

# A Divergence Statistic for Industrial Localization

Tomoya Mori\*, Koji Nishikimi† and Tony E. Smith‡

January 31, 2005

## Abstract

In this paper, we propose a statistical index of industrial localization based on Kullback-Leibler divergence. This index is particularly well suited to cases where industrial data is only available at the regional level. Unlike existing regional-level indices, our index can be employed to test the significance of industrial localization relative to a hypothesized reference distribution of probable locations across regions. In addition, one can test relative degrees of localization among industries. Finally, as with all Kullback-Leibler divergence indices, our index can be decomposed into components representing localization at various levels of spatial aggregation.

*Keywords:* Industrial localization, Kullback-Leibler divergence, Relative entropy, Large-sample analysis of localization, Spatial decomposition of localization

*JEL Classifications:* C19, L60, R12

---

\*Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan. Email: mori@kier.kyoto-u.ac.jp. Phone: +81-75-753-7121. Fax: +81-75-753-7198.

†Institute of Developing Economies/JETRO, 3-2-2 Wakaba, Mihama-ku, Chiba, 261-8545 Japan. Email: nisikimi@ide.go.jp. Phone: +81-43-299-9672. Fax: +81-43-299-9763.

‡Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: tsmith@ssc.upenn.edu. Phone: +1-215-898-9647. Fax: +1-215-898-5020.

## Acknowledgment

Earlier versions of this paper was presented at the 16th Meeting of the Applied Regional Science Conference at Okayama University, December 1-2, 2002, the conference on the economics of cities at London School of Economics, June 6-8, 2003, seminars at Osaka University and Osaka Prefecture University. We are grateful to conference/seminar participants for their constructive comments. We have benefited greatly from discussions with Gilles Duranton, Vernon Henderson, Kazuhiko Kakamu, Yoshihiko Nishiyama, Henry Overman, Akihisa Shibata, and two anonymous referees. This research has been supported by The Murata Science Foundation, The Grant in Aid for Research (Nos. 09CE2002, 09730009, 13851002) of Ministry of Education, Science and Culture in Japan.

# 1 Introduction

In the past decade, a substantial number of empirical studies of industrial localization have appeared in the literature.<sup>1</sup> These studies suggest that industrial localization is far more ubiquitous than previously believed, and extends well beyond the classical agglomerations of industries exemplified by the information-technology industry in Silicon Valley and automobile industry in Detroit. Moreover, the degree of such localization varies across industries, and often tends to be more subtle than these classical examples. However, most localization indices currently in use provide no clear statistical method for detecting the presence of localization.<sup>2</sup> Hence the central purpose of this paper is to develop an index of localization that does provide such a statistical testing framework.<sup>3</sup>

The index we employ is based on the concept of *Kullback-Leibler divergence* (1951), and is here designated as the *D-index* of industrial localization. In particular, a reference distribution of *complete spatial dispersion* is formulated as a null hypothesis, and the *D-index* for each industry is computed as the Kullback-Leibler divergence between the observed spatial distribution of establishments for that industry and this reference distribution. Since higher values of *D* are taken to convey stronger evidence against complete spatial dispersion, we interpret this to be evidence for localization in that industry. Within this framework it is shown that under the null hypothesis of complete spatial dispersion, the *D-index* is asymptotically normally distributed, and thus provides a natural statistical test for this hypothesis. However, the more relevant question for our purposes focuses on the *relative* magnitude of *D* for different industries. Hence the main application of this asymptotic normality property is to construct tests of the differences between *D*-indices for separate industries, and thereby to conclude that industries with significantly higher *D*-values are significantly more localized.<sup>4</sup>

Our null hypothesis of *complete spatial dispersion* for any industry is operationalized by postulating that all feasible locations for establishments in that industry are equally likely. If the totality of these locations is designated as the *economic area* within the given geographic area (see footnote 29), then this hypothesis is formalized as a *uniform probability distribution over economic area*, representing the probable location of a randomly sampled establishment within that industry. Unlike the more traditional approach of using industrial aggregates as reference distributions [e.g., Ellison and Glaeser (1997) and Krugman (1991)],<sup>5</sup> this hypothesis of complete spatial dispersion is more directly related to existing theories of economic agglomeration in terms of spatial proximity.<sup>6</sup>

