10

Measuring Spatial Concentration

The measurement of inequalities has long been of interest to economists, especially to those trying to evaluate inequalities across individual incomes (Sen 1973) or the degree of concentration in a given sector (Scherer 1980). In chapter 1, we presented some stylized facts regarding the spatial concentration of economic activities. However, our assessment of income distributions has been restricted to simple tools, such as numbers, tables, and maps depicting the GDP per capita. Although such tools succeed in conveying a general impression of the scope of spatial inequalities, their limitations are readily apparent. First, while the monotone increase in inequalities that characterizes the nineteenth century is easily captured by simple statistics, as soon as such inequalities become more complex and nonmonotone, we must turn to more sophisticated tools. Second, when the number of regions is large, it is no longer appropriate to simply rank all regions according to a given index of development. Third, extending our analysis to the sectoral level requires specific tools, as both predictions and policy prescriptions are likely to vary markedly across sectors. Last, with the ambition of *making better use of all available information* (modern databases often distinguish between more than 500 sectors) comes the need to move beyond simple maps and tables that cannot capture this breadth of information appropriately.

Geographers and economists alike have sought to develop indices that capture inequality across industries, time, and space. It will become readily apparent that the issue is more complex than it seems at first glance. Although some indices have become standard, the *ideal index* remains to be discovered. Section 10.1 provides a set of properties that should be met by an ideal index, which gives us a benchmark by which to judge existing indices. The subsequent sections introduce the main approaches that are applied in the literature. The first approach is based on indices used for measuring inequality across individuals, such as the Gini index, an approach that can also be used to evaluate the industrial concentration of firms belonging to the same sector. This will allow us to find out how spatial inequalities imply new and specific constraints. Recently, attempts have been made to account for these constraints. An example of primary importance is the seminal work of Ellison and Glaeser (1997), which has generated a new family of indices allowing for more relevant comparisons of spatial concentration across industries. In the last section, we discuss the approach proposed by Duranton and Overman (2005), which constitutes an important step forward in conceptualizing and measuring spatial concentration. Based on the average distance between plants, this approach frees the analysis from the need to use spatial classifications, thereby reducing the corresponding biases.

10.1 The Properties of an Ideal Index of Spatial Concentration

Any empirical tool, ranging from the most basic to the most sophisticated, rests on a specific set of assumptions. This is true of linear regressions that presuppose that the error term has a very specific structure, but it is also true of more descriptive apparatus, such as inequality indices. It is, therefore, important to understand the implications of the assumptions made, and to compare them with the desirable properties that an ideal index should have. Some assumptions are less suggestive than others; as we will see, even the most obvious assumptions are not always satisfied.

To start with, given that most studies are carried out at the sector level, the first property that must be satisfied is as follows.

Property 10.1. Measures of spatial concentration should be comparable across industries.

In more concrete terms, this amounts to saying that we must be capable of comparing the degree of concentration in the automobile industry with, say, that in the chemicals industry. More generally, it means verifying whether a comparison of spatial concentration using a classification that describes a few sectors with spatial concentration using another classification that distinguishes more sectors is possible.

For example, is comparing the spatial concentration of poultry farmers with that of agriculture possible and meaningful? Although this property may seem obvious, it is nevertheless not satisfied by the simplest indices. The difficulty lies in the existence of differences regarding the size distribution of firms belonging to the same sector, a characteristic that is often called its "industrial concentration" (Scherer 1980). By affecting the average size, and hence the total number, of plants (or branches or stores), the degree of concentration of a sector impacts on its spatial concentration. It is, therefore, *highly desirable to distinguish industrial concentration from spatial concentration*, the latter having to be independent of the distribution of the activity among plants. As will be seen later on, this is only possible when data at the plant level are available or, at the very least, when the total number of plants in the sector under scrutiny is known.

The spatial counterpart of property 10.1 is as follows.

Property 10.2. Measures of spatial concentration should be comparable across spatial scales.

When this property holds, it allows for the meaningful comparison of spatial concentration across countries, or across different levels of spatial scales as, say, whether an activity is more concentrated at the national than the regional level. Somewhat surprisingly, property 10.2 has been more readily grasped than property 10.1, even though they are symmetric. For example, geographers have long pointed out that the number of regions considered in different countries is likely to have an influence on the comparison of their degree of regional concentration. Even though this problem has often been raised, only the most recent indices address this issue head-on.

Two other properties that deal with the definition of spatial units and sectors should also be satisfied. The first is as stated below.

Property 10.3. Measures of spatial concentration should be unbiased with respect to arbitrary changes to spatial classification.

For instance, let us suppose that the ninety-four départements of mainland France are replaced by ninety-four spatial units defined in a different way. In measuring the spatial concentration of a given sector, the index should have the same value under both definitions. This problem was brought to light long ago, and may be attributed to the delineation of borders separating spatial units. For any given geographical area, the underlying economic problem stems from the fact that homogeneous *economic* zones seldom coincide with *administrative* zones: tightly linked economic agents (such as employees and their workplaces, or firms and their subcontractors) are thus often split across different administrative spatial units. Hence, changing the definition of spatial units may result in a significant, but artificial, redistribution of economic activity. In other words, such changes can translate into different measures of concentration even though the degree of "real" agglomeration remains unchanged. More generally, the problems lying behind the difficulties encountered with properties 10.2 and 10.3 are related to *the* *discretization of a continuous space*, which is known as the modifiable areal unit problem (MAUP).[1]

For any given discretization, a related problem is that the standard indices generally do not take into account the *relative position* of spatial units (see Thomas (2002) and our discussion regarding two-region models in chapter 4). And yet proposition 10.3 requires that the index changes value when units are switched around (suppose for instance that you could invert the locations of London and Liverpool; the measured spatial agglomeration of activities in the United Kingdom would change). As a counterexample, the first two families of indices discussed below do not satisfy this property: they take the same values regardless of whether economic activities are located in adjacent or distant regions.

