Full Title:
Localization and industry clustering econometrics:
An assessment of Gibbs models for spatial point processes

Short Title:
Localization and industry clustering econometrics

Stuart Sweeney[1,2] **and Miguel Gómez-Antonio**[3]

[1]*Department of Geography, University of California, Santa Barbara, CA 93106-4060*

[2]*Institute for Social, Behavioral, and Economic Research, UC Santa Barbara*

E-mail address: stuart.sweeney@ucsb.edu

[3]*Department of Public Finance and Tax System, Complutense University of Madrid, Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid.*

E-mail address: mgomezan@ucm.es

keywords: industry clusters, localization economies, spatial point processes.

JEL classifications: C21, L63, R12, R15.

Acknowledgments: We would like to thank the participants in the 55th European Regional Science Association and the 89th Western Economic Association International conferences and three anonymous referees for their valuable comments and suggestions. Miguel Gmez-Antonio acknowledges financial support from the ECO2012-36032-C03-01 research funding and Stuart Sweeney acknowledges support from National Science Foundation award BCS-0454993.

Abstract

The objective of this paper is to assess an approach to statistical modeling of point referenced establishment data that permits inclusion of "environmental" or establishment-specific covariates and specific forms of inter-establishment interaction. Gibbs models are used to decompose the conditional intensity of the spatial point process into trend and interaction components. The trend is composed of access measures (primarily different classes of roads) and three different interaction processes are tested: Geyer, Area interaction, and Strauss hard core. While the models used have proved to be useful in ecology, we are unaware of any applications to establishment or firm data. In empirical application the models yield intuitively appealing results for the trend component, and the ability to specify the interaction component gives deeper insights into inter-establishment spatial dynamics than any previously published methods.

1 INTRODUCTION

While theories that motivate why and under what conditions establishments will co-locate are well established, the ability to capture the richness of the theory and make valid comparisons among different industries in different contexts by empirical models has lagged behind (Ellison et al., 2010). Different methods have been used for testing the determinants of agglomeration, however, existing research is only partially successful in isolating the practical significance and the character of externalities. The wide range of possible factors behind observed clustering patterns makes empirical research difficult.

Recent empirical studies continue to rely on rather rudimentary measures, such as the Location Quotient, spatial Gini, and the Ellison and Glaeser (EG) index. These measures are fine for coarse-resolution studies but they are not well suited to capturing the most salient aspects of industrial clustering for several reasons[1] (see Sweeney and Feser (2004)). If the primary goal of analysis is to understand spatial patterns and behavior at sub-metropolitan scales, to design government policy and public investment, then the broad range of techniques emerging from the spatial point pattern analysis literature will yield insights that are more likely to isolate the true scale of the process. Observed co-location is not enough to conclude that localization economies are driving the observed pattern; co-location may occur without linkages or interaction between proximate firms (Gordon and McCann, 2005; Torre and Rallet, 2005; Yeung et al., 2006). When preliminary tests indicate that firms in a given industry appear to co-locate, it would be useful to identify to what degree clustering is due to features of the environment (such as access to roads or other public amenities) or to direct interaction among firms (such that firms choose to locate near each other).

Models based on Markov ("Gibbs") point processes are used here to better isolate and characterize the nature of firm co-location, as a step toward developing more refined tests of localization economies. Gibbs models of a spatial process can be specified using two components: A first-order effect that captures environmental inhomogeneity and a second-order effect that measures the dependence or interaction between the events.[2] If we take Hoovers 1937 definition[3] as our measurement theory, then the interaction component of the Gibbs model can be interpreted as evidence of the existence of localization economies. It would then take additional qualitative or quantitative analysis to attribute the localization to specific forms of co-operation among firms;

for example, joint production, subcontracting, shared purchases, or the use of the same marketing campaigns, after-sales services or R&D activities. Although that second level of analysis is not discussed in this paper, Gibbs models provide a major step forward in isolating and characterizing the nature of spatial co-location based on observational data.

This paper contributes methodologically to the analysis of industry concentration by testing the simplest non-trivial model to separate the two theorized sources of agglomeration. We introduce a new approach in the industrial cluster literature to continue verifying and exploring results from a different angle, increasing the ratio of rigorous empirical findings to theoretical models in the study of business clusters. This paper is largely concerned with methodology but we also provide a simple empirical model, for the electronics industry in Madrid, a sector where the existence of spillovers is known in advance. Additionally, we include an appendix with simulation results assessing properties of the estimators under different specification errors and for both Euclidean space and network space.

To check for the existence of spillovers [4] three different interaction processes are tested: Geyer saturation, Area Interaction, and Strauss hard core. The models yield sensible results for the trend component that confirm aspects of location behavior known from prior studies. The ability to evaluate alternative interaction specifications provides information about the intensity and range of spatial interaction under specific functional forms. The Geyer-saturation form of interaction is the one that provides the best overall fit to the observed pattern of establishments. While the form of interaction is not fully exploited, future research will benefit by deriving and evaluating mathematical forms of interaction that are most likely to inform economic theory.

The paper proceeds as follows. Section 2 provides a review of techniques used to study industrial clustering with point-referenced establishment data. We conclude that the range of current techniques share a set of limitations that are better addressed using the Gibbs modeling framework. Section 3 contains a thorough explication of Gibbs models including the statistical theory and currently available approaches for model specification testing and diagnostics. The methods are then applied to the electronic sector in Madrid with a complete treatment of model diagnostics and interpretation. The electronic sector is chosen because the presence of spatial clustering and spillovers known to exist. Concluding remarks are presented in Section 5.

2 INFERRING CO-LOCATION FROM SPATIAL POINT PATTERNS: PRIOR APPROACHES

In the domain of industrial cluster analysis, it was only within the past two decades that methods for characterizing spatial point patterns – largely developed for applications in ecology – have been adopted to study industry patterns. To date, there are four approaches that have utilized point geo-referenced data to study industrial clusters in the presence of spatial inhomogeneity: (1) D-function, (2) K-density, (3) M-function, and (4) inhomogeneous K-function. All these methodologies are exploratory and are grounded in a case-control strategy. That is, establishments in one industry or value-chain are defined as the "case" group, and their spatial distribution is compared to the spatial distribution of establishments not in that industry or value-chain (the "control" group). The definition of the control group can also include matching along strata (e.g. firm size) to remove the possible influence of characteristics that are known correlates of location choices. Localization economies will manifest as spatial clustering in the case group that is greater than the spatial clustering in the control group. While the methods differ slightly in approach, they all attempt to answer the same basic question: "Is industry X characterized by more (or less) spatial co-location than a set of control industries?"[5]

A review of the literature in which these methods have been developed, refined, and applied is provided below. This serves to document that, while this literature is fairly new (most of the papers are from the last decade), there has been relatively little innovation in the approach, and the distinction is primarily by field of application or origin, rather than inherent advantages of one method over another.

D-function: this approach is based on Ripley's K function, which is a cumulative function measuring the expected number of points of the pattern within a distance, s, of any given point. If the value of K(s) for the "case" industry is greater than the K(s) for controls it indicates clustering at that scale. The function is evaluated for specific values of s over a plausible domain of interfirm interaction. Statistical significance of clustering (or dispersion) is assessed using resampling. Sweeney and Feser (1998) were first to adapt the D-function to the analysis of industrial clustering following on Barff's 1987 use of K-functions to study establishment locations in Cincinnati, Ohio. Sweeney and Feser

(1998, 2004), Sweeney and Konty (2005), and Feser and Sweeney (2000, 2002a,b) constitute a corpus of work using the D-function to study aspects of U.S. manufacturing (14 metropolitan regions and 12 value chains) and to assess the D-function as an inferential framework (leverage/outlier analysis, unbiased estimation under non-ignorable spatial censoring, comparing point-based measures to area-based measures, employment weights versus employment strata) for industrial cluster analyses. The "control" group under this framework is based on stratified random sampling of industries in the study region to match characteristics (employment size, for example) of the industry in question. Subsequent research using the D-function includes: Marcon and Puech (2003) application to Paris, France; Kosfeld et al. (2011) applied to Germany; Albert et al. (2012) and Casanova and Orts (2011) applied to Spain. Other papers use extensions or approaches closely related to the D-function. Arbia et al. (2008) use bivariate K-functions to study inter-sectoral location of patents innovations within six industrial sectors in Italy. Carlino et al. (2012) developed an approach that formally identifies clusters and yields visual representations of the concentration of firms, capturing the relative shape, size and hierarchy of the concentrations. Arbia et al. (2010) analyze the location dynamics of firms that belong to the Information and Communication Technology sector in Rome. This paper simultaneously analyzes the spatial location of plants together with their temporal trends and their space-time interaction.

K-Density function:Duranton and Overman (2005, 2008) develop and apply an alternative approach based on the kernel-smoothed frequency distributions of pairwise distances between plants. They compare whether the number of plants at a given distance is significantly different from the number that would have been found if the location of the firms was random. They define the K density function (Kd) and analyze different sector aggregation (branches, sectors, industries, and sub-industries) of the manufacturing industries in the U.K. and, subsequently, localization within subgroups of an industry. This approach constructs the control group slightly differently and does not use edge corrections as is common in the spatial statistics literature. The method has enjoyed widespread adoption with more than 450 citations of the original 2005 article. Subsequent recent applications and extensions include Klier and McMillen (2008) for the U.S. auto industry; Nakajima et al. (2012) for the Japanese service sector; and Billings and Johnson (2012) construct an index of industrial specialization and compare it to results from K-densities.

M-function: A third approach, termed the M-function, was proposed by Marcon and Puech (2010), generalizing Ripleys K-function to inhomogeneous space. This approach develops a cumulative function that counts neighboring points up to a chosen distance s. The relative weight of a sector is compared with all industrial activities in a circle of radius s, while also accounting for the size of the sector relative to all activities in the study region. They do not implement the approach for real data, but simulate a series of patterns to demonstrate that Duranton and Overman's Kd and M function give pertinent and complementary information on the spatial structure. The M-function has not yet attracted similar attention in the literature, perhaps because it is relatively more recent and apparently has few advantages over the K-density approach.

Inhomogeneous K-function: It is essentially a generalization of Ripley's K-function to the case of non-stationary point processes. Arbia et al. (2012) is the only known paper that has applied inhomogeneous K-functions to assess spatial concentration of five sectors of high-tech manufacturing in the area of Milan, Italy. Dividing each sector of activity according to firm size, they conclude that small firms derive higher benefits than large firms from network linkages to other firms.[6]

We note that, in addition to research applying the four methods above, there is also a small but growing literature that compares among a subset of the four measures, or compares one of the point-based measures to area-based measures. For example, Sweeney and Feser (2004) compare D-function results (for Atlanta and Los Angeles) to several different area-based indicators, including Ellison and Glaesers index, under different levels of area aggregation. Vitali et al. (2013) and Koh and Riedel (2014) compare results from K-densities to the Ellison and Glaeser index. Funderburg and Zhou (2013) use Los Angeles establishment data and compare results for D-functions and K-densities for a large number of industries. They found that results are remarkably similar for most of the value chains, and their results match with prior published work on Los Angeles (Sweeney and Feser, 2004), and that the primary difference is that the K-density appears to yield slightly more conservative test than the D function.

As noted at the outset of this section, the four methods are broadly similar in approach and – while it is possible to methodologically critique one approach vis-a-vis another – in practical application, the methods will yield broadly comparable findings. Because of their similarity, they

also suffer a common set of limitations related to the case-control framework: (1) The validity of the results relies entirely on the control group definition. As noted in Feser and Sweeney (2002a), the controls should closely conform to the cases on all theoretically important dimensions so that only un-measurable factors related to localization remain. However, those "factors" must be chosen a priori and they are not evaluated within the modeling framework. (2) Because they are not evaluated, we cannot recover effect sizes for what are putatively important theoretical factors, and indeed it is implicitly assumed that those factors affect all industries (or value chains) equally. This means the analysis basically yields a series of isolated univariate comparisons and, while it is possible to claim that first-order effects are controlled, nothing is learned about their relation to the industry under study. (3) The counterfactual in these tests is frustratingly non-specific – simply that "case" industries exhibit more co-location than "control" industries. They are not constructed to test against specific (mathematical) forms of clustering. This means that it is difficult to qualitatively compare results among industries or different study regions. (4) First-order effects and second-order effects may be confounded. If firms in the "case" industry are co-locating because of proximity to a feature of the environment – for example, access to a unique transport hub/terminal or to a port – that is not valued by "control" industries, then the descriptive methods will incorrectly attribute the clustering to interaction among firms.

The control group approach is a blunt tool that only measures the composite tendency of all firms, perhaps stratified by similar size, to agglomerate in urban areas. If the goal of understanding clustering behavior of firms is to design government policy and public investment – to selectively encourage and target sectors where productivity advantages are the greatest – then methods are needed that yield more information than simple one-way (univariate) tests of whether one sector clusters relative to a control group. More information can be gained if we could decompose the location choice into specific features of the urban environment that are valued by a particular industry sector and compare it to other industry sectors. For example, to what degree does the observed location pattern for a particular industry/sector reveal a tendency to locate near a ring road or near other specific types of infrastructure? Similarly, the ability to evaluate alternative forms of interestablishment interaction – as a form of residual/unclassified localization tendency – would allow analyst to develop empirical profiles of different industries beyond the current one-way tests of clustering relative to a control group.

