Sectoral and geographical specificities in the spatial structure of economic activities

Giulio Bottazzi\textsuperscript{a,b,*}, Giovanni Dosi\textsuperscript{a}, Giorgio Fagiolo\textsuperscript{a}, Angelo Secchi\textsuperscript{b}

\textsuperscript{a} Sant'Anna School of Advanced Studies, Pisa, Italy
\textsuperscript{b} Dse, University of Pisa, Italy

ABSTRACT

This work explores the spatial distribution of productive activities in the Italian manufacturing industry. We propose an econometric model which tries to disentangle location-specific from sectoral drivers in the dynamic process of spatial agglomeration. The basic idea is that the former typically apply "horizontally" (i.e. across all industrial sectors), while the latter unfold in the form of non-decreasing dynamic returns to the current stock of installed business units. Three different specifications of the model are tested against Italian data on the location of manufacturing activities, studying the distribution of the number of firms and employees. Our results suggest that different locations exert different structural influences on the distribution of both variables. Moreover, a significant horizontal power of "urbanization", which makes some locations, especially metropolitan areas, more attractive irrespectively of the sector, does emerge. However, after controlling for the latter, one is still left with very significant sector-specific forms of dynamic increasing returns to agglomeration, which vary a lot across different manufacturing activities and which plausibly have to do with sectoral-specific and localized forms of knowledge accumulation and spin-offs.

1. Introduction

This work studies the structure of the statistical distribution of economic activities in the geographical space. In particular, we propose different econometric exercises, based on the stochastic Markov model of firm location developed in Bottazzi et al. (2007), aimed to disentangle two distinct classes of agglomeration drivers: "location-specific" drivers, which cut across different types of economic activities and "technology-specific" drivers, whose effect changes across different production sectors.

The ways economic activities are distributed over geographical space along relatively ordered patterns has been a concern for economic analysis at least since Alfred Marshall. Indeed, the first basic stylized fact of economic geography is that locational patterns, over the whole history for which we have some records throughout the world, tend to be much more clustered than any theory of comparative advantage might predict (cf. Krugman, 1991; Fujita et al., 1999, among many others).

At the same time, the evidence suggests a remarkable inter-sectoral variability in agglomeration structures. This applies across different countries such as the US, France, the UK, Germany and Italy: cf. Ellison and Glaeser (1997). Maurel and...
diverse degrees of “attractiveness” of different locations. So, for instance, there are several locations where business units
belonging to almost all sectors are equally represented. On the contrary, in many other sites, agglomeration occurs only for
business units belonging to a small number of sectors (in some cases, one or two). For example, as discussed in Bottazzi et al.
(2006), in the Italian case a quite large fraction of sectors is not even represented in more than 50% of locations. Moreover,
any measure of agglomeration appears to be quite stable over time, notwithstanding the great variability of agglomeration
observed across locations and a turbulent underlying micro-dynamics with persistent flows of entry, exit, and variation in
the relative sizes of incumbents (Dumais et al., 2002).

Taken together, the foregoing pieces of evidence suggest a general picture characterized by different drivers of agglomera-
tion, which might be economy-wide, location-specific or sector-specific. Acknowledging the heterogeneous nature of the
different agglomeration forces, in this work we investigate the relative role of location-specific mechanisms of agglomeration,
independent of individual sectors and technologies vs. the sector-specific ones which apply across different locations within
similar ensembles of production activities. The idea behind the present analyses is that cross-sectoral differences in agglom-
eration forces ought to be, at least partly, explained on the grounds of underlying differences in the relative importance of
phenomena such as localized knowledge spillovers, intra- vs. intra-organizational learning, knowledge complementarities
fueled by localized labor pooling, innovative explorations undertaken through spin-offs and, more generally, the birth of
new firms.

The proposed econometric exercises are different specifications of the simple stochastic model developed in Bottazzi et al.
(2007). This model is built upon the idea of dynamic increasing returns and shares its general structure and several
hypotheses with the models explored by Arthur (1994), Dosi et al. (1994) and Dosi and Kaniovski (1994). However,
in order to obtain empirically testable predictions, instead of the irreversible pure-birth dynamics characterizing those
models, we consider a Markov dynamics where the reversibility of locational choices by firms entails a notion of stochas-
tic equilibrium (i.e. invariant limit distribution). Bottazzi and Secchi (2007) show that, under rather general hypothesis
about the selection mechanism characterizing a heterogeneous population of agents, this equilibrium is equivalent to the
Ehrenfest–Brillouin urn-scheme (cf. Garibaldi and Penco, 2000 and Garibaldi et al., 2002). Building on this notion of dynamic
equilibrium characterizing the spatial distribution of “productive units”, which can be either plants or unit of employment,
we obtain three different statistical models that we estimate using Italian data, disaggregated by “locational units” and by
sector.

Let us illustrate the intuition behind our analysis borrowing from the “dartboard” metaphor discussed in the seminal
work by Ellison and Glaeser (1997). Suppose that the economic space is a sort of dartboard where darts of different colors
are thrown (i.e. economic activities belonging to different sectors are located). Here, the null hypothesis (i.e. “agglomera-
tion does not matter”) is a distribution of darts on the board solely due to random factors. In departing from pure random-
ness, however, one might observe systematic patterns ultimately due to three different factors. The first one has to do
with the generic attractiveness (or repulsiveness) of some areas on the board: hence, one will systematically find there
more (or less) darts of all colors than what sheer randomness would predict. That is, to trivialize, one will find “more of
everything” in New York as compared to Pisa, irrespectively of any finer pattern of comparative advantage. Second, on
the top of these generic locational patterns, one may observe specific patterns distinctive of any one color (i.e. sectoral
specificities). Finally, the last concerns the different size of different darts (i.e. different degrees of lumpiness of single
investments).

