in economics and in other disciplines (e.g. distribution of areal data on patents and trademarks, diseases or crimes just to mention a few).

The broad literature on inward foreign direct investments (FDI) location choice (surveyed by Arauzo-Carod et al., 2010) has widely documented that multinational firms do not locate randomly in space; but rather they tend to concentrate in few neighbouring regions. If the explanatory variables included in the model do not fully capture the spatial clustering of foreign firms, some spatial correlation remains in the errors, affecting the efficiency and/or the consistency of standard estimators. This may occur because of partially unobservable spatial factors, reflecting the role of regional amenities and regional policies.

[1] NUTS is an acronym for *Nomenclature of Units for Territorial Statistics* which indicates a hierarchical classification of administrative areas used by the official European statistical office (Eurostat). NUTS levels (1-3) indicate decreasing degrees of aggregation.

On the other hand, nonlinearities are also very likely to occur in industrial location analysis. For example, theory suggests that a variable effect of agglomeration economies on firms' location decisions might exist according to the level of agglomeration. In fact, as agglomeration reaches some critical value a congestion effect may eventually kick-in reducing the attractiveness of a given location. To explore this issue, some authors have postulated an inverted-U shaped relation between inward FDI and agglomeration economies and modelled it by inserting a squared term of the agglomeration variable (Arauzo-Carod, 2005; Viladecans-Marsal, 2004). Admittedly, this is only one of several competing parametric restrictions which may capture a nonlinear relation. Indeed, nonlinearities can be better accommodated in a semiparametric framework, where the actual shape of the partial effect can be assessed using smooth functions.

The two aforementioned 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 and due to the lack of FDI data at regional level, we are forced to use as dependent variable the number of new ventures, i.e. a count which takes discrete and non-negative values. A Generalised Linear Model (GLM) framework, assuming a negative binomial distribution for the conditional expectation of the number of new foreign plants in each region, provides a relatively flexible framework for the analysis of such count data. In fact, as it will be discussed in greater details in Section 4, GLMs lend themselves to an extension into the semiparametric framework, by adding smooth functions of covariates in the conditional expectation. This class of models is known as Generalized Additive Models (GAMs) (Wood, 2006a). By including also a *smooth spatial trend surface* - i.e. a nonparametric interaction of latitude (northing) and longitude (easting) - among the regressors, a GAM can be turned into the so-called Geoadditive model (see, *i.a.*, Kammann and Wand, 2003;

Wood, 2003; Fahrmeir and Echavarrìa, 2006; Augustin et al., 2009), which allows one to take unobserved spatial heterogeneity into account.

In sum, we use a Geoadditive Negative Binomial Model (Geo-NB-GAM) and apply it to estimate the determinants of multinational firms' greenfield investment location in the NUTS-2 regions of the Enlarged Europe. To the best of our knowledge only one other work uses an additive semiparametric model to investigate the determinants of new plant creation in Spanish provinces (Arauzo-Carod and Liviano, 2007). However, our study differs from this one since we thoroughly address spatial clustering alongside with nonlinearity, we focus on multinational plants and we take a much broader perspective by studying location in the Enlarged Europe.

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. Section 4 introduces the econometric methodology whereas Section 5 reports the empirical results. Section 6 concludes.

2. Spatial distribution of multinational firms' investments within the European Union

We retrieved our data from fDi Markets, an online database maintained by *fDi Intelligence* (a specialist division of the Financial Times Ltd) which monitors cross-border investments covering all sectors and countries worldwide. Relying on media sources and company data, *fDi Markets* collects detailed information on cross-border greenfield investments (available

since 2003). Data are based on the announcement of the investment and are daily updated[2]. The database is used as the source for foreign investment project information in the World Investment Report by UNCTAD (United Nation Conference on Trade and Development) and in publications by the Economist Intelligence Unit.

We selected 1,930 greenfield investments in the creation of manufacturing plants carried out by both European and non-European multinational firms in the European Union over the 2003-2007 period. For each project, detailed information is available on the investor (name, 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 NUTS-2 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 the lack of data. The dataset included therefore 22 EU member states (249 NUTS-2 regions).

The distribution of the 1,930 greenfield investments in the manufacturing sector is right skewed (Figure 1), with a share of zeros of about 14%, suggesting a high degree of overdispersion in the raw data and reveals a substantial degree of geographical clustering (Figure 2). The latter is all the more evident if we regress the number of investment projects on the smooth interaction between latitude (northing) and longitude (easting), $h\ no, e\ .$[3] Figure 3 plots the geographical components of such a model, showing a *saddle*

[2] A team of in-house analysts search daily for investment projects from various publicly available information sources, including Financial Times newswires, nearly 9,000 media, over 1,000 industry organizations and investment agencies, data purchased from market research and publication companies. Each identified project is cross-referenced against multiple sources and over 90% of projects are validated with company sources. More information is available at http://www.fdimarkets.com/.

[3] Technical details on how to estimate these effects are provided in Section 4.

pattern in the (predicted) spatial distribution of FDI. They are clustered in two peripheral areas: one includes regions belonging to the New Member States (Eastern countries), while the other characterizes some western peripheral areas (mainly Ireland, but also Spain, Portugal and France).

- **Insert Figures 1-3 about here** -

In sum, the exploratory spatial data analysis suggests to model FDI counts bearing in mind, on the one hand, the issues of non-normality, skewness and overdispersion and, on the other hand, the presence of spatial clustering. While accounting for the former issue is rather straightforward using negative binomial models, the latter can be tackled modelling the mean function through a set of covariates intended to capture the spatial clustering. In the next section, we discuss some theoretical hypotheses on the location determinants of foreign ventures which suggest a proper set of explanatory variables needed to model the expected mean function. Should regional observable characteristics not be able to fully account for the spatial clustering, residual spatial dependence would remain in the error term. We will tackle this issue by including smooth spatial trends surfaces. This will also allow us to highlight where the unexplained regional clusters occur and to hypothesize some reasonable unobserved factor behind their formation.

3. Determinants of the spatial distribution of inward foreign investments

The spatial distribution of FDI can be modelled as the result of the interaction between centripetal (or agglomeration) and centrifugal (or competition) forces. Among agglomeration forces, we focus on the role of urbanization externalities, while as for

centrifugal forces we consider the effect of labour costs along with other labour market characteristics[4].

3.1. 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 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 spillovers between adjacent firms. Marshallian externalities are therefore more suitable to explain "small scale" agglomeration phenomena.[5] Unfortunately, we do not have

[4] Admittedly, data availability prevents us to include additional determinants. The role of regional policies aimed at attracting FDI in order to boost regional development has been investigated by several regional studies (see, *i.a.*, Wheeler and Mody, 1992; Head et al., 1999; Crozet et al., 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, openness to FDI and the like) on regions' performance in attracting foreign investors (Basile et al. 2006; Barrios et al., 2008). Finally, EU policies (such as the Structural and Cohesion funds allocated to laggard regions) can also be important factors affecting the attractiveness of a location (Basile et al., 2008). Unfortunately, comparable data on institutional variables are not available for all the regions in our sample.

[5] Economic agglomeration can be considered at different levels of aggregation. On the one extreme, there are "small scale" agglomerations, i.e. agglomerations in small areas of firms operating in the same, very finely defined sector, like highly specialized industrial districts. On the other end, there are "large scale" agglomerations of firms operating in the manufacturing sector that span country boundaries, such as the 'Manufacturing Belt' covering the North-East part of the US and the 'Hot Banana' located between Milan

detailed information on the specific industry where the new plant operates so that we can only count the total number of manufacturing foreign investments in the regions. In turn, we are forced to focus on the role of urbanization externalities[6] which we model using four different variables: 1) the size of the regional market, 2) the degree of sectoral diversification, 3) the overall employment density in the manufacturing sector 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 related to the "*home market effect*". As first noted by Krugman (1980), under increasing returns to scale and in presence of transport costs, the appeal of a region as a production site depends crucially on the size of its domestic market. Firms will locate in the region where they can exploit economies of scale to a greater extent and, eventually, export to neighbouring regions.

