# Localization externalities in the United States

Giulio Bottazzi

Sant'Anna School of Advanced Studies
Institute of Economics
Pisa, Italy.

Ugo M. Gragnolati

University of Strasbourg, BETA
Strasbourg, France.

May 3, 2014

**Abstract**

This paper estimates the determinants of the spatial distributions of firms across a variety of sectors in the United States. Estimation is based on a model of endogenous localization choices characterized by a unique stochastic equilibrium. Firms maximize utility as affected by regional fundamentals, localization externalities, and idiosyncratic preferences. Standard procedures of model selection help to severely discriminate among alternative nested models according, in particular, to their account of externalities. These turn out being present across all sectors. Moreover, their attractive pull tends to be stronger than any other determinant of firm localization.

**JEL codes:** L1, C31, R3.

**Keywords:** Firm location, Heterogeneity, Localization externalities, Nested models, Model selection.

## 1 Introduction

This paper analyzes the determinants of the spatial distributions of firms in the United States through a model of endogenous localization choices and its corresponding estimation method. Both theoretical and empirical aspects contribute to characterizing the analysis that follows.

In the model, firms choose where to settle according to three types of determinants: their idiosyncratic preferences, a common set of exogenous regional features, and localization externalities. Being characterized as a random variable, idiosyncratic preferences bring stochasticity into the system and thus lead to identify a stochastic spatial equilibrium. According to this notion, a spatial equilibrium is attained when the net flows of firms across regions are null. Essentially, this corresponds to identifying a detailed balance condition, thus predicting a distribution of firms across regions. In fact, the dynamical system embedded in the model converges toward a unique ergodic distribution, which is the stochastic equilibrium. Notably, a closed form solution of such distribution is derived for any number of firms and regions, so that the predicted equilibrium can be actually compared with the observed spatial distributions of firms.

Hence, the model is estimated at the level of commuting zones for a variety of manufacturing and service sectors in the United States. This estimation exercise is carried out ensuring that localization externalities are severely tested for and disentangled from other effects. In particular, standard techniques of model selection inform the choice between alternative nested models that differ exclusively for the externality parameter. As it turns out, even after penalizing parametric numerosity, the incorporation of externalities allows to explain systematically more of the observed spatial distributions of firms. Furthermore, the marginal elasticity of localization externalities on the probability for a region to attract extra firms is found to be the main determinant of the spatial distributions of firms across most sectors.

This work is related to a number of contributions in the literature. The baseline theoretical model presented here was originally formulated by Bottazzi and Secchi (2007) as a Markovian version of the urn process introduced by Arthur (1990), with the intention to allow for empirical testing in models with social interactions. The typical identification problems associated to this type of models shape the discussion in Blume et al. (2011), to which the present work actively contributes. Within this broader analytical context, specific attention is devoted to disentangling and measuring the determinants of firm localization, thus linking with a plentiful of studies sharing a similar goal (see Beaudry and Schiffauerova (2009), Head and Mayer (2004), Puga (2010), Rosenthal and Strange (2004) for complementary surveys). Yet, three aspects distinguish this contribution from related studies in applied economic geography. First, the final outcome of the present approach does not consist solely in estimating parameters, but also in predicting the entire spatial distribution of firms in each sector. This represents a major difference from the works of Black and Henderson (1999), Desmet and Fafchamps (2006), Devereux et al. (2004), Dumais et al. (2002), Ellison and Glaeser (1997, 1999), Henderson (2003), Maurel and Sédillot (1999), Rosenthal and Strange (2001). Second, and differently from Duranton and Overman (2005), the effect of externalities are disentangled from the effect of region-specific factors deriving, for instance, from natural or infrastructural advantages. Third, substantial improvements are introduced in the estimation method as compared to related studies by Bottazzi et al. (2007, 2008), in particular by applying maximum likelihood to obtain point estimates and bootstrap resampling to estimate their variance. A similar exercise has already been carried out by Bottazzi and Gragnolati (2012) for Italy, but at a coarser sectoral disaggregation.

The remainder of the paper will be organized as follows. Section 2 presents and discusses the model of firm localization on which the subsequent econometric analysis is based. Section 3 describes the estimation method. Section 4 briefly presents the dataset used to carry out the analysis. Section 5 shows the results of the estimation procedure. Finally, Section 6 summarizes the contributions of this paper and sketches possible future developments.

# 2 A model of endogenous localization choices

Consider a sector with a fixed number of firms $N$ that chose where to locate among $L$ regions. The following model aims at characterizing the occupancy vector of firms across regions, $\boldsymbol{n} = (n_1, \ldots, n_L)$, where $n_l$ is the number of firms located in $l$.[1] At each time step, a firm is selected at random to revise its locational choice according to its preferences. This dynamic could represent both the exit of an incumbent and the entry of a new firm with different preferences, or an exogenous shock to the preferences of an active firm. In any case, the occupancy vector $\boldsymbol{n}$ is revised according to utility maximization on the side of the randomly drawn firm.

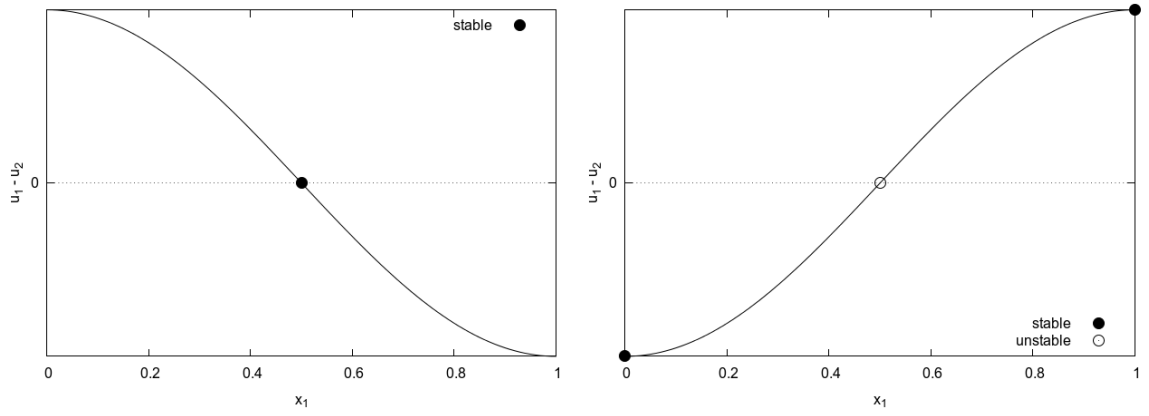
The utility $u_{i,l}(\boldsymbol{\phi}, \boldsymbol{a}, \boldsymbol{n})$ that firm $i$ derives from locating in region $l$ is assumed to depend on three terms. First, the idiosyncratic component $\boldsymbol{\phi} = (\phi_{i,1}, \ldots, \phi_{i,L})$, which is the realization of a random process. Second, the intrinsic attractiveness $\boldsymbol{a} = (a_1, \ldots, a_L)$ given by exogenous regional features. Third, the occupancy vector $\boldsymbol{n}$, which establishes an interdependence between the firm that is choosing and the locational choices made by other firms. In this sense, the present model includes a localized externality. Generally, $\boldsymbol{\phi}$, $\boldsymbol{a}$ and $\boldsymbol{n}$ enter $u_{i,l}$ as vectors since firms can evaluate $l$ as well as its $l - k$ spatial neighbors, thus giving rise to spatial interdependence. However, this aspect of generality will be sacrificed in the following treatment by assuming that $k = 0$, thus abstracting from spatial interdependence. As a consequence, the utility of a generic firm simplifies to the form $u_{i,l}(\phi_{i,l}, a_l, n_l)$.