Ellison and Glaeser (1997, 1999) and Dumais et al. (2002) control for the latter, as captured by the concentration in
plant size distribution, and study the importance of sector-specific agglomeration factors as compared to inter-sectoral,
location-wide, ones (which they call “natural advantage” of a location).1

Our exercise largely shares a similar spirit, albeit with some distinct features. Indeed, we do not “wash out” any lumpiness
effect. We do it partly out of necessity and partly out of choice. The constraint is that given our small spatial units (defined
in terms of local labor mobility basins, typically smaller than most US counties) and our fine-grained sectoral partition, it is
very hard to find the relevant sectoral/spatial breakdown of the data. At the same time, at a conceptual level, it is not entirely
uncontroversial that one should take out the “size effect”. In order to see this, think of, say, five entities located in one particular
place which at some point merge into one. This does not mean that agglomeration has fallen, but rather that whatever
forces driving agglomeration have now been internalized within a single firm. Thus, complementary information may be
usefully obtained by studying, side by side, the agglomeration dynamics in terms of number of firms and of employment
units.

The rest of the paper is organized as follows. After a brief description of our data, in Section 3 we present the basics of
the stochastic model derived from Bottazzi et al. (2007) which constitutes the conceptual framework for the econometric
specifications discussed in Section 4. In Section 5 we test these specifications against data on locational patterns of different
sectors of the Italian manufacturing industry, using both firms and workers as proxy for production units. Finally, Section 6
concludes.

1 Refinements and applications of this basic methodology are in Maurel and Sedillot (1999), Devereux et al. (2004) and Overman and Duranton (2002).
See also the detailed reviews in Combes and Overman (2004) and Ottaviano and Thisse (2004).
2. Data

This research draws upon the “Census of Manufacturers and Services”, a database developed by the Italian Statistical Office (ISTAT) that contains observations about five millions employees and more than half a million business units (BUs). Each observation identifies the location of the employees and of the business units at a given point in time (1996), as well as the industrial sector which they belong to. We consider data disaggregated according to the Italian ATECO classification (which corresponds to the NACE classification system). Among all industries, we focus on the manufacturing segment excluding, however, the sector “16—Tobacco products” which presents a too limited number of business units.

Business units and employees are classified with respect to 784 geographical locations. Each geographical location represents a “local system of labor mobility” (LSLM), that is a geographical area characterized by relatively high internal labor commuters’ flows. LSLMs are periodically updated by multivariate cluster analyses employing census data about social, demographic, and economic variables (see Sforzi, 2000 for details). Table 1 reports for each sector a brief description of the occupancy distribution of employees and business units across sites.

3. A stochastic model of location with dynamic increasing returns

Consider a single-sector economy composed by a fixed number of location, $L$, which can be thought as production sites, and populated by a constant number, $N$, of heterogeneous agents representing different production units. Agents, which are assumed to be boundedly rational profit seekers, have to choose where to locate themselves among the set of available locations. The sequence of locational choices by agents is described as a stochastic process: at each time step an agent is chosen at random to die (i.e. to leave the location where it operates) and, once the exit took place, a new agent enters the economy selecting as productive location the one which maximize his expected utility. The possibility that agents possess heterogeneous preferences and beliefs is introduced by assuming that the expected return associated to different locations posses a common component and an individual, idiosyncratic, one. In turn, the common component is characterized by a constant term which describes the intrinsic “geographic attractiveness” of each locations and by a “social term” which depends on the actual distribution of agents across different locations and captures the strength of agglomeration forces. Bottazzi and Secchi (2007) show that, under rather general assumptions about agents’ preferences structure, their locational choices are, in probability, driven exclusively by the common component of the expected individual return. Assuming a linear form for the social term, the new entrant chooses location $l\in\{1, \ldots, L\}$ with probability:

$$p_l \sim a_l + b_l n_{l,t-1},$$  \hspace{1cm} (3.1)

---

Table 1
Descriptive statistics of the firm occupancy distribution by sector in 1996

<table>
<thead>
<tr>
<th>Sector</th>
<th>Statistics of the occupancy distribution</th>
<th>Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Business units</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>Mean</td>
</tr>
<tr>
<td>15 Food products</td>
<td>75420</td>
<td>96.2</td>
</tr>
<tr>
<td>17 Textiles</td>
<td>36217</td>
<td>46.2</td>
</tr>
<tr>
<td>18 Apparel</td>
<td>49782</td>
<td>63.5</td>
</tr>
<tr>
<td>19 Leather products</td>
<td>25451</td>
<td>32.5</td>
</tr>
<tr>
<td>20 Wood processing</td>
<td>50662</td>
<td>64.6</td>
</tr>
<tr>
<td>21 Pulp and paper</td>
<td>5268</td>
<td>6.7</td>
</tr>
<tr>
<td>22 Publishing and printing</td>
<td>28183</td>
<td>36.0</td>
</tr>
<tr>
<td>23 Coke, refined petroleum and nuclear fuel</td>
<td>825</td>
<td>1.1</td>
</tr>
<tr>
<td>24 Organic and Inorganic chemicals</td>
<td>7593</td>
<td>9.7</td>
</tr>
<tr>
<td>25 Rubber and plastic products</td>
<td>14626</td>
<td>18.7</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>30709</td>
<td>39.2</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>4034</td>
<td>5.1</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>94771</td>
<td>120.9</td>
</tr>
<tr>
<td>29 Industrial machinery and equipment</td>
<td>42984</td>
<td>54.8</td>
</tr>
<tr>
<td>30 Office machinery</td>
<td>592</td>
<td>0.7</td>
</tr>
<tr>
<td>31 Electrical machinery</td>
<td>17312</td>
<td>22.1</td>
</tr>
<tr>
<td>32 Radio, TV and TLC devices</td>
<td>9773</td>
<td>12.5</td>
</tr>
<tr>
<td>33 Precision instruments</td>
<td>28280</td>
<td>36.1</td>
</tr>
<tr>
<td>34 Motor vehicles and trailers</td>
<td>2261</td>
<td>2.9</td>
</tr>
<tr>
<td>35 Other transport equipment</td>
<td>4514</td>
<td>5.8</td>
</tr>
<tr>
<td>36 Furniture</td>
<td>59627</td>
<td>76.1</td>
</tr>
<tr>
<td>37 Recycling</td>
<td>2061</td>
<td>2.6</td>
</tr>
</tbody>
</table>
where \( n_{l,t-1} \) is the number of agents present in that location at the end of the previous time step. The coefficient \( a_l \) represents the geographical attractiveness of location \( l \) and captures the gain that an agent on average expects by choosing to locate its activity in a given site irrespectively of the choices of other agents. This coefficient might be interpreted as controlling for intrinsic exogenous geographical factors (e.g., cost of inputs, infrastructures, etc.). Conversely, the parameter \( b_l \) represents the social term and measures the strength of agglomeration economies in a given location: it is the amount by which the advantages obtained by locating in a certain site increases as a function of the number of agents already located there due for instance to technological factors and externalities of various types. A larger value of the parameter \( b_l \) implies that the incentive for an agent to locate in that site increases faster with the number of agents that have already settled there.