Sectoral diversification within regions (measured by the median of the sectoral specialization indexes[7]) 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.

---

and London. According to the New Economic Geography literature (Ottaviano and Puga, 1998), large scale agglomerations are mainly explained by the occurrence of urbanization externalities.

[6] It is worth mentioning that empirical evidence suggests that urbanization economies outweigh industry-specific localisation economies (Guimarães et al., 2000, Arauzo-Carod and Liviano, 2007).

[7] The specialisation index (or location quotient) is defined as $S_{si} = (E_{si} / \sum_s E_{si}) / (\sum_i E_{si} / \sum_i \sum_s E_{si})$, where E denotes employment, s the sector and i the region. The median of S_{si} is a measure of the number of sectors in which a region shows a revealed comparative advantage: a high median indicates that a region has a

Moreover, a diverse industrial structure protects a region from volatile demand and offers the possibility to switch between input substitutes.

Urbanization economies are not only driven by the degree of diversity of the economy but also by the overall density of economic activity. Therefore, we expect that regions with higher *employment density in manufacturing* (measured as total manufacturing employment per square km) attract more foreign investments. However, 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 foreign investors if, *ceteris paribus,* agglomeration economies overcome congestion costs. Therefore, a nonlinear effect of employment density on the number of 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-Marsal, 2004; Arauzo-Carod, 2005). Although it is the easiest way to deal with such a nonlinearity in a parametric framework, this is only one of several 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.

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

comparative advantage in a large number of sectors and it is therefore diversified whereas a low median means that a region is specialized (De Benedictis and Tamberi, 2004).

facilitating market access. The ensuing increase in private returns to investments makes locations with better infrastructure provisions more attractive for both domestic and foreign investments.

3.2. Labour market characteristics

The role of labour market characteristics as a determinant of inward FDI and new plant creation is well established (Friedman et al., 1992). We follow previous literature by specifying the regional labour market characteristics using three different variables: the average *wage* (measured by the total regional compensation to labour divided by the total number of employees in the region), *labour availability* (approximated by the regional unemployment rate) and *human capital* endowment (measured by the share of regional population aged 24 or more holding a tertiary education degree). The impact of wage and unemployment rate is not univocal, however. Lower wages may in fact attract 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 higher unemployment rates either as a measure of a large labor supply, which would attract firms, or as an indicator of a relatively rigid labor market, which would discourage them from investing in the region. In sum, the effect of basic labor market conditions may be in principle characterized by nonlinearities which should be properly accounted for when modeling foreign investors' location decisions.

4. Modelling regional inward foreign investment counts: a geoadditive negative binomial model

The empirical literature on foreign firms' location choice usually appeals to discrete-choice models (conditional, nested and mixed logit models) based on the Random Utility

Maximization (RUM) framework. Decision probabilities are therefore modelled in a partial equilibrium setting where foreign firms maximize profits subject to uncertainty that derives from unobservable characteristics. However, the use of discrete choice models is often hindered by the large dimension of the choice set (as in our case, where the choice set includes 249 regions) which makes estimation very burdensome. An alternative modelling strategy aggregates data at the elementary choice level by counting the number of times a given alternative is chosen (i.e. the number of investments in each region) leading to discrete, non-negative integer valued dependent variables (so-called count data). Provided that the regressors are choice-specific and either common to all decision makers or to groups of them, the likelihood function of the Poisson regression model is the same as the one of the conditional logit (up to a multiplicative constant) (Guimarães et al., 2003 and 2004). Thus, under these circumstances, the Poisson regression model can be thought as directly derived from a RUM process.

The Poisson regression model can be accommodated as a special case of the Generalized Linear Model (GLM) framework (McCullagh and Nelder, 1989). Let y, be the dependent variable, X a $k\times1$ vector of explanatory variables and β a $k\times1$ vector of regression parameters. The canonical smooth monotonic *link function* in the Poisson GLM is

$$g(\mu) = \log(\mu) = \eta = X'\beta; \qquad \mu = E(y); \qquad y \sim Poi(\mu) \qquad (1)$$

resulting in a log-linear relationship between mean and linear predictor. A characteristic of the Poisson regression model is the assumption of equidispersion, that is $\mu = E(y) = Var(y)$.

In practice, however, the classical Poisson regression model is often of limited use in a regional location analysis since empirical inward FDI counts typically exhibit

overdispersion, i.e. $Var\ y > E\ y$. A way of dealing with over-dispersed count data is to assume a negative binomial (NB) distribution for y which can arise as a gamma mixture of Poisson distributions. One parameterization of its probability density function is

$$P\ Y = y \mid \mu, \theta\ = \frac{\Gamma\ y+\theta}{\Gamma\ \theta\ \cdot y!} \cdot \left(\frac{\mu}{\mu+\theta}\right)^{y} \cdot \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \qquad (2)$$

with θ an index of overdispersion following a gamma distribution with parameters *a* an *b*, $\Gamma\ a,b$. The variance function is now $V\ \mu\ = \mu + \mu^2\theta^{-1}$. Note that, for large θ, the NB model approaches the Poisson model.[8]

An important practical feature of GLM is that they can be fit using a form of Iteratively Re-weighted Least Squares (IRLS). McCullagh and Nelder (1989) prove that the IRLS algorithm leads to maximum likelihood (ML) estimates.

A large number of studies on regional inward FDI counts have used the standard NB regression model with cross sectional data (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) or random effects extensions of NB regression for panel data (Blonigen, 1997; Basile, 2004; Basile et al., 2006).[9]

[8] Although NB models capture overdispersion quite well, they are not always sufficient for modelling excess zeros. To overcome this problem, zero-augmented models that incorporate a second model component capturing zero counts have been proposed. Zero-inflation models (Lambert, 1992) are mixture models that combine a count component and a point mass at zero. Hurdle models (Mullahy, 1986) combine instead a left-truncated count component with a right-censored hurdle component. Applications of these models to FDI location analyses are in Tadesse and Ryan (2004), Basile (2004), Tomlin (2000) and Iannizzotto and Miller (2002).

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

A limit of this recent literature on inward FDI counts is the assumption that all regions obey a common linear specification of the location model, disregarding likely nonlinearities reflecting spatial heterogeneity in the behaviour of economic agents. In particular, as stated above, we cannot disregard possible threshold effects in the impact of agglomeration externalities on regional attractiveness.

Nonlinearities can be addressed in different ways. Some authors have proposed polynomial expansions up to a cubic function within a GLM approach (see, in the context of industrial location, Arauzo-Carod, 2005, and Viladecans-Marsal, 2004). Although rather easy to implement, this solution might introduce severe multicollinearity. In this paper we adopt a different solution, namely the Generalized Additive Model (GAM), taking advantage of their recent extension to handle Negative Binomial responses (Thurston et al., 2000).

The GAM framework (Hastie and Tibshirani, 1990) extends the GLM by introducing nonlinear smooth functions of the covariates as additive components of η:

$$\eta = g\ \mu\ = X^{*}\beta^{*} + f_1\ x_1\ + f_2\ x_2\ + f_3\ x_3, x_4\ + \ldots \qquad (3)$$

$$\mu = E\ Y \qquad Y \sim negbin\ \mu, \theta \qquad \theta \sim \Gamma\ a, b$$

where $f_j\ .$ are unknown smooth functions of the covariates, X^{*} is a vector of strictly parametric components and β^{*} is the corresponding parameter vector.