We find a criterion similar to property 10.3 with respect to the industrial dimension.

Property 10.4. Measures of spatial concentration should be unbiased with respect to arbitrary changes to industrial classification.

As seen above, the carving up of spatial units is arbitrary so that borders may separate regions with strong economic ties. Likewise, in defining a limited number of sectors, the industrial classification may also arbitrarily separate closely related economic activities. In particular, some related activities will inevitably be separated, while conversely others are likely to be grouped together despite marked differences. Furthermore, the precision of any given industrial classification often depends on the sector at hand. For instance, existing classifications typically distinguish between more items in the manufacturing sector than in services. This is another artificial source of difference between measures of concentration. Drawing from the idea of proximity in physical space, it may prove promising to consider the technological proximity that exists between industries by creating a measure of "technological distance." A generalized distance that would account for both spatial and technological distances could then be used when evaluating the spatial concentration of a sector.

The last two desirable properties are related to the possible existence of statistical criteria that enable us to test for the presence of spatial concentration.

Property 10.5. Measures of spatial concentration should be carried out with respect to a well-established benchmark.

[1] See Francis et al. (forthcoming) for a detailed analysis of the MAUP and Briant et al. (2007) for a detailed empirical assessment.

One benchmark that naturally comes to mind, and underlies a number of existing indices, is the uniform distribution. However, when studying the spatial distribution of a given sector, it seems more relevant and fruitful to use the overall distribution of activities. Although this is rarely done in practice, using a benchmark grounded in a specific economic model is also likely to lead to more consistent indices. In particular, such an approach would allow one to investigate whether the observed distribution differs from that derived from a specific theoretical framework.

Finally, regardless of the way the benchmark is defined, being able to determine whether the observed distribution is *significantly* different from its benchmark appears to be crucial. Furthermore, when do two estimators of spatial concentration differ significantly across areas, periods, or industries? This leads us to the last property an ideal index should satisfy.

Property 10.6. The measure should allow one to determine whether significant differences exist between an observed distribution and its benchmark, or between two situations (areas, periods, or industries).

Without these types of statistical tests, concentration indices have little value. This is because we are unable to determine whether we are dealing with high or low concentration, or whether there is even any spatial concentration at all.

10.2 Spatial Concentration Indices

10.2.1 The Gini Index

The most popular index for measuring inequality is undoubtedly the Gini index. It was originally used to evaluate inequalities across personal incomes (Sen 1973). In our context, it will be used to evaluate the spatial concentration of a given sector in terms of some given magnitude such as employment, production, or value-added. Let x_r^s be the level of the magnitude under consideration (e.g., employment) in sector $s = 1, \dots, S$ and in region $r = 1, \dots, R$. As with all the indices presented in this section, the Gini index is based on how regional shares of sector s, denoted λ_r^s, are distributed across regions:

$$\lambda_r^s = \frac{x_r^s}{x^s},$$

where $x^s \equiv \sum_{r=1}^{R} x_r^s$ is the total employment level in sector s.

Let us start with a graphical interpretation of this index, which will convey its intuitive meaning most readily. The main idea is to sort regions in ascending order by their degree of specialization in sector s (as measured by λ_r^s) and to draw what is known as the *Lorenz curve*. The x-coordinate corresponding to a point on this curve represents the fraction n/R of the n regions with the lowest employment shares in sector s. The y-coordinate corresponds to the cumulative share of these n regions in total employment, i.e.,

$$\lambda_{r(n)}^s = \sum_{r=1}^{n} \lambda_r^s.$$

If employment levels in sector s were uniformly distributed across all regions, each region would have $1/R$ of total employment, in which case the Lorenz curve would be given by the 45° line. As soon as the spatial distribution is not uniform, the Lorenz curve lies below the 45° line. In other words, the region with the lowest share of employment in sector s has a share of employment smaller than $1/R$; the first two such regions have a combined share that is smaller than $2/R$, and so on. In this case, a more unequal distribution translates to having greater levels of employment concentrated in a small number of large regions (and therefore lower levels in smaller regions): the greater the inequality, the more the Lorenz curve departs from the 45° line. The Gini index is given by the area that lies between the Lorenz curve and the 45° line (which needs to be multiplied by two for the upper bound of the index to be equal to one). The index ranges from zero, when the distribution of employment in the sector is uniform, to one, when all employment is concentrated in a single region.

Under the normalization $\lambda_{r(0)}^s = 0$, the *Gini index* is formally defined by

$$G^s = 1 - \sum_{n=1}^{R} \frac{1}{R}[\lambda_{r(n-1)}^s + \lambda_{r(n)}^s],$$

with each term in the sum corresponding to twice the area of the trapezoid situated below the Lorenz curve and delimited by the $(n-1)$th and the nth regions. This Gini index is called *absolute* because it uses the uniform distribution as a benchmark: each region is assigned the same weight $1/R$.