3 METHODOLOGY: GIBBS MODELS OF INTRAURBAN INDUSTRIAL LOCATION

Gibbs models have several advantages over the methods reviewed in the previous section and future research on intra-urban industrial location, generally, and localization economies, specifically, should adapt and extend the basic framework presented here. While the approach does not solve all of the challenges inherent in empirical analysis of localization, the models yield a far richer set of results than prior methods, and move from the simple case-control hypothesis testing to complete model specification and validation that forms the basis for most empirical research in economics and regional science. If studying a set of industries that benefit from proximity to features of the built environment, and potentially benefit from proximity to other firms in the same or a different sector, then there is a need for statistical models that can provide estimates of those functionally different components of location behavior. The specific advantage of Gibbs models is to provide a regression framework that takes point-referenced data as input, and allows for separate estimation of effect sizes on components of the trend ("first-order effects") and specific representation of the interaction ("second-order effects").[7] Each industry has different operational characteristics and locational requirements, so the magnitude of effect sizes for covariate effects in the 'trend' and the form and magnitude of interaction should differ among industry sectors. Gibbs models constitute an appropriate approach that can solve the identification problem, isolate the existence of spillovers, and at the same time determine which covariates are most effective at describing the location pattern for each sector.

We loosely frame the analysis in terms of urban economics theory. The parametric trend component can be considered a function of geographically referenced environmental covariates informed by bid rent theory. The resulting parameters reflect the importance of access to features in the urban environment. For example, this may include distances to roads or other transportation infrastructures, the presence of specific community characteristics (occupational profiles of residents), or proximity to location-specific public amenities provided by local governments. The trend can also capture some forms of local spillovers – for example, distance to a university for knowledge intensive industries. The firm-to-firm interaction structure is theoretically driven to the extent that the specific functional form encodes relevant economic concepts such as increasing then decreasing

returns to external scale. We provide a discussion of the economic basis of the interaction structures in section 3.1.4.

Below a review of Gibbs process theory, estimation, and model evaluation/validation is provided, followed by a description of the data used in our case study.

3.1 Gibbs models

A finite Gibbs process, X, has probability density f(x) defined with respect to the standard Poisson process of the form:

$$f(x) = exp(V_0 + \sum_{x \subset X} V_1(x) + \sum_{\{x,y\} \subset X} V_2(x,y) + \dots) \tag{1}$$

where V_k are kth-order potentials with V_0 a normalizing constant, V_1 capturing spatial trend[8], and $V_k \geq 2$ capturing interaction. In using this approach, it is assumed that the observed spatial distribution of firms is a realization of the point process, (1).

In section 4.3 several methods are presented for evaluating whether this assumption is reasonable. The natural way to proceed would be to define the likelihood and then derive estimators for the unknown parameters of the trend and interaction components of the process density, $f(x)$. However, it is possible to directly determine the probability density for only some elementary spatial point processes. As explained in Turner (2009) the density function $f(x)$ can be written as $\alpha g(x)$, where $g(x)$ is expressed in terms of model parameters and statistics calculated from the observed pattern of points X. The normalizing constant α must have the property that $f(x)$ integrates to 1,

$$\alpha = \left[\int_{\mathcal{X}} g(\mathcal{X}) d\mu(\mathcal{X}) \right]^{-1}$$

where $\mathcal{X}$ is the space of all point patterns in a bounded region W of space and μ is an appropriate measure on $\mathcal{X}$. Because there is no natural ordering in two or more dimensions, it is typically impossible to analytically derive the normalization constant α for even the simplest (non-Poisson) Gibbs process. This is particularly the case for models involving interaction because α will be a multiple integral function of model parameters.

It is possible to develop a class of estimable Gibbs models by working with the Papangelou conditional intensity function, $\lambda(u, x)$. The Papangelou conditional intensity function of a process is related to the probability density function by,

$$\lambda(u, x) = \frac{f(x \cup \{u\})}{f(x \setminus \{u\})}$$

where $x \cup \{u\} = x$ if $u \in x$ and $x \setminus u = x$ if $u \notin x$. The models become tractable because the normalizing constant α, cancels out of the expression because it is common to both the numerator and denominator. Looking toward the model interpretation under this reformulation,$\lambda(u, x)du$ is the probability of observing a point u of the process in a small neighborhood du of u, conditional upon the rest of the process x.

Thus, the finite Gibbs process can be expressed instead as a conditional intensity[9],

$$\lambda(u, x) = exp(V_1(u) + \sum_{x \subset X} V_2(u, x) + \sum_{\{x,y\} \subset X} V_3(u, x, y) + \dots) \quad \forall \quad u \notin x \tag{2}$$

The trend component, $V_1(u)$, depends only on the spatial location u, and reflects spatial inhomogeneity in the process. Covariates, z, can be incorporated in the model trend,

$$V_1(u) = \alpha + b_1 z_1(u) + b_2 z_2(u) + \cdots + b_n z_n(u) = z(u) b^T$$

and could include any of the accessibility concepts informed by urban economic theories of intra-urban industrial location.

Higher order potentials, V_k, are included to capture interaction. In applied modeling, it is assumed that interaction can be suitably approximated using only the first, $V_k(.)$, of the $k > 2$ potentials.

The interaction component provides another set of rich specification choices. Only three of the functional forms of interaction that have appeared in applied work (Mateu, 2002) are reviewed here: Strauss hard core, Geyer saturation, and Area/penetrable spheres. The interaction forms may have both canonical parameters and irregular parameters. Canonical parameters are estimated as part of equation (2) directly, whereas irregular parameters are estimated in a separate step; details will

be provided in the model estimation subsection.

3.1.1 Strauss hard core interaction

The simplest form of interaction models use a pairwise interaction process on W with trend b_θ and interaction function h_θ, resulting in the Papangelou conditional intensity:

$$\lambda_\theta(u, X) = b_\theta(u)\Pi_{i=1, x_i \neq u}^{n(X)} h_\theta(x_i, x_j) \tag{3}$$

Note that this is discontinuous at the data points x_i. Also, to ensure that the conditional intensity is well defined and integrable, interaction among pairs of points must be symmetric, $h_\theta(x_i, x_j) = h_\theta(x_j, x_i)$ (Baddeley and Turner, 2000). Pairwise interaction models are primarily used for modeling repulsive processes and, therefore, would appear to be of little value for the study of localization. However, there is one form–the Strauss hard core model–that allows for attractive interaction.

The Strauss hard core interaction function is defined as:

$$h_\theta(x_i, x_j) = \begin{cases} 0 & \text{if} \quad 0 \geq |x_i - x_j| < r_1 \\ \gamma & \text{if} \quad r_1 \geq |x_i - x_j| < r_2 \\ 1 & \text{if} \quad |x_i - x_j| \geq r_2 \end{cases}$$

Irregular parameters are r_2 (the radius of circle) and r_1 (the hard core distance). Values of $\gamma > 1$ yields a clustered process in the range, $[r_1, r_2)$, and a Poisson process (no interaction) beyond pairwise distances of r_2.

It would appear that the Strauss hard core model should provide a sensible means of modeling attraction, however, in practice it is highly unstable. As noted by Hjort et al. (1994, see Møller's comment) and Geyer and Thompson (1995), the Strauss hard core process is a poor model for clustering due to the following "phase transition property": For positive values of the interaction parameter, except for a narrow range of values, the distribution will either be concentrated on point patterns with one dense cluster of points or on "Poisson-like" point patterns.

3.1.2 Geyer saturation interaction

Geyer (1999) derived this model as a modified Strauss process in which the total contribution to the potential from each point is trimmed to a maximum value (d) to effectively control the size of the clusters. Geyer's saturation process has interactions of infinite order. The conditional intensity $\lambda(u, x)$ for the Geyer point process for $u \notin x$ is:

$$\lambda_\theta(u, X) = b_\theta(u)\gamma^{min\{d, N_x(u)\}} \tag{4}$$

where b_θ controls the intensity of the point process X, γ is the interaction parameter, d is the saturation threshold (an upper bound on the contribution to the conditional intensity of any single point), and $N_x(u)$ is the number of neighbors in X of the point uThe interaction parameter has interpretations, $\gamma < 1$ indicating repulsion, $\gamma > 1$ clustering, and $\gamma = 1$ the Poisson case.

3.1.3 Area ("penetrable sphere") interaction

The area interaction process, also known as the Widom-Rowlinson "penetrable sphere model", is constructed to allow for well-behaved attractive processes. It has conditional intensity,

$$\lambda_\theta(u, X) = b_\theta(u)\gamma^{-[A(X\cup\{u\})-A(X)]} \tag{5}$$

where b_θ controls the intensity, $A(x)$ is the area of the union of discs of radius r centered at x_i, and γ is the interaction parameter. The difference $A(X \cup \{u\}) - A(X)$ is the area of that part of the disc of radius r centered on u that is not covered by discs of radius r centered at the other points $x_i \in X$. The process is well behaved and the density function is integrable for all values of $\gamma > 0$ and for all compact $W \subset R^2$. It reduces to a Poisson process when $\gamma = 1$, produces cluster when $\gamma > 1$, and exhibits inhibition for $0 < \gamma < 1$.

3.1.4 Economic interpretation of interaction specifications

The three alternative forms of interaction evaluated in this paper have been used in the applied statistics literature but not in economic applications. Each form of interaction has an implicit economic interpretation even if it was not developed with industry location in mind. All of the

interpretations provided here assume that the first order effect is specified approximately correctly – in the sense that in all econometric modeling we strive to include the important covariates. Also, we assume that increased economic benefit is synonymous with increased probability (or conditional intensity) as measured by the interaction term.

The Strauss hard core model is the most peculiar in economic terms. The model stipulates that benefits to co-location are strictly physically bounded within the distance range $[r_1, r_2)$ and increase at rate γ for each new close-neighbor entrant. This runs counter to economic intuition in two ways. First, there are constant marginal returns to external scale (co-location). This is exactly what causes the tipping behavior in the process when γ is positive and not close to zero. Second, one would expect returns from co-location to decline somewhat smoothly with distance of separation. The rigid range of interaction distance might be indicative of some kind of club or district such as a free trade zone such that within distance range $[r_1, r_2)$ there is benefit and outside it there is none.

The Geyer saturation model improves on the hard core model in two ways. The saturation term, d, limits the amount of localization benefit that a firm can accrue from having other firms locate nearby. In practice, this means that diminishing returns arrive all at once after a ceiling is reached. If the saturation parameter is 3, then the model is stipulating that there are localization benefits related to 1, 2, and 3 "close neighbors" with marginal return proportional to γ – but after the saturation threshold any additional firms locating within r confer no additional benefit. Second, there is no 'hard core' boundary r_1.

For the area interaction model, the localization benefit accrues in proportion to the area of overlap defined by circles of radius r on pairs of points. Thus the radius r is again a threshold beyond which no benefit is present and would link to some idea of physical limits on the ability to interact. The additional dependence that scales with the area of overlap resonates nicely with notions of labor pooling. Shared areas of overlapping circles of size r could be interpreted as shared access to occupations within a commutershed, for example.

3.2 Estimation

Model estimation leverages existing software implementations for estimating generalized linear (or additive) models. To do this, the conditional intensity is written in log-linear form:

$$\lambda(u,x) = exp\{\varphi^T b(u) + \theta^T S(u,x)\} \tag{6}$$

where φ and θ are the canonical parameters and may be vectors of any dimension corresponding to the dimension of the vector-valued statistics $b(u)$ and $S(u,x)$, respectively. The first term $\varphi^T b(u)$ is the trend component of the conditional intensity and the second term $\theta^T S(u,x)$ is the interaction component and may include embedded irregular parameters that must be set prior to estimation.

The likelihood function for inhomogeneous spatial patterns is computationally expensive because of the increase in parameter dimensionality and complexity of simulation. Baddeley and Turner (2000), building on work by Berman and Turner (1992), have developed and encoded computationally feasible maximum pseudo-likelihood estimation (MPLE)[10] for conditional intensity models of the form (6). While MPLE is less efficient than MLE, it is adequate in many practical applications and it is encoded in R as part of the Spatstat package (Baddeley and Turner, 2005). Spatstat does not provide the variance-covariance matrix for models fitted by MPLE unless they are Poisson point processes because for other generating processes the asymptotic variance-covariance matrix cannot be calculated as the inverse of the observed Fisher information. Spatstat also includes an alternative estimator devised by Huang and Ogata (1999) that improves on MPLE, approaches the efficiency of MLE, and yields an estimate of the asymptotic variance-covariance matrix for the canonical parameters in the point process model. A Monte Carlo estimate of the Fisher information matrix is calculated using the results of the original fit.[11]

In our application, the models with Geyer interaction are fitted by the procedure of Huang and Ogata (1999), while the models with Strauss hard core and Area interaction are estimated by the method of maximum pseudo-likelihood. When the Huang-Ogata method was implemented for the latter two models, the estimated coefficients exploded so only the maximum pseudo likelihood procedure could be implemented. The approach described in Coeurjolly and Lavancier (2013) and Coeurjolly and Rubak (2013) was followed to provide t-statistics. The specification diagnostics described in the next section provide a rich set of alternative approaches to assessing the validity

and fit of both the trend and interaction components. These are complex models and the nature of model building and assessment necessarily moves beyond simple inferential tests on model parameters.