Before we illustrate how this model can be used to build empirically testable specifications, two remarks are in order. First, notice that the new “entrant” may well choose a location different from the one where “death” occurred. Thus the model is designed to capture both genuine entry of new agents and the reversibility of locational decisions of incumbents which might exit from one site just to select another one elsewhere. Second, in this model one may refer to events of birth and death as concerning both firms (more precisely business units) and employment opportunities (i.e. the appearance and disappearance of employment units). In both cases the assumption that entry rates are positive, constant and equal to exit rates can be justified on an empirical ground. Indeed the share of firms (employees) belonging to a given sector which enter and leave a given location in a relatively short period of time (e.g. a year) is typically much larger than the net growth of industry size, so that the time-scale at which spatial reallocations occur is generally quite short.3 Similar considerations apply to employment turnover whereby one observes quite high gross turnover even in presence of low net variations.4

In our model, the dynamics implied by the foregoing rules for entry and exit is equivalent to a finite Markov chain whose state space is the set of all the possible distributions of the \( N \) agents across the \( L \) locations. In particular, it can be shown (cf. Bottazzi and Secchi, 2007) that the assumptions of zero net-entry together with the reversibility of individual locational decisions and the constant impact of any single decision on the state of the system (implied by the Eq. (3.1)) guarantee that the evolution of locational choices is an ergodic process that allows for non degenerate limit distributions. Moreover, Bottazzi and Secchi (2007) show (cf. Proposition 3.4) that the process governing the evolution of the industry admits a unique long-run equilibrium (i.e. a unique invariant limit distribution) so that a probability \( \pi \) is assigned to each possible configuration \( \mathbf{n} = (n_1, \ldots, n_L) \) where \( n_l \) is the number of agents in location \( l \). This limit distribution \( \pi(\mathbf{n}; \mathbf{a}, \mathbf{b}) \) is analytically characterized as a function of the set of parameters of the model, the \( L \)-tuples of the geographic attractiveness \( a = (a_1, \ldots, a_L) \) and of the agglomeration strength \( b = (b_1, \ldots, b_L) \) of the \( L \) different locations. By varying the relative strength of geographical attractiveness and of agglomeration positive feedbacks this model is able to reproduce a rich variety of different patterns of spatial concentration. At one extreme, when agglomeration forces are very low, different locations attract on average a number of agents that is proportional to their geographical attractiveness, \( a_l \). At the other hand, when agglomeration forces are very strong this model implies the emergence of highly polarized distribution, where few locations capture the great majority of agents.

To sum up, the dynamics governing the model does generate sharp empirically testable implications, in terms of the probability of finding the economy in a given state \( \pi(\mathbf{n}; \mathbf{a}, \mathbf{b}) \). Notice that this equilibrium (limit) distribution does not necessarily depict a long-run (limit) state associated to some ‘old’ or ‘mature’ industry. Since each entry/exit decision made by any one firm constitutes one time-step in the model, the invariant distribution describes the state of the system after a sufficient large number of spatial reallocation events have taken place (which may well imply a relatively short period of real time). Invariant distributions can then be directly compared with cross-section empirical data as far as they describe a system which is, on average, near its stochastic equilibrium state.

4. Testable instances of the model

The most general version of the model described in the previous section does contain a quite large number of free parameters. More precisely, one has to deal with two parameters for each location \( l \): its geographic attractiveness \( a_l \) and the local strength of agglomeration \( b_l \). In order to estimate such a model against empirical observations time series data on the number of firms in every single location would be required. Unfortunately, we do not have such information. Indeed, in the following we apply the model to a dataset, described in Section 2, which contains only one observation per location per industrial sector. This forces us to explore less general models containing a lower number of parameters. Consequently, in estimating our model on empirical data, we will mainly employ the marginal distribution of the number of firms in a given location. The probability to find \( n \) out to \( N \) firms in a location characterized by coefficients \((a, b)\), \( \pi(n; a, b, N, L) \), can be easily obtained from the limit distribution \( \pi(\mathbf{n}; \mathbf{a}, \mathbf{b}) \) (Bottazzi and Secchi, 2007).

---

3 For a detailed comparative cross-country overview concerning firms turnover cf. Bartelsman et al. (2005). On the Italian case, see, e.g. Quarterly Reports by Unioncamere, “Movimprese: Dati Trimestrali sulla Nati-Mortalità delle Imprese”, Uffici Studi e Statistico Camere di Commercio, Italy, various years, available on line at the url: http://www.starnet.unioncamere.it. Clearly the extent to which the assumption of zero net entry is realistic depends on the level of aggregation. At higher level of disaggregation one should in fact allow for (possibly endogenous) entry-exit processes with positive or negative net entry flows.

In what follows we present different instances of our general model, starting with a simple (and, as we will see, utterly unrealistic) example, characterized by “homogeneous” space and constant returns to agglomeration, and progressively introducing more general models that differentiate locations and sectoral dynamics.