Another possibility involves comparing the distribution of sectoral employment with that of total employment, in order to determine the extent to which a given sector is more, or less, concentrated than the economy as a whole. This can be easily accomplished by replacing the x-coordinate values of the Lorenz curve: instead of using intervals of identical size $(1/R)$ for each region, as done with the uniform distribution, the intervals now have a varying length that corresponds to the total employment share of each region, which is given, for region r, by

$$\lambda_r = \frac{x_r}{x},$$

where $x_r = \sum_{s=1}^{S} x_r^s$ denotes the total employment in region r and $x = \sum_{s=1}^{S} x^s = \sum_{r=1}^{R} x_r$ the total employment in the area under study. What is called the *relative* Gini index uses an alternate Lorenz curve. Specifically, regions are now sorted in ascending order of their specialization with respect to their total size (as measured by λ_r^s/λ_r). Then we denote by $\lambda_{r(n)} = \sum_{r=1}^{n} \lambda_r$ the sum of the shares of total employment of the n least specialized regions in the sector under consideration. The shares $\lambda_{r(n)}$ are now used as the x-axis. Unlike the absolute index, where intervals are given by $1/R$, the relative index uses intervals of variable size given by $\lambda_{r(n)} - \lambda_{r(n-1)}$, which is merely the share in the total employment of the nth region. Formally, the relative Gini index equals twice the area that lies between the 45° line and this new Lorenz curve:

$$G^s = 1 - \sum_{n=1}^{R} \lambda_r[\lambda_{r(n)}^s + \lambda_{r(n-1)}^s]. \qquad (10.1)$$

Unfortunately, both the relative and absolute Gini indices only satisfy a very limited number of the ideal index's properties. For example, they do not allow us to adequately compare industries having different market structures (property 10.1): a limitation that served as the catalyst for the development of a new wave of indices presented in the next section. Examining variations over time can also be biased by the fact that the total number of firms in a country varies over time, even if this variation is uniform across spatial units. It should also be clear that comparing different zones is problematic because they typically differ in their number of regions (property 10.2). For example, splitting a region into two smaller ones changes the ordering of regions, and thus modifies the Gini index.

When the first two properties are not satisfied, it follows that the third and fourth are also violated. On the other hand, the benchmark underlying both the absolute and the relative Gini indices is well-defined (property 10.5): the benchmark is the uniform distribution in the absolute index and the actual distribution of total activity in the relative index. However, to date, no statistical tests have been proposed for determining whether observed values depart significantly from their benchmark values (property 10.6).

It should be emphasized that indices measuring concentration across regions have a natural counterpart, i.e., absolute and relative *specialization indices* that measure the industrial structure of regions. While spatial concentration determines whether a given sector is more or less concentrated across regions, specialization determines whether a particular region accommodates a more or less equal distribution of all sectors. For example, we can construct a Gini specialization index to measure sector s employment shares within a given region r:

$$\mu_r^s = \frac{x_r^s}{x_r}.$$

Again, we sort the sectors in ascending order of their weight in region r, and construct a Lorenz curve by generating intervals on the x-axis that correspond either to $1/S$ (as with the absolute index) or to each sector's share of total employment (as with the relative index). The *Gini specialization index* for region r is given by twice the area between this new Lorenz curve and the 45° line, so that an expression similar to (10.1) holds. Given that indices of both concentration and specialization are founded on the same axioms, it must be that they share the same advantages and limitations.

10.2.2 The Isard, Herfindhal, and Theil Indices

Other indices that share more or less the same characteristics as the Gini index have been proposed in the literature. They are subject to the same drawbacks as the Gini index. They can also be defined by reference to the uniform distribution or the distribution of total activity. In what follows, our benchmark is the total employment distribution, thus making these indices comparable to the relative Gini index. Replacing λ_r with $1/R$ in the expressions given below allows one to obtain the corresponding absolute indices.[2]

1. The *Isard index*, which regained popularity through Krugman (1991c), consists of a measure of concentration based on the absolute distance

between the actual and benchmark employment distributions:

$$I^s = \frac{1}{2}\sum_{r=1}^{R} |\lambda_r^s - \lambda_r|.$$

2. The *Herfindhal index* is the weighted sum of the square of each region's sectoral employment share:

$$H^s = \frac{1}{R}\sum_{r=1}^{R} \lambda_r \left(\frac{\lambda_r^s}{\lambda_r}\right)^2,$$

which reduces to the standard expression $H^s = \sum_{r=1}^{R}(\lambda_r^s)^2$ for the absolute index, where $\lambda_r = 1/R$.

Note that both the Isard and the Herfindhal indices have an upper bound of 1 (when all firms belonging to sector s are located within the same region), while the lower bound is the inverse of the smallest region's share for the former, and the inverse of the number of regions for the latter. As the range of values of these indices depends on the spatial scale and the way regions are defined, they clearly violate the first four properties stated in section 10.1.

3. The idea of *entropy* is borrowed from physics, where it is used as a measure of disorder. It was subsequently used in economics as a measure of concentration/dispersion, and is closely related to the logit and CES models (Anderson et al. 1992, chapter 3).

The entropy indices are defined by

$$E^s(\alpha) = \frac{1}{\alpha^2 - \alpha}\left[\sum_{r=1}^{R} \lambda_r \left(\frac{\lambda_r^s}{\lambda_r}\right)^{\alpha} - 1\right], \qquad (10.2)$$

where α is a parameter which, when less than (respectively, greater than) 1, assigns more weight to observations corresponding to the lower (respectively, higher) tail of the distribution.

