XXXVII CONFERENZA ITALIANA DI SCIENZE REGIONALI

TESTING FOR LOCALIZATION WITH RELATIVE ENTROPY MEASURES

Roy Cerqueti[1], Eleonora Cutrini[2]

ABSTRACT

This paper aims to give statistical significance to the measurement of localization and spatial concentration through relative entropy measures. This work is in line with research on the simulation of confidence intervals for the Ellison and Glaeser index (Cassey and Smith, 2014) and the comparable advancements in the context of distance-based methods that have been developed primarily for absolute indices (Duranton and Overman, 2005; Marcon and Puech, 2003, 2010) and, more recently, for relative indices (Lang, Marcon and Puech, 2014).

First, we provide a simple site-selection theoretical model to describe the location scenarios. The adopted quantitative tool is the probabilistic theory of urns. Second, we introduce a battery of tests to evaluate whether the observed values related to the spatial distribution of industries are significantly different from regularity. Null hypotheses are identified through a Montecarlo procedure. Tests are constructed on the basis of the developed theoretical model. We apply this method to the European manufacturing economy and we found a significant overall localization, whereby significant geographical concentration is evident for low-tech industries usually considered as localized because of Marshallian external economies (i.e. wearing apparel, textiles, publishing, printing) as well as for small-scale knowledge intensive industries (e.g office machinery and computers, radio, television and communication equipment) and industries characterized by internal scale economies (such as motor vehicles, basic metals, chemicals, other transport equipment). The decomposition analysis allows to ascertain that geographical concentration is almost always statistical significant between national borders. Instead, the within-country components of the concentration measures are statistically significant only in five industries –office machinery, motor vehicles, other transport equipment, basic metals, publishing and printing.

JEL codes: C43, C46, C02, L60, R12

Keywords: localization, statistical testing, urn model, relative entropy

[1] University of Macerata, Department of Economics and Law, Via Crescimbeni, 20, I-62100, Macerata, Italy, e-mail: roy.cerqueti@unimc.it.

[2] University of Macerata, Department of Law, Piaggia dell'Università, 2, I-62100, Macerata, Italy, e-mail: eleonora.cutrini@unimc.it (corresponding author).

Introduction

It is widely recognized that industrial clustering is a key stylized fact of economic geography. Firms tend to collocate because of exogenous natural advantages, knowledge spillovers, input-output linkages and/or labor market pooling.

During the last decade, the development of New Economic Geography models was inevitably accompanied by a rising interest in measuring localization of economic activities. Ideally, geographical concentration should be assessed in continuous space to avoid the checkerboard problem related to arbitrary spatial units (See Arbia, 2001 for details on aggregation and scale problems as two different manifestations of the arbitrariness of geographical boundaries). Despite these prominent advantages, distance-based measures may be not the "first-best" metrics, not only because they require micro-geographical data (which are still not always available), but also because, when geographical boundaries are expressions of meaningful and active economic institutions instead of artificial border bias, it may be important to know the spatial scale at which clustering occurs to inform policy makers. This can be certainly appropriate for the European area where national sovereignty has not been completely discarded throughout the process of economic integration. Some authors propose the use of relative entropy indices (Mori et al., 2005; Brülhart and Traeger, 2005; Bickenbach and Bode, 2008, Cutrini, 2009, 2010), which have distinct advantages over the standard concentration measures constructed over the discrete space. The most relevant one is their decomposability. This feature allows authors to decompose the inequality analysis across different spatial and sectoral scale in order to identify the contributions of individual regions (sectors) to the overall localization of economic activity.

Moreover, relative entropy measures allows for an integrated analysis of localization where the complexity of the real location patterns are reduced to two characteristics dimensions (concentration/specialization), so that the spread of economic activities across space is mirrored by the structural differences between geographical units (equivalence between regional specialization and industrial concentration) (Cutrini, 2009).

Relative measures should be preferred over absolute measures to gauge and test for the spatial structure of economic activities (see also Lang et al., 2014 on this view) as long as the aim of the analysis is to assess the strength of industry-specific localization economies. In fact, the two types of measures (absolute and relative) differ fundamentally from each other in terms of what is considered to represent no localization, or noconcentration. While for absolute measures, the reference has generally been chosen to be the uniform distribution, in the case of relative measure the regional distributions of an industry is compared to the regional distribution of total economic activity. This is the logic underlying the Ellison and Glaeser index (Ellison and Glaeser, 1997) and the entropy measures of localization. Even in the context of distance-based methods, the main contributions have originally focused on absolute versions (Marcon and Puech, 2003; Duranton and Overman, 2005), but they have recently been refined to control for the distribution of the economic activity as a whole (Lang et al., 2014).

Before entering the details of our contributions, some further premises are needed.

This paper aims to add a statistical testing procedure to entropy-based measure of spatial concentration. In this respect, it is in line with research on the simulation of confidence intervals for measures constructed over the discrete space such as the Ellison and Glaeser index to test the null hypothesis of no concentration (Cassey and Smith, 2014).

Our approach is also related to recent advancements in the context of distance-based methods for absolute indices (Duranton and Overman, 2005; Marcon and Puech, 2003, 2010) and relative indices (Lang et al., 2014) in that they also seek to detect significant localization patterns over the continuous space.

In this paper, the testing procedure developed to evaluate whether observed spatial distribution of industries is significantly different from regularity is embedded in a micro-founded model of location.

Tests are specific for industrial sectors (concentration), and consider the situations of overall localization, spatial concentration within countries and between countries. The proposed approach is based on the

presence of economic and geographical factors leading to a not uniform distribution of companies and employees across regions. More precisely, we rely on the Ellison and Glaeser's approach (EG henceforth) to derive a meaningful null hypothesis of random choice. In the EG approach it is introduced the idea that in the case of random firms' location choice, the observed employment distribution will not be perfectly regular. It may well be that even when firms choose independently from each other their production sites, the spatial distribution of employment is not stochastic because of the industrial structure. Hence, a certain degree of "dissimilarity" of employment distribution of a manufacturing sector relative to the reference (e.g total manufacturing employment) will be perfectly consistent with the independent location model of firms.

Another novelty characterizes this work with respect to the related literature. Micro-foundation is achieved through a theoretical probabilistic model based on urns theory. The process of regional selection by companies and employees is formalized as a stepwise process of extraction of colored balls from an urn. Different colors correspond to different regions, and the distribution of colored balls in the urn is modified at each step by random insert/removal of balls from the urn so as to mimic the utility-based constrained optimization problem faced by firms and employees. In so doing, the factors leading to the preference of a company/employee for a specific region are taken into account.

The null hypotheses of the tests are defined by introducing critical thresholds, whose values are endogeneized in the application on the basis of empirical data and identified through a Montecarlo procedure. Specifically, the application of the theoretical tests are performed on a sample of NUTS3 regions for 16 countries and for the reference year 2007 (see section 5.1 for the details on the data). For each industrial sector (concentration), 10,000 resample of the available data are drawn on the basis of their empirical distribution. Then, the relative entropies of the resampled series are computed, and the 90%, 95% and 99% percentiles of the entropies distributions are assessed. This gives the critical values of the tests at the related confidence levels.

To the best of our knowledge, this is the first contribution dealing with the development of localization tests based on relative entropies. Furthermore, we also provide a theoretical micro-foundation of the proposed testing procedure.

The paper is organized as follows: Section 2 reviews the related literature. Section 3 introduces a sequential location model of firms and employees. Section 4 presents the testing procedure through Montecarlo simulation. Section 5 includes an application based on regional data. Section 6 offers some concluding remarks.