As mentioned in the previous section, the interaction structures often have irregular parameters in addition to canonical parameters. The irregular parameters are estimated using profile likelihood and there is some degree of iteration between specifying the covariates in the trend and using profile likelihood to search over a grid of irregular parameter values while holding the set of trend covariates fixed.[12]

3.3 Model diagnostics and validation

Recent work by Baddeley et al. (2005, 2011) exploits the GLM framing of (6) to adapt and extend model diagnostics and validation measures from GLMs to the point process setting. The value of their work is in their definition of innovations and residuals in a spatial point process setting. The residuals are expressed:

$$R_{\hat{\theta}}(B) = n(x \cap B) - \int_B \hat{\lambda}(u, x)du$$

where $B \subset W$ and $n(x \cap B)$ is the number of observed points in B, and the second term is the cumulative estimated conditional intensity in B. Once residuals are defined, it is possible to develop measures and visual diagnostics that can be used to assess the fit of the model to the data. Separate measures are required to assess the trend and interaction components of the models.

Two diagnostics were used to validate the trend of a fitted model: a lurking variable plot of each covariate and a contour plot of the smoothed residuals. The lurking variable diagnostic for covariate, z, plots the cumulative sum of residuals over the range of z against z. It is possible to incorporate weights but the focus here is on the raw residuals. Over the full domain of z the residuals should sum to zero, and the expected value is zero at each point of the cumulative sum. Positive (negative) deviations of the measure indicate under- (or over) prediction and the relative importance of those deviations can be assessed against two standard deviation error limits at each value of z.[13] The smoothed residual plots provide a more general sense of where in the domain of W a model is under- or over-predicted. The smoothed raw residuals compare a kernel-smoothed value of the point pattern to the kernel-smoothed parametric estimate of the conditional intensity.

Again, over the domain of W the raw residuals should sum to zero and a model that fits the data should yield a residual plot that presents a relatively flat surface of under- and over-prediction dominated by small deviations from zero.

Two specification diagnostics are also used to validate the inter-point interaction component of the models: the QQ-plot and the G-compensator. The QQ plot compares the empirical quantiles of the smoothed residual field to the corresponding expected empirical quantiles of the residuals under the fitted model estimated by Monte Carlo simulations. If the pattern is more clustered than the model, then the empirical distribution of the smooth residuals should have heavier tails on the left-hand side than the reference distribution; if the observed pattern is more inhibited than the model, then the empirical distribution will have lighter tails on the right-hand side (see Baddeley et al. (2005)).

Baddeley et al. (2011) propose a global diagnostic of interaction connected to the score test. The numerator of the test is the difference between the usual nonparametric estimates of the nearest-neighbour distance distribution, G-function, based on the data alone and the compensator of the G-function. The compensator is the expected value of the G-function under the estimated model. Both measures should be approximately equal if the model fits the data. The denominator of the score test is the surrogate standard deviation (Poincaré standard deviation) residual G-hat function under the fitted model. The bands in the graphs are approximate point-wise critical values for the score test based on fixed r, as the exact null distribution of the standardized residuals is not known.[14] This diagnostic is conservative for small distances because the Poincaré variance is a substantial underestimate of the true variance. For small distances there are small-sample effects so that a normal approximation to the null distribution of the standardized residuals is inappropriate.[15]

Assessment of the overall fit of models that include interaction requires use of methods that are not specific to trend or interaction. Note that, because canonical parameters are all estimated together, a change in the interaction specification may alter the estimated coefficients and standard errors of the trend. Similar to other regression contexts, Akaike Information Criterion (AIC) can be used to evaluate the information loss and compare models that are based on the same data but have non-nested specifications. However, in the tests reported here, the AIC is based on pseudo-likelihood and the penalty includes the number of canonical and irregular parameters in the

model. Another common approach in the spatial point process literature is to compare a summary function (such as the L-function – a transformation of the K-function to ease interpretation) to the L-function based on point pattern realizations of the fitted model. A final crude but useful assessment of fit is to compare the average number of points produced by realizations of the fitted model to the points in the observed data. While all these methods provide some information, if they indicate lack of fit there is no information to identify whether it is a problem with the trend or with the interaction. This is precisely why the measures and diagnostics based on residuals are of value.

4 APPLICATION: ELECTRONICS SECTOR IN MADRID, SPAIN

The test data used in this paper is from the electronics sector in Madrid, Spain. This sector was selected for several reasons: (1) by concentrating on a single narrowly defined sector– thus isolating the particular production characteristics required – the problem of unobserved inhomogeneity is reduced; (2) regional growth theories predict that clustering will be particularly strong among high-technology or knowledge-intensive sectors and technology plays a very important role in the electronics sector; and (3) several studies using different methodologies primarily based on representative surveys from Madrid's electronics – firms conclude that there is a high degree of interaction and local interconnectedness in the electronics sector [16]. We can thus evaluate our model results using an industry that has already been the focus of extensive research and that is known to have a high degree of inter-firm interaction.

4.1 Data

The test data utilized here is from the *Statistics Institute of Comunidad de Madrid*. The basic input to our model requires the point locations of establishments and a set of environmental covariates. The primary goal for this paper is to demonstrate the feasibility and utility of Gibbs models for studying industry location. Therefore relatively few covariates are used in the models reported here and our future publications will focus on models with a more fully developed trend component.[17] One covariate that measures the distance to the center and four covariates related to accessibility are included[18]: (1) distance to ring road M40, (2) distance to ring road M-50, (3) distance to a radial-

type road (R-2 to R-5), and (4) distance to a motorway-type road (A-1 to A-6). The covariate *distance to the center* is a proxy measure of a firm's access to business services and their labor pool. In Madrid, a more central location increases access to clientele, to public transportation (and therefore labor pools), and to the centrally located public bureaucracies that regulate and generate contracts (Estevan, 1988; Suarez-Villa and Rama, 1996). Access to major roads and arteries should decrease transport costs for shipping to external markets while also minimizing journey-to-work commuting times for employees Ryan (2005). The hypothesized sign for the coefficients is negative, reflecting a diminishing probability of finding a firm as distance increases. The closer a point is located to a road or to the city center, the higher probability of finding a firm.

The point coordinates are derived from geo-coded addresses of establishments in the 2002 database. Although the geo-coding was not conducted personally, it is assumed that the data contain no systematic error. The database includes a size category for each establishment and industrial sector. Sector 32 is the focus here, defined as "Manufacture of electronic equipment, manufacturing equipment and radio, television and communication devices" in the National System of Economic Activities (CNAE) classification.

Another important aspect of the data is the selection of an observation window. The window is restricted to the central core of Madrid (see Figure 1) for two reasons: (1) Our primary intent was to inform the process of intraurban industrial location, and (2) In restricting the analysis to a regular shaped subset of the region, the edge correction calculations become less computationally onerous. In the initial analysis there was some experimentation with various simplified versions of the Madrid regional boundaries, however, due to the more complex window geometry the computation times increased substantially, especially for the area interaction model.

We expect that the strength of spillovers may vary with the size of a firms employment (Barff, 1987; Sweeney and Feser, 1998; Duranton and Overman, 2005; Arbia et al., 2012). To evaluate whether effects in the trend and interaction differ with employment size, three strata were introduced: 1-4 employees, 5-19 employees, and 20 or more employees. The results for the full sector and separate models for each of the employment size classes are presented below.

4.2 Model Fit and Validation

An important diagnostic of the model's overall fit is to whether simulations from the fitted models produce approximately the same number of points as the observed data. Simulated realizations from fitted models with the Geyer or Area interaction terms are roughly equal to the number of points in the observed data (see Figure 2, columns 1 and 2). The model with the Strauss hard core interaction clearly fails this fit criteria with simulations yielding a single, dominant cluster with a gross excess of points (see Figure 2, column 3). The failure of the Strauss hard core models was expected given that the model is known to fit poorly for processes with moderate to strong interaction (Hjort et al., 1994; Geyer and Thompson, 1995; Turner, 2009). The overall fit can also be assessed using the AIC (Tables 1-4). The AIC is lowest for the Area interaction models but the Geyer interaction models are a close second.

Assessment of the fitted trend is based on lurking variable plots and smoothed residual plots. The lurking variable plots (Figure 3) are for the subset of covariates that are significant in one of the three size-class models.[19] For both the Geyer and the Area interaction specifications, the true spatial trend can be approximated by the specified trend. Across the full domain of the covariate the sum of the residuals tend to zero and most of the graphs do not exceed the conservative plus/minus two standard deviation envelopes based on the inhomogeneous Poisson model. As shown in Figure 4, the smoothed residuals diagnostic presents a flat surface with small deviations from zero in the models with Geyer or Area interaction terms.

Validation and fit of the interaction term is based on the QQ-plot (see Figure 5) and the G-compensator diagnostic (see Figure 6). The Geyer interaction appears to fit well as indicated in the QQ-plot; note that the empirical distribution of the smoothed residuals lies inside the Monte Carlo simulated envelopes for the expected quantiles under a Geyer interaction. The model with the Area interaction term, however, suggests that for the large-establishment sample the pattern is less clustered than the fitted model, as the lower tail is heavier and beyond the lower edge of the envelope. In the G-compensator diagnostic, seen in Figure 6, the standardized residuals exceed two for small distances in the models with Geyer interaction term. However, this result is consistent because the test is conservative for small distances as the Poincaré variance is a substantial underestimate of the true variance. For small distances there are small-sample effects so

that a normal approximation to the null distribution of the standardized residuals is inappropriate.[20] After a distance of 2 kilometers the standardized residuals lie inside the envelopes showing positive values, suggesting that the data are slightly more clustered than the model. According to this diagnostic the model with the area interaction term is the one that obtains a better fit to the data.

Overall the presented diagnostics indicate that both the model with the Geyer saturation term and the model with the Area interaction term correctly capture the dependence on the covariates and the interaction between the establishments. In terms of global fit, the model with the Area interaction component outperforms the model with Geyer interaction term as measured by AIC. However, simulations from the fitted model with Geyer interaction yields patterns with close to the same number of observations, whereas the area interaction severely underestimates the pattern perhaps indicating that the trend is off by some scaler effect. Another comprehensive assessment of both trend and interaction, is to compare the empirical L-function to the L-function based on simulated patterns from the fitted model. Results are shown in Figure 7 and the model with Geyer interaction again outperforms the area interaction. Overall, the Geyer model seems to be the best of the three models.

4.3 Coefficient Interpretation

The coefficient estimates are similar to those in a standard regression setting but in this case allow us to assess the direction, magnitude, and statistical significance of effects of proximity to features/covariates in the environment on the spatial intensity of the industrial location process. The coefficient estimates from equations 3, 4, and 5, re-parameterized in logarithmic form according to equation 6, are provided in Tables 1-4. While estimates are provided for all three of the interaction specifications, we interpret those from the Geyer interaction model given the assessment of fit and validation in the previous section.

The results indicate that covariate effects differ depending on the employment size class of the industry. The covariate distance to the city center is significant at the 1% level in the models for the small- and medium-sized class of firms. This variable was not included in the model for the large-firms sample because the intra-metropolitan distribution of the electronic industry follows a dichotomous pattern based on employment size (Suarez-Villa and Rama, 1996; Rama et al., 2003); small producers within the region tend to be concentrated near the center of the city, while large

firms are principally located in the peripheral industrial areas.

The covariate *distance to a radial-type road* (R) is significant at the 1% level in the small- and medium-sized samples and at the 5% level for the large establishments model. In a cross-section analysis it is not possible to determine the causal relation between a firm's location and road construction considering the endogenous nature of a firm's location decision and public infrastructure investment decisions. The fact that the road map used is from 2004, implicitly point towards the direction of causality being that firms anticipate the benefit from being in the proximity of a radial road. An interaction term was included in the specification to account for the spatial variation in marginal effect for the sample of small-sized firms. Traveling along a radial road, the probability of finding a small-sized establishment decreases the closer one gets to the city center. This highlights the importance of a central location for small firms, as it is not possible to be located in close proximity to both the center and a radial road since the roads start far from the city center. The covariate that captures the interaction between the distances to the center and radial roads (center*road) is significant at the 5% level.

Our results suggest that access to national markets are a dominant feature of location decisions (Rama and Ferguson, 2007) as reflected in the significance of the radial roadways in all the models. It is interesting to note that the covariate *distance to a freeway* (A) is only significant for the smaller class of establishments at the 5% level, while the covariate *distance to a radial road* is significant for all the models. The highest concentration of firms is located in the northern and eastern districts, near to where the radial -R-2 and R-3- roadways were designed. The ring roads and radial roads begin at a considerable distance from the city center, while the motorways begin very close to the city center. As central locations are of greatest importance to small firms, the combination of being close to the city center and having easy access to the regional and national market can only be attained by locating close to a motorway (A1-A6).

For large establishments the only covariate that became significant at the 1% level is *distance to ring road* (M-40). This may reflect the importance of external markets in the location decision. Almost entirely concentrated in Madrid, telecommunication equipment manufacturing probably contributed to the development of large establishments in the periphery of the city and near the M-40 ring road, facilitating shipment from the metropolitan area to other regions of Spain. To some extent, zoning restrictions control the locations of Madrid's large firms, with most firms located

in the eastern and northern municipal districts, and to a lesser extent in the southern districts, where less advanced industries and working-class residential areas are traditionally located. The covariate M50 is not significant in the models; this is attributed to its recent construction and that it was not perceived to be advantageous.[21] This result is repeated for the models when each class of establishment is analyzed separately.

