JOURNAL OF REGIONAL SCIENCE, VOL. 47, NO. 4, 2007, pp. 753–774

MEASURING THE LOCALIZATION OF ECONOMIC ACTIVITY: A PARAMETRIC APPROACH*

Paulo Guimarães
University of South Carolina, Columbia, SC 29208. E-mail: guimaraes@moore.sc.edu

Octávio Figueiredo
Universidade do Porto and CEMPRE, 4200-464 Porto, Portugal.
E-mail: octavio@fep.up.pt

Douglas Woodward
University of South Carolina, Columbia, SC 29208. E-mail: woodward@moore.sc.edu

ABSTRACT. The index proposed by Ellison and Glaeser (1997) is now well established as the preferred method for measuring the localization of economic activity. In this paper we develop an alternative localization measure that is consistent with the theoretical framework originally proposed by Ellison and Glaeser. Our measure follows directly from the Random Utility (Profit) Maximization (RUM) location decision model. Because the distributional assumptions in our model are fully compatible with RUM, we are able to offer a more efficient measure of industry clustering.

1. INTRODUCTION

Localization is widely recognized as a source of increasing returns for geographically concentrated firms in particular industries. For more than a century, scholars have examined why and how these localization economies—internal to the local industry, but external to the firm—explain the spatial concentration of economic activity. Casual empiricism suggests that there is a marked tendency for industries to localize, i.e., to concentrate over and above overall economic activity. Alfred Marshall's classic examples included cutlery (Sheffield) and jewelry (Birmingham) in 19th century England. Contemporary examples abound, from the automotive industry in Michigan and semiconductors in California, to the footwear cluster of northern Italy and telecommunications in Finland. Yet, how general and strong is the tendency of an industry to agglomerate in local areas? Can we measure the centripetal force of localization economies?

The debate reignited by new theories of economic growth, geography, and international trade (with their emphasis on the industry-specific external

*This research was supported with a grant from FCT (POCTI/ECO/47624/2002).
Received: July 2005; revised: August 2006; accepted: January 2007.

Blackwell Publishing, Inc., 350 Main Street, Malden, MA 02148, USA and 9600 Garsington Road, Oxford, OX4 2DQ, UK.

economies) has again brought these questions to the fore of regional science. To date, Ellison and Glaeser (1997) represents the most rigorous attempt to tackle the problem of measuring localization. Based on a Random Utility Maximization (RUM) model of location, the authors proposed an index that accounts for the amount of spatial concentration of an industry that can be related to that industry's specific spillovers (along with natural advantages of the regions). Subsequently, their work spawned a significant number of studies and rapidly emerged as the standard approach to measuring the localization of economic activity.

In this paper, we review the Ellison and Glaeser (EG) index and contend that the link between the RUM framework and this index can be strengthened. Like Ellison and Glaeser, we use the model of industrial location based on McFadden's RUM framework, but propose an alternative measure that closely ties to the theory behind their original work. Ours is a more efficient measure of localization because it relies on a set of specific distributional assumptions that are fully compatible with the RUM framework.

The rest of the paper consists of four sections. The next section reviews received measures of spatial concentration and attendant problems. In Section 3, we take a more in-depth look at the EG index and develop our alternative method for measuring localization. Section 4 provides an illustration using data for Portuguese industries. Section 5 offers a summary and possible extensions of this paper.

2. MEASURING SPATIAL CONCENTRATION

Researchers have no shortage of tools for measuring the geographical concentration of economic activity. Most prominent are Hoover's (1937) location coefficient and the Gini coefficient, for example, as applied by Krugman (1991). These measures quantify the discrepancy between the distribution of regional employment in a particular industry against the regional distribution of overall employment. But, are these measures able to capture the concept of localization? An obvious problem is that they are sensitive to the levels of concentration within the industry. Take as an example two industries that have identical measures for the Gini index. The first industry is composed of many independent firms, all equally sized and located in a single region, while the second industry is composed of just one firm operating a large establishment. The first case agrees with the notion of spatial external economies, which explains the clustering of all firms in that industry. These external economies—knowledge spillovers, labor market pooling, and input sharing—are well known to regional scientists (for a good review see Kim, Barkley, and Henry, 2000). But for the second industry, with one large establishment, it is obvious that external economies are not a valid explanation to justify the observed spatial concentration. In this case, geographic concentration is entirely explained by internal economies of scale.

Another problem is that these measures do not account for inherent randomness in the underlying location decisions. Firms may exhibit some level of spatial concentration by chance. This idea can be explained by appealing to the balls and urns example often used in statistics. If one has, say, 10 urns (regions) and 10 balls (firms) and drops the balls at random into these urns, then even though all urns are equally probable, it is very unlikely that we will observe exactly one ball in each urn. Some clustering will necessarily occur and that is perfectly compatible with the idea that the balls were thrown at random (the firms' decisions were random). The original indices are not able to control for this type of clustering.

It should be apparent, then, that these indices do not accurately measure an industry's degree of localization. The index proposed in Ellison and Glaeser (1997) overcomes these limitations. Like the Gini coefficient, the EG index attempts to measure the tendency of one industry to agglomerate in relation to the general tendency of all industries to agglomerate. Unlike its predecessors, however, it accounts for the inherent discreteness (lumpiness) that will be observed if location decisions are driven by chance alone. Notably, the EG index is rooted in the location choice model of Carlton (1983), which in turn is based on McFadden's random utility (profit) framework—the workhorse for empirical research on industrial location (e.g., Bartik, 1985; Coughlin, Terza, and Arromdee, 1991; Friedman, Gerlowski, and Silberman, 1992; Head, Ries, and Swenson, 1995; Guimarães, Figueiredo, and Woodward, 2000, and Guimarães, Figueiredo, and Woodward, 2004).

Following in the spirit of the EG index, other researchers have measured and compared levels of industry localization. Based on a model that emphasizes spillovers as the force leading to localization of industries, Maurel and Sedillot (1999) constructed a measure that is similar to the EG index. By comparing the two formulas, they show that the difference between the indices has an expected value of zero. Also noteworthy is the work of Devereux, Griffith, and Simpson (2004). These authors showed that the EG index can be conveniently approximated by the difference between indices of spatial employment concentration and industrial concentration. In turn, Duranton and Overman (2005) have proposed a different way to measure spatial concentration. Their approach draws directly on advanced methods of spatial statistics. They treat space as continuous and compute their measurements based on the Cartesian distances between each pair of plants. Treating space as continuous has an inherent appeal, but their approach lacks a theoretical underpinning (such as RUM) and requires precise information (often unavailable) on the exact location of each business unit.

The new wave of literature prompted by Ellison and Glaeser (1997) has already generated a substantial amount of applied work. Beyond the ongoing research in the United States (Ellison and Glaeser, 1997; Dumais, Ellison, and Glaeser, 2002; Holmes and Stevens, 2002), recent studies characterizing industry localization can be found for France (Maurel and Sedillot, 1999 and Houdebine, 1999), Belgium (Bertinelli and Decrop, 2002), U.K. (Devereux et al., 2004), and Spain (Callejón, 1997). Common to all studies is the finding that the majority of industries are localized.