The most common version corresponds to the value $\alpha = 1$. By using l'Hôpital's rule, we obtain the following expressions:

$$\lim_{\alpha \to 1} E^s(\alpha) = \sum_{r=1}^{R} \lambda_r \lim_{\alpha \to 1} \frac{(\lambda_r^s/\lambda_r)^{\alpha} - 1}{\alpha^2 - \alpha} = \sum_{r=1}^{R} \lambda_r \lim_{\alpha \to 1} \frac{(\lambda_r^s/\lambda_r)^{\alpha} \ln(\lambda_r^s/\lambda_r)}{2\alpha - 1},$$

which yield the *Theil index*,

$$E^s(1) \equiv T^s = \sum_{r=1}^{R} \lambda_r^s \ln \frac{\lambda_r^s}{\lambda_r}. \qquad (10.3)$$

When $\alpha = 2$, we get

$$E^s(2) \equiv C^s = \frac{1}{2}\left[\sum_{r=1}^{R} \lambda_r \left(\frac{\lambda_r^s}{\lambda_r}\right)^2 - 1\right],$$

[2] In Bailey and Gatrell (1995), the reader can find a discussion of *spatial autocorrelation indices*, which are based on the pioneering work of Moran (1950). They have the important advantage of taking into account the relative position of the areas, i.e., these indices are no longer invariant to permutations of locations. However, an important caveat is that autocorrelation indices do not measure spatial concentration in the same way it has been understood so far. Such indices are more akin to an agglomeration index, as they evaluate the correlation between the value of an economic variable for a given area, and the distance-decay sum of the values of this variable for all the other areas. Unfortunately, this type of index shares the same limitations as all of the other indices presented in this section.

which is equivalent to

$$C^s = \frac{R}{2}\left(H^s - \frac{1}{R}\right).$$

Hence, C^s is equal to the difference between the Herfindhal index and its lowest value. Note that C^s also corresponds to the square of a coefficient of variation and varies from 0 to $\frac{1}{2}(R-1)$.

The most appealing property of entropy measures lies in their separability. For example, we can decompose the degree of concentration of European regions into a degree of concentration *between* countries and a degree of concentration across regions *within* each country. This property is especially intuitive when $\alpha = 2$ because the total variance of a variable with two indices (countries c and regions r) can be decomposed into a "between" and a "within" variance. More generally, for all values of α, we can obtain the following expression:

$$E^s(\alpha) = E^s_{\mathrm{b}}(\alpha) + E^s_{\mathrm{w}}(\alpha),$$

where $E^s_{\mathrm{b}}(\alpha)$ is the level of entropy between countries (disregarding the regional dimension) and $E^s_{\mathrm{w}}(\alpha)$ is a weighted average of the regional entropies within each country. Hence, the ratio $E^s_{\mathrm{b}}(\alpha)/E^s(\alpha)$ may be interpreted as the share of total inequality due to international inequalities, while $E^s_{\mathrm{w}}(\alpha)/E^s(\alpha)$ denotes the share due to interregional inequalities within countries.

Unfortunately, Bourguignon (1979) has shown that, except for $\alpha = 1$, the weights used in the within-entropy depend on the between-entropy, thus weakening the appeal of the separability property. As a result, this decomposition is almost exclusively used for the Theil index ($\alpha = 1$). In this case, the *between component*

$$T^s_{\mathrm{b}} = \sum_{c=1}^{C} \Lambda^s_c \ln \frac{\Lambda^s_c}{\Lambda_c}$$

corresponds to the Theil index (10.3) computed over all countries, the Λs being defined by country exactly as the λs were defined by region. More precisely, the Λs represent country c's share of sectoral and total employment respectively:

$$\Lambda^s_c = \frac{X^s_c}{x^s} \quad \text{with} \quad X^s_c = \sum_{r\in c} x^s_r \quad \text{and} \quad \Lambda_c = \frac{\sum_{s=1}^{S} X^s_c}{x}.$$

As for the *within component*, it is given by the mean of the national Theil indices, weighted by the share of each country in the total employment in sector s:

$$T^s_{\mathrm{w}} = \sum_{c=1}^{C} \frac{X^s_c}{x^s} T^s_c,$$

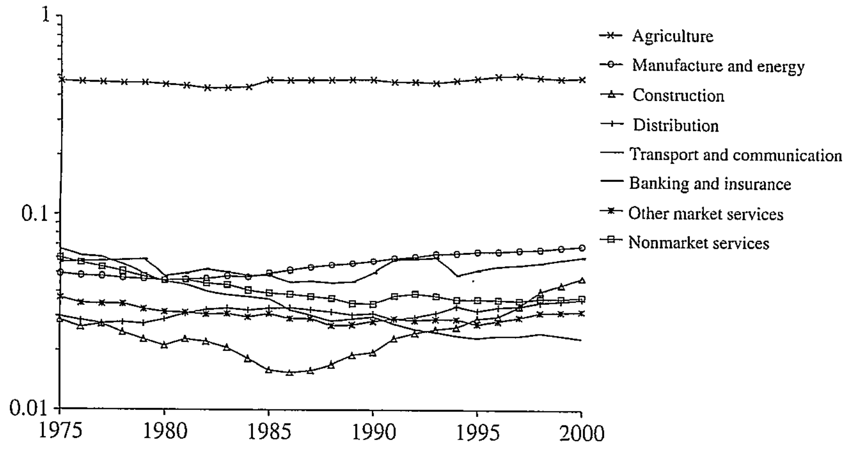

Figure 10.1. Theil indices for eight industries, EU-17, 1975–2000. (Source: Brülhart and Traeger (2005).)

where T^s_c is the Theil index of country c, which is computed only over the regions belonging to this country (see (10.3)) and is given by

$$T^s_c = \sum_{r\in c} \frac{\lambda^s_r}{\Lambda^s_c} \ln \frac{(\lambda^s_r/\Lambda^s_c)}{(\lambda_r/\Lambda_c)}.$$

The Theil index suffers from the same weaknesses as those mentioned above. Still, it is evaluated with respect to a clear benchmark, and, equally importantly, significance tests based on bootstrap methods, have been proposed by Brülhart and Traeger (2005). These authors compute the Theil indices and apply their statistical significance test to eight industries in 236 European regions (NUTS2 or NUTS3) in seventeen Western European countries (EU-15 plus Norway and Switzerland). Figure 10.1 illustrates their results and reveals that agriculture is by far the most spatially concentrated sector with respect to total employment. Moreover, their analysis suggests that the concentration of industry (including energy) has increased regularly since the mid-1980s, whereas the transportation and communications industries have been characterized by dispersion during the last twenty-five years.