The main result of the analysis, and a primary goal of the paper, is to isolate effects related to the spatially varying intensity of the location process from the interaction effect. Recall that the interaction effect can be interpreted as the estimated degree of establishment-to-establishment attractivity. The Geyer interaction term is positive and significant at the 1% level for all employment size class models. The narrow interpretation is that conditional on other covariate values in the trend, the probability of observing a firm at a location is higher if another firm is located nearby. The models also allow us to estimate the radius of interaction using profile likelihood. More broadly, the estimated interaction effects and radii can be interpreted as evidence of the strength and scope of localization economies in the electronics sector. The range of interaction is estimated to be quite small – 1.4 kilometers or less – and with benefits of co-location implied to accrue when there are 2-4 neighboring firms, depending on the employment size. This result is unique to Gibbs models and is not something that could be estimated from the descriptive case-control methods reviewed in the second section of this paper. Possible explanations for the interaction results are related to geographic and professional proximity, as well as to the similar origins of entrepreneurs (Telefónica or Politécnica University), which encouraged the development of stable outsourcing relationships that prompted inter-firm collaboration as a way of minimizing capital risks (see Rama et al., 2003). Finding evidence of strong interaction in this industry confirms results from previous studies of the sector that used different methodologies. Suarez-Villa and Rama (1996) Bayón (2001) Rama et al. (2003) Rama and Calatrava (2002) Overall we observed that Gibbs models prove useful in isolating localization effects using secondary data, allow us to statistically test alternative forms of interaction, and to also recover covariate effects on the intensity of the location process.

5 CONCLUSIONS

The use of Gibbs models as a framework for studying industry co-location, or localization, provides distinct advantages over the largely descriptive approaches that have been dominant in the industry clustering literature. The modeling framework allows us to disentangle first- and second-order effects and thereby to isolate and quantify the key aspects driving localization economies.

Currently popular approaches are generally based on some variant, or derivative, of K-functions, which are limited by their reliance on case-control designs and in that their robustness depends heavily on an appropriate specification of the properties of the spatial process represented by the control group. Further, as long as there are influences that are clearly unique to the sector under evaluation, these approaches will not capture true interaction. In addition to the statistical description of point patterns, and often in combination with it, suitable point process models can be defined and fitted to data. No known papers to date have attempted to fit explicit models to point pattern data that incorporate both spatial inhomogeneity and inter-point interactions to explain the observed pattern of industrial establishments.

Our empirical application hints at the potential of the Gibbs point process models for studying firm location, industry clustering and localization. We have only tested the simplest non-trivial model to separate two sources of agglomeration and have demonstrated that the method produces the expected results. We tested different functional forms for the interaction component, concluding that in general the Geyer saturation model is the one that provides the best fit to the observed firm location pattern. The Area interaction model serves as a good model to explain interaction between the establishments, while the Strauss hard core exhibits clear lack of fit.

Our results are in line with previous research conducted, demonstrating the validity of the employed method to detect clusters. Results indicate that establishments are spatially clustered and that high density is found primarily in areas that are in close proximity to particular classes of roads. A significant interaction was found in all models, indicating the presence of localization economies at scales of 1.4 kilometers or less and in the presence of relatively small numbers of neighbors (2-4). The results show that different covariates play a different role in explaining the trend component for each firm size class. The intra-metropolitan distribution of the electronic industries mainly follows a dichotomous pattern. Small producers within the Madrid region tend

to be concentrated near the center of the city, while large firms are principally located in the peripheral industrial areas.

We conclude that we obtained a reasonable – although not completely ideal – fit even for our simple model; thus, this model can be considered a good starting point for more specific (and complex) analyses. Our future work will use more sophisticated theoretical frameworks in both the trend and interaction components of the model. Regarding the trend component, Gibbs models can be potentially very useful to shed light in the analysis of certain sources of spillovers. Certain knowledge spillovers decay very sharply with distance making it difficult to test for their existence. As noted above, certain forms of knowledge spillovers – such as those measured by proximity to a technical university – can be included in the trend component of Gibbs models. If significant, this would mean that the (conditional) intensity of the point process is higher near universities. Additionally, other source of spillovers, like the existence of urbanization economies, can be contrasted by adding a covariate that captures employment density in the trend. Meanwhile, if the interaction component is still statistically significant, it indicates the presence of other non-classified sources of spillovers and model specification testing will also reveal something about the function form and magnitude of the interaction. To determine and gain deeper insight about the type of firm interaction, a natural extension of the Gibbs model approach is to include marks in the analysis such as the size of the firm or the sector to which the firm belongs, (see Högmander and Särkkä, 1999). Another extension is to specify a type of interaction that captures information at the firm level, at the industry level, and at the county/municipality level, in line with hierarchical models.

A great deal of work remains to rigorously and theoretically evaluate the Gibbs framework for estimation and inference under standard threats to statistical conclusion validity that are typically pursued in econometrics. We provide a few simulation results in the appendix to this paper that provide preliminary and partial evidence that estimation and inference are somewhat robust to omitted variables in the specification of the trend, and that estimation of the trend parameters are robust to a misspecified interaction term. Three critical elements of model specification that need to be refined and evaluated are extensions to include spatial-temporal analysis[22], analysis on a road network, and analysis of cross-industry effects in the interaction term implemented as a marked process. We should also be clear that we are not arguing that Gibbs models should replace the descriptive case-control frameworks currently available. Those approaches can provide impor-

tant insights as part of a comprehensive multi-method strategy to understand industry location processes.

The existence of localization economies has several economic implications especially within economic growth theories. Using the Gibbs modeling framework we can identify whether the observed clustering is driven by localization economies (interation) or is purely a consequence of environmental characteristics such as access to roads. We would like to see this approach replicated for other industries and for other countries with available data, such as U.S., Canada, U.K., France or Germany.

References

Albert, José M, Marta R Casanova, and Vicente Orts. 2012, "Spatial location patterns of spanish manufacturing firms*," *Papers in Regional Science*, 91(1), 107–136.

Arbia, Giuseppe, Giuseppe Espa, Diego Giuliani, and Andrea Mazzitelli. 2010, "Detecting the existence of space–time clustering of firms," *Regional Science and Urban Economics*, 40(5), 311–323.

Arbia, Giuseppe, Giuseppe Espa, Diego Giuliani, and Andrea Mazzitelli. 2012, "Clusters of firms in an inhomogeneous space: The high-tech industries in Milan," *Economic Modelling*, 29(1), 3–11.

Arbia, Giuseppe, Giuseppe Espa, and Danny Quah. 2008, "A class of spatial econometric methods in the empirical analysis of clusters of firms in the space," *Empirical Economics*, 34(1), 81–103.

Baddeley, Adrian, Ege Rubak, and Jesper Møller. 2011, "Score, pseudo-score and residual diagnostics for spatial point process models," *Statistical Science*, 26(4), 613–646.

Baddeley, Adrian and Rolf Turner. 2000, "Practical maximum pseudolikelihood for spatial point patterns," *Australian & New Zealand Journal of Statistics*, 42(3), 283–322.

Baddeley, Adrian and Rolf Turner. 2005, "Spatstat: An R package for analyzing spatial point patterns," *Journal of Statistical Software*, 12(6), 1–42, URL `www.jstatsoft.org`, ISSN 1548-7660.

Baddeley, Adrian, Rolf Turner, Jesper Møller, and M Hazelton. 2005, "Residual analysis for spatial point processes (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5), 617–666.

Barff, Richard A. 1987, "Industrial clustering and the organization of production: a point pattern analysis of manufacturing in Cincinnati, Ohio," *Annals of the Association of American Geographers*, 77(1), 89–103.

Bayón, Susana López. 2001, "Características de la subcontratación electrónica en españa: evidencias empíricas," *Documentos de trabajo (Universidad de Oviedo. Facultad de Ciencias Económicas)*, 246, 1–28.

Berman, Mark and Rolf Turner. 1992, "Approximating point process likelihoods with GLIM," *Applied Statistics*, 41(1), 31–38.

Billings, Stephen B and Erik B Johnson. 2012, "A non-parametric test for industrial specialization," *Journal of Urban Economics*, 71(3), 312–331.

Bonneu, Florent and Christine Thomas-Agnan. 2011, "A unified framework for measuring industry location characteristics based on marked spatial point processes," *Les Journées de Méthodologie Statistique*, URL `http://jms.insee.fr/files/documents/2012/887_2-JMS2012_S10-4_BONNEU-ACTE.PDF`, online Proceedings, 24-26 January 2012.

Carlino, Gerald A, Robert M Hunt, Jake K Carr, and Tony E Smith. 2012, "The agglomeration of R&D labs," Tech. rep., Federal Reserve Bank of Philadelphia Working Paper.

Casanova, Marta R and Vicente Orts. 2011, "Assessing the tendency of spanish manufacturing industries to cluster: Colocalization and establishment size," Tech. Rep. 2011-03, Instituto Valenciano de Investigaciones Econmicas, S.A.

Coeurjolly, Jean-François and Frédéric Lavancier. 2013, "Residuals and goodness-of-fit tests for stationary marked Gibbs point processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2), 247–276.

Coeurjolly, Jean-François and Ege Rubak. 2013, "Fast covariance estimation for innovations computed from a spatial Gibbs point process," *Scandinavian Journal of Statistics*, 40(4), 669–684.

Diggle, Peter J, Amanda G Chetwynd, Roland Häggkvist, and Sarah E Morris. 1995, "Second-order analysis of space-time clustering," *Statistical Methods in Medical Research*, 4(2), 124–136.

Duranton, Gilles and Henry G Overman. 2005, "Testing for localization using micro-geographic data," *The Review of Economic Studies*, 72(4), 1077–1106.

Duranton, Gilles and Henry G Overman. 2008, "Exploring the detailed location patterns of uk manufacturing industries using microgeographic data*," *Journal of Regional Science*, 48(1), 213–243.

Ellison, Glenn, Edward L Glaeser, and William R Kerr. 2010, "What causes industry agglomeration? Evidence from coagglomeration patterns," *American Economic Review*, 100(3), 1195–1213.

Estevan, Antonio. 1988, *La industria electrónica en la Comunidad de Madrid*, Comunidad de Madrid. Dirección General de Economía y Planificación.

Falck, Oliver, Michael Fritsch, and Stephan Heblich. 2008, "The apple doesn't fall far from the tree: location of start-ups relative to incumbents," Tech. rep., Ifo Institute – Leibniz Institute for Economic Research at the University of Munich.

Feser, Edward J and Stuart H Sweeney. 2000, "A test for the coincident economic and spatial clustering of business enterprises," *Journal of Geographical Systems*, 2(4), 349–373.

Feser, Edward J and Stuart H Sweeney. 2002a, "Spatially binding linkages in manufacturing product chains," in P. Brown and R. McNaughton (eds.), *Global Competition and Local Networks*, Ashgate, pp. 111–129.

Feser, Edward J and Stuart H Sweeney. 2002b, "Theory, methods and a cross-metropolitan comparison of business clustering," in Philip McCann (ed.), *Industrial Location Economics*, Edward Elgar Publishing, pp. 222–259.

Funderburg, Richard G and Xiaoxue Zhou. 2013, "Trading industry clusters amid the legacy of

industrial land-use planning in southern California," *Environment and Planning A*, 45(11), 2752–2770.

Geyer, Charles J and Elizabeth A Thompson. 1995, "Annealing Markov chain Monte Carlo with applications to ancestral inference," *Journal of the American Statistical Association*, 90(431), 909–920.

Gordon, Ian R and Philip McCann. 2005, "Innovation, agglomeration, and regional development," *Journal of Economic Geography*, 5(5), 523–543.

Hjort, Nils Lid, Henning Omre, Marianne Frisén, Fred Godtliebsen, Jon Helgeland, Jesper Møller, Eva B Vedel Jensen, Mats Rudemo, and Henrik Stryhn. 1994, "Topics in spatial statistics [with discussion, comments and rejoinder]," *Scandinavian Journal of Statistics*, 21(4), 289–357.

Högmander, Harri and Aila Särkkä. 1999, "Multitype spatial point patterns with hierarchical interactions," *Biometrics*, 55(4), 1051–1058.

Hoover, Edgar M. 1937, "Spatial price discrimination," *The Review of Economic Studies*, 4(3), 182–191.

Huang, Fuchun and Yosihiko Ogata. 1999, "Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models," *Journal of Computational and Graphical Statistics*, 8(3), 510–530.

Klier, Thomas and Daniel P McMillen. 2008, "Evolving agglomeration in the US auto supplier industry*," *Journal of Regional Science*, 48(1), 245–267.

Koh, Hyun-Ju and Nadine Riedel. 2014, "Assessing the localization pattern of German manufacturing and service industries: A distance-based approach," *Regional Studies*, 48(5), 823–843.

Kosfeld, Reinhold, Hans-Friedrich Eckey, and Jørgen Lauridsen. 2011, "Spatial point pattern analysis and industry concentration," *The Annals of Regional Science*, 47(2), 311–328.

Marcon, Eric and Florence Puech. 2003, "Evaluating the geographic concentration of industries using distance-based methods," *Journal of Economic Geography*, 3(4), 409–428.

Marcon, Eric and Florence Puech. 2010, "Measures of the geographic concentration of industries: improving distance-based methods," *Journal of Economic Geography*, 10(5), 745–762.

Mateu, Jorge. 2002, "Statistical procedures for spatial point pattern recognition," *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, 26(1), 29–59.

Møller, Jesper and Rasmus P Waagepetersen. 2007, "Modern statistics for spatial point processes*," *Scandinavian Journal of Statistics*, 34(4), 643–684.