4.1. Model 0: homogeneous locations without agglomeration effects

Let us start with the simplest model where the agglomeration strength parameter is set to zero in any location and all locations possess the same geographic attractiveness \( a \),

\[ a_l = a > 0, \quad b_l = 0 \quad \forall \ l. \]

This case represents as a sort of “null hypothesis” benchmark whereby neither spatial specificities nor agglomeration processes play any lasting role. In this extreme setup, firms choose locations totally at random. The limit distribution \( \pi(n; a, b) \) will then be multinomial, while the probability to find \( n \) firms in any given location reduces to

\[ \pi(n; a, 0, N, L) = \binom{N}{n} \left( \frac{1}{L} \right)^n \left( 1 - \frac{1}{L} \right)^{N-n}, \tag{4.1} \]

that is a binomial distribution with \( N \) trials and probability \( 1/L \). Therefore, in a homogeneous-space model without agglomeration economies, the stationary distribution does not depend on the common geographic attractiveness \( a \). The intuition is that since the asymptotic occupancy of a location is driven by its relative attractiveness rather than its absolute one, when the value of the common parameter \( a \) is the same, the locations are all and always equally attractive. Consequently, given the full symmetry of the model, the marginal distribution is the same for all locations.

4.2. Model 1: homogeneous locations with agglomeration effects

Next, let us consider a model where locations are homogeneous and share the same positive geographic attractiveness \( a \), but one allows for agglomeration economies in the form of a positive and industry-wide agglomeration parameter \( b \)

\[ a_l = a > 0, \quad b_l = b > 0 \quad \forall \ l. \]

As in the Model 0 discussed above, also in this case all locations are identical with respect to the geographic attractiveness and the model is perfectly symmetric. The marginal distribution of the number of firms in a location \( \pi(n) \) does not depend on the particular chosen location and can be shown (Bottazzi et al., 2007) to follow a Polya distribution

\[ \pi(n; a, b, N, L) = \binom{N}{n} \frac{\Gamma'(La/b)}{\Gamma'(La/b+N)} \frac{\Gamma'(a/b+n)}{\Gamma'(a/b)} \frac{\Gamma'(L-1)a/b+N-n)}{\Gamma'(L-1)a/b} \tag{4.2} \]

where \( \Gamma' \) is the Gamma function. In this case the marginal distribution (4.2) depends on the total number of firms \( N \), the total number of locations \( L \) and the ratio of the two parameters \( a \) and \( b \). As an illustration, we report in Fig. 1 the Polya distributions for different values of the parameter \( b \). All distributions are computed according to (4.2), by setting \( a = 1 \) and with the same values for the parameters \( N \) and \( L \) (the latter values are chosen to be similar to the ones found in the subsequent empirical analyses). As shown, for small values of the parameter \( b \) the Polya distribution is similar to the Binomial distribution, with a positive modal value and its well-known “bell” shape. When the parameter \( b \) increases, the mode of the distribution moves

![Fig. 1. Polya marginal distributions for different values of \( b \). All distributions are computed for \( N = 20000, L = 784 \), and geographic attractiveness \( a = 1 \). Note that values for \( N \) and \( L \) are set to be similar to values empirically found in our subsequent analyses.](image-url)
towards \( n = 0 \) and the upper tail becomes noticeably fatter. In tune with the intuition on the properties of agglomeration economies, an increase in the agglomeration strength parameter \( b \) yields a stronger “clusterization” of firms, i.e. a large number of firms in few locations (hence the fat tail), leaving, at the same time, more locations empty (hence the modal value of zero).

### 4.3. Model 2: heterogeneous locations with agglomeration effects

Let us now relax the assumption of homogeneity among locations and consider different geographic attractiveness \( a_l \) for each different location \( l \). The strength of the agglomeration economy is still represented by an industry-specific parameter \( b \), equal for all locations. In this case locations do, in general, differ and are characterized by their specific attractiveness parameter \( a_l \). As it happens to Model 1, also in this case the marginal distribution of the number of firms in a location with geographic attractiveness \( a_l \) can be shown to follow a Polya distribution, given by

\[
\pi(n; a, A, b, N, L) = \frac{\Gamma(A/b) \Gamma(a/b + n) \Gamma((A-a)/b + N - n)}{\Gamma((A-a)/b) \Gamma((A-a)/b)},
\]

where \( A = \sum_{l=1}^{L} a_l \) (cf. Bottazzi et al., 2007). The marginal distribution in (4.3) depends, for a given location with attractiveness parameter \( a_l = a \), on the total number of firms \( N \), the total number of locations \( L \), the global parameter \( b \) and the location-specific parameters \( a_l \) through their sum \( A \).

### 5. Empirical analysis

To recall, the model presented in Section 3 describes the localization pattern of a single sector economy wherein the number of firms is kept constant and the economy is governed by a steady entry/exit process capturing both the flow of firms to and from the industry, and a reallocation process by incumbents across locations. As mentioned, the empirical flows in and out industries are quite high. Hence it is not implausible to assume that the actual observations tell us something about the underlying invariant distribution \( \pi(n; a, A, b) \). Of course, this does not rule out the possibility that in the long-term the nature and intensity of agglomeration drivers may well change. Such longer-term modifications may be captured by corresponding changes in the \( a \) and \( b \) coefficients (eventually detectable by comparing estimates across, say, different decades). However, since our database contains information on one single year, we can only compute the occupancy value for a given location and a given sector at a given point in time. This means that neither a direct verification of the dynamic process described in Section 3 nor a maximum-likelihood estimation of the equilibrium distribution in \( \pi(n; a, A, b) \) are possible. We have therefore to resort to some derived statistics in order to fit our models, exploiting the rich cross-sectional information which derives from the presence of multiple sectors.

Let \( n_j \) be the number of BUs in LSLM \( j \) operating in sector \( j \), where \( 1 \leq j \leq 22 \) and \( 1 \leq l \leq 784 \) (cf. Section 2). Denote with \( N_j = \sum_l n_j \), the total number of BUs operating in location (LSLM) \( j \) and with \( N_j = \sum_l n_j \), the total number of BUs belonging to the \( j \)th sector.