In the next section, we put forward an alternative localization index that has a number of distinct advantages. Essentially, we develop a measure of industry clustering, accounting for what Kim et al. (2000) called establishment concentration. In contrast, Ellison and Glaeser and much of the applied work has focused on employment concentration. Nevertheless, our index falls closely in line with the definition of localization put forth by Ellison and Glaeser. In their original paper, Ellison and Glaeser (1997, p. 1) referred to localization as "levels of concentration beyond those which would be observed if firms had chosen the locations of their plants in a completely random manner." Our index strengthens the link between location choice theory and the localization index. Thus, our establishment concentration index is rooted in microeconomic theory. It also has compelling statistical advantages over previous work, as demonstrated later in the paper.

3. LOCALIZATION INDICES

Location models based on the RUM framework provide an explanation for the spatial distribution of an industry. Idiosyncratic factors aside, firms choose locations that yield the highest profits. Abstracting from dynamic considerations, we can use the RUM theoretical framework to justify the observed geographic concentration of industries at one particular point in time. Ellison and Glaeser (1997) used this approach. Yet, as we shall see, the RUM framework can be used to derive an index that yields more precise estimates of the extent to which industries are localized.

The EG Index

To motivate our approach, let us first consider the derivation of the EG index. Assume at the outset that the economy is divided into J geographical units (regions). Also, we take as our reference a given industry which has n_j plants located in each region j. Thus, $n = \sum_{j=1}^{J} n_j$ represents the total number of existing plants in our reference industry. Next, we briefly sketch how the EG index is obtained taking as a reference their model of "natural advantages." If firm i chooses to locate in region j then its profits will consist of

$$\log \pi_{ij} = \log \bar{\pi}_j + \varepsilon_{ij} \tag{1}$$

where $\bar{\pi}_j$ is a nonnegative random variable reflecting the profitability of locating in area j for a typical firm in the industry. In this formulation of the model, Nature introduces the randomness in $\bar{\pi}_j$ by selecting for each region the characteristics that make it unique (their natural advantages). ε_{ij} is a random disturbance. If we assume that ε_{ij} is an identically and independently distributed random term with an Extreme Value Type I distribution[1] then, conditional on a realization of $\bar{\pi}_j$, we can apply McFadden's (1974) result to obtain,

$$p_{j|\bar{\pi}} = \frac{\exp(\log \bar{\pi}_j)}{\sum_{j=1}^{J} \exp(\log \bar{\pi}_j)} = \frac{\bar{\pi}_j}{\sum_{j=1}^{J} \bar{\pi}_j}$$

which denotes the probability of a firm locating in region j. Thus, p_j is obtained from the Random Utility (Profit) Maximization approach of Carlton (1983) that, as mentioned earlier, gives support to the most recent studies of industrial location. To derive their index, Ellison and Glaeser (1997) introduced two parametric restrictions regarding the expected value and variance of p_j. They assume that the distribution of $\bar{\pi}_j$ is such that

$$E(p_j) = x_j$$

and

$$V(p_j) = \gamma x_j(1 - x_j) \tag{2}$$

where x_j may be thought of as the probability of a firm locating in region j in the absence of any region-specific advantages for that industry. Thus, the larger the discrepancy between x_j and p_j, the larger the influence that these region specific effects (say, natural advantages) play in the location decisions of firms in that industry. That difference is captured by the parameter γ (which we will refer to as the EG parameter) which belongs to the unit interval. It is easy to see that if $\gamma = 0$ then the location probabilities are fixed at x_j and we can conclude that there is no spatial concentration in excess of what we would expect to occur (what Ellison and Glaeser called the dartboard model). If, however, $\gamma > 0$, then the actual location probabilities of the industry will differ from x_j and in the limit, when $\gamma = 1$, each p_j has the largest variance and becomes a Bernoulli random variable. Hence, in the limit, all the investments for that industry would be located in a single region.

Ellison and Glaeser (1997) also showed that the γ parameter may be derived from an alternative model that emphasizes industrial spillovers as the force leading to "excessive concentration." In any case, the theoretical motivation one uses is irrelevant because the two models are observationally equivalent and lead to the same functional form for the index, the practical implication being that we can not readily distinguish the two sources of "excessive" geographic concentration (natural advantages and industrial spillovers).

To estimate γ for a particular industry they let x_j denote area j's share of total manufacturing employment. Here, the idea is that the model should on an average reproduce the overall distribution of manufacturing activity. In a next step they considered the following "raw concentration index" of employment:

[1] In the past, this distribution has been referred to by other names such as Weibull, Gumbel, and double-exponential (Louviere, Hensher, and Swait, 2000).

$$G_E = \sum_{j=1}^{J}(s_j - x_j)^2$$

where, s_j denotes area j's share of employment in that industry and the x_j's are as described above. Now, taking the expected value of G_E they obtain a function of γ and the authors use that relation to propose an estimator for γ. Their proposed estimator for γ (the EG index) is then

$$\hat{\gamma}_{EG} = \frac{G_E - \left(1 - \sum_{j=1}^{J} x_j^2\right) H_E}{\left(1 - \sum_{j=1}^{J} x_j^2\right)(1 - H_E)}$$

where H_E is the employment-Herfindhal index for the industry and the expected value of G_E is replaced by its actual value.

The above index provides an employment-weighted measure that primarily reflects the localization externalities felt by the larger plants and discounts the small ones. This can lead to some peculiar situations. Consider the following example. There are 10 equally probable regions and 10 plants in a given industry. Suppose that nine of those plants have one employee and all locate in the same region (say region A), while the other plant, with 36 employees locates in a different region (say region B). A simple computation of the EG index for this example gives exactly zero. We believe that classifying this industry as nonlocalized is not correct. Indeed, localization economies (along with natural advantages) seem to influence the location decisions of the majority (90 percent) of the plants. It could be argued that the EG index reflects the fact that the large establishment (concentrating 80 percent of employment) is not clustered. But consider a different situation where the regional distribution of employment agrees perfectly with the expected location probabilities and the large firm has a dimension of six employees (that is, the x_j's for region A and B are 0.6 and 0.4, respectively). In this new situation the EG index is exactly -0.25, a value that is normally interpreted as indicating that the industry is nonlocalized. Nevertheless, it is hard to understand why "non-localized" applies to that industry. The small plants are concentrated in excess of what we would expect to occur and they represent 60 percent of the industry employment. These problems may be overcome if one ignores the influence of employment and works directly with establishment counts. In the next section, we derive an unbiased estimator for γ relying solely on counts.