Nakajima, Kentaro, Yukiko Saito, and Iichiro Uesugi. 2012, "Measuring economic localization: Evidence from Japanese firm-level data," *Journal of the Japanese and International Economies*, 26(2), 201–220.

Rama, Ruth and Ascension Calatrava. 2002, "The advantages of clustering: The case of spanish electronics subcontractors," *International Journal of Technology Management*, 24(7), 764–791.

Rama, Ruth and Deron Ferguson. 2007, "Emerging districts facing structural reform: The Madrid electronics district and the reshaping of the Spanish telecom monopoly," *Environment and Planning A*, 39(9), 2207–2231.

Rama, Ruth, Deron Ferguson, and Ana Melero. 2003, "Subcontracting networks in industrial districts: The electronics industries of Madrid," *Regional Studies*, 37(1), 71–88.

Ryan, Sherry. 2005, "The value of access to highways and light rail transit: Evidence for industrial and office firms," *Urban studies*, 42(4), 751–764.

Scholl, Tobias and Thomas Brenner. 2011, "Testing for Clustering of Industries: Evidence from micro geographic data," Tech. rep., Philipps University Marburg, Department of Geography.

Smith, Tony E. 2004, "A scale–sensitive test of attraction and repulsion between spatial point patterns," *Geographical Analysis*, 36(4), 315–331.

Suarez-Villa, Luis and Ruth Rama. 1996, "Outsourcing, R&D and the pattern of intra-metropolitan location: The electronics industries of Madrid," *Urban Studies*, 33(7), 1155–1197.

Sweeney, Stuart H and Edward J Feser. 1998, "Plant size and clustering of manufacturing activity," *Geographical Analysis*, 30(1), 45–64.

Sweeney, Stuart H and Edward J Feser. 2004, "Business location and spatial externalities: Tying concepts to measures," in Michael F Goodchild and Donald G Janelle (eds.), *Spatially Integrated Social Science: Examples in Best Practice*, New York: Oxford University Press, pp. 239–262.

Sweeney, Stuart H and Kevin J Konty. 2005, "Robust point-pattern inference from spatially censored data," *Environment and Planning A*, 37(1), 141–159.

Torre, Andre and Alain Rallet. 2005, "Proximity and localization," *Regional Studies*, 39(1), 47–59.

Turner, Rolf. 2009, "Point patterns of forest fire locations," *Environmental and Ecological Statistics*, 16(2), 197–223.

Vitali, Stefania, Mauro Napoletano, and Giorgio Fagiolo. 2013, "Spatial Localization in Manufacturing: A Cross-Country Analysis," *Regional Studies*, 47(9), 1534–1554.

Yeung, Henry Wai-chung, Weidong Liu, and Peter Dicken. 2006, "Transnational corporations and network effects of a local manufacturing cluster in mobile telecommunications equipment in China," *World Development*, 34(3), 520–540.

Notes

[1]These indices suffer from a number of important aggregation issues that result from using a fixed areal support. One aggregation issue, known as the modifiable area unit problem (MAUP) is that conclusions reached when the underlying data are aggregated to a particular set of boundaries (say counties or municipalities) may markedly differ from conclusions reached when the same underlying data are aggregated to a different set of boundaries (say MSAs or regions). The problem becomes more severe as the level of aggregation increases. EG index assumes that the effect of plant i location on plant j profit depends only on whether they are in the same area, not on the distance between different areas. So the location decision process depends heavily on the definition of subareas. EG can only tests if spillovers are accrued when firms locate in the same geographic unit. However, in practice, spillovers would likely have an effect that declines more smoothly and provides some benefit to locating in nearby areas as well; more so when geographic subunits are small and not homogeneous.

[2]Gibbs models can be further decomposed into higher-order interactions but estimable models use only the first two components (Møller and Waagepetersen, 2007).

[3]Following Hoover (1937) the agglomeration of a particular industry after "controlling" for that of general manufacturing is referred to as localization

[4]In this paper we use the term spillovers quite broadly to refer to technological spillovers, gains from interfirm trade, the effect of local knowledge on the location of spinoff firms, etc., essentially, any forces that lead firms to choose locations near other firms in the industry. However Gibbs models are a valuable tool to detangle different types of spillovers.

[5]Other papers in the literature like Smith (2004) deal with the second order moment of the distribution, where patterns are treated as a realization of some underlying bivariate point process on S. The paper deals with attraction/inhibition of two populations, two different point patterns, and describes the spatial extent or the different scales of the interaction more accurately than other approaches. However our focus here is on univariate point processes.

[6]We are aware of researchers that are developing new test of industrial localization, like Scholl and Brenner (2011), Bonneu and Thomas-Agnan (2011), and Falck et al. (2008), however those paper are not published in any journal so we do not include them in our brief survey and assessment.

[7]This is in contrast to cluster models based on Cox processes where all of the inhomogeneity is loaded on the stochastic trend term. Comparisons between different functional forms of interaction are also possible within this framework.

[8]The trend is the intensity and is the analogue of the expected value of a random variable.

[9]The density $f(x)$ can be expressed as conditional intensity $\lambda(u, x)$ provided that the process has the property of heredity; that is, for any probability density of a finite process X in a bounded region W in R_d it requires that, $f(x) > 0 \rightarrow f(y) > 0$ for all $y \subset x$

[10]The algorithms for fitting point process models to point pattern data are included in the R package spatstat. The accuracy of the algorithm depends on how many additional dummy points are available and if they are sufficiently dense near the established data points.

[11]The appropriate maximum pseudo-likelihood estimator developed by Huang and Ogata (1999) and the performance of the MLE, MPLE, and AMLE (appropriate maximum pseudo-likelihood) are approximately the same in the cases of weak interaction, while the AMLE clearly improves the MPLE in the strong interaction cases of the parameter values. AMLE is remarkably concentrated around the MLE in all cases. From these it is clearly seen that the log-likelihood values of the AMLE are very close to those of the MLE.

[12]When estimating the model with a Geyer interaction specification two irregular parameters must be estimated: the saturation threshold and the interaction radius of influence. A small set of integer values were used for the saturation parameter (1 to 8) and interaction radius (0.2 to 3) and a combination that maximized the profile pseudo likelihood was selected. In the model with a Strauss hard core interaction specification, the irregular parameter to be estimated is the hard core distance and the interaction radius; whereas for the Area interaction component, the interaction radius must be estimated.

[13]Note, however, that the variance is based on an inhomogeneous Poisson process so the error

limits are only approximate and are likely to be conservative bounds relative to an inhomogeneous model with non-Poisson interaction (see Baddeley et al., 2005)

[14]We standardized the bands (with a value of 2) in order to compare the results of the three models in the same plot.

[15]The standardized residuals are highly irregular due to discretization effects in the computation and the inherent non-differentiability of the empirical statistic (see Baddeley et al., 2011).

[16]Reviews of the history of Madrids electronic industry can be found in Suarez-Villa and Rama (1996), Rama et al. (2003), and Rama and Ferguson (2007).

[17]Distance to the labor pool, transportation hubs (particularly the airport), industrial property price, crime, district characteristics, zoning, property exposure, etc.

[18]In Madrid the main roads are classified in ring roads (M) that surrounds the city, the motorways (A) that connect Madrid with the main Spanish regions, and the radial roads (R) that are toll roads of recent construction built to reduce traffic congestion out of the city. Radial road construction planning started during the early 1990s and was included in the 2000-2007 Ministry of Development Infrastructure plan. The road network refers to the year 2004, we assume that firms that value locations in proximity to these infrastructures had incentives to locate in its proximity prior to their inauguration, once the 1998 Land Act enabled sites to be valued according to their expected value. Nevertheless, we cannot ensure the assumption and further research would be needed to confirm the directionality of this relation

[19]The complete set of lurking variable plots is available from the authors. The full set is not provided here because of space constraints and to maintain visual clarity in the published figure.

[20]The standardized residuals are highly irregular due to discretization effects in the computation and the inherent non-differentiability of the empirical statistic (see Baddeley et al., 2011).

[21]Only some sections had been constructed in 2002.

[22]The only spatio-temporal Point Pattern Analysis approach implemented to date to the analysis of industrial location is the D function in Arbia et al. (2010), that basically compares the space-

time K function of the observed spatio-temporal point pattern to a theoretical pattern that has the same temporal and spatial property as the original data but no space-time interaction (Diggle et al., 1995).

List of Tables

List of Figures

Variable	Geyer	Area Interaction	Strauss / Hard Core
Intercept	-1.452 **	-3.373 **	-1.152 **
	(0.265)	(0.317)	(0.275)
Center	-0.060 **	-0.006	-0.079 **
	(0.018)	(0.022)	(0.018)
Road (R)	-0.095 **	-0.020	-0.062
	(0.037)	(0.056)	(0.049)
Road (A)	-0.052 ·	-0.019	-0.050 *
	(0.030)	(0.021)	(0.021)
Road (M40)	-0.018	-0.023	0.008
	(0.019)	(0.019)	(0.018)
Road (M50)			
Center * Road (R)	0.002	0.000	0.001
	(0.002)	(0.003)	(0.003)
Interaction	0.601 **	4.259 **	0.605 **
	(0.049)	(0.231)	(0.047)
Radius	0.8	0.8	0.85
Saturation	3		
Hard Core			0.11
AIC	970	846	1053

Table 1: Gibbs models for all electronics establishments, core Madrid

Variable	Geyer	Area Interaction	Strauss / Hard Core
Intercept	-0.811	-2.030 **	-0.709
	(0.541)	(0.574)	(0.581)
Center	-0.141 **	-0.095 *	-0.151 **
	(0.038)	(0.041)	(0.042)
Road (R)	-0.161 **	-0.125	-0.151 ·
	(0.060)	(0.088)	(0.087)
Road (A)	-0.120 *	-0.078	-0.098 ·
	(0.059)	(0.054)	(0.056)
Road (M40)	0.010	0.001	0.013
	(0.036)	(0.039)	(0.040)
Road (M50)	0.001	-0.013	0.008
	(0.031)	(0.032)	(0.034)
Center * Road (R)	0.006 *	0.005	0.006
	(0.003)	(0.004)	(0.004)
Interaction	0.354 **	3.060 **	0.692 **
	(0.064)	(0.393)	(0.128)
Radius	1	1	0.85
Saturation	4		
Hard Core			0.11
AIC	624	590	633

Table 2: Gibbs models for employment 1-4, core Madrid

Variable	Geyer	Area Interaction	Strauss / Hard Core
Intercept	-1.457 **	-2.519 **	-1.280 **
	(0.426)	(0.353)	(0.273)
Center	-0.119 **	-0.060 *	-0.150 **
	(0.028)	(0.023)	(0.026)
Road (R)	-0.142 **	-0.121 *	-0.113 **
	(0.038)	(0.049)	(0.043)
Road (A)	-0.045	-0.035	-0.021
	(0.059)	(0.047)	(0.050)
Road (M40)	0.040	0.009	0.071
	(0.034)	(0.037)	(0.044)
Road (M50)			
Center * Road (R)			
Interaction	0.521 **	3.461 **	1.001 **
	(0.129)	(0.397)	(0.136)
Radius	1.4	0.9	0.65
Saturation	2		
Hard Core			0.11
AIC	511	476	526

Table 3: Gibbs models for employment 5-19, core Madrid

Variable	Geyer	Area Interaction	Strauss / Hard Core
Intercept	-3.350 **	-4.998 **	-3.206 **
	(0.278)	(0.294)	(0.352)
Center			
Road (R)	-0.056 *	-0.014	-0.063 ·
	(0.026)	(0.036)	(0.038)
Road (A)			
Road (M40)	-0.078 **	-0.015	-0.112 **
	(0.022)	(0.024)	(0.025)
Road (M50)	0.032	0.004	0.047 *
	(0.027)	(0.029)	(0.020)
Center * Road (R)			
Interaction	0.748 **	4.129 **	1.940 **
	(0.085)	(0.423)	(0.231)
Radius	1	1.8	0.75
Saturation	3		
Hard Core			0.11
AIC	440	398	421