For each sector \( j \), we can build the occupancy frequency \( f_j(n) \), counting the number of locations that contain exactly \( n \) firms operating in sector \( j \). For instance, \( f_j(0) \) is the number of locations that contain no firms of sector 3, \( f_j(1) \) is the number of locations that contain exactly 1 firm operating in sector 3, and so on. The formal definition is

\[
f_j(n) = \sum_{l=1}^{L} \delta_{n_j,n}
\]

where \( \delta_{n_j,n} \) is the Kronecker delta. From (5.1) it is obvious that\(^5\)

\[
\sum_{n=0}^{+\infty} f_j(n) = L \quad \forall j \in \{1, \ldots, 22\}.
\]

In Fig. 2 we show, as an example, the occupancy frequencies in four different sectors. Sectoral specificities are striking: both the shape of the distributions and the scales on the x and y axis are, indeed, very different. For instance, consider the ATECO 20 sector (wood processing): there are few locations which do not contain any firm belonging to this sector and the majority of locations contains 10–20 firms operating in it. In the case of the ATECO 28 sector (fabricated metal products) the picture changes. Here the number of empty locations is quite large, around 380, i.e. more than 50% of the total. For this sector, a location with 15 firms is a “crowded” one, and indeed \( n = 15 \) belongs to the upper tail of the frequency distribution. For sector 20 (wood processing), on the other hand, observing around 100 firms in a location is a quite common event. In general, a

\(^5\) Note that one sets infinity as the upper bound of the summation even if, clearly, such a summation stops with the number of firms in the most populated location. For instance, if sector 3 has a location with 5000 firms and no locations with a larger occupancy, we get \( f_3(5000) = 1 \) and \( f_3(n) = 0 \), \( \forall n > 5000 \) so that the summation effectively stops at 5000.
Fig. 2. Occupancy frequency in four different sectors. The largest locations have been removed in order to better focus on the behavior of the distributions near the origin.

frequency distribution with an high modal value around 0 and long tails represent a sector where the majority of firms is clustered in few places and the remaining locations are basically empty. On the contrary, a "bell-shaped" distribution is associated with a sector where the large part of the firms is evenly distributed in a relatively large number of locations.

In the rest of this section we will use the empirical occupancy frequency, defined in (5.1), to study the degree of agreement of the empirical data with the theoretical models presented in Section 4. Indeed, if \( \pi_n \) is the marginal distribution derived from a theoretical model and associated with a given sector, say \( j \), the theoretical prediction for the occupancy frequency is \( \pi_n N_j \).

Since the support of the empirical occupancy frequency is in general large, due to the presence of few extremely populated locations and many (almost) non-populated ones, instead of using each occupancy number we consider occupancy classes (analogous to the often-used size classes) defined, for each sector, as the number of locations having a number of firms belonging to that same sector inside a given range. We define classes with ranges following a geometric progression

\[
C_k = [2^k - 1, 2^{k+1} - 1], \quad k = 0, 1, 2, \ldots
\]

Table 2

<table>
<thead>
<tr>
<th>Class</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_0 )</td>
<td>0</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>1–2</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>3–6</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>7–14</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>15–30</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>31–62</td>
</tr>
<tr>
<td>( C_6 )</td>
<td>63–126</td>
</tr>
<tr>
<td>( C_7 )</td>
<td>127–254</td>
</tr>
<tr>
<td>( C_8 )</td>
<td>255–510</td>
</tr>
<tr>
<td>( C_9 )</td>
<td>511–1022</td>
</tr>
<tr>
<td>( C_{10} )</td>
<td>1023–2046</td>
</tr>
<tr>
<td>( C_{11} )</td>
<td>2047–4094</td>
</tr>
</tbody>
</table>
and we report in Table 2 the first 12 occupancy classes as an example. The frequency of the different occupancy classes $f_j(C_h)$ for $h \in \{1, \ldots, 12\}$ can then be easily computed from (5.1). We have

$$f_j(C_h) = \sum_{n \in C_h} f_j(n),$$  

where the sum spans over the integers belonging to each class range.

5.1. Model 0

Let us start with the simplest benchmark provided by Model 0, described in Section 4.1, where all locations are assumed to be homogeneous and the agglomeration strength is set to zero ($b = 0$). In this case no estimation procedure is necessary. Indeed, the marginal distribution only depends on the number of locations $L$ and the number of firms $N$ operating in the sector (see Section 4.1). For each sector $j$ we can obtain a theoretical prediction for the class frequency directly from (4.1). One has

$$f_{j}^{th}(C_h) = \sum_{n \in C_h} L \pi(n; N_j, L)$$  

where $L = 784$ (i.e., the number of LSLM contained in our database).

Fig. 3 plots the empirical class frequency (5.3) together with the theoretical prediction (5.4) for two representative sectors. The observed agreement is basically nil and it is the same for all sectors. The theoretical frequency is proportional to the binomial distribution, and thus displays a bell-like shape with almost all the weight distributed in few central classes. This pattern, however, is never observed in empirical data. Note that this negative result is indeed an important one in that it falsifies any notion of random attribution of business units over a homogeneous space with null returns to agglomeration (see also Rysman and Greenstein, 2005).