Provided with a vector of utilities $\boldsymbol{u_i} = (u_{i,1}, \ldots, u_{i,L})$, the randomly drawn firm chooses to locate in the region $\tilde{l_i}$ such that

$$\tilde{l_i} = \underset{l \in 1,\ldots,L}{\operatorname{argmax}} \left\{ u_{i,l}(\phi_{i,l}, a_l, n_l) \right\} \ .$$

If the region $\tilde{l_i}$ is different from the previously occupied by firm $i$, this choice generates a modification of the occupancy vector $\boldsymbol{n}$. As discussed below, the dynamics entailed by this model imply notionally different types of equilibrium depending on $\phi$ being or not degenerate. When $\phi$ is degenerate, the model is entirely deterministic, thus resounding with the vast literature relying on the assumption of a representative agent. When $\phi$ is non-degenerate, instead, firm heterogeneity brings stochasticity into the model and leads to identify a different type of equilibrium.

[1] Clearly, the number of firms in a sector is not generally constant in reality. However, the variation in the number of active firms is typically an order of magnitude smaller than the gross number of entries and exits (see Bartelsman et al., 2005). Therefore, the spatial distribution of firms can be safely assumed to be driven by relocations, rather than by genuine entry.




(a) (b)

Figure 1: Equilibria in a two-location deterministic model

## 2.1 Equilibrium notion

To understand the notion of stochastic equilibrium inherent to the present model, it useful to illustrate first how the equilibrium would be defined in some simplified deterministic examples.

In the simplest case, $\phi$ is degenerate and externalities are null. Therefore, the vector of utilities $\boldsymbol{u} = (u_1(a_1), \ldots, u_L(a_L))$ is constant for all firms. The only long-run equilibrium in this baseline case consists in having all firms located in the same region, in which $a$ will be maximum. Things get more complex as soon as externalities are added while neglecting the term $a_l$, so that $\boldsymbol{u} = (u_1(n_1), \ldots, u_L(n_L))$. For an equilibrium to be reached in this case, the firm that is called to make a choice must select the same region that is already occupying, as no stochastic shock to preferences enters $\boldsymbol{u}$. This condition is met only if none of the alternative regions have a utility higher than the utility of the current region. Therefore, those regions whose utility at equilibrium is strictly lower than the utility of some other region will be empty. At the same time, all non-empty regions should have the same utility. Formally, $\boldsymbol{n}^*$ is an equilibrium occupancy if and only if there exists a constant $c$ such that for any $l$

$$\begin{cases} u_l(\boldsymbol{n}^*) = c & \text{if } 0 < n_l^* < N \ , \\ u_l(\boldsymbol{n}^*) \leq c & \text{if } n_l^* = 0 \ , \\ u_l(\boldsymbol{n}^*) \geq c & \text{if } n_l^* = N \end{cases}$$

Notice that, if there exists a region $l$ for which $n_l^* = N$, then for any other region $m \neq l$ it will be $n_m^* = 0$. In this occurrence of full agglomeration, the previous statement simply means that the utility of the region in which firms agglomerate, $u_l(n_l^*)$, is higher than the utility of any other region, $u_m(n_m^*)$. This equilibrium is locally stable if there exists $\epsilon > 0$ such that for any initial condition $\boldsymbol{n}'$ abiding to $||\boldsymbol{n}' - \boldsymbol{n}^*|| < \epsilon$, the system will converge to $\boldsymbol{n}^*$. In general, the existence, stability and number of deterministic equilibria depend on the set of functions $\boldsymbol{u}(\boldsymbol{n})$, as exemplified in Figure 1 for the case $L = 2$.

For instance, the utility of locating in a given region could decreases with the number of

firms already placed there. This shape of $u_l$ could derive from several mechanism, such as the progressive reduction of output price resulting from increased of local competition, or the progressive increase in production cost due to the exhaustion of the local labor pool. Whichever the reason, the difference of utilities $u_1 - u_2$ could then follow a behavior as in Figure 1a, where $x_i$ is the fraction of firms located in region $i$. As long as $x_1 < 1/2$, then only a firm located in region 2 would migrate, thus increasing $x_1$. Conversely, as long as $x_1 > 1/2$, only firms located in region 1 would migrate, and thus decrease $x_1$. Hence, the only stable equilibrium is when firms are symmetrically distributed across the two regions, that is $x = 1/2$.

Alternatively, the utility of locating in a given region increases with the number of firms already placed there. This could happen, for instance, because of an increase in demand when consumers prize diversity, or for the existence of localized technological spillovers which increase productivity. Whichever the reason, the difference of utilities $u_1 - u_2$ could then follow a behavior as in Figure 1b. Now the only stable equilibria are represented by fully agglomerated economies, with all firms concentrated either in region 1 or in region 2, whereas the symmetric equilibrium $x = 1/2$ is unstable.

The analysis of these kinds of deterministic equilibria has produced a vast theoretical literature, typically sharing the assumption of a representative agent (see Bottazzi and Dindo, 2013, Forslid and Ottaviano, 2003, Krugman, 1991, Venables, 1996, among others). The present model resounds with that literature in considering the case for positive externalities. At the same time, however, it also moves away from that approach by describing firms as heterogeneous, thus assuming that the process $\phi$ is non-degenerate. Moreover, here regions are generally considered to be asymmetric in their exogenous features, as captured by the term $a_l$.

Once preference heterogeneity is allowed for through a non-degenerate $\phi$, the characterization of equilibrium changes. Generally, the new entrant never replaces the exiting firm in the same region, because the two normally hold different preferences. But order can still emerge *in probability*. If for any regions $m$ and $l$, the incentives for a firm to move from $m$ to $l$ are exactly counterbalanced by the incentives for another firm to move from $l$ to $m$, then the average number of firms in each region converges to a fixed point. In turn, this defines a spatial distribution $\pi(\boldsymbol{n}; \boldsymbol{a}, b)$, which constitutes the stochastic equilibrium of the present model.

The formal derivation of such equilibrium relies on the fact that choices are not driven by the stochastic component *in probability*. In fact, the probability $p_l$ for location $l$ to be selected by the randomly drawn agent is given by

$$p_l = \text{Prob}\{c_l + \phi_l \geq c_j + \phi_j \; \forall j \neq l \mid \boldsymbol{c}, F(\phi)\} \; ,$$

where $c_l(a_l, n_l)$. Following either Yellott (1977, Th.6 §4) and Luce (1959, Axiom 2.1) or Raouf Jaibi and ten Raa (1997), such probability can be proved to be

$$p_l = \frac{c_l}{\sum_{j=1}^{L} c_j} \; , \tag{1}$$

provided that the upper tail of $F(\phi)$ decays faster than in an exponential distribution (see Bottazzi and Secchi, 2007, Propositions 2.1 and 2.2). Exploiting the fact that $p_l$ depends solely

on the common term $c_l$ and not on the stochastic one, it is then possible to assess the dynamical system entailed by the model.