Table 4: Gibbs models for employment 20+, core Madrid

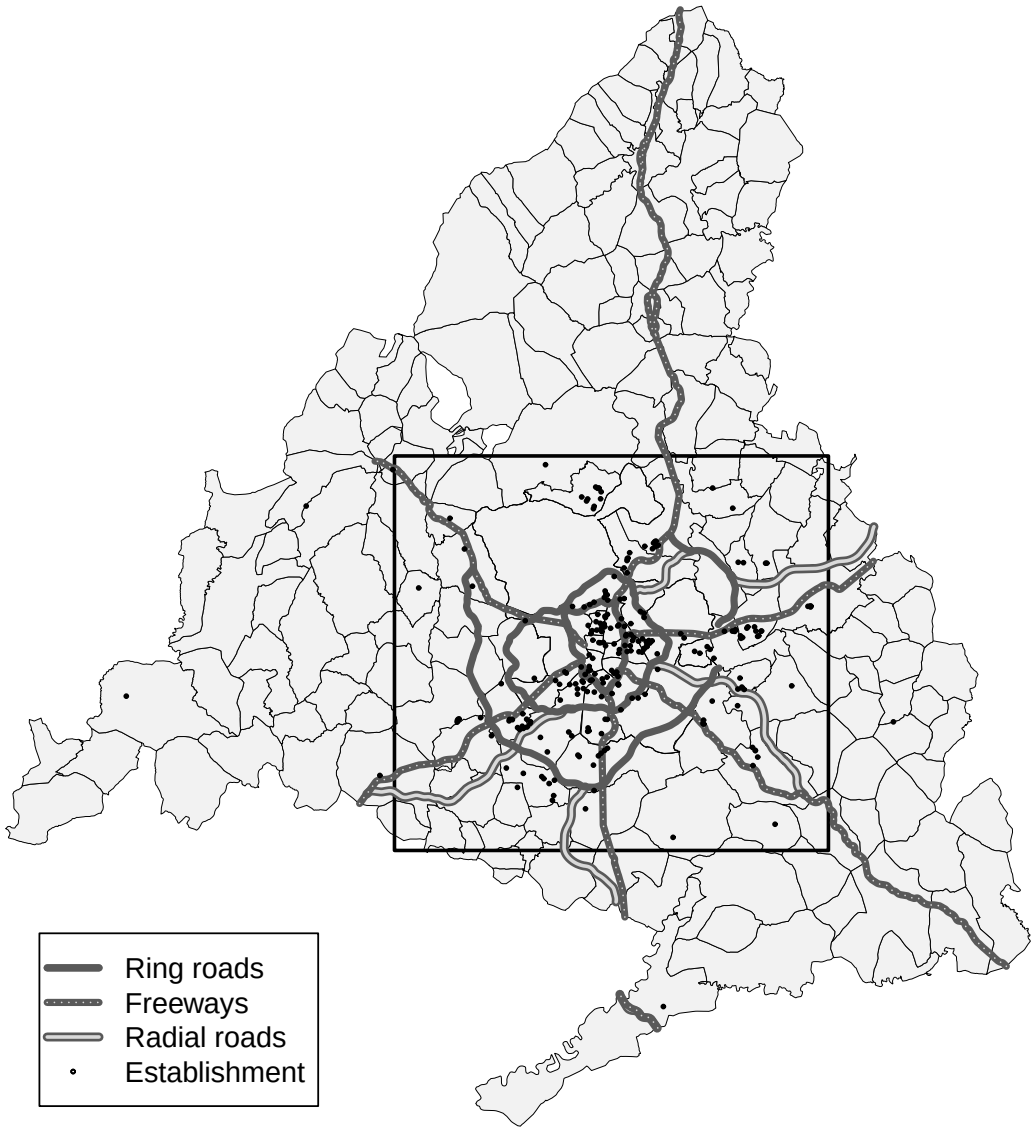

Figure 1: Municipalities/districts of Madrid with overlay of study area, roads, and establishment locations.

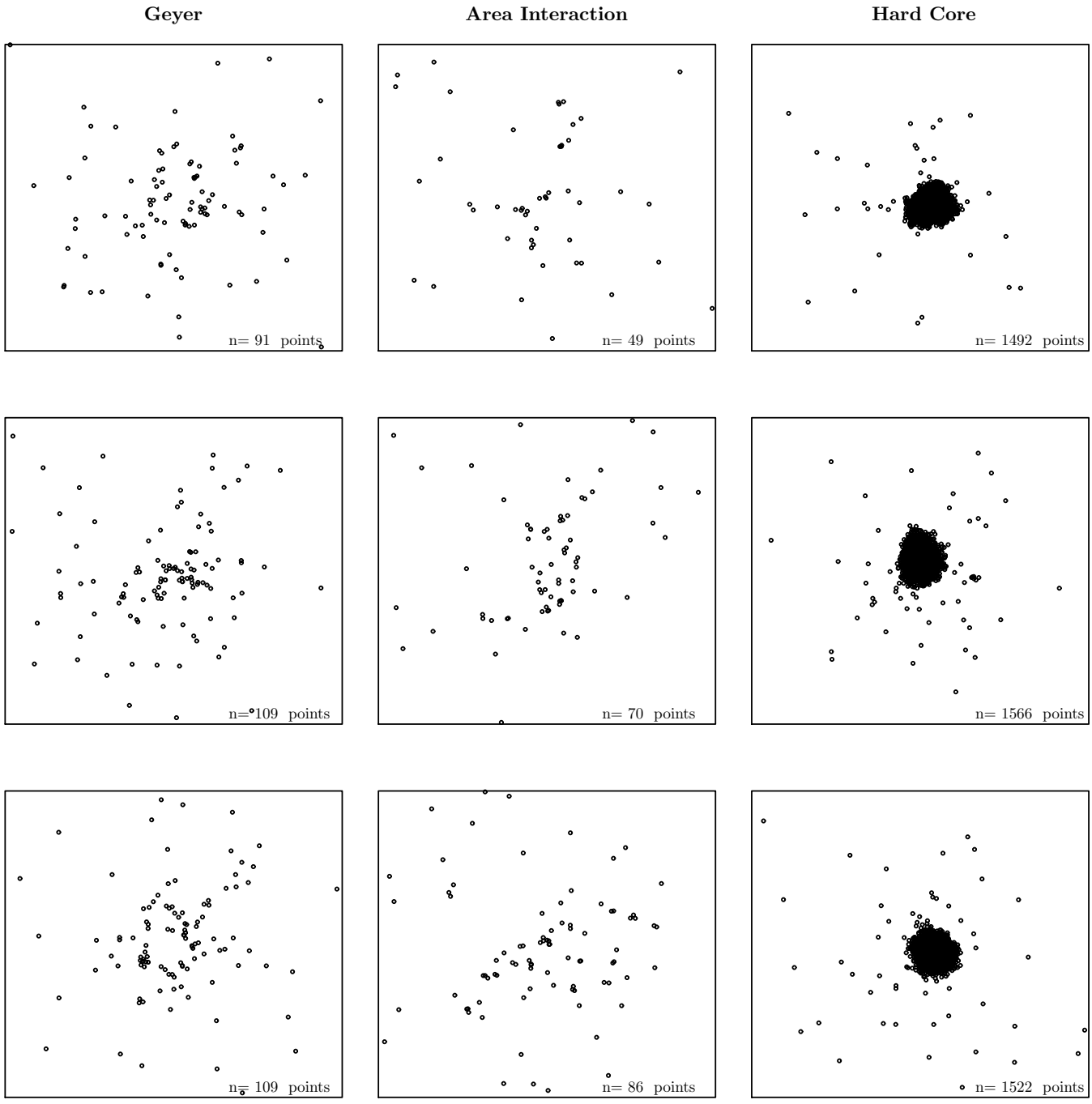

Figure 2: Three sets of simulated realizations for fitted models (employment size 1-4, obs=102)