1. Testing for localization: some insights from the literature

The spatial distribution of economic activity has been the focus of growing research interest and new statistical tools were recently introduced. In this context, we may distinguish two main lines of methodological advancements: one is related to the measures for empirical research over the discrete space (Ellison and Glaeser approach and entropy-based measures) and the other deals with measurement of the geographical concentration over the continuous space (Marcon and Puech, 2003, 2010; Duranton and Overman, 2005; Lang et al., 2014 among others).

Among the measures constructed over the discrete space and after the Ellison and Glaeser approach, entropy-based methods played a noteworthy role thanks to the decomposition properties by subgroups (Brülhart and Traeger, 2005; Bickenbach and Bode, 2008) and because the recently developed indices of overall localization (as the L-index) allowed to retain the symmetry between concentration and specialization, in an integrated approach and along distinct and meaningful geographical scales (Cutrini, 2009, 2010).

Ellison and Glaeser (1997) have the merit to introduce a measure of spatial distribution of manufacturing employment which control for the industrial structure, and which is embedded in a location model. Successive empirical studies (Maurel and Sédillot, 1999, Devereux et al., 2004) have adopted the "dartboard

approach" to investigate the spatial distribution of manufacturing industries in France and United Kingdom, respectively.

A further major advantage of the EG-index is that it is possible to test for the statistical significance of the index and indicate whether a sector's distribution of activity across locations is significantly concentrated or dispersed (Ellison and Glaeser (1997) and Maurel and Sédillot (1999) [3]. For the purpose of our analysis it is important to remind that Ellison and Glaeser (1997) suggested that the theoretical perfect regularity (the null hypothesis of no concentration and no dispersion) will not be reached for two reasons: randomness properly speaking, that is to say stochasticity, and industrial concentration, that is to say the non-independence of employment distribution (firms choose their location as a whole, but what we usually observe is just employment).

Recently, Cassey and Smith, 2014 show how the ad hoc thresholds suggested by Ellison and Glaser (1997) can flaw the interpretation of results based on the EG index and simulate confidence intervals that can be used for statistical testing.

The identification of a null hypothesis and the procedure for testing significant localization is also a main feature of the distance-based approach, ever since the original contributions that have been implemented to evaluate the spatial distribution of firms in France and United Kingdom, respectively (Marcon and Puech, 2003; Duranton and Overman, 2005). Although, the significant spatial concentration of an industry was referred to a departure from a theoretical random distribution of the same industry. Marcon and Puech, 2003 construct the counterfactual with no-localization being referred to as a random distribution of the same industry in all the sites occupied by the industry, while Duranton and Overman, 2005 project the distribution of establishments in the industry in the wider area of all the possible locations of the manufacturing industry (the aggregate economic activity)[4]. Following this procedure manufacturing industries are detected as dispersed (at some distance) if their degree of concentration at that distance is lower than the random generated confidence interval. In this manner, spatial point patterns do not control for the distribution of the economic activity as a whole. This is the main reason why Lang et al., 2014 have recently introduced a relative density function.

This feature – to control for the spatial distribution of overall manufacturing - is instead typical of the application of relative entropy measures, extensively adopted in the literature because they are also endowed with desirable decomposability properties across spatial scales and across sectors. Mori et al. (2005) measure topographic concentration for Japan using a dissimilarity entropy measure (a discrete Kullback-Leibler divergence) to compare the distribution of establishments over the distribution of economic land area, while Brülhart and Traeger (2005), Cutrini (2009, 2010) and Bickenbach et al. (2008) apply the decomposition of relative entropy indices in the context of the analysis of spatial concentration of manufacturing industries in Europe. Relative concentration indices (as the raw G-index of Ellison and Glaeser) measure the discrepancy between the spatial distribution of one industry and that of the aggregate activity selected as a benchmark (e.g. aggregate manufacturing).

[3] It is worth noting that the version of the index proposed by Maurel and Sédillot (1999) does not allow to control for the economic size of each region since they simply calculate G as a difference between the spatial concentration of the industry and the spatial concentration of the aggregate. Therefore in Maurel and Sédillot (1999) G is simply a difference between two level of absolute concentration (and do not measure relative concentration which is instead the G in Ellison and Glaeser). In this way they can not control for the dissimilarity in the spatial distribution of the two economic activities. Therefore, comparison across industries may lead to awkward interpretations from the standpoint of localisation, since in Maurel and Sedillot (1999) it may occur that two industries A and B (with the same industrial structure) have the same degree of geographical concentration even if the allocation of employment across regions in industry A span exactly proportionally to total employment while industry B is mostly located in the less economically advanced regions.

[4] Concentration of a sector is detected comparing the Kernel density function of bilateral distances between pairs of establishment to a "theoretical" random distribution of an equivalent set of firms over a wider space, namely all sites (postcode level) where at least one manufacturing establishment is located. They simply reshuffle locations where at least one establishment of the aggregate economic activity exists.

Relative entropy-based measures have attracted considerable interest among academics and practitioners. More recently, Bickenbach et al. (2010) specifically addressed the discrepancy between absolute and relative entropy measures, while Haedo and Mouchart (2013), suggest how relative entropy measures of overall localization, specialization and concentration can be viewed through the lens of a stochastic independence approach, that is in the context of the statistical analysis of contingency tables. Recent methodological advancements include their characterization on the basis of the axiomatic principles implicitly assumed by regional economists when using them to quantify the spatial concentration of a sector (continuity, symmetry, weak scale invariance, location division property, group division property, type I and type II independence) (Alonso-Villar and Del Río, 2013).

Although relative entropy measures are becoming frequently used to measure industrial concentration and regional specialization (e.g. Bagoulla and Péridy, 2011; Vechiu and Makhlouf, 2014; Stierle-von Schultz and Stierle, 2013, Gokan, 2010, Evans, 2010 among many others), to date, to the best of our knowledge, the literature lacks of a testing procedure for the empirical analysis of localization based on relative entropy measures embedded in a location choice model.

2. A sequential location model of firms and employees

Testing for the presence of significant localization requires defining a null hypothesis of no localization (or random localization or stochastic localization) referring to a theoretical model of location choice. As in the early contribution in international economics and location theory, we basically refers to the "no localization" scenario as a configuration where the spatial spreading of an economic activity does not diverge to a "theoretical case" in which simple location characteristics are at work (Ohlin, 1933; Hoover 1936). We also assume that history matters and the spatial configuration of an industry tends to be 'locked in' trough circularity after initial ex-ante possibility of multiple locations (Arthur, 1994; David, 2000). We thus assume that clustering and establishments' location is path-dependent.

Disproportionality in the spatial distribution of employment in one industry, relative to the employment in the overall manufacturing activity is considered as a sign of agglomeration, due to supply-side factors that influence the industrial structure such as the presence of scale economies (Marshallian externalities or internal increasing returns, or a combination of the two concomitant forces) or demand-side factors related to the distribution of final demand (i.e. market potential).

This section contains a theoretical model for the distribution of the employees in a set of regions.

In particular, after the identification of the distribution of firms across regions, we consider the process used by the employees to spread across the same regions. It is indeed reasonable that such a settlement procedure is conditioned to the spatial distribution firms.

We consider S industrial sectors forming the entire manufacturing system. We aim at modeling and discussing different types of relationships between the distribution –at a regional level- of the employees and of the firms, both for each sector and for the overall manufacturing activity. The distribution of employees and firms is described as a dynamic settlement process through a urn model. We enter the details.

We consider a urn containing a finite number of colored balls. We denote such colors as $C_1, C_2, \ldots, C_R$. The R regions can be viewed as the colors of the balls in the urn.

For both cases of employees and firms, the process of region selection is modeled by a sequential extraction of a ball from a suitably defined urn, and the color of the extracted ball represents the selected region.