5.2. Model 1

Next, we investigate the relevance of agglomeration economies by considering Model 1, described in Section 4.2, in which we allow for a non-zero agglomeration strength parameter $b > 0$. In this case, the marginal distribution of the model, defined in (4.2), depends on the parameters ratio $a/b$. This means that the model is insensitive to re-scaling, by a common factor, of both the locational geographic attractiveness $a$ and the agglomeration strength $b$. Without losing in generality, in the following analysis we set $a = 1$, and, for each sector, we estimate the best fit by varying the parameter $b$. For this purpose, we use the $\chi^2$ statistics with the occupancy classes $C_h$ as categories. For each sector, starting from the marginal distribution in (4.2), we can build the observed class frequencies $f_j(C_h)$ and also the theoretical class frequencies as

$$f_{j}^{th}(C_h) = \sum_{n \in C_h} N_j \pi(n; N_j, L, 1, A, b).$$  

We then consider the $\chi^2$ statistics

$$\chi^2_j(b) = \sum_{h} \frac{(f_j(C_h) - f_{j}^{th}(C_h))^2}{f_{j}^{th}(C_h)}$$  

Fig. 3. Occupancy class frequencies computed on observed data (white bars) and estimated using Model 0 (gray bars).
defined, for each sector \( j \), as a function of the parameter \( b \). Finally, we estimate the sector-specific optimal value \( b_j^\star \) according to

\[
b_j^\star = \arg\min_{b \in \mathbb{R}^+} \chi_j^2(b).
\]  

(5.7)

The resulting \( b_j^\star \) for different sectors are reported in Table 3 together with the average absolute deviation (AAD) that represents a measure of the agreement between the empirical and the theoretical frequencies and is defined as

\[
\text{AAD}_j = \frac{1}{|K_j|} \sum_{h} |f_j^h(C_h) - \hat{f}_j^h(C_h)|,
\]

(5.8)

where \( K_j \) denotes the number of classes in sector \( j \). From Table 3, it is apparent the high degree of sectoral heterogeneity in both the strength of the agglomeration forces and in the ability of Model 1 to reproduce empirical distributions in different sectors. This is well illustrated by Fig. 4 showing, for six different sectors, the theoretical class frequencies obtained using (4.2) with the estimated value \( b_j^\star \) (gray bars). Visual inspection of these plots reveals that the degree of accordance with the data dramatically improves as compared with Model 0. In particular, the agreement with empirical frequencies (white large bars) is, in general, good in the central part of the distribution while the fit remarkably worsens at the two extremes: in some sectors (for instance ATECO sectors 20, 26, 28 and 36) Model 1 largely overestimates the number of locations with a huge number of firms but, at the same time, predicts that all other locations are essentially empty. Conversely, a small value of \( b \) alone. Indeed if the value of \( b \) is large, Model 1 predicts the existence of several locations with a huge number of firms but, at the same time, predicts that all other locations are essentially empty. Conversely, a small value of \( b \) accounts for a large number of locations with few firms, but reduces the probability of finding occupancies close to zero. This difficulty can be partly tackled with Model 2, wherein different locations are allowed to have different geographic attractiveness, so that the observed clusterization of a large number of firms in a single location can be explained by the presence of a relative high geographic attractiveness, even if the sector is characterized by a mild value of the agglomeration parameter \( b \).

The drawback of Model 2 as presented in Section 4.3, however, rests in its large number of parameters: one should specify the value of the sector-specific parameter \( b \) and the value of the parameter \( a \) for each location. Hence, one cannot hope to

Table 3

<table>
<thead>
<tr>
<th>Sector</th>
<th># of firms</th>
<th>Model 1</th>
<th>Model 2, all sites</th>
<th>Model 2, no metropolis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( b ) AAD</td>
<td>( \beta )</td>
<td>( b ) AAD</td>
</tr>
<tr>
<td>15 Food products</td>
<td>75420</td>
<td>62751</td>
<td>0.060</td>
<td>0.000</td>
</tr>
<tr>
<td>17 Textiles</td>
<td>36217</td>
<td>32043</td>
<td>0.050</td>
<td>0.000</td>
</tr>
<tr>
<td>18 Apparel</td>
<td>49782</td>
<td>38137</td>
<td>0.040</td>
<td>0.000</td>
</tr>
<tr>
<td>19 Leather products</td>
<td>25451</td>
<td>19791</td>
<td>0.030</td>
<td>0.000</td>
</tr>
<tr>
<td>20 Wood processing</td>
<td>50662</td>
<td>42322</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>21 Pulp and paper</td>
<td>5267</td>
<td>3698</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>22 Publishing and printing</td>
<td>28183</td>
<td>16402</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>23 Coke, refined petroleum and nuclear fuel</td>
<td>825</td>
<td>617</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>24 Organic and Inorganic chemicals</td>
<td>7593</td>
<td>4941</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>25 Rubber and plastic products</td>
<td>14626</td>
<td>11324</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>30709</td>
<td>25140</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>4034</td>
<td>3010</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>94771</td>
<td>74340</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>29 Industrial machinery and equipment</td>
<td>42984</td>
<td>33157</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>30 Office machinery</td>
<td>592</td>
<td>311</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>31 Electrical machinery</td>
<td>17312</td>
<td>11906</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>32 Radio, TV and TLC devices</td>
<td>9773</td>
<td>6546</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>33 Precision instruments</td>
<td>28280</td>
<td>18713</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>34 Motor vehicles and trailers</td>
<td>2261</td>
<td>1619</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>35 Other transport equipment</td>
<td>5614</td>
<td>3500</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>36 Furniture</td>
<td>59627</td>
<td>46460</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>37 Recycling</td>
<td>2061</td>
<td>1568</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ a ] Values smaller than 10^{-4} are reported as 0.0.
obtain the values of all these parameters from a $\chi^2$ minimization procedure, as in (5.6), undertaken on each sector separately. Indeed, in our case the number of parameters is equal to the number of observations plus one (i.e. 785). In order to overcome this problem, we exploit the double disaggregation (by sector and by location) of our database.

First of all, let us make the following

Assumption 1 (Urbanization effect). The geographic attractiveness $a_{j,l}$ of location $l$ for firms operating in sector $j$ is proportional to the number of firms located in $l$ belonging to all the sectors except $j$

$$a_{j,l} \sim \alpha_j + \beta_j \frac{N_{j,l}}{N_{j,.}}$$  \hspace{1cm} (5.9)

where, with the usual notation

$$N_{j,l} = \sum_{i \neq j} n_{i,l}$$

$$N_{j,.} = \sum_l N_{j,l}.$$

Fig. 4. Occupancy class frequencies computed on observed data (white bars) and estimated using Model 1 (gray bars) and Model 2 (black bars).
Fig. 5. Scatter plot of the $b$ and $\beta$ parameters estimated from model 2 for different sectors with (left panel) and without (right panel) the metropolitan areas (estimates are based on the number of firms).