An EG Index Based on Plant Count Data

Defining the "raw concentration index" as

$$G_C = \sum_{j=1}^{J}\left(\frac{n_j}{n} - x_j\right)^2$$

and proceeding in a fashion similar to Ellison and Glaeser (see Appendix A), we derive the following alternative estimator for γ:

$$\hat{\gamma}_C = \frac{nG_C - \left(1 - \sum_{j=1}^{J} x_j^2\right)}{(n-1)\left(1 - \sum_{j=1}^{J} x_j^2\right)}$$

The above expression is very similar to that of the EG index. It replaces the Herfindhal index by $1/n$ and the "raw concentration index," G_E, is replaced by its counterpart expressed in terms of counts of plants, G_C. Like the estimator proposed by Ellison and Glaeser this estimator for γ is also, by construction, unbiased. Most notably, it has a smaller variance. To see this consider the ratio of the variances,

$$\frac{V(\hat{\gamma}_{EG})}{V(\hat{\gamma}_C)} = \left(\frac{n-1}{n(1-H_E)}\right)^2 \frac{V(G_E)}{V(G_C)}$$

If all the plants have the same dimension, then H_E assumes its smallest value of $1/n$. In this case $\hat{\gamma}_{EG} = \hat{\gamma}_C$ and thus $V(\hat{\gamma}_{EG}) = V(\hat{\gamma}_C)$. If, however, $H_E > 1/n$, then $V(\hat{\gamma}_{EG}) > V(\hat{\gamma}_C)$ (for a formal proof see Appendix B).

Again, a simple example will give some intuition on how much employment affects the variance of the $\hat{\gamma}_{EG}$. Suppose that we have three regions and four firms. Admit that ($x_1 = x_2 = x_3 = 1/3$). Table 1 lists the distribution of $\hat{\gamma}_{EG}$ for several employment configurations under the null hypothesis that $\gamma = 0$. For the employment configurations we assume several limiting cases: all firms have identical employment ($H_E = 1/4$); one firm is very large (H_E tends to 1); two firms are very large (H_E tends to 1/2); and there are three very large firms (H_E tends to 1/3). As we can see from Table 1, when compared with the "all equal" case (the case where $\hat{\gamma}_{EG} = \hat{\gamma}_C$) the variance of the $\hat{\gamma}_{EG}$ may be up to six times larger than that of the $\hat{\gamma}_C$. Thus, we argue that, also from a statistical point of view, it is preferable to work directly with counts of plants and ignore the confounding influence of plant size.

An Alternative Method for Measuring Localization

As we have just seen, by restricting attention to counts of plants, it is possible to derive an alternative estimator for the γ parameter that is also unbiased but more efficient. Additional gains in efficiency would be achieved if one identifies a statistical distribution for the counts of plants that is compatible with all the parametric assumptions laid out by Ellison and Glaeser (1997). When discussing the theoretical model of natural advantages, Ellison and Glaeser (1997) resort to a particular multivariate distribution—the Dirichlet—to illustrate their idea that the location probabilities are random. However, in Ellison and Glaeser (1997) that assumption about the location probabilities is devoid of

TABLE 1: Limiting Distributions of $\hat{\gamma}_{EG}$ for Different Employment Configurations

$\hat{\gamma}_{EG}$	All equal	1 large	2 large	3 large
−0.5	0	8/27	18/27	6/27
−0.25	12/27	0	0	0
0	6/27	12/27	0	18/27
0.25	8/27	0	0	0
0.5	0	6/27	0	0
1	1/27	1/27	9/27	3/27
$V(\hat{\gamma}_{EG})$	1/12	1/6	1/2	1/6

context—the authors do not offer any explanation how that distribution could result from the RUM framework or even what the implication of that assumption would be for the overall distribution of counts of plants. In the following we show how this could be achieved. We build upon the RUM framework to derive the final distribution of counts of plants in the context of Ellison and Glaeser's (1997) model of natural advantages. The distribution that we obtain for counts of plants is also compatible with a spillover-based motivation. As we will see, this distribution contains a parameter that satisfies the same theoretical requirements as the EG index—thus providing a natural alternative for measuring an industry's degree of localization.

To show the derivation of our index let us start by rewriting expression (1) as

$$\log \pi_{ij} = \log \tilde{\pi}_j + \eta_j + \varepsilon_{ij}$$

where we let $\tilde{\pi}_j$ be the expected profitability of locating in region j for a typical firm in an industry and η_j is a regional random effect that picks the unobservable advantages of that region for that particular industry. The other random term, ε_{ij}, is as defined earlier. Now, conditional on the η_j's and again drawing on McFadden's (1974) result we can write

$$p_{j|\boldsymbol{\eta}} = \frac{\exp(\log \tilde{\pi}_j + \eta_j)}{\sum_{j=1}^{J} \exp(\log \tilde{\pi}_j + \eta_j)} = \frac{\tilde{\pi}_j \exp(\eta_j)}{\sum_{j=1}^{J} \tilde{\pi}_j \exp(\eta_j)} \tag{3}$$

Assume that the $\exp(\eta_j)s$ are i.i.d. gamma distributed with parameters $(\frac{1-\gamma}{\gamma}x_j, \frac{1-\gamma}{\gamma}x_j)$ and consequently with expected value of 1 and a variance of $\frac{\gamma}{1-\gamma}x_j^{-1}$.[2] In this case, we know from Mosimann (1962) that the $(p_1, p_2, \ldots, p_J)$ are Dirichlet distributed with parameters $(\frac{1-\gamma}{\gamma}x_1, \frac{1-\gamma}{\gamma}x_2, \ldots, \frac{1-\gamma}{\gamma}x_J)$. The expected values and variances for the marginal distribution of the Dirichlet (the Beta distribution) are

[2]As noted in Ellison and Glaeser (1994), the assumption of a gamma distribution generalizes the Chi-squared based examples provided in Ellison and Glaeser (1997).

$$E(p_j) = \frac{\tilde{\pi}_j}{\sum_{j=1}^{J} \tilde{\pi}_j} = x_j$$

and

$$V(p_j) = \gamma x_j(1 - x_j)$$

To obtain an estimator for γ, we make use of the information on counts of plants. Conditional on the regional random effects, the location probabilities, $p_{j|\boldsymbol{\eta}}$, are fixed and the joint probability of observing a particular spatial distribution of plants $(n_1, n_2, \ldots, n_J)$ follows a multinomial distribution:

$$P(n_1, n_2, \ldots, n_J \,|\, n, \boldsymbol{\eta}) = n! \prod_{j=1}^{J} \frac{p_{j/\boldsymbol{\eta}}^{n_j}}{n_j!}$$

We are interested in finding the unconditional distribution of counts of plants, and thus we need to "integrate out" the random effects. Mosimann (1962) showed that the distribution that results from compounding the multinomial distribution with the Dirichlet has a closed form. Following Mosimann, the unconditional multivariate distribution for the counts of plants is given by:[3]

$$(4) \qquad P(n_1, n_2, \ldots, n_J \,|\, n) = \frac{n!\Gamma(\gamma^{-1} - 1)}{\Gamma(\gamma^{-1} + n - 1)} \prod_{j=1}^{J} \frac{\Gamma\left[(\gamma^{-1} - 1)x_j + n_j\right]}{\Gamma\left[(\gamma^{-1} - 1)x_j\right]n_j!}$$

Maximization of the above expression provides a maximum-likelihood estimator for γ. By the properties of maximum-likelihood, we know that this estimator, $\hat{\gamma}_{DM}$, is consistent and asymptotically efficient. Unlike the EG index, which often produces estimates that lack interpretation in the context of their theoretical model, our estimator has the desirable property of generating estimates that always belong to the unit interval.[4] However, our estimator may be biased in small samples.[5] Because under the null hypothesis that

[3]In the statistical literature, this distribution is known as the compound multinomial or Dirichlet-Multinomial distribution (Johnson, Kotz, and Balakrishnan, 1997). Note that in the following expression $\Gamma()$ denotes the gamma function.