2.1 Firms

We consider a set of K firms.

The probability that the k-th firm locates in the r-th region coincides with the probability that the k-th individual extracts from the urn U a ball with color C_r. Such a probability increases as the ratio between the

number of balls with color C_r and the total balls contained in U does. Moreover, the heterogeneity among the firms and the presence of agglomerative effects is modeled through the modification of the urn configurations at each firm's location decision. In particular, we consider a multistage procedure, in which the k-th stage is associated to the selection of the region by the k-th firm.

At the initial stage, the colors are assumed as being identically distributed in the urn. This means that the first extraction from the urn (i.e.: the location choice of the first firm) is implemented according to a uniform distribution over the available regions. The uniform hypothesis of the initial configuration of the urn stands for the irrelevance of natural advantages, which are the only mean to let an empty region be preferable than another empty one and are not considered here.

What could be relevant for the firms is the distribution of the firms itself, in the sense that the k-th firm's choice might be affected also by the location of the previous *k-1* firms. In particular, the k-th firm might decide to be located in a specific region r for several reasons:

- presence in the region r of firms of the same sector to take advantage of industry-specific knowledge spillovers, skilled workers in the local labour market, and within-industry input-output linkages;
- presence in the region r of firms operating in ancillary manufacturing activities to exploit complementarities with upstream and downstream firms along the value chain and knowledge spillovers across sectors;
- presence in region r of infrastructures and services leading to a more profitable business environment (urbanization economies)

So, in general, the uniform distribution of the colors in the urn is an invalid condition in each step after the first one.

At the first stage, the firm labeled with 1 implements the selection of the region. In the context of colored balls in the urn, the individual labeled with 1 extracts a ball from U.

The drawn color corresponds to the region where the 1-st firm locates. So, the number of the firms in the R regions after the 1-st drawn from the urn is a vector $(n_1^{(1)}, n_2^{(1)}, \dots, n_R^{(1)})$, where the subscript indicates the color/region. Of course, if the extracted ball has color C_r, then it results $n_r^{(1)} = 1$ and $n_r^{(1)} = 0$ for each $r \neq r$.

After the extraction, the drawn ball is reinserted in the urn. At the same time, some other colored balls are inserted in or removed from the urn. At the end of this procedure, the number of the balls with colors $C_1, C_2, \dots, C_R$ is $C_1^{(2)}, C_2^{(2)}, \dots, C_R^{(2)}$, respectively. The resulting new configuration of the balls in the urn changes –in general- the probability of extracting a ball with a specific color in the next step, according to a utility maximization criterion (see below the details on this for the general case of the k-th step). This new urn is the one used by the firm labeled with 2 for the ball drawn procedure.

Now, the firm labeled with 2 takes a ball from the urn. The number of the firms in the regions after the 2-st drawn from the urn is $(n_1^{(2)}, n_2^{(2)}, \dots, n_R^{(2)})$, where the subscript indicates the color/region. In accord to the previous step, the nonempty regions are all the ones selected in the 1-st and 2-nd step, and we have $n_1^{(2)} + n_2^{(2)} + \cdots + n_R^{(2)} = 2$.

The procedure goes on according to this rule: at the k-th stage, the firm extracts from a urn such that the number of balls with colors $C_1, C_2, \dots, C_R$ is $C_1^{(k)}, C_2^{(k)}, \dots, C_R^{(k)}$, respectively.

As already mentioned above, the probability $p_r^{(k)}$ of taking a specific color/region C_r from the urn is the relative number of the balls with color $C_r^{(k)}$ in the urn, so that

$$p_r^{(k)} = \frac{C_r^{(k)}}{\sum_{r=1}^{R} C_r^{(k)}}, \quad \forall\, k = 1,2,\dots,K \text{ and } r = 1,2,\dots,R. \qquad (1)$$

In general, $p_r^{(k)} \neq p_r^{(k-1)}$, for each $k = 1,2,\dots,K$ and $r = 1,2,\dots,R$.

After the k-th drawn, the number of firms in the regions is $(n_1^{(k)}, n_2^{(k)}, \dots, n_R^{(k)})$.

The balls addition/removal procedure implemented at each stage –and the associated probability distribution $p^{(k)} = (p_1^{(k)}, p_2^{(k)}, \dots, p_R^{(k)})$- has economic reasoning.

In fact, the k-th firm maximizes its utility $u_k: D \to \mathbb{R}$, where D is the set collecting all the discrete probability distributions of the type $p = (p_1, p_2, \ldots, p_R)$ over the set $\{1,2,\ldots,R\}$, where

$$p_r \in [0,1]; \forall r = 1,2,\ldots,R \text{ and } \sum_{r=1}^{R} p_r = 1.$$

The meaning of such probabilities in our settlement process is $p_r = Prob(\text{to draw a ball of color } C_r)$.

After the utility maximization procedure, the colored balls are added (removed) in (from) the urn, so that the distribution of the balls in the urn is in accord to the utility maximizing probability distribution, which is exactly $p^{(k)} = (p_1^{(k)}, p_2^{(k)}, \ldots, p_R^{(k)})$ as in formula (1).

The specific shape of the utility function depends also on the sectors of activities of the previously located firms and on the locating one. We assume that firms are shared among S sectors of activity and denote the sectors by $s = 1,2,\ldots,S$.

To better explain how the settlement process works, an illustrative example is needed.

Example 1.

Consider five regions ($R = 5$) and suppose that eleven firms have already chosen their regional location. So, the 12-th firm must now extract a ball from the urn and locate (i.e.: $k = 12$).

Let us also suppose that the located firms belong to four sectors (i.e.: $S = 4$). Specifically, three of them are in the Textiles, two of them are in the Motor vehicles, six of them are in the Machinery and one in the Basic metals.

The regional distribution of the firms is as in Figure 1.

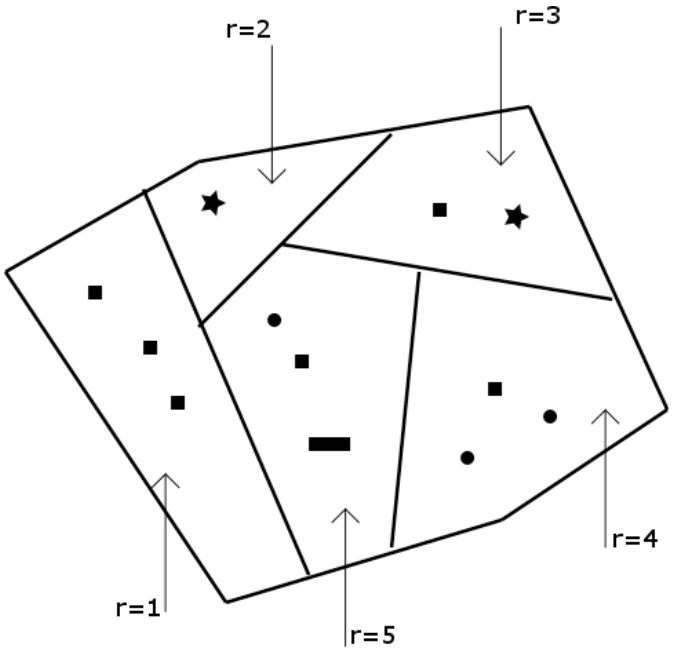

Figure 1. Situation at the beginning of the 12-th step, describing the localization of the eleven firms among the five regions. Different symbols stand for different sectors of activities. Circles represent the Textiles, stars are associated to the Motor vehicles, square is Machinery while the rectangle indicates the Basic metals.