Each univariate 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 $\mathbf{S}_j$ are positive semidefinite matrices, 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 (Wood, 2003, 2006a). The penalized spline base-learners can be extended to two or more dimensions to handle interactions. Specifically, Wood (2006a, Section 4.1.5) recommends to use thin-plate regression splines for smooth interactions of quantities measured in the same units (such as the spatial coordinates) and tensor products for smooth interactions of quantities measured in different units.

Given the bases for each smooth term, model (3) can be re-written as a GLM, $g(\mu) = X^{*}\beta$ where some columns of X include X^{*} and the others represent the basis functions evaluated at the covariate values, while β contains β^{*} and all the smooth coefficient vectors β_j. The model is estimated by minimizing the penalized deviance:

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

with respect to β, where λ_j are positive smoothing parameters and $D(\beta) = 2\phi(l_s - l(\beta))$ is the model deviance, with l the model log-likelihood, l_s the saturated log-likelihood and ϕ the scale parameter.

Penalized ML estimation is performed by Penalized Iteratively Re-weighted Least Squares (PIRLS). Let $\hat{\beta}$ and $\hat{\mu}$ be current best estimates, pseudo values, $z = g'(\hat{\mu})(y - \hat{\mu}) + X\hat{\beta}$, and weights, $W = V(\hat{\mu})^{-1} g'(\hat{\mu})^{-2}$, are firstly computed. Then, the weighted penalized least square problem of minimizing $\left\| \sqrt{W}(z - X\beta) \right\|^2 + \sum_j \lambda_j \beta' \mathbf{S}_j \beta$ w.r.t. β is solved to obtain the updated $\hat{\beta}$ and $\hat{\mu}$. Iterating these two steps to convergence leads to the penalized ML estimates.

Given λ_j, therefore, computing the estimates $\hat{\beta}_\lambda$ is straightforward. But λ_j are unknown and must be estimated. Wood (2006a, 2008) has proposed different methods for automatic and integrated smoothing parameters selection. In the first one, called

"*performance iteration*", a grid search provides estimates for λ_j at each PIRLS step based on Generalised Cross Validation (GCV). In this case, the parameter θ is chosen in order to ensure that the estimate of the scale parameter is as close as possible to 1, the value that the scale parameter should have. With the second method, termed "*outer iteration*", the scale parameter is set to 1 and λ_j and θ are chosen by minimizing the Akaike Information Criterion (AIC). In this case, the PIRLS scheme is iterated to convergence for each trial set of smoothing parameters and AIC scores are only evaluated on convergence; optimization is then "outer" to the PIRLS loop. Finally, within the "*outer iteration*" method, it is possible to resort to Restricted Maximum Likelihood (REML), instead of the AIC, due to the possibility of rewriting the penalized GAM as a generalized mixed model (Wood, 2006b; Ruppert et al. 2003; Kammann and Wand, 2003). Simulation results suggest that the REML method offers some improvement in mean-square error performance relative to GCV and AIC in most cases (Wood, 2011).[10]

Both GLM and GAM are based on the assumption of spatial independence. As shown in Section 2, instead, investments from multinational firms tend to cluster over space and this does not necessarily reflect the distribution of observed explanatory variables. Spatial autocorrelation in the residuals may indeed occur (and should be controlled for when specifying a location model) when unobserved heterogeneity is spatially correlated. For example, a number of factors related to culture, policy actions (incentives, corporate taxes and other institutional characteristics) and various forms of amenities can affect the regional attractiveness. Unfortunately, these factors are often either

[10] Estimation of GAMs can be performed with the *R* software using different packages, such as *gam* developed by Hastie and Tibshirani, *gamlss* developed by Stasinopoulos, Rigby and Akantziliotou and *mgcv* developed by Wood. In the present paper we have estimated NB-GAMs using *mgcv*. Wood (2006a, ch. 5) thoroughly describes how to use *mgcv* and its main differences with *gam*.

unobservable or cannot be properly measured, especially in samples composed of many geographical units (as in our case). Insofar as these variables are spatially correlated, the residuals will be spatially correlated too.

The issue of unobserved heterogeneity can be addressed within the GAM framework by incorporating the spatial location as an additional covariate, that is by including the bivariate smooth term of latitude (*northing*) and longitude (*easting*), $h\ no,e$, in model (3), thus generating what is known in the literature as the Geostatistical Additive Model (in our case Geo-NB-GAM):[11]

$$\eta = g\ \mu\ = X^{*}\beta^{*} + f_1\ x_1\ + f_2\ x_2\ + f_3\ x_3, x_4\ + \ldots + h\ no,e \qquad (5)$$

$$\mu = E\ Y \qquad Y \sim negbin\ \mu,\theta \qquad \theta \sim \Gamma\ a,b$$

Following Wood (2003), we estimate the spatial term, $h\ no,e$, using a low-rank thin plate regression spline.[12]

Although geoadditive models are widely used in environmental studies and in epidemiology (see, *i.a.*, Kelvyn and Wrigley, 1995; Kammann and Wand, 2003; Augustin et al. 2009), they are rarely considered for modelling economic data and, to the best of our knowledge, this is the first application to the location of industrial plants.

[11] An alternative method to control for spatially autocorrelated unobserved heterogeneity is the application of spatial econometrics tools, such as spatial lag, spatial Durbin and spatial error models (Anselin, 1988; LeSage and Pace, 2009). Examples of implementations of these methods to FDI location models are in Coughlin and Segev (2000), Blonigen et al. (2007), Baltagi et al. (2007, 2008). All these articles adopt a parametric approach, thereby neglecting possible nonlinearities. A semiparametric extension of the spatial lag model in presence of Gaussian responses has been proposed in Basile (2009) and Basile and Girardi (2010). Unfortunately, the properties of semiparametric spatial lag models with non-Gaussian errors have not been explored.

[12] We refer to Wood (2003) for a technical description of this smoother.

It is worth mentioning that a variety of regional economic phenomena give rise to count data and can thus borrow this methodology. First and foremost, regional innovation output is usually measured by counts, such as patents or the number of innovative firms[13]. Other regional counts may be related to migration flows (the number of im/emigrants into/from each region), or some survey-related information (such as the number of households/firms with a broadband connection and other similar socio-economic aspects). So far, in this literature, scholars have often resorted to transformation of the dependent variable, expressing as a share of some denominator or taking logs, and applied standard spatial econometric techniques, but this is clearly an unsatisfactory modelling strategy.

Our modelling strategy could be extended by including "spatially varying coefficients", i.e. parameters which differ not over the range of the covariates but over space. To make this extension, terms like $x_j f_4\ no,e$ should enter the GAM. We submit that this extension would be interesting, but it also requires some theoretical underpinnings as to which term x_j should be chosen to have a spatially varying effect on the response variable.[14] The same kind of parameter instability can also be addressed by the Geographically Weighted Regression (GWR), but this methodology does not easily extend to overdispersed data.

[13] Patents at the regional level are now rather easily available for both the EU (Regio dataset) and OECD (OECD Patent dataset). The number of innovative firms is instead provided for all EU countries from the Community Innovation Survey and similar information is available from many non-EU countries.

[14] For a theoretical comparison of spatially varying coefficients and GWR models, see Waller et al. (2007). For an application of GWR in the field of economics see Bivand and Brunstad (2006).