[4]The range of estimates of the EG index can be very wide. To see this, consider the following two extreme cases. First, admit that $G_E = 0$. Now, $\hat{\gamma}_{EG} = -H_E(1 - H_E)^{-1}$, and thus the EG estimate is always negative. Next, suppose that all sectorial employment is concentrated in one region, say region 1. Then, $\hat{\gamma}_{EG} = 1 + \frac{2\left(\sum_{j=1}^{J} x_j^2 - x_1\right)}{(1 - \sum_{j=1}^{J} x_j^2)(1 - H_E)}$. If all regions are equiprobable, $\hat{\gamma}_{EG} = 1$, but if the regions are not equiprobable and H_E is close to 1 then $\hat{\gamma}_{EG}$ may produce nonsensical estimates taking very high values that are either negative (if $x_1 > \sum_{j=1}^{J} x_j^2$) or positive.

[5]$\hat{\gamma}_{DM}$ must have an upward bias when the true value is zero. As pointed out by one of the referees, this may be an undesirable property if one is constructing localization indices to be used in the left-hand side of a regression.

$\gamma = 0$, the Dirichlet-Multinomial distribution collapses to a standard multinomial distribution we can use a likelihood ratio test to examine this hypothesis.[6]

Our derivation of the Dirichlet-Multinomial based index (OM index) offers an explanation for localization based on natural advantages. As in Ellison and Glaeser (1997), we could also provide an interpretation motivated by spillovers. To see this, let y_{ij} denote Bernoulli random variables equal to one with probability x_j, indicating whether a potential firm locates in area j. The number of investments in region j is:

$$n_j = \sum_{i=1}^{n} y_{ij}$$

The existence of spillovers implies that the individual investment decisions are correlated. Suppose that all Bernoulli variables are equally correlated and denote the (nonnegative) correlation coefficient between any pair of random variables by γ. Now

$$E(n_j) = nx_j$$

and

$$\begin{aligned} V(n_j) &= \sum_{i=1}^{n} V(y_{ij}) + \sum_{i \neq l}^{n} cov(y_{ij}, y_{lj}) \\ &= nx_j(1 - x_j)[1 + (n-1)\gamma] \end{aligned}$$

Examination of the previous expressions reveals that these are precisely the expected value and variance for the marginal distributions of the Dirichlet-Multinomial distribution in (4).[7] Thus, as in Ellison and Glaeser (1997), interpretation of $\hat{\gamma}_{DM}$ is compatible with a pure natural advantage model, a pure spillover model, or a combination of both.

4. APPLICATION: LOCALIZATION OF PORTUGUESE MANUFACTURING INDUSTRIES

Data

The availability of detailed establishment and employment information by industry and region allowed us to apply our approach to Portugal. Our source of data is the *Quadros do Pessoal* database. The *Quadros do Pessoal* is a yearly survey collected by the Ministry of Employment for all the existing companies

[6]Under the null hypothesis, the value of the log-likelihood function is simply $\log n! + \sum_{j=1}^{J} n_j \log x_j - \log n_j!$. Because we are testing a value which is in the boundary of the set of admissible values for γ_{DM}, we follow the suggestion in Self and Liang (1987) and adjust the level of significance of the chi-squared statistic accordingly.

[7]The marginal distribution of the Dirichlet-Multinomial is the Beta-Binomial distribution (Johnson et al., 1997).

operating in Portugal (except family businesses without wage earning employees).[8] Our sample covers 45,350 plants surveyed in 1999. Using this source, we tallied the number of plants as well as employment for each *concelho* (county) in continental Portugal.[9] We rely on the three-digit (103 industries) classification of the Portuguese Standard Industrial Classification system (CAE).[10] Using the 275 Portuguese *concelhos* as the spatial units of analysis, we estimated for each industry the Dirichlet-Multinomial-based measure of localization given by (4).

To implement our model, we wrote the likelihood function using the statistical software Stata (version 8) and that package's standard numerical maximization routine (a modified Newton-Raphson algorithm). Convergence was fast, with a small number of iterations (less than 10 for most cases). As in Ellison and Glaeser (1997), we let the expected location probabilities for each *concelho* be approximated by its share of total manufacturing employment.

Localization of Portuguese Industries

We computed the localization index $\hat{\gamma}_{DM}$ for each of the three-digit SIC industries at the *concelho* level.[11] For 19 industries, $\hat{\gamma}_{DM}$ was zero or not statistically different from zero at the 95 percent level of confidence. For the remaining 81 industries (81 percent of the 100 industries analyzed) we find evidence of "excess of concentration" ($\gamma > 0$). Therefore, a high percentage of Portuguese manufacturing industries appear to be localized. This result corroborates similar evidence for others countries. Ellison and Glaeser (1997) found that 446 out of 459 four-digit SIC industries in the United States were localized ($\hat{\gamma}_{EG} > 0$). Based on a test of statistical significance, Maurel and Sedillot (1999) found that 77 percent of the 273 four-digit French industries display "excess of concentration." A high percentage of U.K. industries were also found localized by Devereux et al. (2004).

As previously observed for others countries, the localization index displays a very skewed distribution, the majority of industries showing slight levels of localization. This pattern is displayed in Figure 1, where we show a histogram of $\hat{\gamma}_{DM}$ at the *concelho* level for the 100 three-digit SIC industries.

Tables 2 and 3 provide information for individual industries. In Table 2, we list the 22 sectors for which $\hat{\gamma}_{DM}$ is significantly different from zero and above the industry average. Among them, we find a large number of traditional sectors for which localization is associated with the historical specialization

[8]For a thorough description of this database see, for example, Cabral and Mata (2003).

[9]The *concelho* is an administrative region in Portugal. In recent years, some new *concelhos* have been created by the incorporation of parts of existing *concelhos*. To maintain data compatibility, we used the spatial breakdown of 275 *concelhos* that was still valid in 1997. These have an average area of 322.5 square kilometers.

[10]Revision 2 of the CAE.

[11]Our dataset contains information for 100 three-digit SIC industries. For SICs 231, 233, and 300, the *Quadros do Pessoal* data set did not report any plant in 1999.