Suppose now that the 12-th firm belongs to the Textiles. The firm is aware about the location configuration presented in Figure 1. Then, it ranks regions in accord to its preferences. An example of plausible reasoning could be the following: regions $r = 1,2,3$ are free of firms coming from the Textiles sectors. Then, placing in one of such regions could not be a good choice. However, $r = 1$ is better than the other two, since the presence of a higher number of firms is for sure associated to a higher level of infrastructures, hence facilitating the business activity. Besides, the existing firms are operating in a sector,

which may be important for the firm, because of the possible complementarities with an upstream sector producing specialized machinery for textile products. Nevertheless, region $r = 4$ should be a better choice for the possibility to exploit Marshallian economies. Region $r = 5$ has also some attractive features due to the presence of a firm from the Textiles. In fact, the new established firm can also take advantage of the already established business environment in the Textiles sector fostered by the existing firm (e.g. presence of skilled labor, and/or specialized services for the textile industry). Said this, $r = 5$ is perceived to be better than $r = 1,2,3$.

To conclude, the 12-th firm decides to locate in a region different from $r = 1,2,3$. In a preference scale, it decides to assign mark 10 to the preferred region $r = 4$ and, by comparison, mark 7 to $r = 5$ and mark 5 to $r = 1$. Regions $r = 2$ and $r = 3$ has mark 0. Such a score leads to a probability distribution $p^{(12)} = \left(p_1^{(12)}, p_2^{(12)}, p_3^{(12)}, p_4^{(12)}, p_5^{(12)}\right) = \left(\frac{5}{22}, 0, 0, \frac{10}{22}, \frac{7}{22}\right)$.

The probability distribution $p^{(12)}$ may be viewed as solution of a utility-based constrained optimization problem. The utility function is suitably defined as $u_{12}: \boldsymbol{D} \to \mathbb{R}$, so that

$$u_{12}\left(p^{(12)}\right) = u_{12}\left(\frac{5}{22}, 0, 0, \frac{10}{22}, \frac{7}{22}\right) = \max_{p \in \boldsymbol{A}} u_{12}(p),$$

where

$$\boldsymbol{A} = \{p = (p_1, p_2, p_3, p_4, p_5) \in [0,1]^5 : \sum_{r=1}^{5} p_r = 1$$

is the admissible region.

The removal/addition procedure is then in accord to the optimizing probability $p^{(12)}$. Specifically, if we suppose that, for example, the number of the colored balls in the urn at the 11-th step was configurated as $(C_1^{(11)}, C_2^{(11)}, C_3^{(11)}, C_4^{(11)}, C_5^{(11)}) = (0,3,2,14,25)$, then the (not unique) urn for the 12-th extraction could be created by: adding two balls of color C_1, removing all three balls of color C_2, removing all two balls of color C_3, adding six balls of color C_4 and removing eleven balls of color C_5. The resulting configuration of the colors is $(C_1^{(12)}, C_2^{(12)}, C_3^{(12)}, C_4^{(12)}, C_5^{(12)}) = (10,0,0,20,14)$, and the probability distribution associated to the extraction of a colored ball from the urn is exactly $p^{(12)}$.

At this point, the 12-th extraction takes place. Consistently with the above example, consider that the drawn ball is of color C_4. Then, we have $(n_1^{(12)}, n_2^{(12)}, n_3^{(12)}, n_4^{(12)}, n_5^{(12)}) = (3,1,2,\mathbf{4},3)$, where the bold indicates the change with respect to the location of Figure 1.

The regional distribution of the firms at the end of the 12-th step is illustrated in Figure 2.

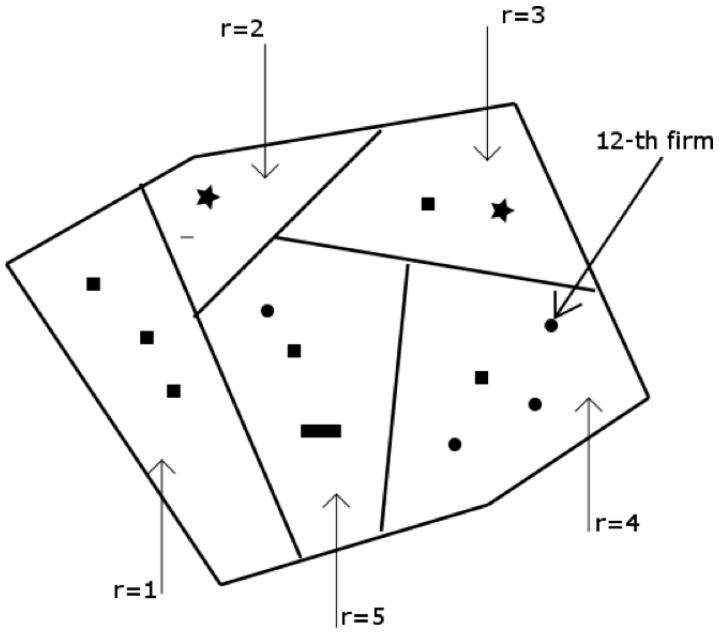

Figure 2. Situation at the end of 12-th step. The location of the 12-th firm is highlighted. Symbols are those of Figure 1.

The process of location of firms in the regions stops at the K-th stage.

The final distribution of the K firms in the R regions will be denoted by $p = (p_1, p_2, \ldots, p_R)$, and it is the outcome of how firms located during the K steps. Hence, it can be derived in a natural way from the number of firms in the regions at the end of the K-th drawn from the urn, namely $(n_1^{(K)}, n_2^{(K)}, \ldots, n_R^{(K)})$, as follows:

$$p_r = \frac{n_r^{(K)}}{K}, \forall r = 1, 2, \ldots, R.$$

It is possible to derive the final regional distribution of the firms belonging to a specific sector by the settlement procedure implemented above.

At each step of the location procedure, we identify the sector of the located firm so that $(n_1^{(s,K)}, n_2^{(s,K)}, \ldots, n_R^{(s,K)})$ is the vector collecting the number of the firms of sector s in the R regions. The final distribution of the firms belonging to sector s is then a vector $p^{(s)} = (p_1^{(s)}, p_2^{(s)}, \ldots, p_R^{(s)})$,

where

$$p_r^{(s)} = \frac{n_r^{(R,s)}}{\sum_{r=1}^{R} n_r^{(R,s)}}, \forall r = 1, 2, \ldots, R \text{ and } s = 1, 2, \ldots, S.$$

The final distribution of firms normally does not follow a uniform law because of path dependence and the location factors described above (i.e. industry-specific Marshallian economies, complementarities between upstream and downstream firms across sectors, and urbanization economies).

2.2 Employees

We now deal with the employees' distribution, and assume a manufacturing system populated by L employees.

Also in this case, we assume that an employee locates in the r-th region with the same probability of extracting a ball with color C_r from a urn.

Our grounding assumption is that firms belonging to the same sector s have the same number of employees, say L_s. So, the number of employees of sector s in region r is proportional to the number of firms of sector s in region r, and the proportionality factor is exactly L_s. This means that the spatial distribution of

the firms and the spatial distribution of employees of sector s should coincide. Deviations from this hypothesis is due to a different site selection process of economic agents (firms vs employees) and thus occurs for three main reasons:

a) the structure of preferences of workers is usually different from those of entrepreneurs. For examples workers may assign higher attractiveness to some regions (e.g. metropolitan regions for the easier access to urban amenities, regions with higher wages).

b) specialized workers may choose to migrate/settle to those regions with a higher concentration of sectors that use intensively those skills they are endowed with.

c) the heterogeneity of firm size across regions, which could be different from the typical industrial structure of the sector.