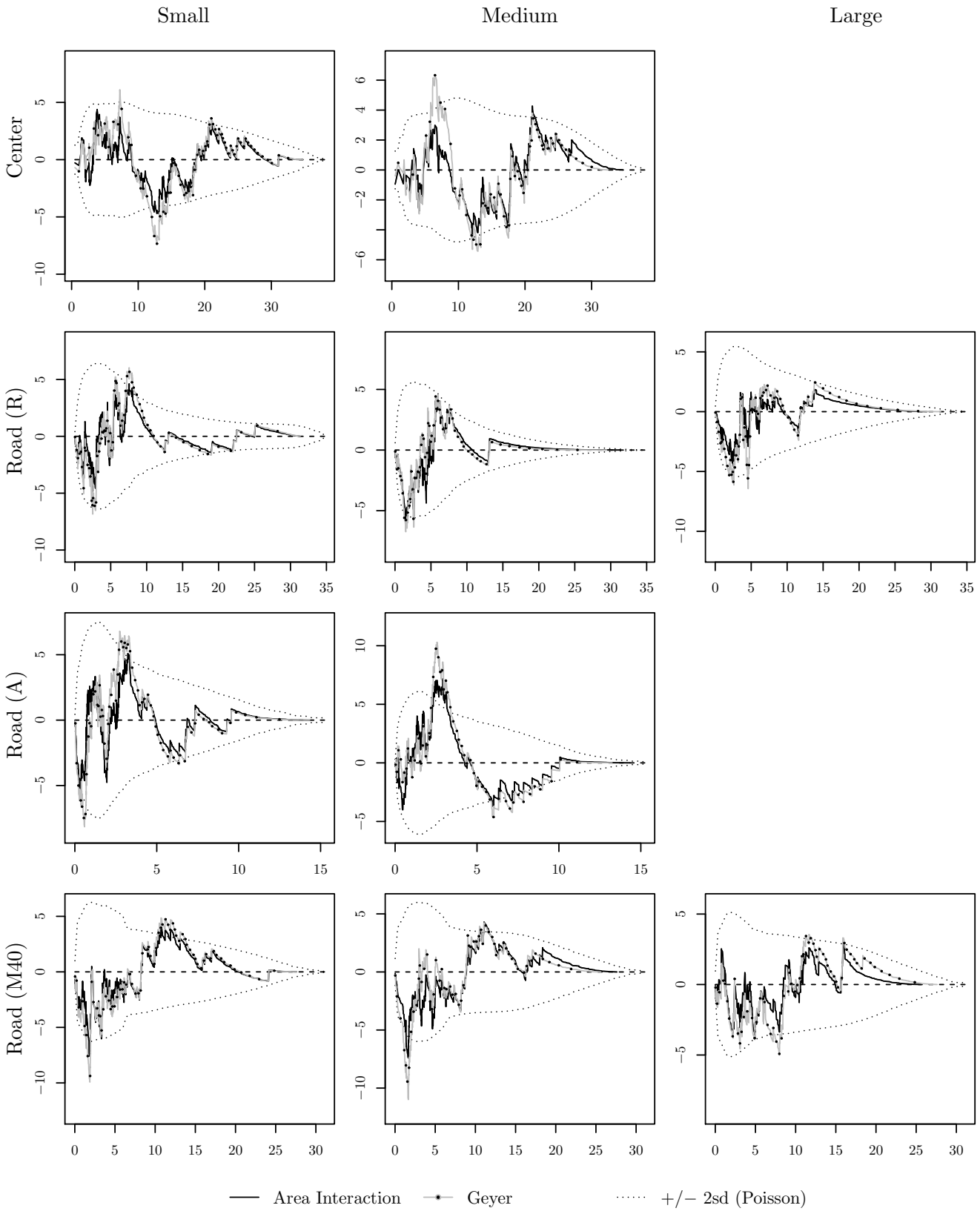

Figure 3: Lurking variable plots, raw residuals, all establishments

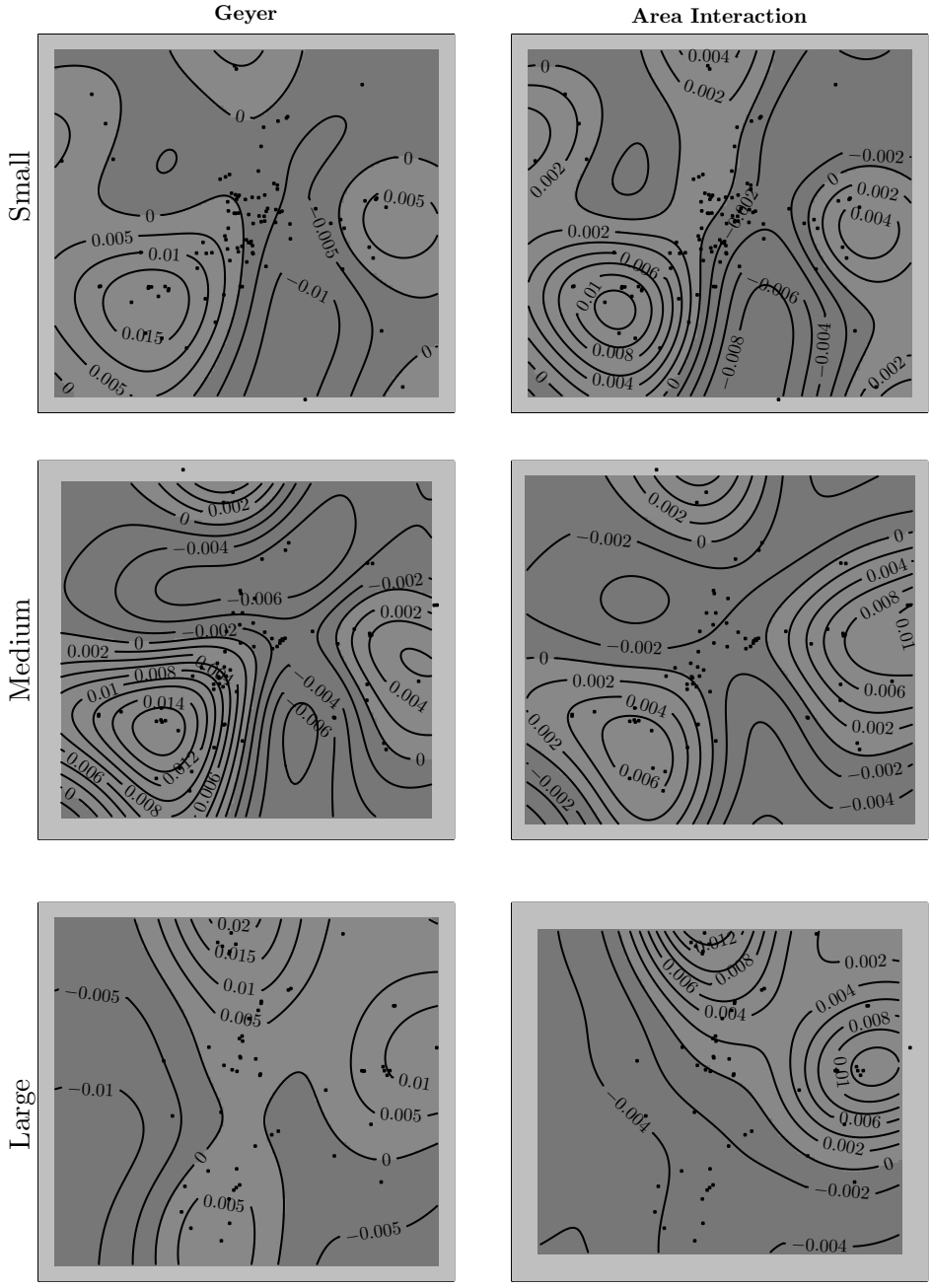

Figure 4: Raw residuals, core Madrid

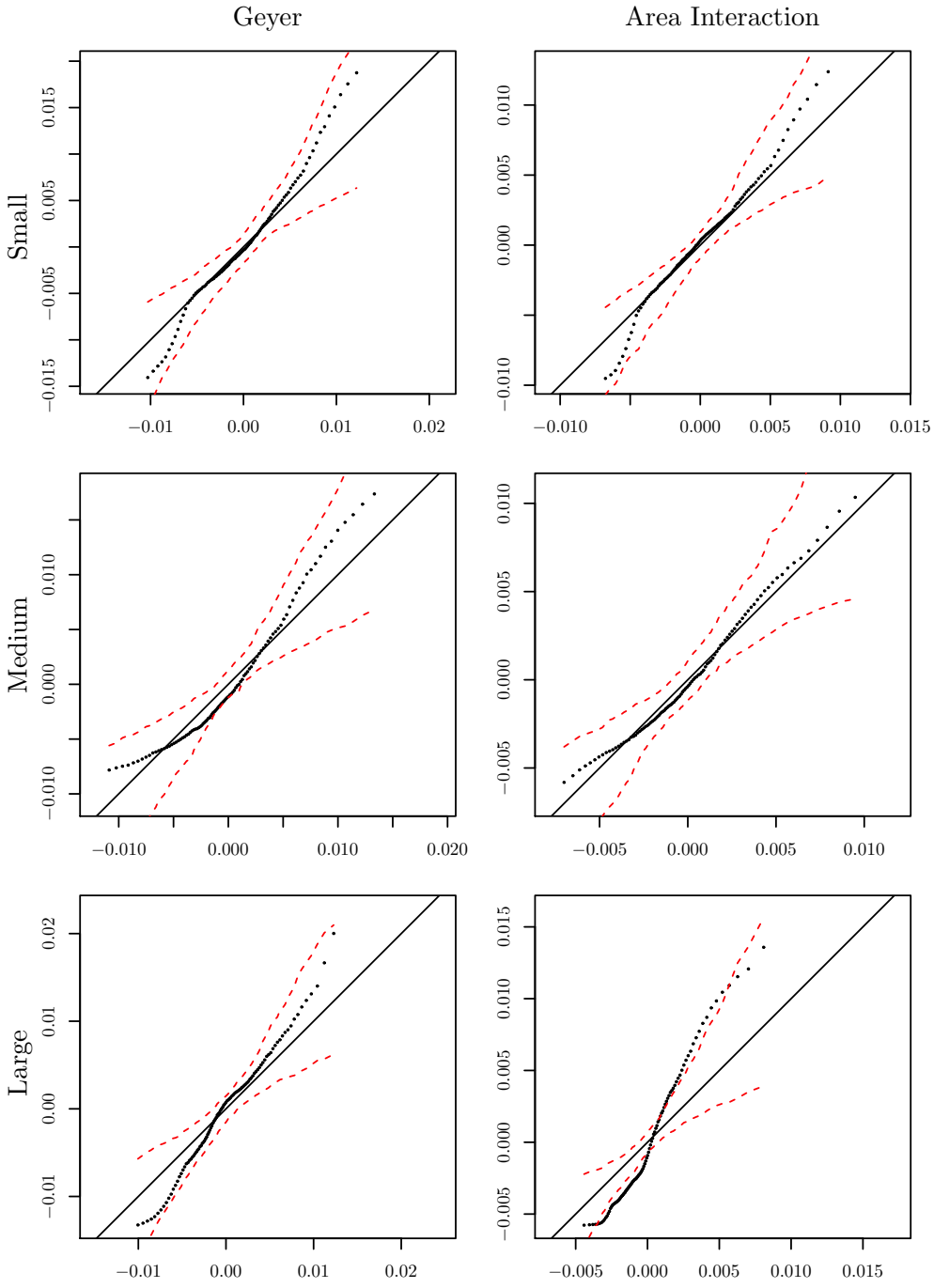

Figure 5: QQ plots

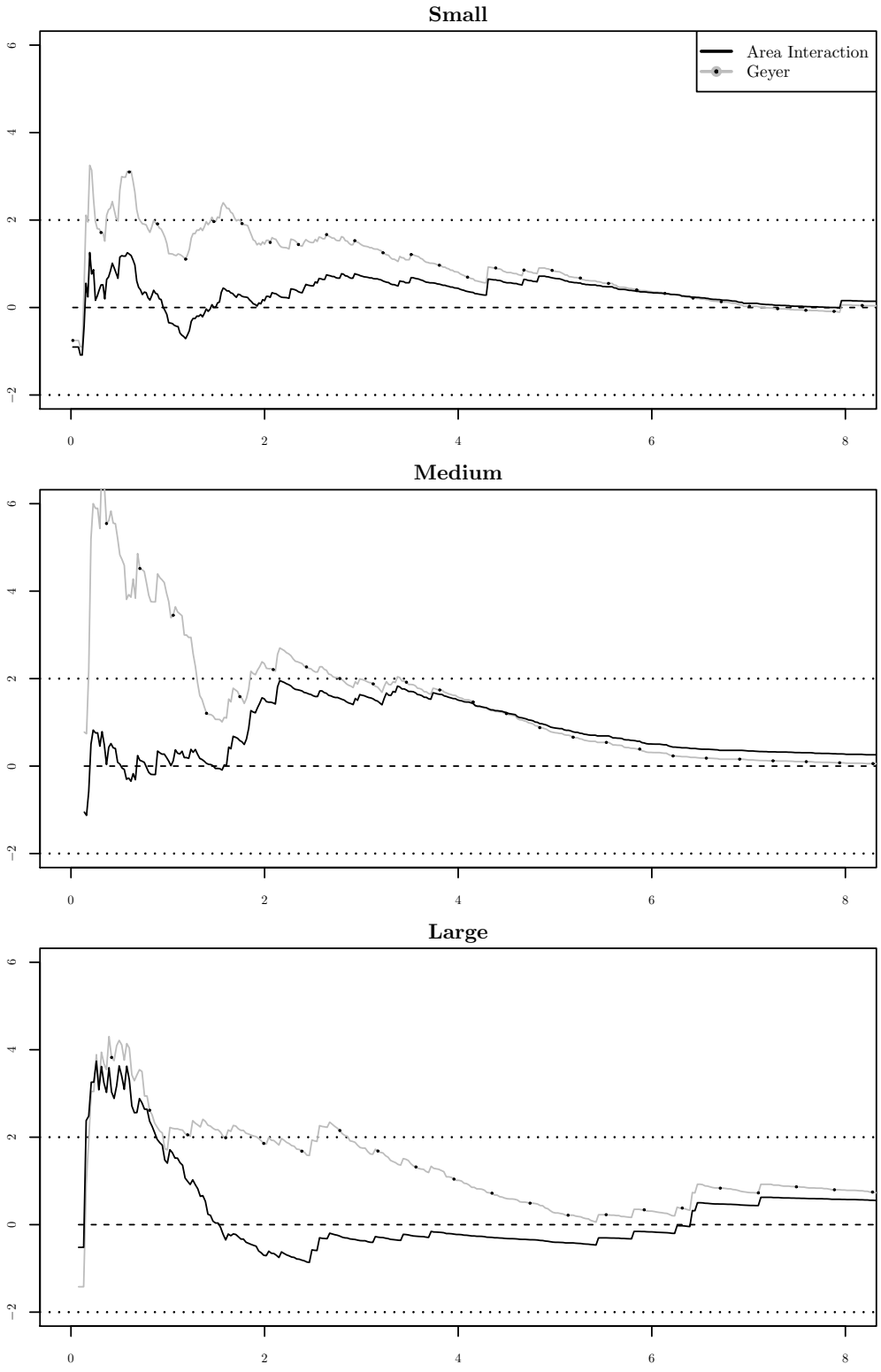

Figure 6: G compensator

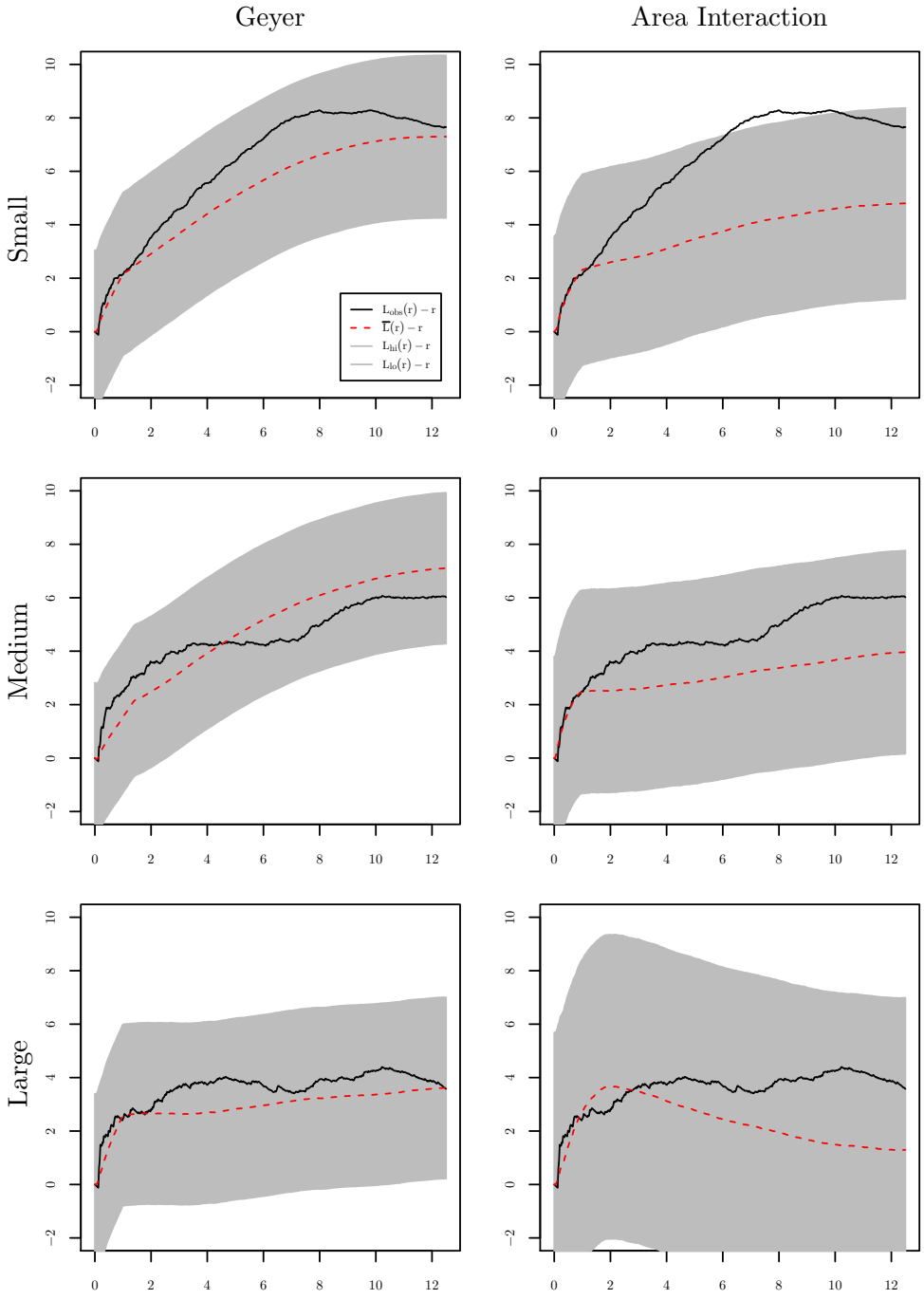

Figure 7: L-function for inhomogeneous Geyer and Area Interaction

Appendices

The main body of the paper provides an empirical demonstration with a full set of diagnostics and interpretation. As an addendum, we include here a small set of simulation results where the data generating process is known, allowing us to evaluate alternative estimators and the performance of the models under different specification errors. The first set of results is for a planar point process and relies on functions that are available in *spatstat*. The second set of results, for point processes constrained to a network, required some extension of *spatstat* functionality.

A Simulation results for planar point processes

The essence of our argument in this paper is that Gibbs models provide a means for making inferences about both spatial covariates in the trend and the mathematical structure of the interaction in studies of industrial clustering. As with any econometric application, the validity of those inferences depends on theoretical model assumptions being satisfied. Also, with the availability of alternative estimators in *spatstat*, we wanted to provide some sense of the degree to which the Huang-Ogata (HO) estimator improves on the computationally faster maximum pseudo-likelihood (MPL) estimator.

The planar results are based on realizations from an inhomogeneous Gibbs process with Geyer interaction. We use the same boundary domain as we use for the empirical study, and we include two spatial covariates – the ring road (M40) and motorways type roads (R.A). Specifically, realizations are from a Geyer(1,3) with trend $exp(-2-0.1R.A-0.1M.40)$ and the degree of interaction γ=2; see Figure A.1. Table A.1 reports the mean and standard errors of parameter estimates from models fit to 1000 simulated point patterns, when MPL is evaluated, and 250 simulated point patterns when HO is evaluated. Each simulated point pattern is a realization of the process with the trend and interaction as defined above.