Finally, it is well known that Kullback-Leibler divergence indices are decomposable with respect to partitions of the sample space.<sup>7</sup> In particular, the *D-index* can be decomposed with respect to the choice of geographical units. This provides a way of measuring the spatial extent of industrial localization, thereby suggesting the most appropriate geographic units for industrial localization analysis. While the decomposability result itself is not new, its application to the localization index is new.<sup>8</sup>

The rest of the paper is organized as follows. In Section 2, we begin by defining our  $D$ -index for industrial localization and highlight some of its major characteristics. In Section 3, we discuss large sample properties of the  $D$ -index, and develop its asymptotic normality properties. In Section 4, these results are applied to the case of Japan. Here the  $D$ -index is computed for Japanese industries and compared with the traditional index of Ellison and Glaeser (1997). The paper concludes with a brief discussion of directions for further research.

## 2 Measure of localization

In this section, we propose our new index of industrial localization and discuss its major aspects. We start by laying out the formal framework for our analysis in Section 2.1. The  $D$ -index is then defined in Section 2.2, and its interpretation in terms of likelihood ratios is given in Section 2.3. Finally, the decomposition properties of the  $D$ -index are developed in Section 2.4.

### 2.1 Basic setting

Consider a finite set of *industries*,  $i \in \mathbf{I} = \{1, \dots, I\}$ , located within a set of *regions*,  $r \in \mathbf{R} = \{1, \dots, R\}$ . Suppose that  $N_{ir}$  denotes the total number of establishments of industry  $i$  in region  $r$ . Then the objective of the present analysis is to characterize the degree of geographic concentration or dispersion of establishments among industries, as exhibited by the *establishment location pattern*  $(N_{ir} : ir \in \mathbf{I} \times \mathbf{R})$ . The approach adopted here is to treat this establishment location pattern,  $N_i = (N_{ir} : r \in \mathbf{R})$ , of each industry  $i$  as a random sample of size  $n_i = \sum_r N_{ir}$  from a larger statistical population of potential establishments. If  $p_{ir}$  denotes the probability that a randomly sampled establishment from industry  $i$  will be located in region  $r$ , then our interest focuses on the relative degrees of spatial concentration exhibited by each of these  *$i$ -establishment distributions*,  $p_i = (p_{ir} : r \in \mathbf{R})$ ,  $i \in \mathbf{I}$ . It should be noted that the random sampling assumption above implies that the locational decisions of individual establishments are treated as *statistically independent events*. Of course this is at best an approximation to the actual dynamics of successive locational decisions by establishments.<sup>9</sup>

To characterize the observed location pattern of establishments, we begin by formulating a probability model of “complete spatial dispersion” and then consider the deviations of each distribution,  $p_i$ , from this benchmark model. Here it is postulated that a completely dispersed distribution for industry  $i$  is one in which randomly sampled  $i$ -establishments are equally likely to be located anywhere within the *economic area*,  $a$ , of the given regional system. If  $a_r$  denotes the economic area of region  $r$ , then under complete spatial dispersion the probability that a randomly sampled  $i$ -establishment will be located in region  $r$  is

given by the fraction:

$$p_{0r} = a_r/a \quad , \quad r \in \mathbf{R}. \quad (1)$$

Hence, this hypothesis of *complete spatial dispersion* is summarized by the probability distribution

$$p_0 = (p_{0r} : r \in \mathbf{R}) \quad (2)$$

which we now adopt as a benchmark against which to compare all  $i$ -establishment distributions,  $p_i$ .

It should be noted at this point that alternative null hypotheses are possible in which the benchmark distribution is, for example, based on the regional fractions of total establishments or employment levels for all industries. The implications of these alternative choices are discussed more fully in Appendix A below.