As noted in Section 3, the geographic attractiveness coefficient $a$ controls for all geographical factors that are not related with the sector under study. We can think to all such factors as both exogenous “geographical” and infrastructural ones, but also general demand-induced externalities or market proximity effects.

The linear relation in (5.9) depends on two sectoral parameters $\alpha_j$ and $\beta_j$. The parameter $\beta_j$ represents a measure of the overall “pull” exerted by all business units from all other sectors on the locational decision of firms belonging to sector $j$. Parameter $\beta_j$ captures what we call “urbanization effect”: the overall installed base of production units in a particular location brings about a stronger attractive strength in sectors with a higher value of $\beta$.

The stationary distribution of Model 2 depends only on the ratios $a_l/b$ so that, again, we can rescale all the parameters $a$ and $b$ by the same factor without affecting the distribution. In order to obtain values for $b$ comparable with the ones found when estimating Model 1, where we assumed $a = 1$, we impose\(^6\) the further requirement that the average value of $a$ is 1, i.e.

$$\frac{1}{L} \sum_l a_{j,l} = 1$$

so that (5.9) reduces to a one parameter relation

$$a_{j,l} = 1 + \beta_j \left( \frac{N_{j,l}}{N_j} - \frac{1}{L} \right).$$

(5.10)

Substituting (5.10) in the marginal distribution (4.3) one can compute the theoretical prediction for the occupancy class frequency

$$f_{j}^{th}(C_h) = \sum_{n \in C_h} \sum_l \pi(n; N_{j,l}, L, \beta_j, A, b).$$

(5.11)

Notice that in (5.11) a summation over $l$ is required since different locations now possess different geographic attractiveness and, consequently, are characterized by different marginal distributions.

Finally, following the same approach described in the previous section, one can obtain an estimate for $(b, \beta)$ as

$$(b^*_j, \beta^*_j) = \arg\min_{b, \beta \in \mathbb{R}^+} \chi^2_j(b, \beta),$$

(5.12)

where $\chi^2$ is defined as in Eq. (5.7).

Let us start by noting that moving from Model 1 to Model 2, one observes an unambiguous improvement of the ability of model to reproduce the empirical observations: this is clear from visual inspection of Fig. 4, where one observes a very good accordance of predicted frequencies (black bars) with the observed ones and is confirmed by the reduction in the average absolute deviation (AAD) reported in Table 3 together with the point estimates. Indeed, Model 2 seems able to overcome, at least in first approximation, the inability of the previous one to capture the tails behavior of the empirical distributions.

The distribution of estimates $(b^*_j, \beta^*_j)$ is illustrated in Fig. 5 (left panel). A rather striking feature of the plot is the apparent polarization between a group of sectors which shows a nearly exclusive impact of “urbanization effect” and a second group wherein sector-specific agglomeration effects dominate. In that, the attribution of individual sectors to the two groups turns

\(^6\) This assumption is made only for comparability purposes and does not affect our results in any respect.
Table 4
Summary statistics of estimates from Models 1 and 2, by sector (estimates are based on the number of employees)

<table>
<thead>
<tr>
<th>Sector</th>
<th># of employees</th>
<th>Model 2, all sites</th>
<th>Model 2, no metropolis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All sites</td>
<td>β</td>
<td>b^2</td>
</tr>
<tr>
<td>15 Food products</td>
<td>434515</td>
<td>728.85</td>
<td>0.03</td>
</tr>
<tr>
<td>17 Textiles</td>
<td>345318</td>
<td>0.00</td>
<td>6.45</td>
</tr>
<tr>
<td>18 Apparel</td>
<td>346387</td>
<td>791.87</td>
<td>0.00</td>
</tr>
<tr>
<td>19 Leather products</td>
<td>230543</td>
<td>398.67</td>
<td>4.81</td>
</tr>
<tr>
<td>20 Wood processing</td>
<td>170294</td>
<td>0.00</td>
<td>1.57</td>
</tr>
<tr>
<td>21 Pulp and paper</td>
<td>85424</td>
<td>0.00</td>
<td>7.72</td>
</tr>
<tr>
<td>22 Publishing and printing</td>
<td>175012</td>
<td>680.44</td>
<td>2.87</td>
</tr>
<tr>
<td>23 Coke, refined petroleum and nuclear fuel</td>
<td>24147</td>
<td>0.00</td>
<td>16.21</td>
</tr>
<tr>
<td>24 Organic and Inorganic chemicals</td>
<td>209242</td>
<td>843.29</td>
<td>0.31</td>
</tr>
<tr>
<td>25 Rubber and plastic products</td>
<td>198401</td>
<td>0.00</td>
<td>6.07</td>
</tr>
<tr>
<td>26 Non-metallic mineral products</td>
<td>250824</td>
<td>756.22</td>
<td>0.06</td>
</tr>
<tr>
<td>27 Basic metals</td>
<td>136123</td>
<td>872.11</td>
<td>0.96</td>
</tr>
<tr>
<td>28 Fabricated metal products</td>
<td>621642</td>
<td>594.54</td>
<td>0.80</td>
</tr>
<tr>
<td>29 Industrial machinery and equipment</td>
<td>554105</td>
<td>290.72</td>
<td>3.71</td>
</tr>
<tr>
<td>30 Office machinery</td>
<td>18609</td>
<td>1083.65</td>
<td>15.05</td>
</tr>
<tr>
<td>31 Electrical machinery</td>
<td>205797</td>
<td>821.10</td>
<td>0.15</td>
</tr>
<tr>
<td>32 Radio, TV and TLC devices</td>
<td>103161</td>
<td>556.44</td>
<td>2.77</td>
</tr>
<tr>
<td>33 Precision instruments</td>
<td>129448</td>
<td>660.57</td>
<td>0.81</td>
</tr>
<tr>
<td>34 Motor vehicles and trailers</td>
<td>185748</td>
<td>0.00</td>
<td>15.09</td>
</tr>
<tr>
<td>35 Other transport equipment</td>
<td>100780</td>
<td>0.00</td>
<td>11.59</td>
</tr>
<tr>
<td>36 Furniture</td>
<td>309911</td>
<td>0.00</td>
<td>4.66</td>
</tr>
<tr>
<td>37 Recycling</td>
<td>8327</td>
<td>2.30</td>
<td>5.71</td>
</tr>
</tbody>
</table>

* Values smaller than 10^{-4} are reported as 0.0.