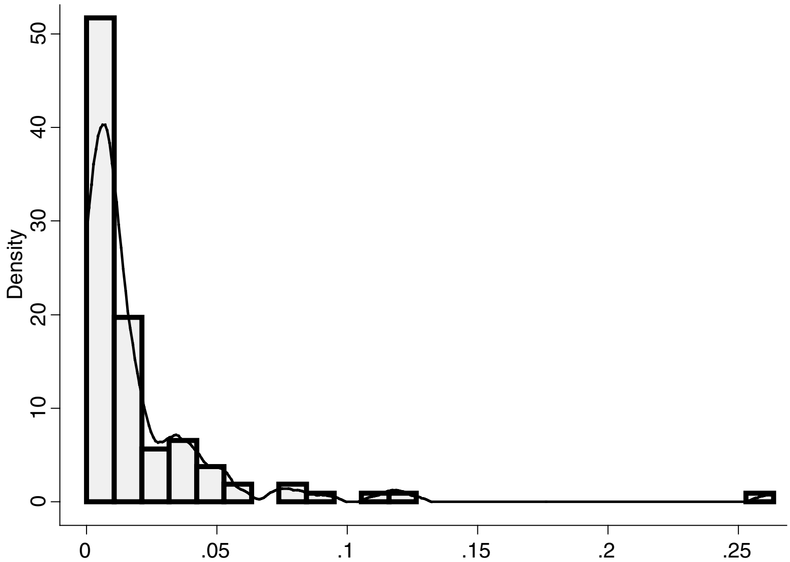

FIGURE 1: Histogram of $\hat{\gamma}_{DM}$ at the "Concelho" Level.

of Portuguese regions (e.g., tannery, jewelry, textiles, and footwear industries). Again, this pattern coincides with evidence for other countries that suggests that typically traditional industries are highly localized.[12] Table 2 also shows that several more technologically advanced industries (such as fabrication of pharmaceutical products, artificial and synthetic fibers, automobiles, and measuring and controlling devices) exhibit higher than average levels of localization. As could be expected, shipbuilding and industries that process sea products are also among the most localized industries. Access to a natural resource, the sea, explains firms' clustering in these last sectors.

Table 3 displays the group of nonlocalized sectors (i.e., those for which $\hat{\gamma}_{DM}$ is zero or non-significant). For this last group it is important to distinguish our measure of localization from a simple measure of geographic concentration. While some of these sectors (such as tobacco and aircraft and space vehicles fabrication) are highly concentrated in space, this concentration is almost entirely explained by industrial concentration, and thus by returns to scale rather than natural advantages or external economies associated with firms' clustering.

Comparison with the EG Index

Using Data on Portuguese Industries. We now use the data on Portuguese manufacturing industries to compare our estimates of localization with those

[12]See Ellison and Glaeser (1997), Table 4, Maurel and Sedillot (1999), Tables 1 and 2, Devereux et al. (2004), Tables 4 and 5, and Krugman (1991), Appendix D.

TABLE 2: Geographic Concentration, by Most Localized Industries According to $\hat{\gamma}_{DM}$

3-digit SIC industry (Portuguese CAE-Rev2)	$\hat{\gamma}_{DM}$	Number of plants	Rank		
			$\hat{\gamma}_{DM}$	$\hat{\gamma}_{C}$	$\hat{\gamma}_{EG}$
363- Musical instruments	0.263	8	1	5	4
354- Motorcycles and bicycles	0.122	45	2	3	1
191- Leather tanning and finishing	0.114	110	3	1	5
232- Petroleum refining	0.090	13	4	8	14
223- Gravure printing	0.078	16	5	24	34
362- Jewelry and related products	0.074	561	6	2	7
244- Pharmaceutical products	0.056	101	7	9	27
296- Arms and ammunition	0.055	7	8	20	n.s.
334- Optical, photographic, and cinematographic instruments	0.049	28	9	17	n.s.
263- Ceramic wall and floor tile	0.047	57	10	27	25
341- Automobiles	0.046	15	11	30	n.s.
351- Shipbuilding and repairing	0.046	155	12	15	28
332- Measuring, analyzing, and controlling instruments	0.039	29	13	23	n.s.
172- Broadwoven fabric mills	0.038	256	14	10	11
193- Footwear	0.037	1932	15	7	17
173- Dyeing and finishing textiles	0.037	275	16	16	21
323- Radio and television apparatus (reception)	0.035	28	17	36	3
247- Artificial and synthetic fibers	0.034	12	18	n.s.	8
171- Yarn spinning mills	0.034	226	19	18	23
192- Other leather products	0.031	244	20	25	30
152- Sea products processing	0.031	106	21	21	19
176- Knit fabric mills	0.027	284	22	14	15

Note: *n.s.*—not significantly different from zero at 95 percent confidence.

provided by the EG index ($\hat{\gamma}_{EG}$) and the alternative EG index based on plant counts ($\hat{\gamma}_{C}$). If we first look at the extent of localization across the 100 three-digit sectors, we find very similar results for the two measures based on plant counts. Indeed, there is statistical evidence that 82 percent and 81 percent of the industries are localized, according to $\hat{\gamma}_{C}$ and $\hat{\gamma}_{DM}$, respectively. The corresponding statistic for the original EG index ($\hat{\gamma}_{EG}$) is substantially lower (68 percent).[13]

Figure 2 displays the box-whisker plots for the three measures. To increase readability, the graph omits a few extreme (high) values for each one of the distributions. Clearly, all distributions show the same pattern of skewness with increasing interquartile ranges. Nevertheless, as we anticipated, our proposed

[13]For $\hat{\gamma}_{EG}$ and $\hat{\gamma}_{C}$ we used the test shown in Maurel and Sedillot (1999), which assumes a normal distribution for the estimator. For $\hat{\gamma}_{DM}$, we performed the likelihood ratio test indicated before.

TABLE 3: Geographic Concentration, by Nonlocalized Industries According to $\hat{\gamma}_{DM}$

3-digit SIC industry (Portuguese CAE-Rev2)	$\hat{\gamma}_{DM}$	Number of plants	Rank		
			$\hat{\gamma}_{DM}$	$\hat{\gamma}_{C}$	$\hat{\gamma}_{EG}$
242- Agricultural chemicals	0.000	10	82	n.s.	16
365- Games and toys	0.000	29	83	n.s.	40
160- Tobacco	0.000	2	n.s.	n.s.	n.s.
268- Miscellaneous nonmetallic mineral products	0.000	29	n.s.	n.s.	18
271- Primary iron industries	0.022	13	n.s.	n.s.	2
272- Iron and steel pipes and tubes	0.014	15	n.s.	n.s.	n.s.
273- Other iron and steel primary industries	0.000	30	n.s.	n.s.	n.s.
274- Primary Nonferrous industries	0.003	81	n.s.	72	48
283- Steam generators	0.030	8	n.s.	n.s.	12
294- Machine tools	0.005	70	n.s.	68	n.s.
297- Household appliances, n.e.s.	0.003	71	n.s.	87	n.s.
311- Electrical motors, generators, and transformers	0.002	83	n.s.	81	n.s.
314- Electric batteries and related products	0.008	11	n.s.	n.s.	n.s.
322- Radio and television apparatus (emission)	0.014	23	n.s.	n.s.	n.s.
333- Controlling devices for manufacturing	0.006	15	n.s.	n.s.	n.s.
335-Watches, clocks, and clockwork operated devices	0.019	15	n.s.	n.s.	n.s.
353- Aircraft and space vehicles	0.014	13	n.s.	n.s.	n.s.
364- Sporting goods	0.000	13	n.s.	n.s.	n.s.
372-Recycling of Nonmetalic products	0.004	70	n.s.	n.s.	54

Note: *n.s.*—not significantly different from zero at 95 percent confidence.

measure of localization (labeled as DM index in the figure) has a much smaller total variance (across and within sectors) when compared with the EG index ($\hat{\gamma}_{EG}$) and the alternative EG index based on counts ($\hat{\gamma}_{C}$). Because the relation between the total variance of the estimators provides information about the relation between the variance of the estimators within sectors, Figure 2 also offers some indirect evidence that potential gains in efficiency could be obtained by using $\hat{\gamma}_{DM}$ instead of $\hat{\gamma}_{C}$ and $\hat{\gamma}_{EG}$.[14]

[14]Note that the total variance of $\hat{\gamma}$ can be decomposed according to: $V(\hat{\gamma}) = E_s[V(\hat{\gamma} \mid s)] + V_s[E(\hat{\gamma} \mid s)]$, where s stands for sector and $V_s[E(\hat{\gamma} \mid s)]$ is (asymptotically) identical for the three estimators.