The first comparison we make is between the MPL and HO estimators and we assume the irregular parameters and interaction structure are known. Both the HO and MPL provide comparable unbiased estimates of the coefficients on the spatial covariates, however the MPL estimates of the intercept and degree of interaction (γ) are moderately biased. Even with the bias, the crude inferential task of assessing whether the interaction is repulsive ($\gamma < 0$), non-existent ($\gamma = 0$), or

attractive ($(\gamma > 0)$ would still be correct. Since the HO estimator appears to be superior we focus on it for the remainder of the planar simulations. The next evaluation (column 3) is based on realizations of a process with the same trend as above but without the interpoint interaction. The HO estimator correctly returns a coefficient of approximately 0 on γ, and the estimates in the trend remain unbiased and as efficient as the first HO estimation (column 1) even with what amounts to the inclusion of an irrelevant variable in the model (a Geyer(1,3) interaction term).

As a next step we evaluate omitted variable bias. It is almost always the case that some covariates will be missing in empirical applications simply because of the limits of what is measured (and measured well) relative to prevailing theories of the process. The results in columns 4 and 5 of Table A.1 indicate the effects of excluding one of the spatial covariates but including the other. In each case there is mild bias in the estimate for the remaining spatial covariate, but the main impact is on biased estimation of the intercept and interaction coefficient. While this small example is far from conclusive, the upward biased interaction coefficient could result in incorrectly concluding that there is significant interaction when no interaction is present. In our case, the crude inferential task concerning repulsive, attractive, or zero interaction remains the same even with omitted variables.

A final focus for the planar simulation is on the use of profile likelihood for the irregular parameters. We evaluate first the use of profile likelihood when the trend is correctly specified (column 6) and then omitted one or the other spatial covariate (columns 7 and 8). For the profile likelihood we used a grid search over the parametric space defined by a radius in the domain $\{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ and saturation in the domain $\{1, 2, 3, 4, 5\}$. When the mean is correctly specified (column 6), and also when the covariate R.A is omitted (column 7), profile likelihood correctly selects the irregular parameter values of r=1 and saturation=3. When R.M40 is omitted the saturation parameter is incorrectly set to 4.

On balance the estimation framework appears to be fairly robust to a few specification errors. Also, there is nothing uniquely problematic about the estimation of Gibbs models that isn't also problematic in other spatial econometric estimation. That is, estimates will always suffer from some degree of omitted variable bias even as we strive to include as many of the theoretically relevant covariates that are available. The ability to additionally evaluate alternative specification of the interaction terms is unique to this framework and the small set of simulations we provide show that the combination of HO with profile likelihood does provide a solid basis for estimation and

inference.

B Simulation results on a road network space

While the scale-dependence and point resolution permitted in planar point process models resolves the main issues present in working with spatially aggregate data, the approach still abstracts from reality in allowing for "firms" to locate anywhere in space. In fact, establishments are constrained to locate along road networks and may be excluded from some areas because their industrial process is incompatible with zoning regulations. More refined models would have a view towards spatial point processes on the domain of road networks.

The development of network processes in *spatstat* is new and has limited functionality. It is currently possible to estimate parameters of in homogenous Poisson processes on a network using MPL as the estimator (the HO method has not been implemented yet). An evaluation of that estimator is provided in column 1 of Table A.2. The mean and standard errors of parameters are based on 1000 simulated points patterns where point realizations are from a process with the true trend parameters as shown. MPL recovers the parameters with no apparent bias. Note that the covariates in this case are distances from the road types shown (R.A and R.M40) measured along the network. There is currently no mechanism to simulate inhomogenous network processes with interaction in *spatstat*. It is possible to approximate the process by using masking to only allow a domain for points realization within short distance from the road network. The second column of Table A.2 shows the results of using MPL to estimate the parameters of the trend and interaction constrained to realizations on a network; see Figure A.2. The current estimation in *spatstat* uses Euclidean distances, not network distances, in the construction of the interaction covariate. We note that while the results are biased, this is also characteristic of MPL estimation.

While the use of Gibbs models constrained to a road network is not fully available, approximate methods shown here indicate that the models are feasible. We also note that the move to networks is only one of several issues along the path towards more refined and realistic models of firm location. Other refinements needed include directionality along a network and perhaps models that are based on time distances rather than physical distance.

	True	1	2	3	4	5	6	7	8
(Intercept)	**-2**	-1.976	-1.440	-2.033	-2.305	-2.998			
		(0.189)	(0.147)	(0.223)	(0.173)	(0.170)			
R.A	**-0.1**	-0.101	-0.112	-0.099		-0.093			
		(0.022)	(0.025)	(0.045)		(0.021)			
R.M40	**-0.1**	-0.099	-0.115	-0.099	-0.109				
		(0.012)	(0.014)	(0.022)	(0.012)				
r	**1**	1	1	1	1	1	1	1	1
							(0)	(0)	(0)
sat	**3**	3	3	3	3	3	3	3	4
							(0)	(0)	(0)
Interaction	**0.6931**	0.683	0.514	0.045	0.719	0.855			
		(0.047)	(0.029)	(0.167)	(0.049)	(0.049)			

Table A.1: Simulation experiments on a plane. 1=Huang-Ogata (HO) approximate ML; 2=Maximum Pseudolikelihood (MPL); 3=HO fitted to simulated data with trend but no interaction; 4=R.A omitted from the trend; 5=R.M40 omitted from the trend; 6=irregular parameters from PL using the correct trend; 7=irregular parameters from PL using the trend missing R.A; 8=irregular parameters from PL using the trend missing M.40

	1		2	
	True	MPL	True	MPL
(Intercept)	**-0.1**	-0.1007	**-0.75**	-1.0602
		(0.1259)		(0.2006)
f.ra	**-0.02**	-0.0201	**-0.03**	-0.0273
		(0.0049)		(0.0054)
f.m40	**-0.05**	-0.0502	**-0.03**	-0.0271
		(0.0031)		(0.0028)
Interaction			**0.405**	0.3568
				(0.0448)

Table A.2: Simulation experiments on a network. Model 1 is an inhomogeneous Poisson process on a network (no interaction). Model 2 is a inhomogeneous point process on a network with Geyer(1,3) interaction.

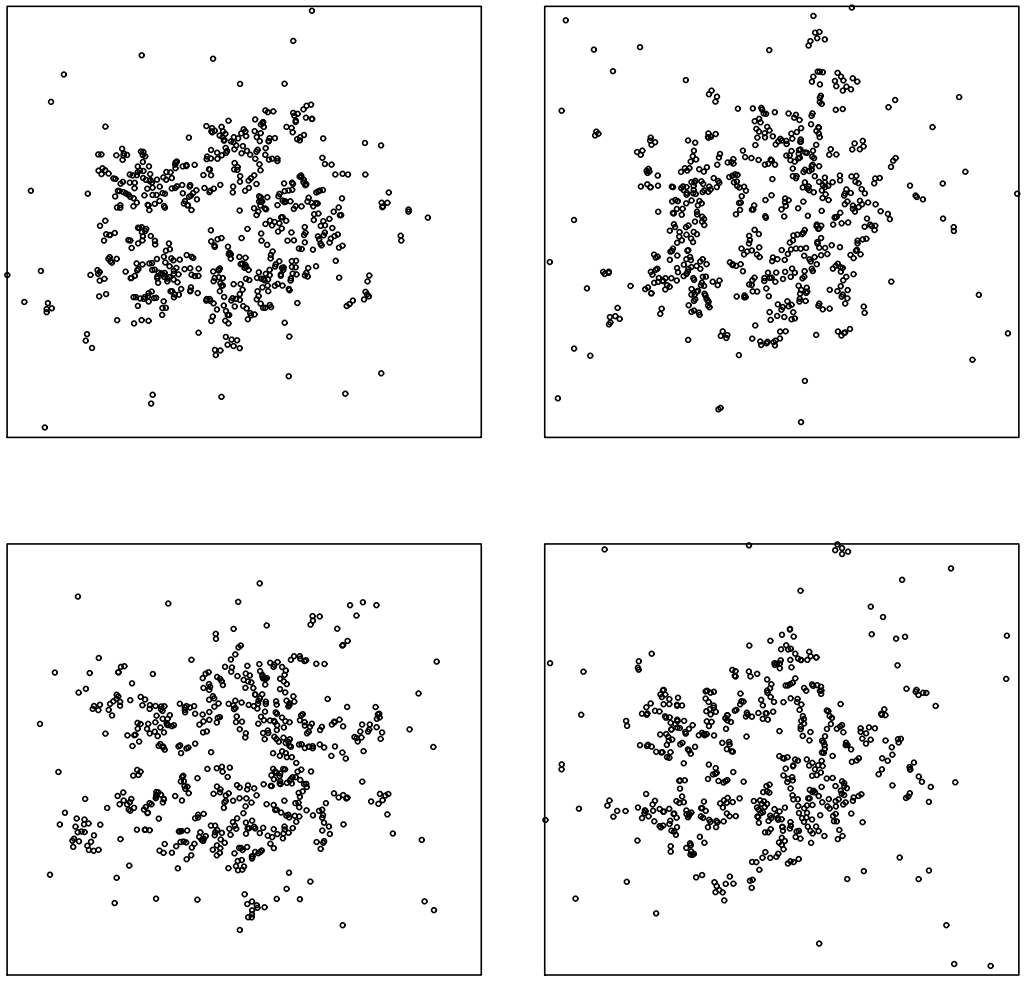

Figure A.1: Simulation on a plane from inhomogeneous Geyer(1,3) process with trend $exp(-2 - 0.1R.A - 0.1M.40)$ and γ=2 ($log(2) \approx 0.6931$)

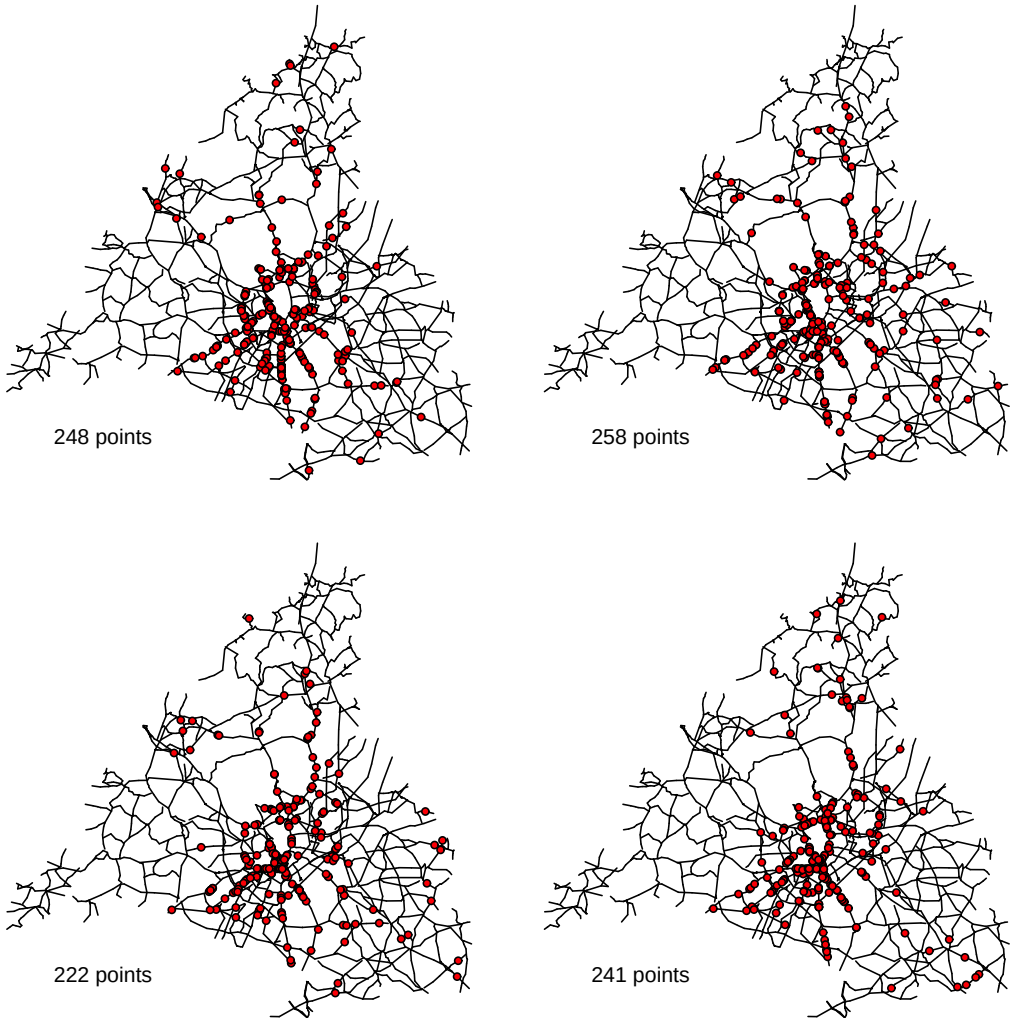

Figure A.2: Simulation on a network from inhomogeneous Geyer(1,3) process with trend $exp(-0.75 - 0.03R.A - 0.03M.40)$ and γ=1.5 ($log(1.5) \approx 0.4055$)