It is worth noting the unique nature of the construction industry, whose initial dispersion was later reversed to concentration. As discussed in the following section, however, intersectoral or intertemporal comparisons based on this index may be biased on account of differences across sectors in their degree of industrial concentration.

In order to illustrate the merits of the separability property in the case of the Theil index, figure 10.2 presents variations from 1982 to 1996 of

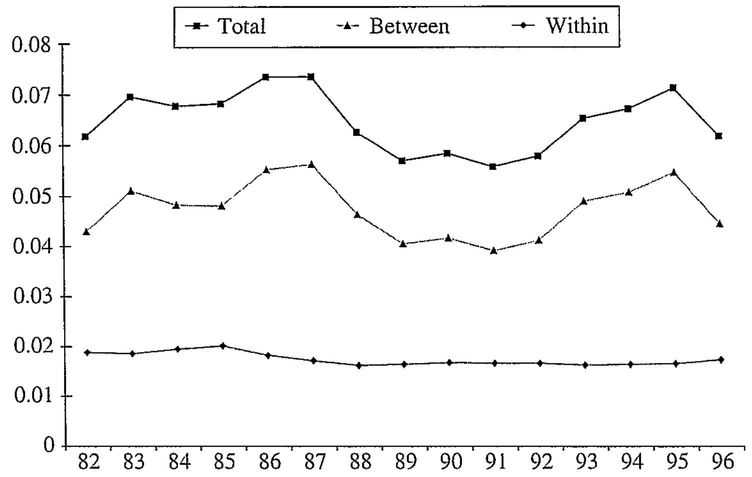

Figure 10.2. Decomposition of the Theil index for total employment, EU-12, 1982–96. (Source: Combes and Overman (2004).)

the overall concentration of total employment across European regions between and within countries.

This shows that short-run variations in the regional distribution of activity in Europe is due primarily to between-country variations, the within-country concentration remaining very stable over time. In the long run, overall spatial concentration varies little. At this stage, it is hard to say whether the above-mentioned differences really spring from changes in plant locations or from changes in the market structure (given by the number and size of firms) of the industries under consideration. We now go on to ponder such questions.

10.3 Indices Accounting for Industrial Concentration

Ellison and Glaeser (1997) radically depart from standard measures of spatial concentration by explicitly controlling for industrial concentration. To illustrate their point, they provide the following example: in the United States, 75% of employment in the vacuum-cleaner industry is covered by a mere four plants. Thus, necessarily, at most four regions account for three-quarters of the employment in this industry, which suggests a strong spatial concentration, as defined in the previous section. However, this strong spatial concentration is obviously tied to the fact that employment itself is concentrated in a very small number of plants. Conversely, a sector in which employment is spread across a large number of plants is more likely to be present in a large number of

regions. The novel feature of Ellison and Glaeser's approach arises from recognizing that the existence of a limited number of plants confines employment to a small number of regions, which in turn influences the sector's spatial concentration. This is to be contrasted with the indices discussed above, which treat each employee as if she were to choose her location independently of the choices made by others. Specifically, Ellison and Glaeser compare the degree of the spatial concentration of *employment* in a given sector with the one that would arise if all *plants* in this sector were located randomly across locations. Note that, like the absolute and relative indices of concentration described above, there are two ways to define the weights associated with locations in those random choices. Locations may be given the same weight (in which case the probability that a given plant is located in each of them is the same, which corresponds to the uniform distribution) or different weights (matching, for instance, the total employment or the population share of the location).

Instead of following Ellison and Glaeser, we consider a slightly modified and more intuitive approach due to Maurel and Sédillot (1999). They use the correlation between the location choices of two plants i and j belonging to the same sector as an index of spatial concentration:

$$\gamma^s = \operatorname{corr}(u^s_{ir}, u^s_{jr}),$$

where $u^s_{ir} = 1$ if plant i in sector s is located in region r, and $u^s_{ir} = 0$ otherwise. If $\gamma^s = 0$, location choices are independent, which corresponds to a random distribution of plants across space. If $\gamma^s = 1$, all plants in this sector are located together. If the distribution of economic activity is considered as the benchmark, the probability that a given plant in sector s chooses to be located in region r is given by the relative size of this region with respect to the overall level of economic activity. This amounts to assuming that the u^s_{ir} are nonindependent Bernoulli variables such that $P(u^s_{ir} = 1) = \lambda_r$.

It is fairly straightforward to derive an estimator of the concentration index γ^s from the observed distribution of plants. To this end, we begin by noting that the probability that two plants are located in the same region r is given by

$$P^s_r = E(u^s_{ir}u^s_{jr}) = \operatorname{cov}(u^s_{ir}u^s_{jr}) + E(u^s_{ir})E(u^s_{jr}) = \gamma^s\lambda_r(1 - \lambda_r) + \lambda^2_r,$$

while the probability that two plants are located within the same region is

$$P^s = \sum_{r=1}^{R} P^s_r = \gamma^s\Big(1 - \sum_{r=1}^{R} \lambda^2_r\Big) + \sum_{r=1}^{R} \lambda^2_r.$$

If an estimator $\boldsymbol{P}^s$ of this probability is available, we may derive an estimator for y^s as follows:

$$\hat{y}^s = \frac{\hat{\boldsymbol{P}}^s - \sum_{r=1}^R \lambda_r^2}{1 - \sum_{r=1}^R \lambda_r^2}. \tag{10.4}$$

There are many possible estimators of $\boldsymbol{P}_r^s$ and, therefore, of $\boldsymbol{P}^s$. For $\boldsymbol{P}_r^s$, we could divide the number of plants located in region r by the total number of plants in sector s. Maurel and Sédillot choose to take into account the fact that plants have different sizes and assign a larger weight to the larger plants. Specifically, they use

$$\hat{\boldsymbol{P}}_r^s = \frac{\sum_{i\in r, j\in r, i\neq j} z_i^s z_j^s}{\sum_{i,j,i\neq j} z_i^s z_j^s},$$

where z_i^s is the share of plant i in total employment in sector s.