## 2.2 Definition of the $D$ -index

How can we measure the deviation between the two distributions,  $p_i$  and  $p_0$ ? While there are many possible measures of deviation between distributions,<sup>10</sup> the most natural choice from a statistical viewpoint is the *Kullback-Leibler divergence* between  $p_i$  and  $p_0$ .<sup>11</sup> This divergence measure (also called as the *relative entropy* of  $p_i$  with respect to  $p_0$ ) is defined as follows:<sup>12</sup>

$$D(p_i|p_0) = \sum_{r \in \mathbf{R}} p_{ir} \ln \left( \frac{p_{ir}}{p_{0r}} \right). \quad (3)$$

$D(p_i|p_0)$  is well known to be nonnegative, and to achieve its minimum uniquely at zero when  $p_i = p_0$ . Moreover, its local maxima are achieved precisely at the degenerate distributions in which all  $i$ -establishments are concentrated in a single region (i.e., with  $p_{ir} = 1$  for some  $r \in \mathbf{R}$ ).<sup>13</sup> Hence it is natural to regard values of  $D(p_i|p_0)$  for each industry  $i$  as reflecting its *degree of localization*. Similarly, for any pair of industries,  $i$  and  $j$ , we now regard industry  $i$  as *more localized* than industry  $j$  whenever  $D(p_i|p_0) > D(p_j|p_0)$ . Here we designate  $D(p_i|p_0)$  as the *D-index* of localization and propose to use it as a measure of industrial localization.

But notice that  $D(p_i|p_0)$  is not directly observable. In particular, while the reference distribution,  $p_0$ , can generally be measured with a reasonable degree of accuracy, the establishment-location probabilities,  $p_i$ , are not directly observed. However, the current location pattern,  $N_i = (N_{ir} : r \in \mathbf{R})$ , yields natural sample estimates of these probabilities:

$$\hat{p}_{ir} = \frac{N_{ir}}{n_i} \quad , \quad r \in \mathbf{R}. \quad (4)$$

Moreover, since these are in fact maximum-likelihood estimates (under our random sampling assumptions) it follows from the well-known invariance properties of such estimates that a corresponding maximum-likelihood estimate of  $D(p_i|p_0)$  is given by:

$$D(\hat{p}_i|p_0) = \sum_{r \in \mathbf{R}} \hat{p}_{ir} \ln \left( \frac{\hat{p}_{ir}}{p_{0r}} \right) \quad (5)$$

where  $\hat{p}_i = (\hat{p}_{ir} : r \in \mathbf{R})$ . Since the probability estimates  $\hat{p}_i$  converge *exponentially* to  $p_i$  in probability,<sup>14</sup> it follows that  $D(\hat{p}_i|p_0)$  also converges to its true value,  $D(p_i|p_0)$ , exponentially fast. Hence when total number of establishments,  $n_i$ , is large, these sample values should provide sharp estimates of true divergence.

Finally, to gain some feeling for “degree of localization” implied by the value,  $D_i = D(\hat{p}_i|p_0)$ , for any given distribution of  $i$ -establishments, it is of interest to use simple *core-periphery* distributions as a baseline of comparison. To do so, consider a range of hypothetical situations in which the given nation of size  $a$  is partitioned into only two regions  $\{c, \bar{c}\}$  where all  $i$ -establishments are uniformly distributed in region,  $c$ . If we designate  $c$  as the *core region* for  $i$  (with  $\bar{c}$  denoting the *periphery region*), then one may ask how small (i.e., how concentrated) this core region must be in order to yield the same degree of localization. More precisely, if this hypothetical distribution of  $i$ -establishments is denoted by  $p_i = (p_{ic}, p_{i\bar{c}}) = (1, 0)$ , with corresponding reference distribution,  $p_c = [a_c/a, (a - a_c)/a]$ , where  $a_c$  is the size of  $c$ , then how small must  $a_c/a$  be in order to yield the same  $D$ -index value, i.e., to ensure that  $D(p_i, p_c) = D_i$ . By (5) we see that  $D(p_i, p_c) = 1 \cdot \ln[1/(a_c/a)] + 0 = -\ln(a_c/a)$ , and thus that one must have  $a_c/a = e^{-D_i}$ . So if the given value of  $D_i$  were say 2.3, then this would imply that  $a_c/a = e^{-2.3} = 0.10$ , and hence that the core region for  $i$  could be only 10% the size of the nation. In other words, industry  $i$  would have to be 10 times more concentrated than under a uniform distribution of  $i$ -establishments throughout the nation. Similarly, if  $D_i$  were to equal 4.6, then industry  $i$  would have to be 100 times more concentrated than under uniformity.