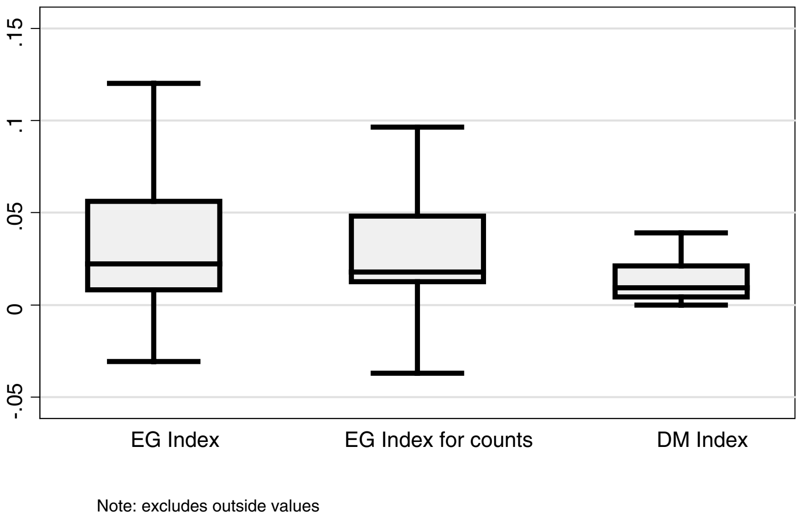

FIGURE 2: Box–Whisker Plots for the Three Localization Indexes.

If we look at the hierarchy of individual industries, we find a significant degree of concordance among the indices based on plant counts. Indeed, the Spearman rank correlation coefficient between $\hat{\gamma}_{DM}$ and $\hat{\gamma}_{C}$ is 0.79. On the other hand, the rank correlation between $\hat{\gamma}_{DM}$ and $\hat{\gamma}_{EG}$ is still positive, but much smaller 0.46.[15] A quick inspection of Table 4 also reveals that the concordance between our index and the EG index based on plant counts is higher and more consistent across quartiles.

Simulation Study. To investigate further the behavior of our proposed measure of localization, we performed a simulation study using different scenarios based on our empirical application. The simulation was implemented as follows. First, we set a value, n, for the total number of firms. Next, we use expression (3) to generate location probabilities for the 275 *concelhos*. In this expression we perturbe the expected location probabilities (each *concelho* share on total manufacturing employment) using the gamma, uniform, normal, and triangular distributions.[16] Next, we randomly allocate the n firms by the 275 *concelhos* using the computed location probabilities. This procedure was repeated a 1,000 times. Each time we apply our estimator to the spatial distribution of plants as well as the EG index for counts.

[15]Both correlation coefficients are statistically different from zero.

[16]When simulating from the gamma, we assumed that the $\exp(\eta_j)$ are random variables following the assumptions that lead to the Dirichlet-Multinomial index. For the other simulations, we let η_j follow a distribution centered in zero with different levels of dispersion.

TABLE 4: Number of Industries by Quantiles

	Quantiles of $\hat{\gamma}_C$				
Quantiles of $\hat{\gamma}_{DM}$	1	2	3	4	Total
1	19	4	2	0	25
2	4	14	5	2	25
3	2	7	11	5	25
4	0	0	7	18	25
Total	25	25	25	25	100

	Quantiles of $\hat{\gamma}_{EG}$				
Quantiles of $\hat{\gamma}_{DM}$	1	2	3	4	Total
1	12	7	3	3	25
2	4	10	7	4	25
3	6	8	8	3	25
4	3	0	7	15	25
Total	25	25	25	25	100

Figure 3 displays a summary of the results for different distributions and values of γ using 16, 75, 208, and 1521 for the total number of firms.[17] The figure reports the ratio of the average mean squared errors (MSE) for our index over the same average for the EG index based on plant counts.[18] The functions shown in the graphs result from the interpolation of the estimates obtained for a grid of values of γ corresponding to increasing levels of dispersion in the random variables. As expected, our index always produces more precise estimates with random gamma effects. Interestingly, for other distributions we consistently find that the performance of our estimator is better for lower values of γ. For the range considered in Figure 3 ($0 \leq \gamma \leq 0.6$), the $\hat{\gamma}_C$ was only able to outperform our estimator when γ exceeded 0.26 and 0.46 for the normal and triangular distributions, respectively.[19] It is also noteworthy that the performance of our index is superior even when we set $\gamma = 0$. For this latter case, we know that $\hat{\gamma}_{DM}$ is slightly biased because it always generates nonnegative estimates. Nevertheless, because it has a much smaller variance, it still is able to outperform the $\hat{\gamma}_C$.

[17] 16, 75, 208, and 1521 are, respectively, the average number of firms by sector in the first, second, third, and fourth quartiles of our dataset for Portugal. To approximate the "true" value of γ we computed the variance of p_j across simulations and used the relationship in Equation (2).

[18] The MSE is given by $\sum(\hat{\gamma} - \gamma)^2 = bias^2 + V(\hat{\gamma})$.

[19] Note that values for the "true" γ above 0.26 are rarely observed in real data. Ellison and Glaeser (1997), for example, considers industries with γ's above 0.05 as being highly concentrated.