Clearly, we have

$$(\lambda_r^s)^2 = \Big(\sum_{i\in r} z_i^s\Big)^2 = \sum_{i\in r, j\in r, i\neq j} z_i^s z_j^s + \sum_{i\in r} (z_i^s)^2.$$

Similarly, by summing across regions, we obtain

$$\begin{aligned} 1 &= \Big(\sum_r \lambda_r^s\Big)^2 = \Big(\sum_r \sum_{i\in r} z_i^s\Big)^2 = \Big(\sum_i z_i^s\Big)^2 \\ &= \sum_{i,j,i\neq j} z_i^s z_j^s + \sum_i (z_i^s)^2 = \sum_{i,j,i\neq j} z_i^s z_j^s + H^s, \end{aligned}$$

where $H^s = \sum_i (z_i^s)^2$ is the Herfindhal index of sector s, which measures the degree of industrial concentration in this sector, disregarding any spatial considerations. Combing these expressions yields

$$\hat{\boldsymbol{P}}^s = \sum_{r=1}^R \hat{\boldsymbol{P}}_r^s = \frac{\sum_{r=1}^R (\sum_{i\in r, j\in r, i\neq j} z_i^s z_j^s)}{\sum_{i,j,i\neq j} z_i^s z_j^s} = \frac{\sum_{r=1}^R (\lambda_r^s)^2 - H^s}{1 - H^s}.$$

By plugging this value into (10.4), we obtain the spatial concentration index of Maurel and Sédillot (1999), denoted by $\hat{y}_{\text{MS}}^s$:

$$\hat{y}_{\text{MS}} = \frac{G_{\text{MS}}^s - H^s}{1 - H^s},$$

where

$$G_{\text{MS}}^s = \frac{\sum_{r=1}^R [(\lambda_r^s)^2 - \lambda_r^2]}{1 - \sum_{r=1}^R \lambda_r^2}$$

is a gross concentration index akin to those discussed in section 10.2. The main difference in using this spatial concentration index over those considered in section 10.2 is that it depends on the industrial concentration index H^s, not just on λ_r^s and λ_r. Instead of ignoring the fact that the distribution of employment in a given sector is conditioned by the way workers are grouped within firms, this fact is now explicitly taken into account. This new index better satisfies property 10.1, which requires comparability across industries, in that what is maybe the most crucial difference across industries, namely their industrial concentration, is explicitly taken into account. Unfortunately, most of the other properties are still violated, apart from property 10.5.

Although the literature reflects the lack of inclination to develop significance tests, any such efforts, based on bootstrap methods for instance, should lead to satisfying property 10.6. It is worth noting that these indices require very detailed (or fine) data, such as plants' sizes. If such data are not available, using the number of plants per industry on a national level, n_s, allows one to make a preliminary correction by assuming that all firms have the same size, which yields $H_s = 1/n_s$.

To conclude, it is worth comparing the index proposed by Ellison and Glaeser (1997) with that from Maurel and Sédillot (1999). The former is based on an Isard-type measure of gross spatial concentration (G_{EG}):

$$G_{\text{EG}}^s = \frac{\sum_{r=1}^R (\lambda_r^s - \lambda_r)^2}{1 - \sum_{r=1}^R \lambda_r^2}$$

from which we similarly obtain

$$\hat{y}_{\text{EG}} = \frac{G_{\text{EG}}^s - H^s}{1 - H^s}.$$

It can be shown that $\hat{y}_{\text{EG}}$ is also an unbiased estimator of y^s. Note the difference with the index of gross concentration used by Maurel and Sédillot (1999), which has the benefit of being derived directly from a probabilistic model of plants' location choices.

Table 10.1 provides a ranking of all the sectors in the two-digit classification for the United States and France, using Ellison and Glaeser's index. The similarities between the two rankings are striking. The two most spatially concentrated sectors are the same in both countries, while three of the four least concentrated sectors in the United States belong to the set formed by the four least concentrated sectors in France. Moreover, the sectoral rank correlation between the two countries is high (0.6).

10.4 The Duranton–Overman Continuous Approach

The additional contribution of the indices presented in the previous section, when compared with those drawn from other fields, is that

Table 10.1. Spatial concentration industries in the United States and France. (Source: Maurel and Sédillot (1999).)

Two-digit industries (U.S. definition)	U.S.A.		France	
	γ	Rank	γ	Rank
Textile mill products	0.127	1	0.036	2
Leather and leather products	0.029	2	0.039	1
Furniture and fixtures	0.019	3	0.008	10
Lumber and wood products	0.018	4	0.012	8
Primary metal industries	0.018	5	0.010	9
Instruments and related products	0.018	6	0.018	5
Transportation equipment	0.016	7	0.000	17
Apparel and other textile products	0.016	8	0.020	4
Miscellaneous manufacturing industries	0.012	9	0.014	6
Chemicals and allied products	0.009	10	0.012	7
Paper and allied products	0.006	11	0.007	11
Electronic and other electrical equipment	0.005	12	0.004	13
Printing and publishing	0.005	13	0.032	3
Fabricated metal products	0.005	14	0.003	14
Rubber and miscellaneous plastics	0.004	15	0.006	12
Stone, clay, and glass products	0.004	16	0.003	15
Industrial machinery and equipment	0.003	17	0.002	16

they explicitly take into account industry-level differences in market structure. These indices are not robust, however, in the way in which geographical areas are defined, and they do not take into account the relative positions of the areas, nor the distances separating them. The latter drawback is a crucial limitation: jobs can be permuted across areas and yet the indices still yield the same values. Moreover, developing methods that account for the distance between areas seems warranted if properties 10.2 and 10.3 are to be satisfied.