### 2.3 Relationship between the $D$ -index and likelihood ratio

From a statistical viewpoint, the appeal of this  $D$ -index is due largely to its interpretation as a limiting form of the log-likelihood ratio for testing the hypothesis,  $p_i = p_0$ . Since we assumed in Section 2.1 that samples are independent, the probability,  $P_i$ , that the employment pattern,  $N_i$ , is realized under any distribution,  $p_i$ , is given by the multinomial probability:

$$P_i(N_i) = \left( \frac{n_i!}{\prod_{r \in \mathbf{R}} N_{ir}!} \right) \prod_{r \in \mathbf{R}} (p_{ir})^{N_{ir}}. \quad (6)$$

Following standard convention, this is reinterpreted to be the *likelihood*,  $L_i(p_i|N_i)$ , of distribution  $p_i$  given the realized pattern  $N_i$ . This likelihood is of course maximized by the sample relative frequency distribution  $\hat{p}_i$  in (4) derived from  $N_i$ . Hence the *relative likelihood* of the hypothesized distribution,  $p_0$ , given  $N_i$  is taken to be given by the *likelihood ratio*:

$$\lambda \equiv \frac{L_i(p_0|N_i)}{L_i(\hat{p}_i|N_i)} = \prod_{r \in \mathbf{R}} \left( \frac{p_{0r}}{\hat{p}_{ir}} \right)^{N_{ir}}, \quad (7)$$

With these definitions, it is seen by taking negative logs that

$$-\ln \lambda = \sum_{r \in \mathbf{R}} N_{ir} \ln \left( \frac{\hat{p}_{ir}}{p_{0r}} \right) \quad (8)$$

and hence that

$$-\frac{\ln \lambda}{n_i} = \sum_{r \in \mathbf{R}} \left( \frac{N_{ir}}{n_i} \right) \ln \left( \frac{\hat{p}_{ir}}{p_{0r}} \right) = D(\hat{p}_i|p_0). \quad (9)$$

Thus our *D-index* can be interpreted as simply the negative log-likelihood ratio normalized by sample size. In particular this implies that divergence,  $D(\hat{p}_i|p_0)$ , from  $p_0$  is larger for those observed patterns,  $N_i$ , that attribute smaller relative likelihood to  $p_0$ . Moreover, this implies from (9), together with the probabilistic convergence of  $D(\hat{p}_i|p_0)$  to  $D(p_i|p_0)$ , that the true divergence value  $D(p_i|p_0)$  is in fact the (*normalized*) *limiting form of the negative log-likelihood ratio*.

Given this correspondence, together with the well known distributional properties of likelihood ratios,<sup>15</sup> it is natural to ask at this point why not simply use these more standard test statistics. The key here is the role of industry size,  $n_i$ , which constitutes the relevant sample size in such testing procedures. Observe in particular from (8) that the negative log-likelihood ratio,  $-\ln \lambda$ , depends *linearly* on the sample sizes. So by doubling the size of an industry, one necessarily doubles the negative log-likelihood ratio, and hence the “weight of evidence” for localization (i.e., against complete dispersion). This would make perfect sense if one were observing a single industry  $i$  growing proportionally over time – where successively larger sample sizes would indeed add strength to the hypothesis that the realized sample distribution,  $\hat{p}_i$ , is close to the underlying statistical population of establishments for industry  $i$ . However, when comparing different industries, this tends to give undue weight to larger versus smaller industries. In particular, ubiquitous industries with large numbers of establishments across all regions can in fact appear significantly more localized than small industries that are concentrated in only a few regions.<sup>16</sup> Hence, it is our view that in order to be comparable across industries, *an index of localization should be independent of sample size*, as is the case for our *D-index*. However, it should also be noted that sample size continues to play a statistical role, and in particular that (as shown in Section 3 below): *larger sample sizes yield tighter confidence bounds on the true value of D*.

## 2.4 Spatial decomposition of the $D$ -index

As with all relative-entropy indices, our  $D$ -index is definable with respect to any finite (measurable) partition of the sample space. Moreover, it is well known that there exists a powerful decomposition relation between the values of such indices for nested partitions of the sample space.<sup>17</sup> It is to be stressed, however, that while decomposability of relative entropy itself is not new, its application to the regional decomposition of localization indices is new.<sup>18</sup> As we have attempted to show, this technique is particularly useful for identifying the geographic structure of localization, and for studying how this structure changes over time.