5. Evidence from parametric and semi-parametric regressions

In this section we present the results of our econometric analysis. We first illustrate the results from a standard GLM and then we highlight to what extent the estimation of a GAM (equation 3), which allow for nonlinearities, and of a Geo-GAM (equation 5) which also controls for spatial dependence, affects the results.

5.1 GLM estimation results

Table 1 reports coefficients and standard errors estimated with parametric GLMs. 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 in both the Poisson and the NB regression models. Since all variables are in logarithms, coefficients can be interpreted as elasticities. First, the positive coefficient associated to *mkt* confirms that foreign firms locate in high demand regions and possibly serve smaller markets via exporting. Second, the expected number of FDI 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, it seems that congestion costs are, on average, more than counteracted

[15] See the appendix for a thorough definition of the variables.

by agglomeration externalities. Finally, high wages seem to discourage foreign investments whereas high regional unemployment and tertiary education attract foreign investors.

- **Insert Table 1 about here** -

Table 1 also reports some diagnostics tests and measures of the goodness of fit. The value of the AIC clearly works in favour of the NB model. The latter seems to perform rather well both against the Poisson model, in that it is able to account for overdispersion[16], and against a Zero-Inflated NB model (ZINB) and a hurdle model, as the null hypothesis of no excess of zeros with respect to the prediction of the NB distribution cannot be rejected.[17]

To assess the extent of spatial autocorrelation in the residuals, we rely on the method proposed by Lin and Zhang (2007) and compute the Moran's *I* statistic using Pearson residuals and distance-based binary spatial weights matrices. Several geographical distance cut-offs, ranging from 320 km (the minimum distance which allows all regions to have at least one neighbour) to 920 km with a step of 50 km, have been adopted. All corresponding spatial weights matrices yield significant values of Moran's *I*, thereby showing the existence of spatial correlation. The highest standardized Moran's *I* value

[16] The over-dispersion test is based on the estimation of the simple nonparametric model $|res| = f(y) + \varepsilon$, where $|res|$ is the absolute value of the residuals of the GLM 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 $|res|$ (Thurston et al., 2000).

[17] 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 not significant, so that the existence of zero-inflation can be excluded. The Wald test for the NB against a hurdle model also points towards the use of the NB model.

occurred in correspondence to the minimum distance and it decreases monotonically with distance. In Table 1 we report the Moran's *I* values and the corresponding p-values obtained with binary spatial weights matrices based on minimum and maximum geographical distance cut-offs (320 and 920 km).

5.2 GAM estimation results

Table 2 summarizes the results of different nonparametric and semiparametric GAMs. As a first step, we estimate a nonparametric model which allows all covariates to have a nonlinear effect. Although we have some theoretical priors about the functional form of the relationship between inward investments and some explanatory variables (see Section 3), we prefer to be agnostic about which covariate should enter the model linearly. Therefore, we specify all terms non-parametrically (using cubic regression splines) and test for which variable the parametric specification is not rejected. This test is based on the Effective Degrees of Freedom (*edf*, henceforth) estimated for each smooth function. If the *edf* is equal to 1, a linear relationship cannot be rejected. In order to assess the robustness of the results of the linearity test, we estimate the model by using all the three different methods for automatic smoothing parameter choice mentioned in Section 4: *i*) GCV, *ii*) AIC and *iii*) REML. The evidence clearly reveals that the *edf* is equal to 1 for *Jacobs, infra* and *ter* regardless of the estimation method, while mixed results emerge for *ur* and *mkt*. As the AIC values of the three models are almost equal, they can be safely considered equivalent in statistical terms. Given the relatively small size of our sample (249 observations), we prefer the most parsimonious strategy, which minimizes the number of non-parametric terms. Therefore, we decide to rely on the REML results and assume a linear parametric form for the effect of *ur* and *mkt*. Finally, since the *edf* is always higher than 1 in the case of *empdens* and *wage*, these two variables undoubtedly enter the model nonlinearly.

- **Insert Table 2 about here** -

Based on this evidence, we use the REML criterion to estimate two semiparametric models with linear terms for all covariates but *wage* and *empdens*. In the first model, we do not introduce any control for unexplained spatial clustering, whereas the second model introduces the smooth interaction between latitude and longitude (specified using a thin plate regression spline). As measured by the AIC, the fit of the more parsimonious model is not worse than the fully nonparametric specifications. Thus, we are reassured that some terms could enter linearly. At the same time, the GAM has a much lower AIC than the GLM (presented in Table 1), suggesting that allowing for nonlinearity in *empdens* and *wage* improves the model fit. It is also worth noticing that the Geo-NB-GAM encompasses all the other models. Moreover, GAM residuals are still spatially autocorrelated whereas only in the Geo-NB-GAM residuals are purged from spatial dependence (see the results of the Moran's *I* tests). In other words, even controlling for a number of key regional characteristics and allowing for some nonlinearities, we are not able to explain all of the actual clustering of foreign investments in European regions. This suggests that some unobserved characteristics may still affect the location patterns. We will return on this issue later in this section.

The upper part of Table 2 reports the estimated parameters for the parametric terms and their standard error. It is worth noting that whereas GAM parameters (Column 5 of Table 2) do not differ significantly from those obtained through GLM (Table 1), the Geo-NB-GAM specification, which purges the residuals from spatial dependence, yields a slight change in the magnitude of some coefficients. In particular, the effects of market size (*Mkt*) and unemployment rate (*Ur*) increase, while the coefficients on tertiary education (*Ter*) and transport infrastructure (*Infra*) slide and become non-significantly different from zero. This

is consistent with the idea that unobserved spatial heterogeneity in previous estimates may have biased the estimated coefficients.

The middle part of the Table reports χ^2- tests for the overall significance of the smooth terms for *wage* and *empdens*, as well as their *edf*. Low values of the χ^2-test imply a high probability that the estimated smooth term is not different from zero, while as stated above *edf* is a measure of its nonlinearity. Results support the hypotheses that both *wage* and *empdens* are significant determinants of FDI location and that their relationship with the number of new investments is nonlinear. The spatial trend surface, $h\ no, e\ $, is also highly significant, suggesting the presence of an unexplained spatial concentration in FDI location.

Figures 4 and 5 show the smoothed partial effects of *wage* and *empdens* on the expected number of foreign investments estimated through the Geo-GAM. The shaded areas highlight the 95% confidence intervals. The wage-plot shows that, *ceteris paribus,* regions with low average labour costs tend to attract more FDI. The effect of a wage drop appears to be slightly nonlinear and lower for very high wage regions, consistently with the idea that in high-end areas the wage rate does not only captures labour cost, but also proxies for its quality, so as an increase in wages may not discourage multinationals after all. However, it is also worth noticing that the upturn for high values of wages is driven by the single highest value.

As for *empdens,* the graph in Figure 5 shows that the expected number of inward FDI 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 more dense industrial activity can exert a positive externality which promotes the location of foreign firms but, when the level of agglomeration becomes too

high, congestion costs kick-in and gradually reduce the attractiveness of a region to foreign investors. However, in our sample, congestion costs never overcome the magnitude of the positive externality so as the relationship between employment density and location of foreign plants is nonlinear but it 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 function does not show such a declining pattern.

- **Insert Figures 4 and 5 about here** -

Finally, Figure 6 displays the map plot of the geographical component in the geoadditive model. The plot shows that, after controlling for the most relevant variables and allowing for nonlinearities, there are still some unexplained clusters of foreign investments in Ireland and, to a lesser extent, in the North of England. This evidence might be interpreted in the light of the quite effective policies to attract foreign investors that some national and regional institutions have adopted. Due to the lack of comparable information across countries and regions on policies towards foreign investors, we cannot account explicitly for the different effectiveness of such policies. Nonetheless, the spatial trend surface partially captures these unobserved regional characteristics, in turn providing useful insights for the interpretation of the economic phenomenon and making the model less prone to misspecification problems. Finally, comparing Figures 3 and 6, it emerges that the strong clustering in Eastern European countries, evident in Figure 3, has almost vanished in Figure 6, even though there are some pockets of high predicted values. This suggests that this cluster was mainly driven by explanatory variables included in our model, *in primis* the wage rate.