Building on earlier works developed by geographers (Bailey and Gatrell 1995), Duranton and Overman (2005) go much further by discarding any geographical classification and by basing their approach on the actual distances separating plants. As a result, properties 10.2 and 10.3 are both satisfied. They can even refine their conclusions by specifying the spatial scale on which the concentration is strongest. This calls for very precise data that give an accurate measure of the distance between plants. Duranton and Overman have access to plants' locations in the United Kingdom on the basis of their postal codes, which gives a precision level on the order of 100 m. With their data set, they are also able to work on a very detailed level, i.e., 234 sectors. This enables Duranton and Overman to draw maps such as those in figure 10.3, in which four industries are considered and where each point stands for a plant having more than ten employees.

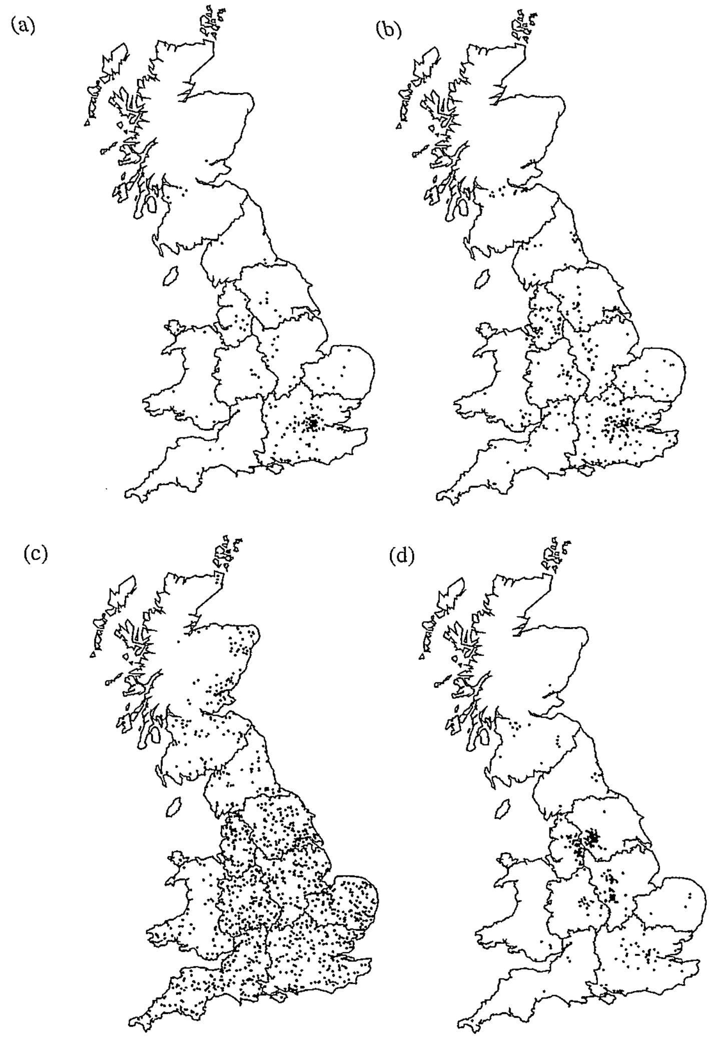

Figure 10.3. Maps of the distribution of plants in four industries in the United Kingdom: (a) basic pharmaceuticals; (b) pharmaceutical preparations; (c) other agricultural and forestry; (d) machinery for textile, apparel, and leather production. (Source: Duranton and Overman (2005).)

The starting point for Duranton and Overman's approach is to count the number of plants that are separated by a given distance. By plotting

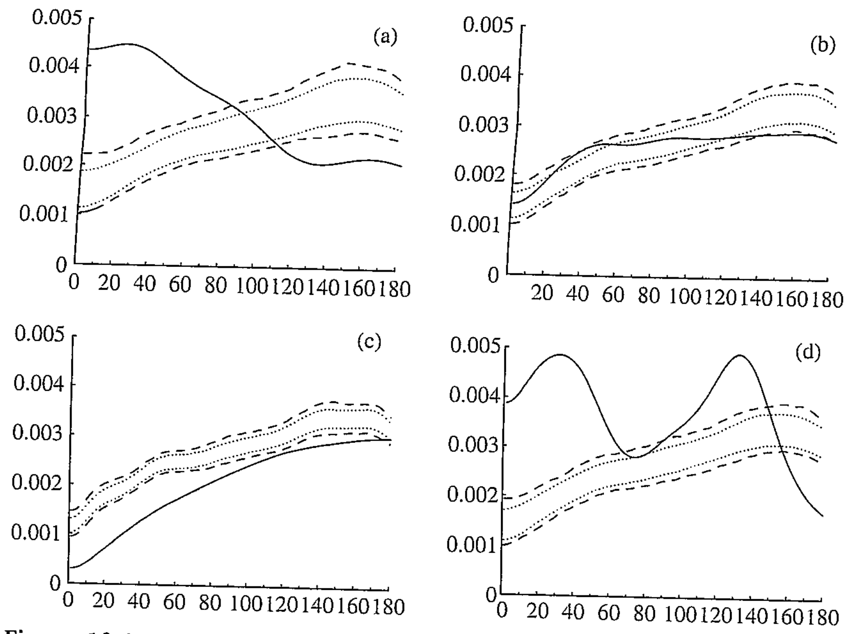

Figure 10.4. Densities and confidence intervals for four industries in the United Kingdom: (a) basic pharmaceuticals; (b) pharmaceutical preparations; (c) other agricultural and forestry; (d) machinery for textile, apparel, and leather production. (Source: Duranton and Overman (2005).)