Fig. 6. Scatter plot of the β and β̃ parameters estimated from model 2 for different sectors with(left panel) and without(right panel) the metropolitan areas
(estimates are based on the number of employees).

out to be somewhat puzzling (for example “17—Textiles” and “18—Apparel” appear to belong, counterintuitively, to the former group). Such a puzzling evidence, in fact, may be largely the outcome of a sort of “horizontal pull” of metropolitan areas which tend to exert what we could call a more-of-everything effect (including more of the activities which are traditionally associated with sector-specific agglomeration phenomena, such as the mentioned Textiles and Apparel). In fact by removing the metropolitan areas the picture significantly changes: cf. Fig. 5(right panel) and Table 3. When they are present, agglomeration effects tend to be mostly of a sector specific nature. Note that, even in those sectors where β is positive, urbanization tends to explain a relatively small part of the inter-site variation in locational intensities.

In the previous analyses agglomeration has been measured by considering only the number of firms present in each location, and not their (relative) size. Further precious information, stemming from firm size distribution in different locations, may be obtained by estimating our model on employment data. That is, instead of using data on the number of firms belonging to any given sector that are present in each location, we can apply the model to the number of firm employees, per location and per sector. In this case agglomeration also captures the effect of increasing returns and internalization of productive activities within single firms. So, for example, in employment-based estimates the strength of agglomeration

8 Rough but illustrative evidence comes from the low goodness of fit of the linear relation between the share of activity of a particular sector in a given location and the share of the overall, excluded that sector, economic activity located there.
of a location with say one firm with a thousand employees is taken to be equivalent to another one with 100 firms of 10 employees each (which of course would not be the case in the previous estimation procedures).

The estimates of the \( b_j, \beta_j \) are presented in Table 4 and illustrated in Fig. 6 (left panel). Again the analysis of the universe of locations tend to be affected by the rather special agglomerative pull of metropolitan areas. If one excludes them, the picture, Fig. 6 (right panel), is relatively similar to the one stemming from firms locational patterns. Sectoral agglomeration effects seem to dominate.\(^9\) And, of course, given the somewhat expansive notion of agglomeration, the estimates now capture also the effects of the location patterns of few but large firms (cf. for example, the sector “34—motor vehicles”).

6. Conclusions

The aim of this work has been to offer rather general and empirically applicable formal tools able to assess the importance of agglomeration phenomena distinguishing between their location-wide and sector-specific drivers. Indeed, despite its simple structure, the model generates testable implications on the whole shape of the distribution of firms locational and employment choices in any given sector, a notable improvement over the majority of existing models which only provide simple structure, the model generates testable implications on the whole shape of the distribution of firms locational and employment choices irrespectively of the characteristics of sectors. This pull is horizontal in the sense that it tends to join together all activities.

However, when few big urban centers are excluded from the analysis the impact of horizontal agglomeration forces appears reduced so that we are able to detect also the important role played by the very history of locational decision within each sector. In several cases a form of dynamic increasing returns is observed, such that the number of production units belonging to one particular sector of production in a given point in time influences the probability that an additional unit will be located there. The strength of this effect seems to vary a good deal across sectors.

As such our findings suggest that beyond the location-wide externalities typically emphasized by the new economic geography models, the agglomeration dynamics is driven by sector specific mechanisms, plausibly related to localized forms of knowledge accumulation, spin-offs and formation of new firms. Such conclusions are indeed strengthened by the application of the model to the dynamics of employment: again, sector specific agglomeration forces appears to be powerfully at work, sometimes closely resembling the agglomeration profiles of (district-type) firms and some other time internalized within the employment strategies of relatively fewer but bigger firms.

The foregoing model can be extended in different ways. First, one might explicitly take into account interdependencies between locations and industries. In its present version, our model does not include the possibility that firms locational choices may be influenced by the choices made by firms belonging to different sectors, possibly located in neighboring regions. One might think to an extended version of the model where locations are positioned over some metric space, e.g. a two-dimensional lattice, and firms decisions (entry and exit) are somewhat correlated in space. Similarly, one might introduce urbanization economies whose advantages spill over to neighboring regions (\(\) unlike being concentrated in a given region). Second, one might try to explicitly consider the time dimension, and directly estimate the evolution predicted by the model instead of exclusively rely on the equilibrium distribution. Third, it would be interesting to obtain a testable specification of the model that considers the contemporaneous location of both business units and workers, so that a more proper account of the relative importance of agglomeration and scale economies could be derived. Finally, starting from the present analysis and the refinements suggested above, an interesting challenge might involve the development of an evolutionary inspired theoretical approach (Frenken and Boschma, 2007; Bottazzi and Dindo, 2008) which could provide a more reliable interpretative framework and make more effective economic and policy implications possible.

Acknowledgements

We gratefully acknowledge the support to the research by the Italian Ministry of Education, University, and Research (MIUR, Project EFIRB03GD), the Scuola Superiore Sant’Anna (Grant ERIS03GF), the European Commission (CO3 STREP Project, Contract No. 012410, NEST) and The University of Manchester. Part of the research that has lead to this work has been undertaken inside the activities of the DIME Network of Excellence, sponsored by the European Union. Comments by Thomas

\(^9\) Also in this case the goodness of fit of the relation discussed in footnote 8 is relatively low.
Brenner, Yannis Ioannides, Karl Schlag, Alan Kirman and two anonymous referees have helped along the various revisions of the work.

References