Such system does not keep track of the position of each single firm over time. More simply, the state of the system is the occupancy vector $\boldsymbol{n} = (n_1, \ldots, n_L)$, while the number of possible configurations corresponds to the number of ways in which $N$ firms can be distributed in $L$ regions. Hence, the number of possible configurations is $\binom{N+L-1}{N}$. This system is equivalent to a finite Markov chain with transition probability defined by

$$P(\boldsymbol{n}' \mid \boldsymbol{n}) = \begin{cases} \frac{n_l}{N} \frac{q_{l'}(\boldsymbol{n}-\delta_l)}{\sum_{m=1}^{L} q_m(\boldsymbol{n}-\delta_l)} & \text{if } \boldsymbol{n}' = \boldsymbol{n} - \delta_l + \delta_{l'} \\ 0 & \text{otherwise} \end{cases}$$

where $\delta_l$ is the vector of length $L$ with the $l$-th entry equal to 1 and all other entries equal to zero and $q_l = \exp(u_l)$ (see Bottazzi and Secchi, 2007, Propositions 3.2). The stochastic equilibrium of this dynamical system will then correspond to the detailed balance condition

$$\Pr\{\boldsymbol{n} + \boldsymbol{\delta}_l - \boldsymbol{\delta}_m | \boldsymbol{n}\} \pi_{\boldsymbol{n}} = \Pr\{\boldsymbol{n} | \boldsymbol{n} + \boldsymbol{\delta}_l - \boldsymbol{\delta}_m\} \pi_{\boldsymbol{n}+\boldsymbol{\delta}_l-\boldsymbol{\delta}_m} , \tag{2}$$

where $\boldsymbol{\delta}_i$ is a $L \times 1$ vector with value 1 in the $i^{\text{th}}$ element and 0 otherwise. As a remark, equation (2) expresses a notion of "collective equilibrium" in the sense that it guarantees the stationarity of $\pi(\boldsymbol{n})$ in the aggregate, while individual turbulence might take place.

## 2.2 Preference structure and closed form solution

The dynamics of the system generated through entry and exit depend on the structure of $u_l(a_l, n_l, \phi_l)$. In particular, a linear form will be assumed here:

$$u_{i,l} = a_l + b n_l + \phi_{i,l} \quad b \geq 0 , \tag{3}$$

where $c_l = a_l + b n_l$. Given this preference structure, the dynamical system can be proved to converge to a unique stationary distribution $\pi(\boldsymbol{n}; \boldsymbol{a}, b)$ that is not affected by initial conditions. Therefore, the system is ergodic (see Bottazzi and Secchi, 2007, Proposition 3.4 and Appendix A.3.). Clearly, $\pi(\boldsymbol{n}; \boldsymbol{a}, b)$ reveals how many regions host $\boldsymbol{n} = (0, 1, ..., N)$ firms for given values of $\boldsymbol{a} = (a_1, ..., a_L)$ and $b$.

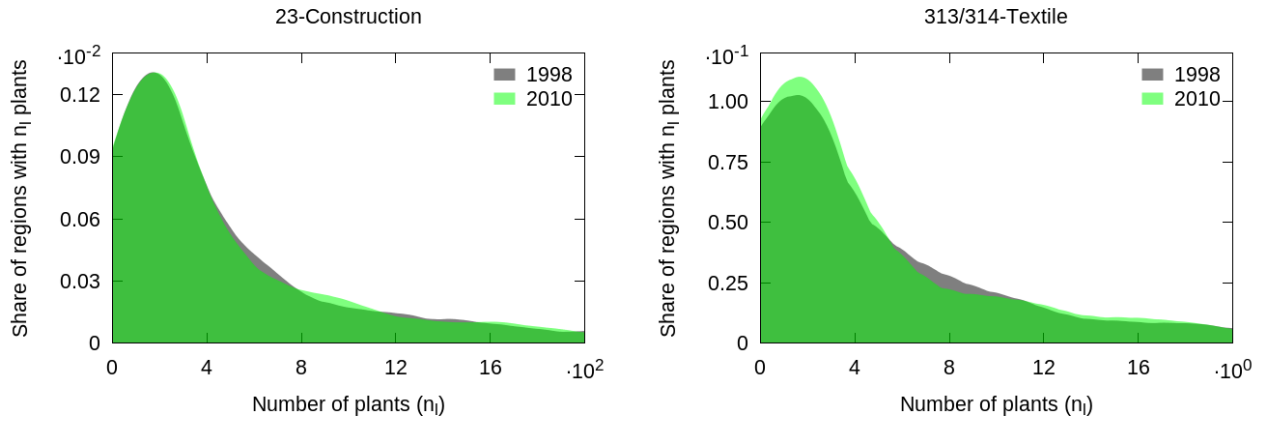
In general, $\pi(\boldsymbol{n}; \boldsymbol{a}, b)$ is characterized by the Polya form

$$\pi(\boldsymbol{n}; \boldsymbol{a}, b) = \frac{N!\Gamma(A/b)}{\Gamma(A/b + N)} \prod_{l=1}^{L} \frac{1}{n_l!} \frac{\Gamma(a_l/b + n_l)}{\Gamma(a_l/b)} , \tag{4}$$

where $A = \sum_l a_l$. In the specific case of null externalities, however, equation (4) reduces to a multinomial form:

$$\pi(\boldsymbol{n}; \boldsymbol{a}, b = 0) = N! \prod_{l=1}^{L} \frac{1}{n_l!} \left(\frac{a_l}{A}\right)^{n_l} . \tag{5}$$

In the following exposition, equations (4) and (5) will be referred to as the "Polya model" and




(a) (b)

Figure 2: Spatial distributions of plants across US commuting zones in two exemplifying sectors.

the "Multinomial model". But they actually constitute the varying prediction of a single model of firm localization for varying values of $b$. In this sense, the Polya and Multinomial models are nested and their performances can be compared with standard procedures of model selection (see Section 3.2).

The estimation of the model will take place under the assumption that the observed spatial distribution of firms is an equilibrium. It is easy to find empirical support to back this assumption. As an example, Figure 2 overlays the spatial distributions of two representative sectors in 1998 and 2010. The impression of stationarity that one has from visual inspection of Figure 2 is indeed confirmed by Kolmogorov-Smirnov tests. Across all sectors, the hypothesis of a null difference between spatial distributions is never rejected when comparing 1998 with 2010. Therefore, assuming that the observed spatial distributions are an equilibrium is generally unproblematic at the empirical level. In parallel, the model is designed so that each time step corresponds to a localization choice rather than to an actual time unit. Consequently, convergence to the equilibrium distribution occurs after a sufficient number of choices have taken place, and this might well correspond to a relatively short real time.