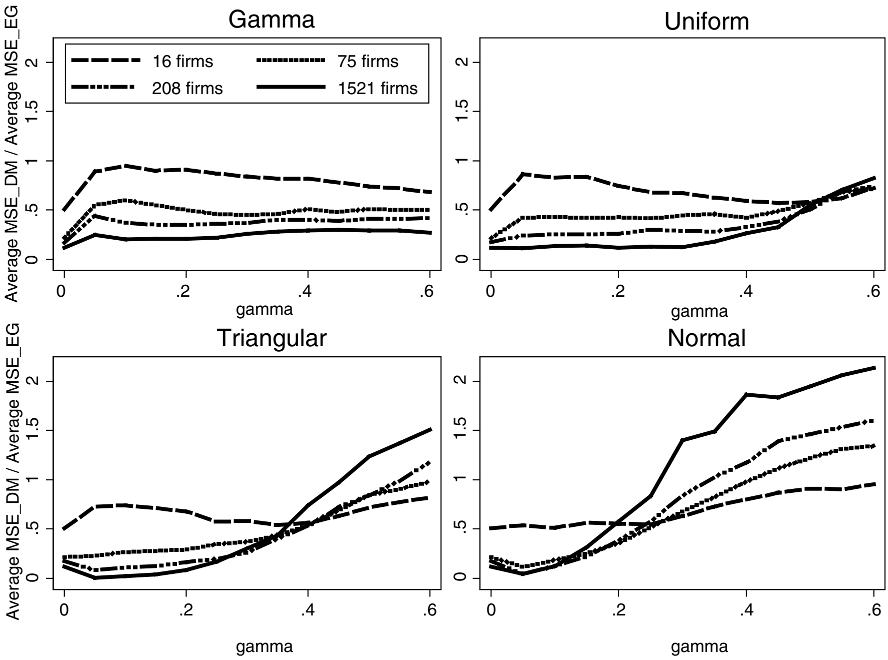

FIGURE 3: Average Mean Squared Errors (MSE) for 1,000 Runs.

5. SUMMARY AND EXTENSIONS

Regional scientists have long emphasized the pivotal role of localization, which encompasses firm decisions to exploit natural advantages and external spatial economies like knowledge spillovers, labor market pooling, and input sharing. The index proposed by Ellison and Glaeser (1997) significantly advanced research in this area with a measure of localization more statistically and economically sound than found in the literature up to that point. Subsequently, much applied work has focused on employment concentration. Our index, in contrast, accounts for location decisions and intra-industry establishment clustering. Building on the Random Utility (Profit) model of industrial location, we develop a parametric version of localization consistent with the theoretical construct underlying the original work of Ellison and Glaeser. Because the distributional assumptions in our model are fully compatible with RUM, we are able to offer a more efficient measure that will generate more precise estimates of establishment clustering. One virtue is that establishment (plant) count data are often available for detailed industries without disclosure problems. Thus, our approach can be tested and compared across many regional contexts.

Where data can be obtained, future research could examine the behavior of our index with respect to employment size, as has been explored previously (Holmes and Stevens, 2002, and Barrios, Bertinelli, and Strobl, 2006).

An essential question for studies of agglomeration concerns whether plants located in areas of concentration tend to be larger or smaller. The literature is ambiguous as to the direction of the effect. For example, it has been often asserted that vertical disintegration (with many small, specialized plants) is prevalent in Italian industrial districts (areas of concentration). Knowledge externalities may be crucial for small firms, leading to more clustering. Yet, for the United States, Holmes and Stevens (2002) found that dropping small plants from the sample increases their index measuring concentration. Using Irish data, Barrios et al. (2006) also uncover a positive relation between establishment scale and regional industrial concentration. With the Portuguese data examined in this paper, dropping all plants with less than five employees amounts to deleting 41.4 percent of all plants. In a preliminary investigation, we found that restricting the calculation to plants with five or more employees increases the index for 15 of the 22 industries shown in Table 2. This appears congruent with the Holmes and Stevens (2002) results. A deeper investigation could explore the important issue of plant size and establishment concentration using the plant count index proposed in this paper.

REFERENCES

Barrios, S., Bertinelli, L., and Strobl, E. 2006. "Geographic Concentration and Establishment Scale: An Extension Using Panel Data," *Journal of Regional Science*, 46(4), 733–746.

Bartik, T. 1985. "Business Location Decisions in the United States: Estimates of the Effects of Unionization, Taxes, and Other Characteristics of States," *Journal of Business and Economic Statistics*, 3(1), 14–22.

Bertinelli, L., and Decrop, J. 2002. "Geographical Agglomeration: The Case of Belgian Manufacturing Industry," Working Paper no. 14, Federal Planning Bureau of Belgium.

Cabral, L., and Mata, J. 2003. "On the Evolution of the Firm Size Distribution: Facts and Theory," *American Economic Review*, 93(4), 1075–1090.

Callejón, M. 1997. "Concentratión geográphica de la industria y economias de aglomeratión," *Economia Industrial*, 5, 61–68.

Carlton, D. 1983. "The Location and Employment Choices of New Firms: An Econometric Model with Discrete and Continuous Endogenous Variables," *The Review of Economic and Statistics*, 65(3), 440–449.

Coughlin, C., Terza, J., and Arromdee, V. 1991. "State Characteristics and the Location of Foreign Direct Investment within the United States," *The Review of Economic and Statistics*, 73(4), 675–683.

Devereux, M. P., Griffith, R., and Simpson, H. 2004. "The Geographic Distribution of Production Activity in the UK," *Regional Science and Urban Economics*, 34(5), 533–564.

Dumais, G., Ellison, G., and Glaeser, E. 2002. "Geographic Concentration as Dynamic Process," *The Review of Economics and Statistics*, 84(2), 193–204.

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

Ellison, G., and Glaeser, E. 1994. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," Working Paper no. 4840, NBER.

Ellison, G., and Glaeser, E. 1997. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 105(5), 889–927.

Friedman, J., Gerlowski, D., and Silberman, J. 1992. "What Attracts Foreign Multinational Corporations? Evidence from Branch Plant Location in the United States," *Journal of Regional Science*, 32(4), 403–418.

Guimarães, P., Figueiredo, O., and Woodward, D. 2000. "Agglomeration and the Location of Foreign Direct Investment in Portugal," *Journal of Urban Economics*, 47(1), 115–135.
Guimarães, P., Figueiredo, O., and Woodward, D. 2004. "Industrial Location Modeling: Extending the Random Utility Framework," *Journal of Regional Science*, 44(1), 1–20.
Head, K., Ries, J., and Swenson, D. 1995. "Agglomeration Benefits and Location Choice: Evidence from Japanese Manufacturing Investments in the United States," *Journal of International Economics*, 38(3–4), 223–247.
Holmes, T., and Stevens, J. 2002. "Geographic Concentration and Establishment Scale," *The Review of Economics and Statistics*, 84(4), 682–690.
Hoover, E. 1937. *Location Theory and the Shoe and Leather Industries*. Cambridge, MA: Harvard University Press.
Houdebine, M. 1999. "Concentration géographique des activités et spécialisation des départments français," *Economie et Statistique*, 326–327(6–7), 189–204.
Johnson, N., Kotz, S., and Balakrishnan, N. 1997. *Discrete Multivariate Distributions*. New York, NY: Wiley Series in Probability and Statistics, John Wiley and Sons.
Kim, Y., Barkley, D., and Henry, M. 2000. "Industry Characteristics Link to Establishments Concentration in Nonmetropolitan Areas," *Journal of Regional Science*, 40(2), 231–259.
Krugman, P. 1991. *Geography and Trade*. Cambrige, MA: M.I.T. Press.
Louviere, J., Hensher, D., and Swait, J. 2000. *Stated Choice Methods: Analysis and Application*. Cambridge, UK: Cambridge University Press.
Maurel, F., and Sedillot, B. 1999. "A Measure of the Geographic Concentration in French Manufacturing Industries," *Regional Science and Urban Economics*, 29(5), 575–604.
McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka (ed.), *Frontiers in Econometrics*. NY: Academic Press, pp. 105–142.
Mosimann, J. 1962. "On the Compound Multinomial Distribution, the Multivariate Beta-Distribution, and Correlations Among Proportions," *Biometrika*, 49, 65–82.
Self, S., and Liang, K.-Y. 1987. "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