In the present case, suppose that set of regions,  $\mathbf{R}$ , is partitioned into  $M$  ( $< R$ ) bundles of regions, where  $m^{\text{th}}$  bundle,  $\mathbf{R}_m$ , is comprised of  $R_m$  regions ( $\sum_{m=1}^M R_m = R$ ). Then the conditional probability that an  $i$ -establishment in  $m^{\text{th}}$  regional bundle is located in region  $r \in \mathbf{R}_m$  is given by

$$p_{ir|m} = \frac{p_{ir}}{q_{im}}, \quad m = 1, \dots, M, \quad (10)$$

and, similarly, the conditional probability under the reference distribution is given by

$$p_{0r|m} = \frac{p_{0r}}{q_{0m}}, \quad m = 1, \dots, M \quad (11)$$

where  $q_{im}$  and  $q_{0m}$  are the marginal probabilities that an establishment is located in  $m^{\text{th}}$  regional bundle, i.e.,  $q_{im} = \sum_{r \in \mathbf{R}_m} p_{ir}$  and  $q_{0m} = \sum_{r \in \mathbf{R}_m} p_{0r}$ . Using these relations, we can rewrite  $D(p_i|p_0)$  as follows:

$$\begin{aligned} D(p_i|p_0) &= \sum_{m=1}^M \sum_{r \in \mathbf{R}_m} q_{im} p_{ir|m} \ln \left( \frac{q_{im} p_{ir|m}}{q_{0m} p_{0r|m}} \right) \\ &= \sum_{m=1}^M q_{im} \ln \left( \frac{q_{im}}{q_{0m}} \right) + \sum_{m=1}^M q_{im} \left\{ \sum_{r \in \mathbf{R}_m} p_{ir|m} \ln \left( \frac{p_{ir|m}}{p_{0r|m}} \right) \right\} \\ &= D(q_i|q_0) + \sum_{m=1}^M q_{im} D(p_{i|m}|p_{0|m}), \end{aligned} \quad (12)$$

where  $p_{i|m} = (p_{ir|m} : r \in \mathbf{R}_m)$  and  $p_{0|m} = (p_{0r|m} : r \in \mathbf{R}_m)$ . The first term in the right hand side shows the  $D$ -index among the regional bundles while the second term represents the weighted average of  $D$ -indices within each regional bundle. In other words, the  $D$ -index for all regions can be decomposed into those representing the localization *among* and *within* regional bundles. As in eq.(5), the estimate of  $D$  is obtained by

$$D(\hat{p}_i|p_0) = D(\hat{q}_i|q_0) + \sum_{m=1}^M \hat{q}_{im} D(\hat{p}_{i|m}|p_{0|m}), \quad (13)$$

where  $\hat{q}_{im} = \sum_{r \in \mathbf{R}_m} \hat{p}_{ir}$  and  $\hat{p}_{ir|m} = \frac{\hat{p}_{ir}}{\hat{q}_{im}}$ .

### 3 Large sample properties of the $D$ -index

As was discussed in Section 2.2, the  $D$ -index should provide a sharp estimate of true divergence between  $p_i$  and  $p_0$  when the number of establishments,  $n_i$ , of industry  $i$  is large. But even when industries are large, there is always the question of whether the degrees of localization among industries are *significantly* different. To answer such questions it is necessary to take sample sizes explicitly into account, especially when size differences between industries are great. Hence the objective of this section is to show that the  $D$ -index,  $D(\hat{p}_i|p_0)$ , has an asymptotic normal distribution which allows the sample size,  $n_i$ , to be reflected in an explicit way. We should note here that the basic asymptotic normality property of entropy measures follows from more general results (see footnote 23 below). However, the present sharper form allowing zero values (Theorem 1 below) appears to be new. Moreover, while the confidence intervals and hypothesis tests derived from these asymptotic results are quite standard, the present application of these results to questions of spatial localization is new.<sup>19</sup>

We first establish in Section 3.1 that  $D(\hat{p}_i|p_0)$  is asymptotically normally distributed. In Section 3.2, we then derive the confidence interval for the true level of localization,  $D(p_i|p_0)$ , and illustrate the accuracy of this normal approximation in terms of a few numerical examples. Finally, in Section 3.3, we operationalize procedures for testing both the presence of localization in individual industries and differences in the degree of localization between a pair of industries.