# 3 Estimation

Given a set of $H$ region-specific variables $\boldsymbol{x_l} = (x_l^1, ..., x_l^H)$, these can serve to define the attractiveness or a region via the generic function $g(\boldsymbol{\beta}, \boldsymbol{x_l})$, where $\boldsymbol{\beta}$ are unknown parameters to be estimated. More precisely, equations (4) and (5), yield the following log-likelihood functions for the Polya and Multinomial model respectively:

$$\log \pi = \log N! - \sum_l \log n_l! + \sum_{l=1}^{L} \sum_{k=0}^{n_l - 1} \log(g(\boldsymbol{\beta}, \boldsymbol{x}_l) + k) - \sum_{k=0}^{N-1} \log(G + k) \ , \tag{6}$$

$$\log \pi = \log N! - \sum_{l=1}^{L} \log(n_l!) + \sum_{l=1}^{L} n_l \left( \log g(\boldsymbol{\beta}, \boldsymbol{x}_l) - \log A \right) \ , \tag{7}$$

where $g(\boldsymbol{\beta}, \boldsymbol{x}_l) = a_l/b$ in (6) and $g(\boldsymbol{\beta}, \boldsymbol{x}_l) = a_l$ in (7), while $G = \sum_{l=1}^{L} g_l$. Maximizing (6) and (7) yields the maximum likelihood point estimates $\hat{\boldsymbol{\beta}}$ for each of the two models. These estimates can then be used to quantify the marginal effects of localization externalities on $p_l$ while controlling for the effect of region-specific variables.

In doing so, $p_l$ is actually substituted with its monotone transformation $q_l = -\log(1 - p_l)$, so as to obtain an unbounded measure. According to (1), in the context of the Polya model, the marginal effects associated respectively to region-specific factors and to localization externalities are

$$\frac{\partial q}{\partial \log x^h} = \sum_{l=1}^{L} \frac{\partial q_l}{\partial \log x_l^h} = \sum_{l=1}^{L} \frac{x_l^h \partial_h g_l}{N + G} \,, \tag{8}$$

$$\frac{\partial q}{\partial \log n} = \sum_{l=1}^{L} \frac{\partial q_l}{\partial \log n_l} = \frac{L}{N + G} \,. \tag{9}$$

On the other hand, in the context of the Multinomial model, the only type of marginal effect is given by

$$\frac{\partial q}{\partial \log x^h} = \sum_{l=1}^{L} \frac{\partial q_l}{\partial \log x_l^h} = \sum_{l=1}^{L} \frac{x_l^h \partial_h c_l}{G} \,, \tag{10}$$

as localization externalities are ruled out by definition (i.e. $b = 0$).

While $\partial q/\partial n$ in (9) does not depend on the specification of $g()$, the computation of $\partial q/\partial x^h$ in (8) and (10) needs such function to be specified. Here a log-linear specification is adopted which reads

$$\begin{aligned} g(\boldsymbol{\beta}, \boldsymbol{x_l}) &= \exp\left(\sum_h \beta_h \log(x_{h,l}) + \beta_0\right) \\ &= \prod_h x_{h,l}^{\beta_h} \, \exp(\beta_0) \,. \end{aligned} \tag{11}$$

Indeed, (11) is equivalent to the standard Cobb-Douglas functional form, so that $g_l$ can be seen as the accumulated multiplicative effect of the different variables. As such, it has a probabilistic interpretation: if, on average, the probability to choose region $l$ according to factor $h$ is proportional to $x_{h,l}^{\beta_h}$, and if the effects of the different factors can be assumed as roughly independent, the combined probability of the firm to choose this region is given by expression (11).[2] The $H$ variables taken in consideration are listed in Table 2. Note that all controls are inputed into equation (11) as $z$-scores, so that the resulting estimates are fully comparable.

Given specification (11), it is especially convenient to obtain the corresponding marginal

[2] In the multinomial case, it is assumed that $\beta_0 = 0$ in (11). In fact, in that case, the log-likelihood (7) is invariant for a rescaling factor, i.e. the transformation $a_l \to \lambda a_l$ applied to each $a_l$ leaves the likelihood level invariant. Consequently, leaving $\beta_0$ to be estimated would result in an over-specified model.

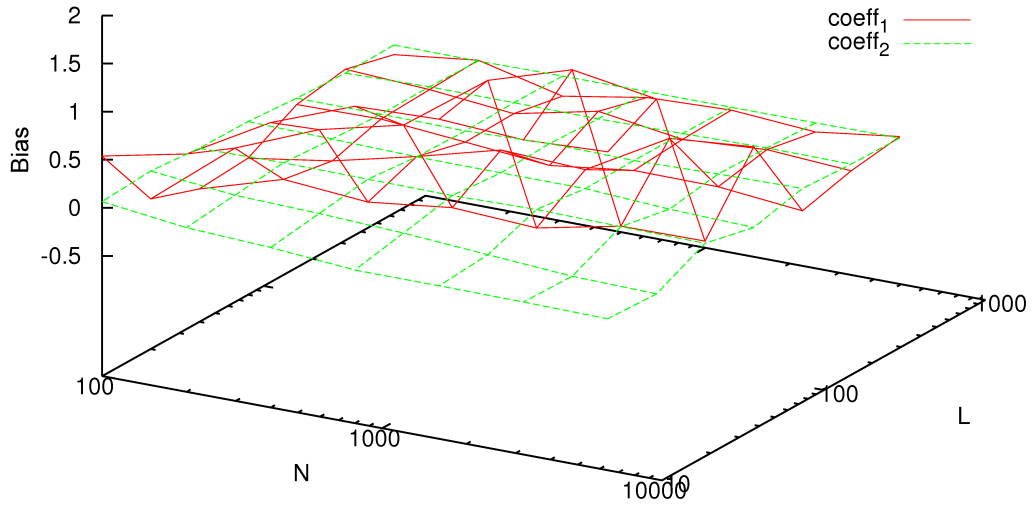



Figure 3: Monte Carlo analysis of coefficient bias.

elasticities

$$\frac{\partial q}{\partial \log x^h} = \hat{\beta}^h \frac{G}{N+G} \,, \tag{12}$$

$$\frac{\partial q}{\partial \log x^h} = \hat{\beta}^h \,, \tag{13}$$

for the Polya and Multinomial model respectively. Indeed, the estimates $\hat{\boldsymbol{\beta}}$ needed to compute (12) and (13) come precisely from their maximum likelihood estimation in (6) and (7).

The variance of these estimates is computed through bootstrap resampling, thus allowing to assess their statistical significance. Also, Figure 3 shows the results of Monte Carlo analysis on the bias of the estimated coefficients for varying values of $N$ and $L$. The asymptotic consistency generally ensured by maximum likelihood estimation is reached in practice for any $L > 100$, even when regressors are strongly collinear. This condition is always respected in the present analysis, as estimation occurs on a dataset composed by 691 regions that cover the entire mainland United States. Therefore, the estimates do not suffer of small sample problems.

Finally, it is worth emphasizing that the present approach is immune to endogeneity problems. As long as the underlying variables of the model are actually in equilibrium, the present method reveals how much each variable contributes to the probability for a location to host a certain number of firms. In fact, any possible co-evolution in time of the variables ceases once equilibrium is reached. As mentioned before and further discussed below, this condition is actually met in the data (see Figure 2 and Figure 4).

## 3.1 Controls

The set of $H$ controls taken in consideration here aims at ensuring that only sector-specific localization externalities are measured through equation (9), while washing away correlated effects. In this sense, the controls listed in Table 2 are meant to capture the other classical drivers of firm localization, namely urbanization economies, natural or infrastructural advantages, and

the attractive pull of vertically-related sectors.[3]

Urbanization economies are controlled for by a mix of variables highlighting the different sources of attractiveness for cities. First, the population of each commuting zone together with its median income level captures the pull of local demand. Second, adding to the equation also the area of each commuting zone serves to control for population density, and thus for the level of generic local labor supply. Third, a measure of local infrastructural endowment takes in consideration the cost-sharing advantages that typically characterize cities with respect to infrastructures (see below for more details on the measurement). Fourth, the Shannon's variety index has been inserted in trial regressions, following Jacobs (1969) on the idea that cities might offer greater inter-sectoral spillovers due to their greater industrial variety.[4] However, this variable has turned out being either statistically or economically non significant for most sectors, and thus it was dropped. One could perhaps presume that also other effects contribute to urbanization economies. If this were the case, a binary variable for metropolitan and/or micropolitan areas should be associated, at least, to a statistically significant coefficient. This, however, never happens in the sectors under scrutiny. Hence, urbanization economies are already largely captured by the set of controls in Table 2.