However, the settlement process is of sequential type: the first employee is assumed to select a region in accord to the initial assumption, since the industrial system is not affected at the beginning by the effects a) - c). After the first step, the economical-geographical activity proceeds and effects a) – c) needs to be taken in full consideration. We want to stress, that the identification of the sector is of paramount relevance at the beginning of the location process.

There are S urns, one for each sector. We denote them as $U_1, U_2, \dots, U_S$.

At the beginning of the location process, each urn contains balls of colors $C_1, C_2, \dots, C_R$, so that the distribution of the colored balls in U_s coincides with the final distribution of the firms of sector s, i.e. $p^{(s)} = (p_1^{(s)}, p_2^{(s)}, \dots, p_R^{(s)})$, for each $s = 1,2,\dots,S$.

Suppose that the first employee (first extraction from the urn) belongs to the sector $s_1 \in \{1,2,\dots,S\}$. Then, she/he extracts her/his region by drawing a ball from the sector-specific urn U_{s_1}. Then, the employee locates, according to the color of the extracted ball. The number of the employees belonging to sector s in the R regions after the 1-st drawn is a vector $(m_1^{(1,s)}, m_2^{(1,s)}, \dots, m_R^{(1,s)})$, for $s = 1,2,\dots,S$.

After the extraction, colored balls are added/removed from each urn (and not only from U_{s_1}). This hypothesis captures the dependence structure among different sectors, which actually may interact.

By adopting the same notation of the previous section, at the end of the addition/removal procedure the number of the balls with colors $C_1, C_2, \dots, C_R$ in urn U_s is $C_1^{(2,s)}, C_2^{(2,s)}, \dots, C_R^{(2,s)}$, respectively, for each $s = 1,2,\dots,S$. These urns are ready for the second extraction (i.e., the second employee), even if only one of the urns will play an active role (the one associated to the sector of activity of the employee labeled with 2).

Indeed, the second employee is assumed to belong to the sector $s_2 \in \{1,2,\dots,S\}$. Then, she/he extracts her/his region by drawing a ball from U_{s_2}, while the remaining urns are not considered.

Recursively: at the l-th stage, the number of the balls with colors $C_1, C_2, \dots, C_R$ in urn U_s is $C_1^{(l,s)}, C_2^{(l,s)}, \dots, C_R^{(l,s)}$, respectively, for each $s = 1,2,\dots,S$. The number of the employees belonging to sector s in the R regions after the l-th drawn is $(m_1^{(l,s)}, m_2^{(l,s)}, \dots, m_R^{(l,s)})$, for $s = 1,2,\dots,S$ and $l = 1,2,\dots,L$.

The procedure stops at the L-th stage, when the last employee extracts and locates in a region.

Also in this case, the addition and removal of colored balls from the urn is due to a utility maximization over the set D of the discrete probability distributions on $\{1,2,\dots,L\}$.

Specifically, the l-th employee, belonging to sector s_l, is assumed to have utility function $u_{is_l}: D \to \mathbb{R}$. The utility maximization over D leads to a removal/addition of colored balls from each urn, and not only from U_{s_l}, to obtain a distribution of colored balls identical to the maximizing probability distribution $p^{(l,s)} = (p_1^{(l,s)}, p_2^{(l,s)}, \dots, p_R^{(l,s)})$, i.e.

$$p_r^{(l,s)} = \frac{C_r^{(l,s)}}{\sum_{r=1}^{R} C_r^{(l,s)}}, \quad \forall\, l = 1,2,\dots,L;\ s = 1,2,\dots,S \text{ and } r = 1,2,\dots,R.$$

As already stated above, all the urns change their configurations because the action of a sector might change the equilibria also in different sectors, and let a region be more or less attractive than another one also for employees belonging to sectors different from s_1.

The final distribution of the employees belonging to sector s over the R regions is a vector $q^{(s)} = (q_1^{(s)}, q_2^{(s)}, \dots, q_R^{(s)})$, where

$$q_r^{(s)} \quad \frac{m_r^{(L,s)}}{L_s}, \quad \forall r \quad 1,2,\ldots,R \text{ and } s \quad 1,2,\ldots,S.$$

By the sectorial analysis we can also infer the distribution of the employees in the overall manufacturing activity. Such a distribution is denoted as $q \quad (q_1, q_2, \ldots, q_R)$, where

$$q_r \quad \frac{\sum_{s=1}^{S} m_r^{(L,s)}}{L}, \quad \forall r \quad 1,2,\ldots,R$$

The deviation between $q^{(s)} \quad (q_1^{(s)}, q_2^{(s)}, \ldots, q_R^{(s)})$ and $q \quad (q_1, q_2, \ldots, q_R)$ can be viewed as the result of a cumulative process in the location of firms and employees that relate to different sources of agglomeration economies (e.g. industry-specific Marshallian economies, intra- and inter-sectoral complementarities between upstream and downstream firms along the value chains, knowledge spillovers).

3. A testing procedure for relative entropy measures

On the basis of the location choice model presented in 2, our method starts by postulating a null hypothesis of absence of relative concentration and then seeks to test this hypothesis by finding a threshold level able to distinguish a localization pattern significantly different from the counterfactual scenario, that is a disproportional share of employment relative to total manufacturing higher than the theoretical distribution.

Testing for localization taking advantage of the distinctive decomposition properties of relative entropy measures naturally involve testing the following null:

$$H0: 0 \leq I < I^*$$
$$H1: I \geq I^*$$

where I is one of the relative entropy statistics of overall localization (L, L^b, L^w), and relative concentration of sector s (T_s, T_s^b, T_s^w) defined and extensively discussed in Cutrini (2009). The threshold I^* represents the critical value, and can be arbitrarily selected, even if a too high or too low value of it could lead to a too narrow rejection or acceptance regions. The procedure of identification of I^* will be clear below, where the application of the localization test on real data will be implemented.

In our approach, the null hypothesis for each sector s represents perfect regularity of the location of employees relative to the spatial distribution of manufacturing employment across the observed units of analysis (EU regions). We advocate the possibility of a non-random spatial distribution of industry employment even when firm location is purely random, because of different sector-specific firm size, in the same vein as the Ellison and Glaeser (1997)'s approach. In fact, it may occur that, since firms locate independently from each other but employees do not, some degree of concentration of employment will naturally emerge when large establishments dominate the industry structure[5]. The test presented here aims to identify a significant localization of employees based on different sources of agglomeration effects, beyond what would arise simply because of the lumpiness of establishments' distribution across space.

Therefore, as for concentration, the null hypothesis thus refers to a counterfactual distribution where each sector employment is distributed across regions almost proportionally as the respective distribution of establishments is. Particularly, it is assumed that, in the absence of any agglomerative forces, for each sector, the distribution of employment is overlapping the distribution of firms.

[5] This is the reason why the raw G index is corrected in the EG's approach. In other words, they can distinguish a localization determined by the lumpiness of establishments' distribution across space, from a localization of employees based on different sources of agglomeration effects.

Hence, bearing in mind the location model presented in section 2, the null hypothesis of perfect regularity is constructed on the basis of two basic assumptions:

(a) uniform distribution of firms controlling for industrial structure and
(b) employment assigned to regions proportionally to firms.

The null hypothesis is rejected when the true value of the relative entropy index J is higher than the threshold J^*. In this case, the divergence between the distribution of employment and the distribution of establishments is considered to be sufficiently high.

Since we do not have individual data, we use aggregate values for the distribution of employment and the distribution of firms: here l_{rs} is the number of employees in sector *s* working in region *r* of country *c* while n_{rs} is the number of establishments in sector *s* settled in region r of country *c*.

We construct the counterfactual distribution of employment, considering that, under the null hypothesis, firms' location follows a uniform law with upper and lower bounds based on the actual distribution of firms across regions. Then a pull of workers l_{rs} is assigned to each region according to the industry-specific average firm size (L_s).