this number against distance, one obtains a simple frequency graph of the distance between plants. The highest peaks correspond to the distance that most often separates any two plants. Let us assume that two peaks are present for a given sector, one at 30 km and the other at 110 km. This means that the spatial distribution of plants is characterized by a cluster of plants separated on average by 30 km, and that the mean distance between clusters is on average 110 km.

However, defining the spatial concentration as the number of plants separated by a *given* distance is unsatisfactory for the following two reasons. First, such an approach arbitrarily carves up distance into discrete intervals depending on which unit of length is used. Second, the measure of distance per se is debatable. For example, is measuring distance as the crow flies the most appropriate method? These two limitations imply that the distance between two plants is subject to measurement errors. With this in mind, it is preferable to smooth out the distribution of distances between plants. Intuitively, we can estimate the number of plants separated by a given distance d, say, 30 km, by taking the number of plants separated by a distance varying from 28 to 32 km and

dividing this by four (corresponding to the distance between these two bounds if the unit of length is in kilometers), instead of the single number recorded in the database for 30 km. However, this type of smoothing remains incomplete as it still ignores the presence of plants separated by 27 or 33 km, and so on. Furthermore, it assigns the same weight to all points, regardless of the distance separating them from the reference distance (30 km in our example). Yet it seems reasonable to give extreme observations less weight and give more weight to, say, those at 29 or 31 km. Satisfying these two properties would require use of a kernel method, which is precisely what Duranton and Overman use. This method applies a weighting scheme that follows the normal distribution to all points lying a given distance from the reference point. Once this density is estimated, curves similar to those presented in figure 10.4 are obtained. In the case of textiles and apparel, the figure illustrates the two peaks discussed above.

The following question has yet to be answered: for every given distance, to what extent does the number of plants observed after smoothing *significantly* differ from the number obtained if their location were chosen randomly, or according to any benchmark distribution (property 10.6)? In contrast to previous indices, Duranton and Overman's approach allows them to address this question. Given the existing number of plants, they randomly assign each of them to one of any possible locations, and then calculate the number of plants separated by any distance. This operation is repeated, say, 1,000 times, which leads to a set of 1,000 values for each distance. They finally construct two-sided confidence intervals containing 90% of these values, i.e., with the upper and lower bounds given by the 95% and 5% percentiles of the generated values, respectively. This procedure generates two smooth curves, as illustrated by dotted lines in figure 10.4. If the number of plants observed after the smoothing procedure exceeds the upper bound of the confidence interval, the sector is said to be *locally concentrated* at the distance under consideration with a confidence level of 95% (as we ignore the 5% of observations situated above the upper limit of this interval). If the number of plants is smaller than the lower limit, the sector is said to be *locally dispersed* at the distance under consideration.

In addition to dealing with local concentration and dispersion, Duranton and Overman also work in a global way by defining the upper limit of the confidence interval in such a way that 95% of the whole set of draws (at any distance) lie below this upper bound. In this case, a sector is said to be *globally concentrated* if its density exceeds this limit at least once after smoothing out. They proceed analogously for the lower bound.

The corresponding curves are illustrated by broken lines in figure 10.4. Note that the basic pharmaceutical industry is locally concentrated for every distance below 80 km and globally concentrated, even if it is locally dispersed for distances exceeding 110 km.

Finally, it is worth noting that the continuous approach, by working with the actual number of plants when estimating confidence intervals, automatically takes into account differences in industrial concentration (property 10.1). By starting at a microscopic level, spatial classification is circumvented, which allows properties 10.2 and 10.3 to be satisfied. Moreover, this approach distinguishes itself from existing indices in that it also captures the firms' relative positions in space. By contrast, using a predefined industrial classification implies that property 10.4 is violated. Consequently, the Duranton–Overman approach may be sensitive to how sectors are defined, and should be extended to account for the technological distance between industries.

10.5 Concluding Remarks

The ideal index of spatial concentration still seems far from reach. Nevertheless, the growing availability of data on a very fine spatial scale has led to finer and more accurate methods than the standard indices of spatial concentration. These new indices account for a number of specificities that characterize spatial data. This type of approach allows us to satisfy several of the properties of an ideal index and, therefore, to better assess its variations over time or across industries. Collecting the required data can often prove very costly, however.

10.6 Related Literature

Marcon and Puech (2003) use an approach that is similar to that of Duranton and Overman (2005). The former were inspired by methods developed in forestry to study the spatial distribution of tree species. The latest version of their index (Marcon and Puech 2005) makes it possible to account for differences in industrial concentration between industries, and it can be modified to obtain a measure of co-location. In the spirit of Ellison and Glaeser's proposal, they can measure the likelihood that a given industry will locate in the same place as another industry. Barrios et al. (forthcoming) used Ellison and Glaeser's index to compare the spatial structure of industry in three small countries: Belgium, the Republic of Ireland, and Portugal. Feser et al. (2005) follow a different approach that rests on local spatial autocorrelation. Finally, we should

mention Mori et al. (2005), who propose an alternative method based on aggregated data but which allows for a number of different tests to be run, such as the test of significant differences between an observed distribution and its reference, or between different industries, as required by property 10.6.