- **Insert Figure 6 about here** -

6. Conclusions

In this paper we propose a geoadditive model for over-dispersed regional count data which allows one to simultaneously address some important methodological issues, such as spatial clustering, nonlinearities and overdispersion. As such, this framework can be used to model a variety of phenomena in both economic and non-economic fields.

We apply this model to analyze location determinants of inward greenfield investments occurred over the 2003-2007 period in 249 European regions. Our paper, therefore, contributes to the extensive literature on industrial location where spatial clustering and nonlinearities have been almost neglected.

Results suggest that multinational firms' location choices are spatially clustered: even controlling for several regional characteristics, such as employment density, market size, Jacobs externalities, human capital, labour cost, unemployment and density of transport infrastructure, the residuals of a semiparametric additive model turn out to be spatially auto-correlated. A Geo-NB-GAM, which incorporates a smooth spatial trend surface, is able to purge the errors from spatial dependence. The advantage of the inclusion of a spatial trend surface term is twofold. On the hand, it makes the model less prone to misspecification problems. On the other hand, it allows us to identify which clusters of FDI, after controlling for the most relevant variables, still remain. Not surprisingly, these clusters occur in those countries (the UK and Ireland) which have introduced several policy actions to attract foreign firms' investments within their regional development strategies.

The flexibility of the semiparametric approach also allows us to appreciate that whereas some regional characteristics have indeed a linear effect on FDI counts, others show important nonlinearities. In particular, in line with theoretical predictions, the effect of agglomeration economies appears to fade as the density of economic activities reaches

some limit value. However, this nonlinear relation does not seem to be well captured by an inverted-U shaped function: it is monotonically positive but the marginal effect decreases as agglomeration rises and, for a significant portion of our sample, the relation is flat. Therefore, no matter how dense the economic activity becomes, our data suggest that congestion (or competition) effects would never overcome positive agglomeration externalities.

To sum up, our results suggest that the use of flexible and general models has some clear advantages over standard parametric GLM traditionally employed in the analysis of the determinants of regional economic phenomena measured by count data. We showed the potentials of Geo-NB-GAM models in the case of industrial location, but the method lends itself to a range of application in the field of regional economics. Therefore, applications of this models to similar circumstances are most welcomed.

Appendix: definition of explanatory variables

- ***Market size***: log of total value added in the region in million Euros at constant 1995 prices (source: Cambridge Econometrics).
- ***Jacobs externalities***: median of the regional specialisation indexes defined as $S_{si} = (E_{si} / \sum_s E_{si}) / (\sum_i E_{si} / \sum_i \sum_s E_{si})$, where E denotes employment (source: Cambridge Econometrics), *s* the sector and *i* the region.
- ***Employment density:*** number of people employed in the manufacturing industry per km^2 (source: Cambridge Econometrics)
- ***Public infrastructure:*** length of highways and other roads network (in kilometres) divided by total population in the region (in thousands) (source: Eurostat).
- ***Tertiary education:*** share of adults (population aged 25-64) with tertiary education (ISCE97 codes 5 and 6) averaged over the 1999-2002 period. For the regions DE41 and DE42 data on tertiary education were available only for the years 2004 and 2005 (source: Eurostat)
- ***Labour cost***. total compensation (in thousand Euros, at constant 1995 prices) divided by the number of employees (in million). For German and UK, data are available only at the NUTS-1 level (source: Cambridge Econometrics.
- ***Unemployment rate***. percentage of unemployed people over total labour force (source: Cambridge Econometrics).

Since we estimated the effect of all these variables over the 2003-2007 period, we used (unless differently stated) all explanatory variables averaged over the 2000-2002.

References

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

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

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

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

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

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

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

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

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

Barry, F., H. Görg, and E. Strobl. (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. (2009). "Productivity polarization across regions in Europe: The Role of Nonlinearities and Spatial Dependence." *International Regional Science Review* 31, 92-115.

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

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

Basile, R. and A. Girardi. (2010). "Specialization and Risk Sharing in European Regions." *Journal of Economic Geography*, 5, 645-659.

Bivand, R. and R. Brunstad. (2006): "Regional growth in Western Europe: detecting spatial misspecification using the R environment." *Papers in Regional Science* 85, 277-97.

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

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

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

De Benedictis, L. and M. Tamberi. (2004). "Overall specialization empirics: techniques and applications." *Open Economies Review* 15, 323–346.

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

Fahrmeir, L., and L. O. Echavarrìa. (2006). "Structured additive regression for overdispersed and zero-inflated count data." *Applied Stochastic Models in Business and Industry* 22, 351–369.

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

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

Guimarães, P., O. Figueiredo, and D. Woodward. (2003). "A Tractable Approach to the Firm Location Decision Problem." *Review of Economic and Statistics* 85, 201-204.

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

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

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

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

Iannizzotto, M., and N. J. Miller. (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.

Kammann, E. E., and M. P. Wand. (2003). "Geoadditive Models." *Journal of the Royal Statistical Society, Series C* (Applied Statistics) 52, 1–18.

Kelvyn, J., and N. Wrigley. (1995), Generalized Additive Models, Graphical Diagnostics, and Logistic Regression, *Geographical Analysis* 27, 1-18.

Kogut, B., and S. J. Chang. (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 the pattern of trade." *American Economic Review* 70, 950-959.

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

LeSage, J., and R. K. Pace. (2009). *Introduction to spatial econometrics*, Boca Raton: CRC Press Inc.

Lin, G., and T. Zhang. (2007). "Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky Breast cancer data.", *Geographical Analysis* 39, 293-310.

McCullagh, P., and J. A. Nelder. (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.

Ottaviano G. I. P. and D. Puga. .(1998). "Agglomeration in the Global Economy: A Survey of the 'New Economic Geography'." *The World Economy*, 21(6), 707-731.

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

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

Thurston, S.W., M. P. Wand, and J. K. Wiencke. (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, 460-468.

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

Waller, L. A., L. Zhu, C. A. Gotway, D. M. Gorman and P. G. Gruenewald. (2007). "Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models." *Stochastic Research and Risk Assessment* 21, 573-588.

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

Wood, S. N. (2003). "Thin plate regression splines." *Journal of the Royal Statistical Society*, Series B (Statistical Methodology) 65, 95–114.

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.

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society, Series B* (Statistical Methodology) 73, 3-36.