We run 10,000 Montecarlo simulations for each industry employment distribution, separately[6].

This procedure is adopted for assessing the values of the thresholds J^* introduced for defining the hypothesis of the tests, in all the cases of overall localization and the components between and within countries, for each industrial sector. Specifically, we calculate confidence intervals for each index of localization (and concentration) and by ordering the 10,000 simulated values of the 60 entropy measures (for 3-digit we select the threshold levels (1%, 5%, 10%) from the top of our generated distribution and then record the critical values. Such critical values are the required thresholds. If employment is actually clustering, the real value of the index should be above the thresholds defined as explained above, and we will conclude that this sector is characterized by spatial clustering because of a disproportional share of employment relative to total manufacturing higher than the counterfactual distribution.

4. An illustrative application

This section present the results of the test applied to European manufacturing location pattern in 2007.

4.1 Data

We collected data on existing firms and employment in Europe in 2007 from the Eurostat database. The analysis considers a sample of 133 NUTS 3 level in 16 countries: Austria, Belgium, Bulgaria, Czech Republic, Finland, France, Germany, Hungary, Italy, Norway, Poland, Portugal, Romania, Spain, Sweden, and the UK. The list of the regions is given in the Appendix A. The analysis is restricted to 19 manufacturing industries according to the NACE rev.1.1 three-digit classification (Statistical Classification of Economic Activities in the European Community). In a very limited number of cases, we complete data on the basis of the preceding year (2006), when they were available. In any case, we had to exclude four manufacturing sectors from our analysis (food, beverages and tobacco, leather and footwear, coke, refined petroleum) because of the huge amount of confidential and missing data.

[6] We do not use bootstrapping but rather Monte Carlo: we do not resample the original data with replacement. When we used bootstrapping we found that it leads us to always reject the null hypothesis.

4.2 Results

Before presenting the results of the statistical testing procedure applied to entropy-based measures of spatial concentration we present summary statistics for the average regional firm size for all the industries and total manufacturing (See Table 1).

Figures in Table 1 clearly highlight that the industries with smaller average firm sizes are *wood, recycling, wearing apparel, fabricated metal products, textiles, Medical, precision and optical instruments, watches and clocks, furniture, manufacturing nec, Other non-metallic mineral products, publishing printing*. In these low-scale industries the regional variation in the average firm size is also quite low.

Instead, higher average firm sizes are found in *Motor vehicles* (113 employees in each establishment, on average), *Basic metals* (73), *Chemicals* (52), *Other transport equipment* (61). The empirical evidence is consistent with the idea that the automotive industry and the chemical and steel industry are characterized by increasing returns to scale.

Table 1 Average regional firm size, aggregate regional data by sector

	Obs	**Mean**	**Std. Dev.**	**Lr_ Min**	**Lr_ Max**
Textiles	133	19	22.73	1	125
Wearing apparel	133	15	24.13	0	147
Wood	133	11	14.31	2	82
Paper	133	43	40.83	3	228
Publishing, printing	133	17	58.91	1	652
Chemicals	133	52	64.19	6	474
Rubber and plastics	133	28	22.13	3	117
Other non-metallic mineral products	133	16	12.40	3	69
Basic metals	133	73	74.40	3	410
Fabricated metal products	133	15	19.61	3	97
Machinery and equipment	133	27	31.23	2	169
Office machinery and computers	133	34	106.75	1	1142
Electrical machinery and apparatus n.e.c.	133	39	47.35	2	265
Radio, television and communication equipment	133	38	59.74	1	544
Medical, precision and optical instruments, watches and clocks	133	16	22.88	2	136
Motor vehicles	133	113	179.42	4	1492
Other transport equipment	133	61	84.43	1	720
Furniture; manufacturing n.e.c.	133	13	21.11	1	117
Recycling	133	11	11.17	3	74
Total manufacturing	133	22	31.31	2	192

Table 2 summarizes our results on the level of overall localization and concentration, and its statistical significance for 2007: we report the relative concentration values (T_s) for each industry, industries are grouped according to the technology intensity classification based on OECD, 2011.

Table 2 Geographical concentration of employment in Europe, 2007

	real values	sign		real values	sign
Overall localization	0.210	***			
High-tecnology industries			**Medium-low-tecnology industries**		
Office machinery and computers	0.453	***	Other non-metallic mineral products	0.141	
Radio, television and communication equipment	0.314	***	Basic metals	0.353	***
Medical, precision and optical instruments, watches and clocks	0.175		Fabricated metal products	0.060	
Medium-high-tecnology industries			**Low-tecnology industries**		
Chemicals	0.209		Textiles	0.346	***
Rubber and plastics	0.093		Wearing apparel	0.592	***
Machinery and equipment	0.079		Wood	0.200	
Electrical machinery and apparatus n.e.c.	0.109		Paper	0.161	
Motor vehicles	0.367	***	Publishing, printing	0.333	***
Other transport equipment	0.465	***	Furniture; manufacturing n.e.c.	0.134	
			Recycling	0.241	*

Note: Technology intensity classification based on OECD, 2011

We found higher and significant values for concentration in 9 out of 19 manufacturing industries. With our theoretical model in mind, we suggest that the significant spatial concentration in these industries is the result of the strength of localization economies that is a mixture of advantages accruing to firms from being located close to other establishments in the same industry (industry-specific knowledge spillovers, labour market pooling, sharing of specialized suppliers).

It is worth noting that industries with significant overall concentration are low-tech industries usually considered as localized because of Marshallian external economies (i.e. *wearing apparel, textiles, publishing, printing*), small-scale knowledge intensive industries (e.g *office machinery and computers, radio, television and communication equipment*) but also industries characterized by internal scale economies (*motor vehicles, basic metals, chemicals, other transport equipment*).

For the remaining industries in which concentration is not statistically significant, we suggest that location choices may have been driven mainly by the advantages of inter-industry complementarities and spillovers among firms across sectors, urbanization economies and the associated benefits of diversity and innovation potential.

Table 3 reports detailed results for industrial concentration. We report the actual values of the relative entropy measures of industrial concentration and their geographical components together with the threshold levels to be considered as statistically different from regularity (at the 90%, 95% and 99% levels).

Table 3 Detailed results for industrial concentration measures - three-digit industrial disaggregation (NACE rev. 1.1), 2007

				Critical values (thresholds I^*)		
		real values	Sign	p90	p95	p99
Localization within countries	Lw	0.120		0.164	0.166	0.171
Localization between countries	Lb	0.090	***	0.020	0.021	0.023
High-tecnology industries						
Radio, television and communication equipment	Tw	0.189		0.204	0.213	0.231
	Tb	0.124	***	0.030	0.034	0.041
Medical, precision and optical instruments, watches and clocks	Tw	0.103		0.204	0.212	0.229
	Tb	0.072	***	0.030	0.034	0.041
Office machinery and computers						
	Tw	0.286	***	0.216	0.225	0.245
	Tb	0.167	***	0.032	0.035	0.043
Electrical machinery and apparatus n.	Tw	0.064		0.182	0.190	0.206
	Tb	0.045	***	0.027	0.030	0.037
Medium-high-tecnology industries						
Chemicals	Tw	0.151		0.195	0.203	0.219
	Tb	0.058	***	0.029	0.032	0.039
Rubber and plastics	Tw	0.069		0.194	0.202	0.219
	Tb	0.024		0.028	0.032	0.039
Machinery and equipment	Tw	0.051		0.164	0.171	0.186
	Tb	0.027	**	0.023	0.027	0.033
Other transport equipment	Tw	0.312	***	0.197	0.205	0.223
	Tb	0.153	***	0.029	0.032	0.039