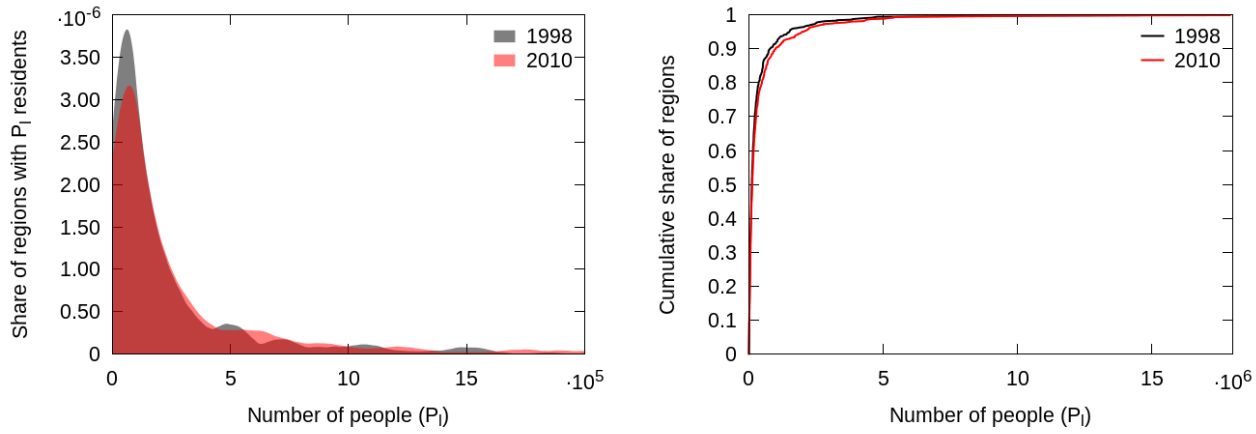
Another subgroup of controls aims at capturing natural and infrastructural advantages. On the one hand, the number of mines, as well as the intensity of cropland, forests, rivers and coasts are meant to proxy natural advantages. On the other hand, each commuting zone is also characterized in terms of its endowment of transport infrastructures. In fact, while partly shaping urbanization economies as described above, transport infrastructures are more generally a strategic assets for trade as they affect transport costs. The local endowment of transport infrastructures is measured as the average distance of the centroid of each commuting zone from the nearest airport, seaport, intermodal facility, highway, state route, and freight railroad. Notably, some alternative regressions have been run considering the distance from each single type of infrastructure, rather than their average.[5] However, the various distances from the different types of infrastructures turn out being rarely significant, both statistically and economically, while some significance is gained by considering their average. This suggests that firms are solving a problem of joint transport cost minimization, possibly because they rely on multiple types of transport infrastructures across virtually all sectors.

An unbiased estimate of sector-specific localization externalities requires also to control for *inter*-sectoral spatial relations. As discussed by Marshall (1890) and Venables (1996), localization choices within a certain sector might be correlated with those of another sector due

[3]For a summary of the different drivers of firm localization considered in the literature, see Beaudry and Schiffauerova (2009), Head and Mayer (2004), Puga (2010), Rosenthal and Strange (2004).

[4]The Shannon variety index for each region is computed as $S_l = \sum_i s_{l,i} \log(1/s_{l,i})$, where $s_{l,i}$ is the share of the $i$-th sector in region $l$.

[5]Also, other trials regressions have been attempted. In particular, the distance from seaports was dropped due to its dependence on the presence of maritime coasts, which enter in another control. However, this variation does not affect the results. Similarly, the infrastructural endowments have been measured also in alternative ways. In particular, instead of considering the distance of the region from the relevant infrastructures, an alternative measure consists in counting the number of kilometers of highways, state routes, railways, as well as the number of airports, seaports and intermodal facilities in each region. In general, statistical significance diminishes with these alternative measures, thus indicating that the one finally adopted here is more accurate in detecting the relevant effects.




(a) (b)

Figure 4: Spatial distribution of population across US commuting zones.

to self-reinforcing mechanisms acting across sectors. Inserting as a control the local share of vertically related sectors is meant to wash away from the estimate of sector-specific externalities the effect *inter*-sectoral dependencies. This control is computed by giving more weight to the sectors that are more strongly related to the one under scrutiny according to input-output tables. Given an input-output matrix $M$, label with $m_{i,j}$ its element expressing the value of the generic input-output relation between sector $i$ and $j$. For sector $i$, the weight of such relation is given by $w_{i,j} = m_{i,j} / \sum_j m_{i,j}$, where $w_{i,j} = 0$ if $i = j$. Labeling as $s_{j,l}$ the share of sector $j \neq i$ in region $l$, the weighted local share of sectors that are vertically integrated with $i$ is given by $v_{i,l} = \sum_j s_{j,l} w_{i,j}$ $\forall i \neq j$. Hence, the variable $v_{i,l}$ changes across sectors due to their different input-output relations. Table 2 reports the cross-sectoral arithmetic average of this variable. Notice that, in principle, weights can be assigned according to either downstream or upstream relationships among sectors. In practice, however, the values of $v_{i,l}$ that result from these alternative weighting procedures display a correlation amounting to $r = 0.998$ ($p$-value $= 0.001$). Therefore, in the present context, either one of them will be referred to in terms of generic vertical integration.

As a remark, it is worth emphasizing that the spatial distributions of these controls display the same type of stability over time that is displayed by the spatial distributions of firms. In particular, Figure 4 shows the spatial distribution of population across the commuting zones of the United States in year 1998 and 2010. This control is particularly relevant since it will turn out being orders of magnitudes more explicative than any other control (see Section 5). In fact, the Kolmogorov-Smirnov test reveals no statistically significant variation over time has occurred in the spatial distribution of population ($p$-value $= 0.240$). This evidence provides further support to the idea that we are actually observing a spatial equilibrium, thus meeting the necessary condition for a sensible application of the present approach.

## 3.2 Model selection and goodness of fit

As already mentioned, the methodology presented here allows to evaluate directly the relative performance of the Polya and Multinomial models. The former differs from the latter by the sole externality parameter $b$, thus placing the analysis in the context of a comparison between nested models. In order to compare the relative performances of the two models it is possible to use the Akaike Information Criterion corrected by finite sample size, $AICc$ (see Akaike, 1974, Hurvich and Tsai, 1989). Formally, this statistics is defined as

$$AICc = 2k - 2\ln(\ell) + \frac{2k(k+1)}{n-k-1} \,, \tag{14}$$

where $n$ is the sample size, $k$ is the number of parameters in each model, and $\ell$ is the maximized value of the likelihood function. Between two alternative models, the one with a lower $AICc$ value is to be preferred as it "dissipates" less information. In this sense, the definition of the $AICc$ prizes the goodness of fit of the model via $\ell$, while it penalizes its parametric numerosity through $k$.