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

Figure 1 – Regional distribution of foreign investments in the European Union

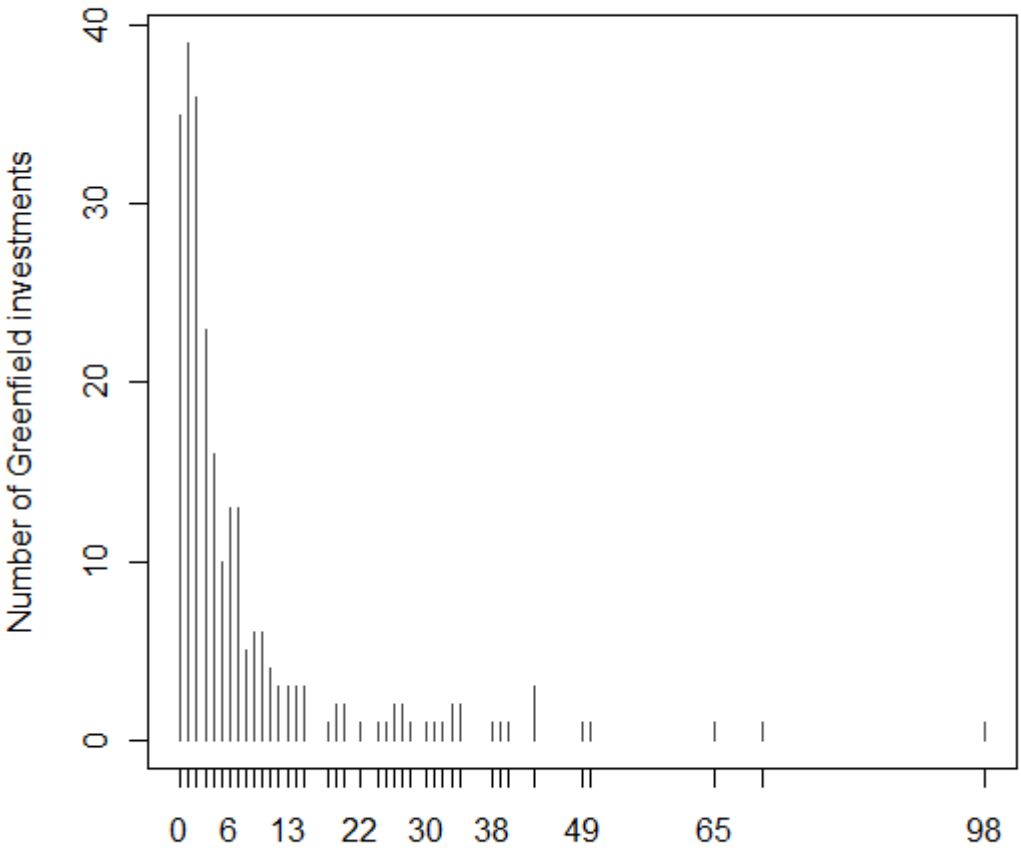
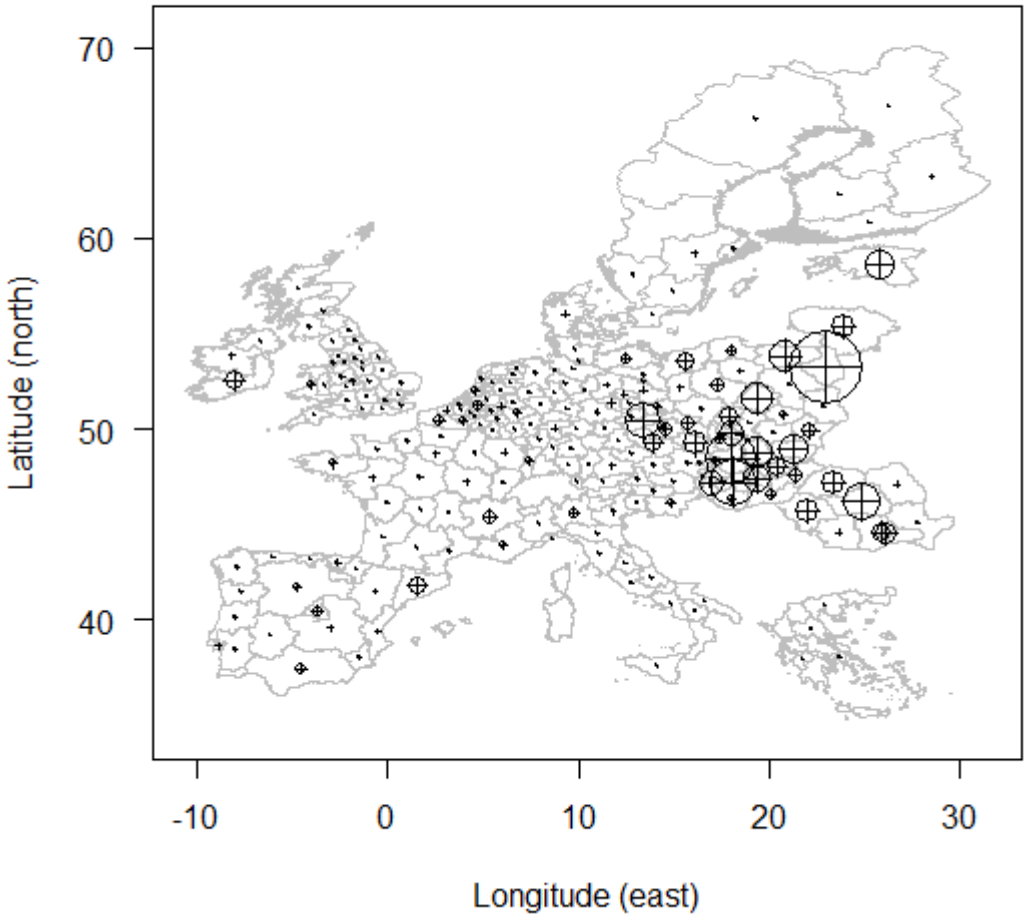
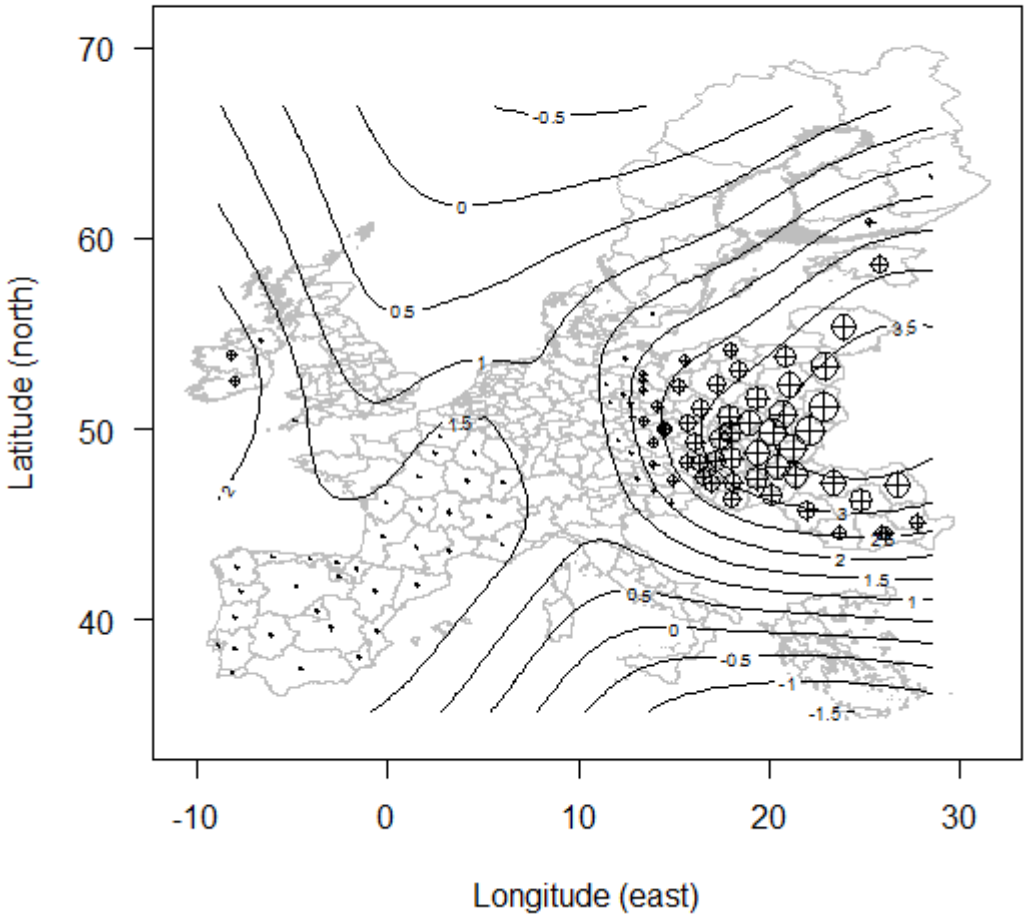