				Critical values (thresholds I^*)		
		real values	Sign	p90	p95	p99
Medium-low-tecnology industries						
	Tw	0.098		0.200	0.208	0.227
Other non-metallic mineral products	Tb	0.043	***	0.030	0.033	0.040
	Tw	0.296	***	0.198	0.207	0.226
Basic metals	Tb	0.058	***	0.029	0.033	0.040
	Tw	0.039		0.159	0.167	0.181
Fabricated metal products	Tb	0.021		0.023	0.025	0.031
	Tw	0.228	***	0.200	0.209	0.227
Motor vehicles	Tb	0.140	***	0.029	0.033	0.041
Low-tecnology industries						
Textiles	Tw	0.186		0.200	0.209	0.226
	Tb	0.160	***	0.029	0.033	0.040
Wearing apparel	Tw	0.109		0.198	0.206	0.222
	Tb	0.483	***	0.029	0.032	0.039
Wood	Tw	0.113		0.206	0.215	0.230
	Tb	0.088	***	0.030	0.034	0.042
Paper	Tw	0.094		0.208	0.217	0.236
	Tb	0.067	***	0.031	0.034	0.042
Publishing, printing	Tw	0.218	***	0.182	0.189	0.206
	Tb	0.115	***	0.026	0.029	0.036
Furniture; manufacturing n.e.c.	Tw	0.067		0.197	0.207	0.224
	Tb	0.067	***	0.029	0.032	0.039
Recycling	Tw	0.089		0.216	0.226	0.245
	Tb	0.152	***	0.032	0.035	0.044

The decomposition method used in the present work to account for the complexity in the spatial concentration and overall localization at the different spatial scales is based on Cutrini (2009). There is evidence that the component of the index related to the case "within countries" is greater than the one associated to "between countries". This outcome is rather expected, and captures the fact that any statistical measure based on spatial aggregates is sensitive to the scale and aggregation problems. In fact, any time we superimpose a finer grid (regions) onto the same set of data, a spatial concentration index will take a higher values as in the case data are aggregated according to national boundaries. Interestingly, in this respect three Low-tech industries are exceptions in our application: *Wearing apparel* (T_w= 0.109; T_b=0.483), *Recycling* (T_w= 0.089; T_b=0.152) and *Furniture* (T_w= 0.067; T_b=0.067).

Let us denote as $T^*(j), T^*w(j), T^*b(j)$ the thresholds – denoted as I^* in the definition of the hypotheses of the test and appearing in the Table as "Critical values"- obtained through the 10000 Monte Carlo simulations at j 90%, 95%, 99% for the case "overall countries", "within countries" and "between countries", respectively. Such thresholds have been computed case by case on the different industrial sectors.

As already stressed above, small values of the indices mean substantial absence of concentration. The analysis of the significance of the statistics for the different sectors provides interesting insights.

In general, it is worth saying that the statistics for "between countries" are more often significant than those of "within countries", and this is true at the level 99%. Specifically, in several cases we have $T^*b(0.99) < Tb < Tw < T^*w(0.90)$. For examples, in some industrial sectors characterized by very different kind of knowledge intensity - *such as Wearing apparel* and *Radio, television and communication equipment* – the overall index and its between-country component are statistically significant (at the 99% level*)* while the spatial concentration in other industries – i.e. *Wood, Paper, Furniture, Other non-metallic mineral products, Fabricated metals, Machinery and equipment, Electrical machinery, Rubber and Plastics, Chemicals, Medical, precision and optical instruments, whatches and clocks* - is significantly different from regularity only at the between country level.

Conversely, *Rubber and plastics* and *Fabricated metal products* are sectors where localization indices are not significant at any level in the three cases *overall, between and within*. The concentration of such industrial sectors is then rather close to a random distribution once controlled for industrial structure, and does not exhibit polarization properties.

The decomposition analysis allows to ascertain that five of the three-digit manufacturing industries result as being significantly clustered both within- and between national borders. In particular, the statistics for the industrial sectors of *Office machinery and computers, Publishing and printing, Basic metals, Motor vehicles and other transport equipment* are significant at the 99% level in all the cases (within, between, overall). We may suggest that, in these cases, both high-scale polarization forces -such as market potential or institutional heterogeneity across national borders- and localization economies acting at shorter distances are important factors explaining location patterns. In particular, as for the agglomeration forces within countries, we advocate that in some sectors –like *motor vehicles*, *transport equipment, basic metals* - the observed relative concentration of employment exceeds the regular scenario and thus it goes beyond the pure effect of internal scale economies while in other industries – like *office machinery and computers, publishing, printing* - knowledge spillovers and other Marshallian economies may act as the main agglomerative forces at short distances.

Conclusions

Even if they are constructed over the discrete space, entropy-based measures are particularly informative because of the possibility to disentangle the relative importance of intra-country locational advantages from cross-country divergence in localization. Moreover, they are easily applicable in practice because the data requirements for computing them are relatively low. Hence, it is of great importance to define critical values to give statistical significance to empirical findings on localization based on relative entropy measures.

To improve the economic interpretation of spatial concentration we have introduced a testing method for entropy-based indices embedded in a micro-founded location model. In the EG approach is introduced the idea that in the case of random location, the observed distribution will not be perfectly regular. Firms locate independently under the null hypothesis, but employees do not. Therefore, even if the null hypothesis represents stochastic location some degree of spatial concentration is present, and this is the reason for identifying critical values of significant localization through counterfactual distributions.

Our method based on entropy measures allows identifying statistically significant departures from a random distribution of economic activity across geographic space.

As an illustrative exercise, we use our method to calculate critical values of significant concentration of manufacturing industries in Europe in 2007. We run 10,000 Montecarlo simulations for each industrial distribution across regions and we determine 10,000 simulated values of the whole set of entropy measures of concentration and localization and their components. We identify critical values of type I error by ordering the 10,000 simulated values of each relative entropy index and selecting those that correspond to the top 5% (or 10%) cases of each generated distribution. Our findings suggest that a significant departure from a regular location scenario does not depend on technology intensity, neither by the presence of internal scale economies. By the mean of the decomposition analysis we found that five of the three-digit manufacturing industries result as being significantly clustered both within- and between national borders. In particular, the

statistics for the industrial sectors of *Office machinery and computers, Publishing and printing, Basic metals, Motor vehicles and other transport equipment* are significant at the 90% level in all the cases (within, between, overall). Moreover, for almost all manufacturing sectors, national boundaries are relevant economic borders from the point of view of industrial clustering.

Unfortunately, this method does not allow distinguishing empirically agglomeration due to Marshallian externalities, from that due to other agglomeration forces like the distribution of final demand (market potential) and/or inter-sectoral linkages with other manufacturing industries or service activities. This should be one promising topic for future research.

Acknowledgements

The authors are grateful to Masahisa Fujita, Jacques-François Thisse, Kristian Behrens, Philipp Ushchev, and the participants at The Fourth International Conference "Industrial Organization and Spatial Economics" Saint-Petersburg, Russia, 2015 and at the GeComplexity conference, Heraklion, Greece 2016 for very helpful comments.

Bibliography

Alonso-Villar O. and Del Río C. (2013), Concentration of Economic Activity: An Axiomatic Approach, *Regional Studies* 47(5), 756-772.

Arbia G. (2001), Modelling the geography of economic activity on a continuous space, *Papers in Regional Science* 80, 411-424.

Arthur, W. B. (1994). Industry location patterns and the importance of history*, vol. Increasing returns and path dependence in the economy,* ch. 4, pp. 49–68. Michigan U. Press, Michigan.