In principle, this procedure of model selection could produce three qualitatively distinct results. On the one hand, it could turn out that the multinomial model outperforms the Polya model in each single sector under scrutiny. This type of result would suggest that a model with positive sector-specific externalities does not reveal any extra information about the spatial distribution of firms, as compared to a model in which such externalities do not exist. On the other hand, also the opposite scenario could take place, one in which the Polya model systematically outperforms the multinomial model. In this case, no economic sector could do without an account of positive sector-specific externalities in order to explain completely the spatial distribution of firms. Still, the strength of externalities relative to other determinants of firm localization would remain to be assessed. Finally, the multinomial model could outperform the Polya model only in some sectors, while the opposite would be true in other sectors. As reported in Section 5, the Polya model outperforms the multinomial model in each single sector of the United States economy in year 2010.

Once a model is selected, its goodness of fit is measured by comparing the predicted and the observed number of plants per region. Formally,

$$\text{Efron-}R^2 = 1 - \frac{\sum_l (n_l - n_l^*)^2}{\sum_i (n_l - \bar{n})^2} \,, \tag{15}$$

where $n_l^*$ is the prediction of the model. This measure is bound in $[0, 1]$ and thus allows for direct comparisons across sectors. Clearly, the absolute performance of the model in explaining the observed variability is an indicator of how much an omitted variables bias could possibly affect the results. A high fit indicates that, whatever the missing controls are, they are unlikely to affect substantively the estimated strength of localization externalities. In fact, this turns out to be actually the case.

# 4 Data

The localization of economic activities is identified with sectoral plant data coming from the U.S. Census Bureau (2010b) County Business Pattern database. On the side of control variables, population and income data come from the U.S. Census Bureau (2010a,c), while the other controls are derived from georeferenced layers coming from the National Atlas (2013) and from the input-output tables of the Bureau of Labor Statistics (2010).

Table 1: Summary statistics on the spatial distributions of a subsample of sectors.

| NAICS-BLS Sector | Total plants | Mean | Std.Dev | Max |
|---|---|---|---|---|
| 1133-Logging | 8254 | 12 | 21 | 186 |
| 213-Support activities for mining | 11866 | 17 | 50 | 684 |
| 23-Construction | 677305 | 980 | 2274 | 27944 |
| 3121-Beverage manufacturing | 4473 | 6 | 28 | 649 |
| 313/314-Textile | 8739 | 13 | 39 | 721 |
| 315-Apparel manufacturing | 7150 | 10 | 115 | 2772 |
| 3254-Pharmaceutical manufacturing | 1935 | 3 | 10 | 150 |
| 3315-Foundries | 1900 | 3 | 8 | 125 |
| 3344-Semiconductor manufacturing | 4390 | 6 | 29 | 505 |
| 44/45-Retail trade | 1324570 | 1917 | 4577 | 61778 |
| 5112-Software publishers | 7224 | 10 | 40 | 409 |
| 512-Motion picture industries | 23875 | 35 | 266 | 6334 |
| 517-Telecommunications | 51462 | 74 | 181 | 2290 |
| 6216-Home health care services | 27208 | 39 | 111 | 1570 |
| 7111-Performing arts companies | 8578 | 12 | 61 | 1089 |
| 722-Food services and drinking places | 576020 | 834 | 2251 | 31893 |

For what concerns spatial disaggregation, all county-level data have been aggregated in commuting zones following standard correspondences (see U.S. Department of Agriculture-ERS, 2010). Therefore, the regions over which plants are spread represent economically meaningful units, as they identify local labor markets. This constitutes a distinctive trait of the present analysis as compared to other studies, which typically resort to merely administrative spatial units. Moreover, aggregating by commuting zones ends up absorbing much of the spatial autocorrelation that generally exists among smaller units such as counties. Also for this reason the present work does not introduce explicitly any spatial lag in the variables of inquiry. By using commuting zones, the number of spatial units amounts to 691 after having dropped Alaska and Hawaii from the analysis.

As for sectoral disaggregation, plants are subdivided according to the Bureau of Labor Statistics (2010) classification. This choice is motivated by the need to have sectoral correspondence between plant data and the input-output matrix that serves to compute the local share of vertically-related sectors, as explained above. With this choice, the present analysis spans

161 sectors ranging from manufacturing and services to agriculture related ones. More than 85% of the national economy is taken in consideration.

As a remark, it is worth signaling that there is an analytic reason to use plant rather that employee data to identify the localization of economic activities. That is, differently from the number of employees, the number of plants settled in a region is an accurate measure of the number of localization choices on the side of firms that have identified such region as preferable. Summary statistics on the spatial distributions of plants are shown for a subsample of sectors in Table 1, which gives an intuition of the strong heterogeneity existing among sectors as well as of the vast variety of economic activities included in the present analysis.

Table 2: Summary statistics on the spatial distributions of control variables.

| Variable by commuting zone | Mean | Std.Dev | Min | Max |
|---|---|---|---|---|
| Population | 443732 | 1144284 | 1006 | 17897500 |
| Area (kmsq) | 19089 | 18057 | 225 | 219265 |
| Median income | 42695 | 8055 | 25912 | 86363 |
| Distance from infrastructures (km) | 138 | 84 | 10 | 391 |
| Number of mines | 0 | 1 | 0 | 19 |
| Cropland (% of area) | 46 | 24 | 0 | 97 |
| Forest (% of area) | 36 | 32 | 0 | 99 |
| Rivers (m per kmsq) | 38 | 23 | 0 | 117 |
| Coasts (m per kmsq) | 18 | 79 | 0 | 824 |
| Local share of related sectors (‰)* | 1 | 3 | 0 | 46 |

* The datum reported here is the cross-sectoral arithmetic average (see text).

# 5 Results

This section reports a summary of the marginal elasticities estimates deriving from the procedure described in Section 3. Both the Polya and the Multinomial models have been estimated on all 161 sectors in the dataset. For the sake of brevity, only the main results are highlighted here, while the full set of estimates can be consulted at the authors' web page.[6]

To begin with, the procedure of model selection described in Section 3.2 ends up delivering a clearcut result. The Polya model outperforms the Multinomial one in each single sector of the United States economy: that is, the $AICc$ values associated to the former are systematically lower than those associated to the latter model. Hence, no sector can do without an account of localization externalities for its spatial distribution to be fully explained.[7] However, the actual strength of externalities remains to be assessed through the estimation of marginal elasticities. Nevertheless, given the outcome of the model selection procedure, it is possible to focus only on the estimates resulting from the Polya model without fearing to lose any relevant information.

[6] See https://sites.google.com/site/gragnolati/results/bograusa_2013.

[7] A partial exception is represented by sector `622-Hospitals`, in which the $AICc$ values of the Polya and Multinomial model are identical.