APPENDIX A: DERIVATION OF THE EG INDEX BASED ON COUNTS OF PLANTS

In the context of the EG model, the number of investments of a given industry in region j, conditional on the total number of investments in the industry, and on the vector of locational probabilities ($\mathbf{p} = p_1, p_2, \ldots, p_J$), follows a binomial law with parameters:

$$E(n_j \mid \mathbf{p}) = np_j$$

$$V(n_j \mid \mathbf{p}) = np_j(1 - p_j)$$

Define a "raw index of concentration" as

$$G_C = \sum_{j=1}^{J} \left(\frac{n_j}{n} - x_j\right)^2$$

Expanding terms we obtain

$$G_C = \frac{1}{n^2}\sum_{j=1}^{J} n_j^2 + \sum_{j=1}^{J} x_j^2 - \frac{2}{n}\sum_{j=1}^{J} n_j x_j$$

and the expected value of the above equation gives

$$E(G_C \mid \mathbf{p}) = \frac{1}{n^2} E\left(\sum_{j=1}^{J} n_j^2\right) + \sum_{j=1}^{J} x_j^2 - \frac{2}{n} E\left(\sum_{j=1}^{J} n_j x_j\right)$$
$$= \frac{1}{n^2}\left(\sum_{j=1}^{J} np_j - np_j^2 + n^2 p_j^2\right) + \sum_{j=1}^{J} x_j^2 - 2\sum_{j=1}^{J} p_j x_j$$
$$= \frac{1}{n} + \frac{(n-1)}{n}\sum_{j=1}^{J} p_j^2 + \sum_{j=1}^{J} x_j^2 - 2\sum_{j=1}^{J} p_j x_j$$

Applying the law of iterated expectations we get

$$E(G_C) = \frac{1}{n} + \frac{(n-1)}{n} E\left(\sum_{j=1}^{J} p_j^2\right) + \sum_{j=1}^{J} x_j^2 - 2E\left(\sum_{j=1}^{J} p_j x_j\right)$$
$$= \frac{1}{n} + \frac{(n-1)}{n}\sum_{j=1}^{J}\left(\gamma x_j - \gamma x_j^2 + x_j^2\right) + \sum_{j=1}^{J} x_j^2 - 2\sum_{j=1}^{J} x_j^2$$
$$= \frac{1+\gamma(n-1)}{n}\left(1 - \sum_{j=1}^{J} x_j^2\right)$$

and, as in Ellison and Glaeser (1997), the estimator for γ is obtained by replacing the $E(G_C)$ by the observed value of G_C and solving for γ. The proposed estimator is:

$$\hat{\gamma}_C = \frac{nG_C - \left(1 - \sum_{j=1}^{J} x_j^2\right)}{(n-1)\left(1 - \sum_{j=1}^{J} x_j^2\right)}$$

APPENDIX B: PROOF THAT $V(\hat{\gamma}_{EG}) \geq V(\hat{\gamma}_C)$

We know that

$$V(\hat{\gamma}_{EG}) = \frac{V(G_E)}{(1-H)^2\left(1 - \sum_{j=1}^{J} x_j^2\right)^2} \tag{B1}$$

with,

$$V(G_E) = 2AH^2 - 2B\sum_{i=1}^{n} z_i^4$$

where z_i is the share of plant i in a particular industry employment and

$$A = \sum_{j=1}^{J} x_j^2 - 2\sum_{j=1}^{J} x_j^3 + \left(\sum_{j=1}^{J} x_j^2\right)^2$$

and

$$B = \sum_{j=1}^{J} x_j^2 - 4\sum_{j=1}^{J} x_j^3 + 3\left(\sum_{j=1}^{J} x_j^2\right)^2$$

It follows that

$$\text{(B2)} \qquad V(\hat{\gamma}_C) = \frac{V(G_C)}{\left(1 - \dfrac{1}{n}\right)^2 \left(1 - \displaystyle\sum_{j=1}^{J} x_j^2\right)^2}$$

and

$$V(G_C) = 2A\frac{1}{n^2} - 2B\frac{1}{n^3}$$

Direct comparison of (B1) and (B2) shows that $V(\hat{\gamma}_{EG}) \geq V(\hat{\gamma}_C)$ if

$$A\left(\frac{H^2}{(1-H)^2} - \frac{1}{(n-1)^2}\right) \geq B\left(\frac{\displaystyle\sum_{i=1}^{n} z_i^4}{(1-H)^2} - \frac{1}{n(n-1)^2}\right)$$

Since $A - B = 2\sum_{j=1}^{J} x_j^3 - 2(\sum_{j=1}^{J} x_j^2)^2$, it follows by Jensen's inequality that $A \geq B$ (Jensen's inequality states that if $\sum_{i=1}^{n} a_j = 1, a_j \geq 0, j = 1, \ldots, n$, and f is convex then, $\sum_{i=1}^{n} a_j f(x_j) \geq f[\sum_{i=1}^{n} a_j x_j])$. To see that $(\frac{H^2}{(1-H)^2} - \frac{1}{(n-1)^2}) \geq (\frac{\sum_{i=1}^{n} z_i^4}{(1-H)^2} - \frac{1}{n(n-1)^2})$ note that this inequality can be written as

$$\text{(B3)} \qquad \frac{H^2 - \displaystyle\sum_{i=1}^{n} z_i^4}{(1-H)^2} \geq \frac{1}{n(n-1)}$$

Since

$$1 - H = 2\sum_{i=1}^{n}\sum_{j \neq i}^{n} z_i z_j$$

and

$$H^2 - \sum_{i=1}^{n} z_i^4 = 2\sum_{i=1}^{n}\sum_{j \neq i}^{n} z_i^2 z_j^2$$

we can replace the above expressions in (B3) to obtain

$$\text{(B4)} \qquad \frac{n(n-1)}{2}\sum_{i=1}^{n}\sum_{j \neq i}^{n} z_i^2 z_j^2 \geq \left(\sum_{i=1}^{n}\sum_{j \neq i}^{n} z_i z_j\right)^2$$

The summations in (B4) have exactly $\frac{n(n-1)}{2}$ terms, and thus we can use Chebyshev's sum inequality to directly prove the relation (Chebyshev's sum inequality states that if $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$ then, $n\sum_{i=1}^{n} a_j b_j \geq [\sum_{i=1}^{n} a_j][\sum_{i=1}^{n} b_j]$). Note that if $H > \frac{1}{n}$ then $\frac{n(n-1)}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} z_i^2 z_j^2 > (\sum_{i=1}^{n}\sum_{j\neq i}^{n} z_i z_j)^2$ and thus (since $A \geq B$) $V(\hat{\gamma}_{EG}) > V(\hat{\gamma}_C)$.