Figure 2 – Map plot of regional distribution of foreign investments in the European Union



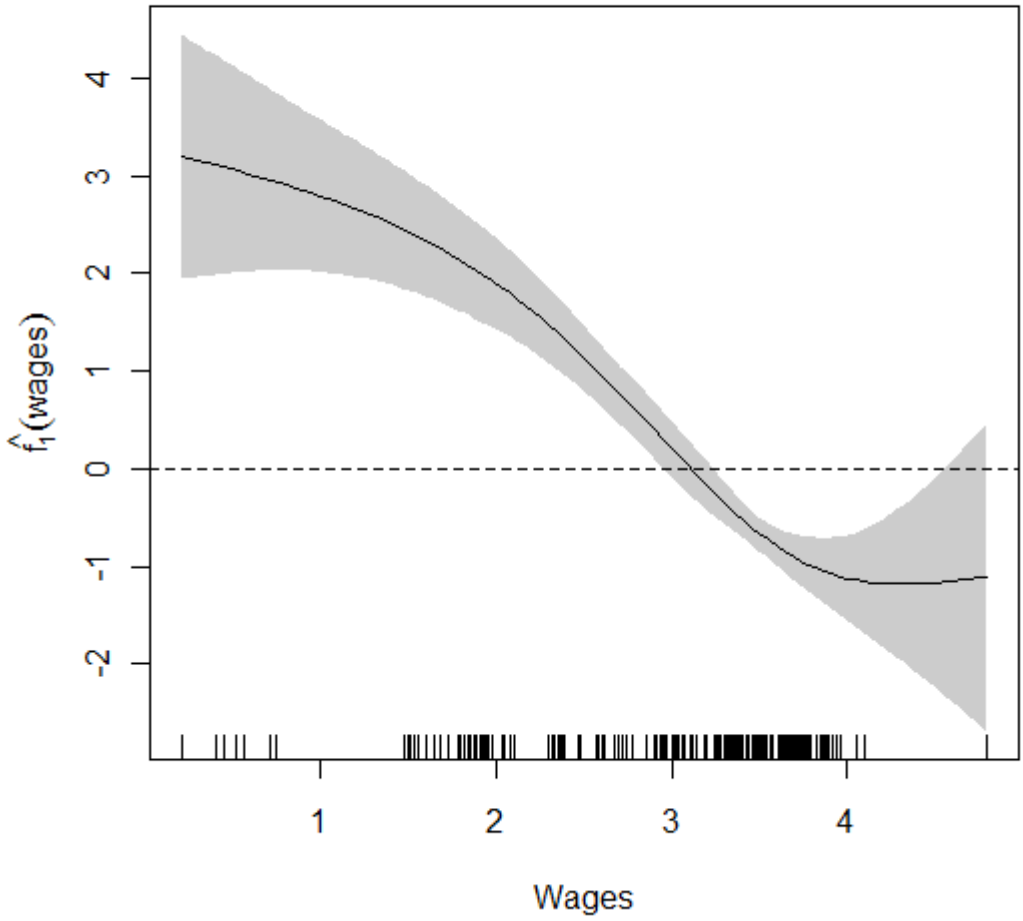
Note: Each circle in the plot, centred at the regional centroid, is proportional to the percentage of foreign investments attracted by the region on the total number of inward FDI in Europe. X- and Y-axis measure degrees of longitude and latitude, respectively..

Figure 3 – Smooth trend surface of foreign investments in the European Union



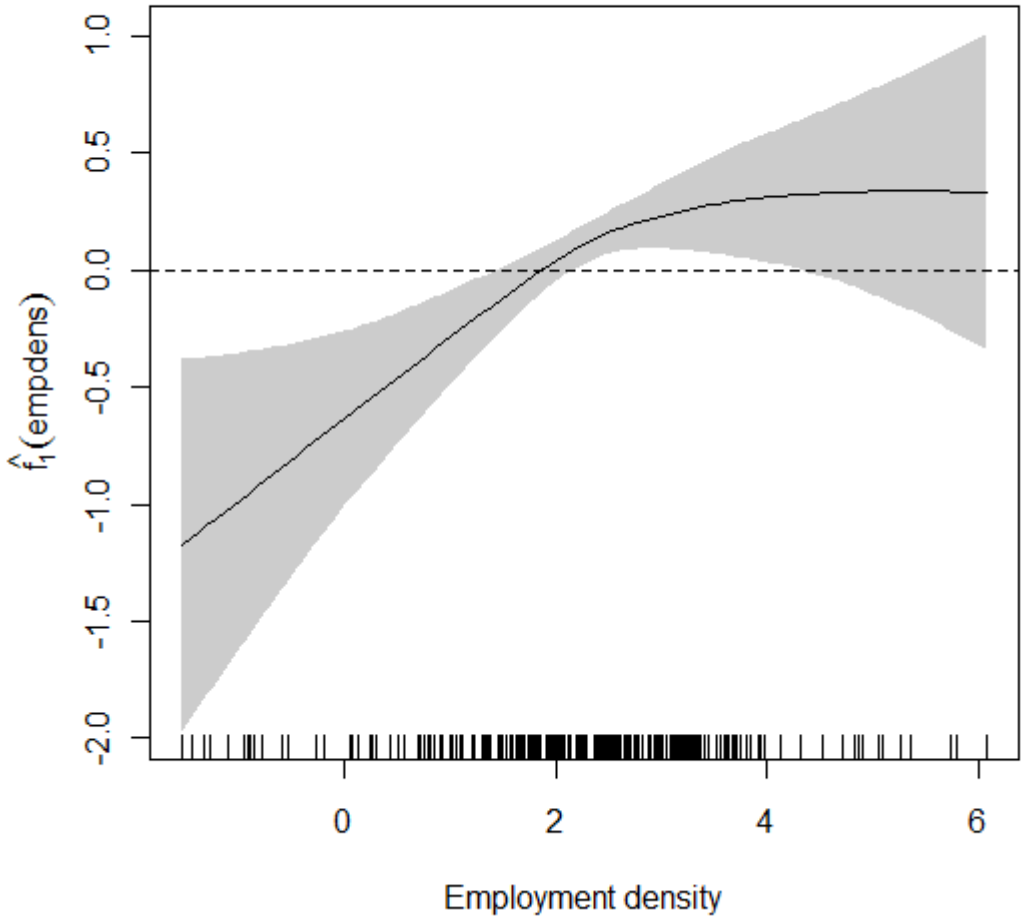
Note: Contour lines are drawn for different values of the predicted log of the number of foreign investments attracted by each region. Each circle in the plot, centred at the regional centroid, is proportional to the same predicted value. X- and Y-axis measure degrees of longitude and latitude, respectively.