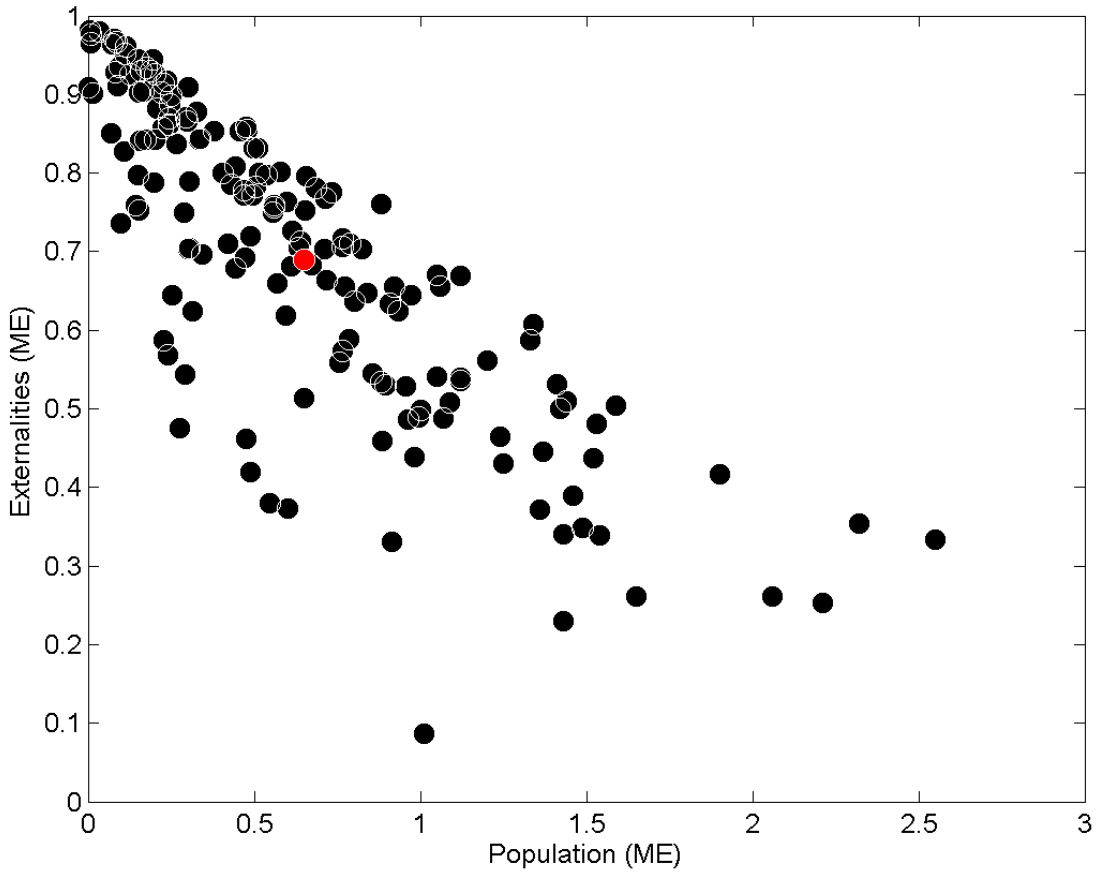



Figure 5: Estimates of marginal elasticities. Each black dot corresponds to a sector, while the red dot is the cross-sectoral arithmetic average. The estimates are obtained from equations (12) and (9).

For what concerns the role of the control variables, another clearcut result emerges. As an average across sectors, the effect of population largely outperforms the effects of any other control variable. More precisely, the marginal elasticity associated to population is the only one being statistically significant at least at 99% level across all sectors, while being also, on average, at least one order of magnitude greater than the elasticities associated to the other controls.[8] For instance, the average magnitude of the marginal elasticity associated to population is 0.65, followed from very far away from the local share of related sectors with an average elasticity of 0.04. Given their weakness, the elasticities related to the other controls are not discussed here.

In fact, the only determinant of firm localization that is comparable in strength to the pull of population are sector-specific localization externalities. Figure 5 illustrates this point by showing the marginal elasticities associated to population in relation with those associated to localization externalities. Each black dot in the figure represents one of the 161 sectors under scrutiny, while the red dot is their cross-sectoral average. With an average magnitude of 0.69, the effect of localization externalities turns out being slightly stronger that the effect of population and predominant in about 60% of the sectors. Moreover, the marginal elasticity

[8]Sector `114-Fishing, hunting and trapping` represents an exception, being the only one in which the marginal elasticity of population is not statistically significant.

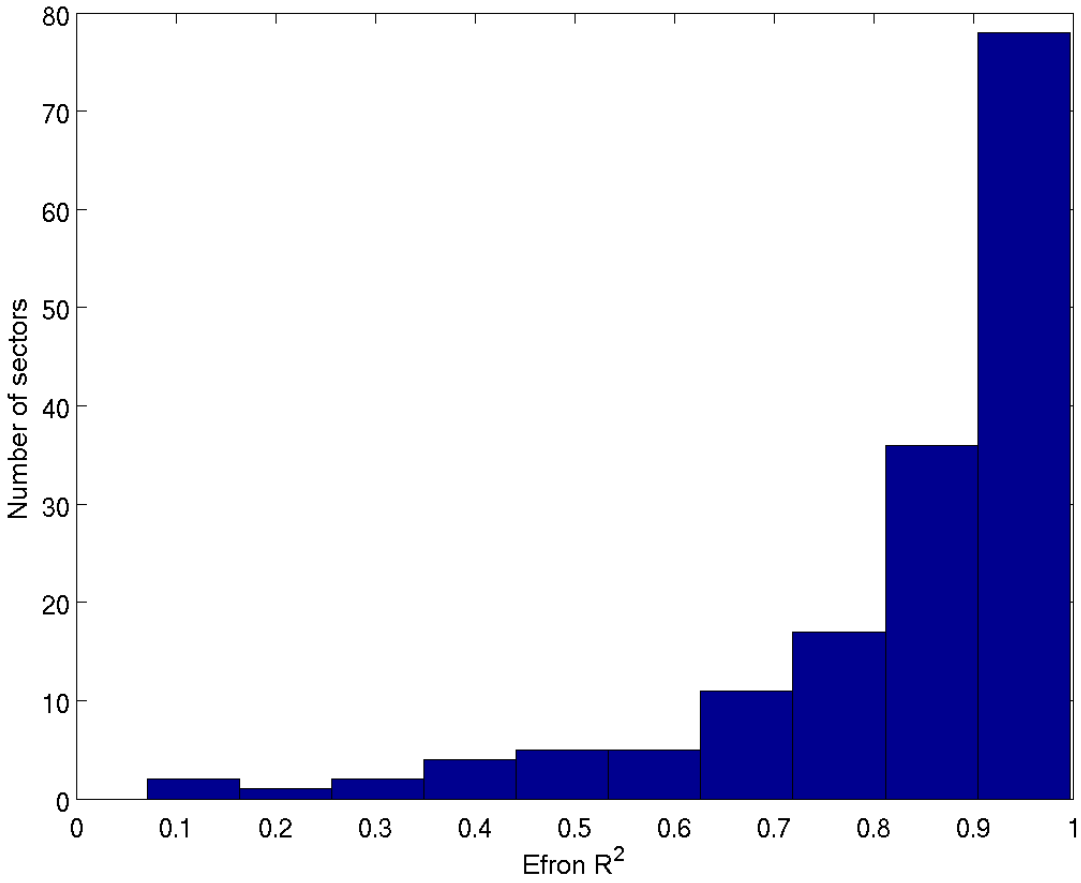



Figure 6: Distribution of the Efron $R^2$ across sectors.

associated to externalities always reaches at least a 99% level of statistical significance across all sectors. Therefore, their average strength is not concentrated in just a few sectors, but rather the opposite: localization externalities diffusely predominate across various economic activities.