Bagoulla C., Péridy N. (2011), Market access and the other determinants of North–South manufacturing location choice: An application to the Euro-Mediterranean area, *Economic Systems* 35, 537-561

Bickenbach, F., and E. Bode (2008), Disproportionality Measures of Concentration, Specialization and Localization. *International Regional Science Review* 31(4), 359-388.

Bickenbach, F., Bode, E., and Krieger-Boden, C. (2010), Closing the gap between absolute and relative measures of localization, concentration or specialization. Kiel Institute for the World Economy, Working Paper 1660.

Brülhart, M., and R. Träger (2005), An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics* 35(6), 597-624.

Cassey, A. J. & Smith, B. O. (2014), Simulating confidence for the Ellison–Glaeser index, *Journal of Urban Economics*, Elsevier, vol. 81(C), 85-103.

Cutrini E. (2009). Using entropy measures to disentangle regional from national localization patterns. *Regional Science and Urban Economics* 39(2), 243-250.

Cutrini, E. (2010), Specialization and Concentration from a Twofold Geographical Perspective: Evidence from Europe. *Regional Studies* 44(3), 315–336.

David, P. (2000). Path dependence, its critics and the quest for 'historical economics'. Oxford and Stanford University.

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

Duranton, G., and H. G. Overman (2005), Testing for Localisation Using Micro-Geographic Data. *Review of Economic Studies* 72(4), 1077–1106.

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

Evans, A. J. (2010), Complex spatial networks in application. *Complexity* 16, 11–19

Gokan, T. (2010), On the usage of the measurements of geographical concentration and specialization with areal data. In: *Spatial Statistics and Industrial Location in CLMV: An Interim Report,* cap. 2, Kuroiwa ed., Chosakenkyu Hokokusho, IDE-JETRO.

Haedo C., and Mouchart M. (2013), A stochastic independence approach for different measures of concentration and specialization, *CORE Discussion Paper N. 25*

Hoover, E.M., 1936. The measurement of industrial localization. *Review of Economics and Statistics* 18(4), 162–171.

Lang, G., Marcon, E. and F. Puech (2014), Distance-based measures of spatial concentration: introducing a relative density function, *Paper presented at the NARSC Annual Conference*, November 2014.

Marcon, E., and F. Puech (2010), Measures of the Geographic Concentration of Industries: Improving Distance-based Methods. *Journal of Economic Geography* 10(5), 745–762.

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

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

Mori, T., Nishikimi, K., Smith, T.E., 2005. A divergence statistic for industrial localization. *The Review of Economics and Statistics* 87(4), 635–651.

OECD (2011), ISIC Rev. 3 Technology Intensity Definition - Classification of manufacturing industries into categories based on R&D intensities, OECD Publishing

Ohlin, B., 1933. Interregional and International Trade. Harvard Economic Studies, vol.39. Harvard University Press, Cambridge, MA.

Stierle-von Schultz U., Stierle M. H. (2013), Regional Specialisation and Sectoral Concentration in an Enlarged EU, INFER WP. 2/2013

Vechiu N., Makhlouf F. (2014), Economic integration and specialization in production in the EU27:does FDI influence countries' specialization?, *Empirical Economics* 46, 543-572

Appendix A

Table A1 List of NUTS 3 regions (3 digit classification)

NUTS3	region	NUTS3	region	NUTS3	region
BE35	Prov. Namur	ES30	Comunidad de Madrid	ITD2	Provincia Autonoma Trento (NUTS 2006)
BG32	Severen tsentralen	FR10	Île de France	ITD3	Veneto (NUTS 2006)
BG34	Yugoiztochen	FR21	Champagne-Ardenne	ITD4	Friuli-Venezia Giulia (NUTS 2006)
BG41	Yugozapaden	FR22	Picardie	ITD5	Emilia-Romagna (NUTS 2006)
BG42	Yuzhen tsentralen	FR23	Haute-Normandie	ITE1	Toscana (NUTS 2006)
CZ01	Praha	FR24	Centre (FR)	ITE2	Umbria (NUTS 2006)
CZ02	Střední Cechy	FR25	Basse-Normandie	ITE3	Marche (NUTS 2006)
CZ03	Jihozápad	FR26	Bourgogne	ITE4	Lazio (NUTS 2006)
CZ05	Severovýchod	FR30	Nord - Pas-de-Calais	ITF1	Abruzzo
CZ06	Jihovýchod	FR41	Lorraine	ITF3	Campania
CZ07	Střední Morava	FR42	Alsace	ITF4	Puglia
CZ08	Moravskoslezsko	FR43	Franche-Comté	ITF5	Basilicata
DE11	Stuttgart	FR51	Pays de la Loire	ITF6	Calabria
DE13	Freiburg	FR52	Bretagne	ITG1	Sicilia
DE21	Oberbayern	FR53	Poitou-Charentes	ITG2	Sardegna
DE30	Berlin	FR61	Aquitaine	HU10	Közép-Magyarország
DEA1	Düsseldorf	FR62	Midi-Pyrénées	HU21	Közép-Dunántúl
DEA2	Köln	FR63	Limousin	HU22	Nyugat-Dunántúl
DEA5	Arnsberg	FR71	Rhône-Alpes	HU23	Dél-Dunántúl
DED1	Chemnitz (NUTS 2006)	FR72	Auvergne	HU31	Észak-Magyarország
DEG0	Thüringen	FR82	Provence-Alpes-Côte d'Azur	HU32	Észak-Alföld
ES11	Galicia	FR92	Martinique (FR)	HU33	Dél-Alföld
ES12	Principado de Asturias	FR94	Réunion (FR)	AT31	Oberösterreich

NUTS3	region	NUTS3	region	NUTS3	region
PL21	Malopolskie	UKD5	Merseyside (NUTS 2006)	NO04	Agder og Rogaland
PL22	Slaskie	UKE3	South Yorkshire	NO05	Vestlandet
PL32	Podkarpackie	UKE4	West Yorkshire	NO06	Trøndelag
PL34	Podlaskie	UKF2	Leicestershire, Rutland and Northamptonshire		
PL41	Wielkopolskie	UKG1	Herefordshire, Worcestershire and Warwickshire		
PL42	Zachodniopomorskie	UKG2	Shropshire and Staffordshire		
PL43	Lubuskie	UKG3	West Midlands		
PL51	Dolnoslaskie	UKH1	East Anglia		
PL52	Opolskie	UKH2	Bedfordshire and Hertfordshire		
PT11	Norte	UKH3	Essex		
PT16	Centro (PT)	UKI1	Inner London		
RO11	Nord-Vest	UKI2	Outer London		
RO12	Centru	UKJ1	Berkshire, Buckinghamshire and Oxfordshire		
RO21	Nord-Est	UKJ2	Surrey, East and West Sussex		
RO22	Sud-Est	UKJ3	Hampshire and Isle of Wight		
RO31	Sud - Muntenia	UKJ4	Kent		
RO32	Bucuresti - Ilfov	UKK1	Gloucestershire, Wiltshire and Bristol/Bath area		
RO41	Sud-Vest Oltenia	UKK2	Dorset and Somerset		
RO42	Vest	UKK3	Cornwall and Isles of Scilly		
FI13	Itä-Suomi (NUTS 2006)	UKL1	West Wales and The Valleys		
FI18	Etelä-Suomi (NUTS 2006)	UKL2	East Wales		
FI19	Länsi-Suomi	UKM2	Eastern Scotland		
SE22	Sydsverige	UKM3	South Western Scotland		
SE31	Norra Mellansverige	UKN0	Northern Ireland (UK)		
UKD2	Cheshire (NUTS 2006)	NO01	Oslo og Akershus		
UKD3	Greater Manchester	NO03	Sør-Østlandet		