Figure 4 – Smooth effects in the semiparametric Geo-GAM. Effect of wages



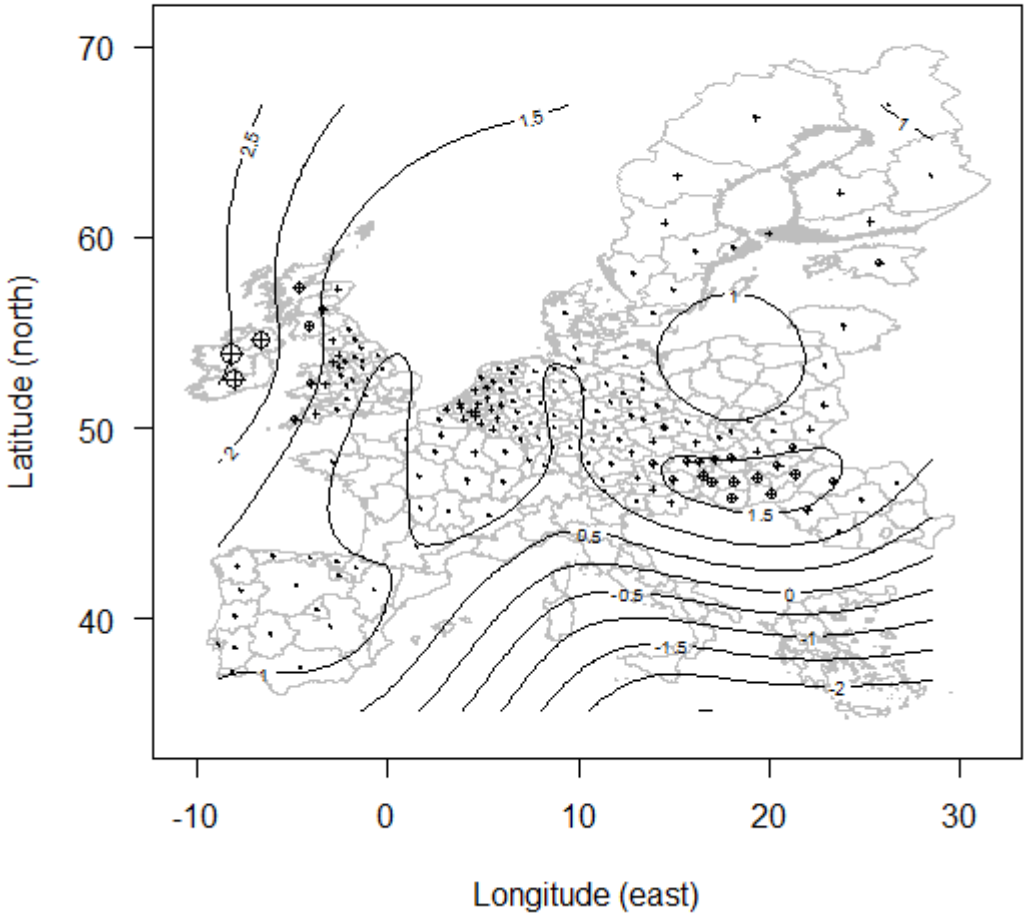
Note: Solid lines represent smooth functions of each term, alongside Bayesian confidence intervals (shaded grey areas) at the 95 percent level of significance. In each plot, the vertical axis displays the scale of the estimated smooth function, while the horizontal ones report the scale of each determinant. Rug plot along the horizontal axis represents observed data.

Figure 5 – Smooth effects in the semiparametric Geo-GAM. Effect of employment density



Note: Solid lines represent smooth functions of each term, alongside Bayesian confidence intervals (shaded grey areas) at the 95 percent level of significance. In each plot, the vertical axis displays the scale of the estimated smooth function, while the horizontal ones report the scale of each determinant. Rug plot along the horizontal axis represents observed data.

Figure 6 – Geographical components of geoadditive model



Note: Contour lines are drawn for different values of the predicted log of the number of foreign investments attracted by each region. Each circle in the plot, centred at the regional centroid, is proportional to the same predicted value. X- and Y-axis measure degrees of longitude and latitude, respectively.

Table 1 – GLM results

	Poisson	**Negative Binomial**
Variables	*Coefficients and std. errors (in parenthesis)*	
mkt	0.349*** (0.037)	0.522*** (0.098)
infra	0.329*** (0.033)	0.305*** (0.087)
Jacobs	1.263*** (0.151)	1.295*** (0.380)
empdens	0.422*** (0.030)	0.351*** (0.075)
wage	-1.232*** (0.046)	-1.576*** (0.146)
ur	0.520*** (0.042)	0.309** (0.130)
ter	0.452*** (0.067)	0.669*** (0.203)
	Diagnostics and goodness of fit (p-values in square brackets)	
AIC	2,078	1,335
Overdispersion	10.390 [0.000]	0.031 [0.860]
Moran's *I* (320 km)	4.805 [0.000]	4.565 [0.000]
Moran's *I* (920 km)	1.518 [0.064]	2.668 [0.004]
θ		1.183
Vuong		0.309 [0.378]
Wald-Hurdle		0.281 [0.998]

Notes: The overdispersion test is based on the *F* test of the overall ("approximate") significance of the smooth term $f(y)$ in the nonparametric regression of absolute residuals, $|res|$, on fitted values, y. $\hat{\theta}$ is the estimated NB shape parameter. Vuong is a non-nested hypothesis test statistic asymptotically distributed $N(0,1)$ under the null that the models ZINB and NB 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.

Table 2 - Nonparametric and semiparametric NB-GAM results

	Performance. iteration (GCV)	**Outer iteration (AIC)**	**Outer iteration (REML)**		
Parametric terms	*Coefficients and std. errors (in parenthesis)*				
Mkt				0.586*** (0.090)	0.665*** (0.093)
Infra				0.256*** (0.080)	0.150 (0.108)
Jacobs				0.938*** (0.342)	0.947*** (0.319)
Ur				0.267** (0.118)	0.485*** (0.135)
Ter				0.737*** (0.180)	0.289 (0.210)
Nonparametric terms	*F test and edf (square brackets)*	*χ^2 test and edf (square brackets)*			
f *mkt*	13.113*** [2.658]	47.432*** [2.810]	42.234*** [1.000]		
f *infra*	8.996*** [1.000]	10.852*** [1.000]	9.859*** [1.000]		
f *Jacobs*	7.164*** [1.000]	8.385*** [1.000]	7.375*** [1.000]		
f *empdens*	7.309*** [2.688]	25.344*** [2.668]	27.383*** [2.855]	27.290*** [2.851]	7.272** [2.261]
f *wage*	33.298*** [3.080]	132.983*** [2.999]	149.513*** [3.089]	149.410*** [3.085]	70.913*** [3.080]
f *ur*	4.735** [1.000]	8.087** [2.547]	5.178** [1.000]		
f *ter*	9.235*** [1.476]	17.350*** [1.426]	18.121*** [1.390]		
f *no,e*					76.079*** [19.057]
	Diagnostics and goodness of fit (p-values in square brackets)				
REML			652.4	656.6	637.5
AIC	1,308	1,306	1,307	1,307	1,244
Overdispersion	0.161 [0.689]	0.277 [0.599]	0.123 [0.726]	0.155 [0.695]	0.699 [0.404]
Moran's *I* (320 km)	4.279 [0.000]	4.072 [0.000]	4.150 [0.000]	4.399 [0.000]	0.645 [0.260]
Moran's *I* (920 km)	2.515 [0.006]	2.493 [0.006]	2.533 [0.006]	2.656 [0.004]	0.110 [0.456]
θ	1.672	2.070	1.867	1.862	3.035

Notes: *F* and χ^2 tests are used to investigate the overall ("approximate") significance of smooth terms. *edf* (effective degrees of freedom) reflect the flexibility of the model. An *edf* = 1 suggests that the smooth term can be approximated by a linear term. The overdispersion test is based on the *F* test of the overall ("approximate") significance of smooth term *f(y)* in the nonparametric estimation of absolute residuals, $|res|$, on fitted values, *y*. $\hat{\theta}$ is the estimated NB shape parameter.