The robustness of this set of conclusions is strengthened by the remarkable goodness of fit of the model. As an average across sectors, the Polya model correctly predicts about 83% of the number of plants in each region. As shown in Figure 6, the Polya model explains more than 70% of the observed variance in 131 sectors out of 161. Among other things, this implies that there is very little room for omitted variables bias in a vast majority of sectors.

# 6 Conclusion

This work has presented a methodology to disentangle and quantify the strength of localization externalities, as well as of other drivers of firm location. Such methodology is based on a discrete choice model in which the idiosyncratic preferences of firms enter as a stochastic term, thus depicting agents as being heterogeneous. The resulting equilibrium is of a stochastic type, as it predicts the probability distribution of firms across regions given some region-specific features that capture the intrinsic attractiveness of each location. Relying on maximum likelihood, it is then possible to estimate the marginal effects of the various determinants on the probability

for a region to attract and extra firm.

When applied on the various economic sectors of the US, this estimation procedure has delivered three main results. On the one hand, our analysis finds empirical justification for the huge stress that economic geography has traditionally put on urbanization economies and localization externalities. Indeed, these two forces turn out being the most important drivers of firms location in terms of relative magnitudes. On the other hand, our analysis finds localization externalities to be widespread across the entire economy, rather than in some particular sectors. In particular, strong localization externalities do not result to be a prerogative of high-tech sectors. In comparative terms, these results for the US are in line with those obtained for Italy in Bottazzi and Gragnolati (2012).

It remains an open issue to better understand the role that knowledge plays as a source localization externalities. Having found their effect to be so strong and widespread in the economy might imply that knowledge spillovers are equally diffused. To tackle this issue more precisely, it might be worthwhile to apply the framework presented here specifically to the localization of innovative activities, and then compare the results with those of other approaches in the literature (see Audretsch and Feldman, 1996, Jaffe et al., 1993, among others). This might well gives some further insights to the understanding of the role of knowledge in the geography of production and innovation.

# References


H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

W.B. Arthur. 'Silicon Valley' locational clusters: When do increasing returns imply monopoly? *Mathematical Social Sciences*, 19(3):235–251, 1990.

D.B. Audretsch and M.P. Feldman. R&d spillovers and the geography of innovation and production. *American Economic Review*, 86(3):630–640, 1996.

E. Bartelsman, S. Scarpetta, and F. Schivardi. Comparative analysis of firm demographics and survival: evidence from micro-level sources in OECD countries. *Industrial and Corporate Change*, 14(3):365–391, 2005.

C. Beaudry and A. Schiffauerova. Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy*, 38(2):318–337, 2009.

D. Black and V. Henderson. Spatial evolution of population and industry in the United States. *American Economic Review*, 89(2):321–327, 1999.

L.E. Blume, W.A. Brock, S.N. Durlauf, and Y.M. Ioannides. Identification of social interactions. *Handbook of Social Economics*, 1:855–966, 2011.

G. Bottazzi and P. Dindo. Gobalizing knowledge: How technological openness affects output, spatial inequality, and welfare levels. *Journal of Regional Science*, 2013.

G. Bottazzi and U.M. Gragnolati. Cities and clusters: economy-wide and sector specific effects in corporate location. *Regional Studies*, 2012. Forthcoming.

G. Bottazzi and A. Secchi. Repeated choices under dynamic externalities. LEM Working Paper Series, September 2007. URL http://www.lem.sssup.it/WPLem/files/2007-08.pdf.

G. Bottazzi, G. Dosi, G. Fagiolo, and A. Secchi. Modeling industrial evolution in geographical space. *Journal of Economic Geography*, 7(5):651–672, 2007.

G. Bottazzi, G. Dosi, G. Fagiolo, and A. Secchi. Sectoral and geographical specificities in the spatial structure of economic activities. *Structural Change and Economic Dynamics*, 19(3): 189–202, 2008.

Bureau of Labor Statistics. Inter-industry relationships (Input/Output matrix), July 2010. URL http://www.bls.gov/emp/ep_data_input_output_matrix.htm.

K. Desmet and M. Fafchamps. Employment concentration across US counties. *Regional Science and Urban Economics*, 36(4):482–509, 2006.

M.P. Devereux, R. Griffith, and H. Simpson. The geographic distribution of production activity in the UK. *Regional Science and Urban Economics*, 34(5):533–564, 2004.

G. Dumais, G. Ellison, and E.L. Glaeser. Geographic concentration as a dynamic process. *Review of Economics and Statistics*, 84(2):193–204, 2002.

G. Duranton and H.G. Overman. Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4):1077, 2005.

G. Ellison and E.L. Glaeser. Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5):889–927, 1997.

G. Ellison and E.L. Glaeser. The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review*, 89(2):311–316, 1999.

R. Forslid and G.I.P. Ottaviano. An analytically solvable core-periphery model. *Journal of Economic Geography*, 3(3):229–240, 2003.

K. Head and T. Mayer. The empirics of agglomeration and trade. *Handbook of Regional and Urban Economics*, 4:2609–2669, 2004.

J.V. Henderson. Marshall's scale economies. *Journal of Urban Economics*, 53(1):1–28, 2003.

C.M. Hurvich and C.L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297, 1989.

J. Jacobs. *The economy of cities*. Vintage, New York, 1969.

A.B. Jaffe, M. Trajtenberg, and R. Henderson. Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598, 1993.

P.R. Krugman. Increasing Returns and Economic Geography. *Journal of Political Economy*, 99(3):483–499, 1991.

H. Luce. *Individual choice behavior*. Wiley, New York, 1959.

A. Marshall. *Principles of economics*. McMillan, London, 1890.

F. Maurel and B. Sédillot. A measure of the geographic concentration in French manufacturing industries. *Regional Science and Urban Economics*, 29(5):575–604, 1999.

National Atlas of the United States of America, January, 15 2013. URL http://www.nationalatlas.gov/atlasftp.html.

D. Puga. The magnitude and causes of agglomeration economies. *Journal of Regional Science*, 50(1):203–219, 2010.

M. Raouf Jaibi and Thijs ten Raa. An asymptotic foundation for logit models. *Regional Science and Urban Economics*, 28(1):75–90, 1997.

S.S. Rosenthal and W.C. Strange. The determinants of agglomeration. *Journal of Urban Economics*, 50(2):191–229, 2001.

S.S. Rosenthal and W.C. Strange. Evidence on the nature and sources of agglomeration economies. *Handbook of Regional and Urban Economics*, 4:2119–2171, 2004.

U.S. Census Bureau. Small Area Income and Poverty Estimates, July 2010a. URL http://www.census.gov/did/www/saipe/data/statecounty/data/2009.html.

U.S. Census Bureau. County Business Patterns, July 2010b. URL http://www.census.gov/econ/cbp/download/10_data/index.htm.

U.S. Census Bureau. Population Estimates, 2010c. URL http://www.census.gov/popest/data/historical/2000s/index.html.

U.S. Department of Agriculture-Economic Research Service (ERS). Measuring rurality: Commuting zones and labor market areas, July 2010. URL http://www.ers.usda.gov/Briefing/Rurality/readings.htm. Last consulted in January 2012.

A.J. Venables. Equilibrium locations of vertically linked industries. *International Economic Review*, 37(2):341–359, 1996.

J.I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15 (2):109–144, 1977.