#### 3.1 Asymptotic distribution of the $D$ -index

To begin, observe that the identity,  $\sum_{r=1}^R p_{ir} = 1$ , implies that there are only  $R - 1$  free parameters in the distribution  $p_i$ .<sup>20</sup> Hence to analyze this distribution it is essential to choose an explicit set of free parameters. Here we simply drop the last parameter,  $p_{iR}$ , and now represent the distribution as an  $(R - 1)$ -dimensional vector,

$$p_i = (p_{ir} : r = 1, \dots, R - 1) \quad (14)$$

where by definition  $p_{iR} = 1 - \sum_{r=1}^{R-1} p_{ir}$ . If we also represent the reference distribution by an  $(R - 1)$ -dimensional vector

$$p_0 = (p_{0r} : r = 1, \dots, R - 1) \quad (15)$$

with  $p_{0R} = 1 - \sum_{r=1}^{R-1} p_{0r}$ , then it follows that  $D(p_i|p_0)$  can be equivalently written in terms of this new parameterization as

$$D(p_i|p_0) = \sum_{r=1}^{R-1} p_{ir} \ln \left( \frac{p_{ir}}{p_{0r}} \right) + \left( 1 - \sum_{r=1}^{R-1} p_{ir} \right) \ln \left( \frac{1 - \sum_{r=1}^{R-1} p_{ir}}{1 - \sum_{r=1}^{R-1} p_{0r}} \right). \quad (16)$$

Moreover, for any fixed  $n_i$ , there is also a linear dependency between the numbers of establishments  $(N_{ir} : r = 1, \dots, R)$ . Hence, in a similar manner, we now drop the last number,  $N_{iR}$ , and represent the establishment frequencies for industry  $i$  by an  $(R - 1)$ -dimensional random vector,

$$N_i = (N_{ir} : r = 1, \dots, R - 1) \quad (17)$$

where again by definition  $N_{iR} = n_i - \sum_{r=1}^{R-1} N_{ir}$ . In these terms, it follows from our random-sampling assumption that the random  $(R - 1)$ -vector,  $N_i$ , is *multinomially distributed* with mean vector,  $n_i p_i$ , and  $(R - 1)$ -square covariance matrix:

$$\text{cov}(N_i) = n_i [\text{diag}(p_i) - p_i p_i'] \quad (18)$$

where  $\text{diag}(p_i)$  is the diagonal matrix with elements  $p_i$ .<sup>21</sup> By the multinomial extension of the normal approximation to the binomial, it is well known that the corresponding  $(R - 1)$ -vector of probability estimates,

$$\hat{p}_i = \left( \hat{p}_{ir} = \frac{N_{ir}}{n_i} : r = 1, \dots, R - 1 \right) \quad (19)$$

converges in law to an  $(R - 1)$ -variate normal distribution [see, e.g., Wilks (1962)], and in particular that

$$\sqrt{n_i} (\hat{p}_i - p_i) \xrightarrow{L} N [0, \Sigma(p_i)] \quad (20)$$

where

$$\Sigma(p_i) = \text{diag}(p_i) - p_i p_i'. \quad (21)$$

In addition, it is well known that for any totally differentiable function,  $g$ , if the gradient  $\nabla g(p_i)$  is nonzero at the true distribution,  $p_i$ , then the corresponding estimate,  $g(\hat{p}_i)$ , of  $g(p_i)$  is also asymptotically normally distributed [see, e.g., Rao (1973)], and in particular that

$$\sqrt{n_i} [g(\hat{p}_i) - g(p_i)] \xrightarrow{L} N [0, \sigma^2(p_i)] \quad (22)$$

with covariance given by

$$\sigma^2(p_i) = \nabla g(p_i)' \Sigma(p_i) \nabla g(p_i). \quad (23)$$

In the present case,  $D(p_i|p_0)$  is a totally differentiable function<sup>22</sup> of  $p_i$  with gradient given by

$$\nabla D(p_i|p_0) = \left( \nabla_r D(p_i|p_0) = \frac{\partial D}{\partial p_{ir}} : r = 1, \dots, R - 1 \right) \quad (24)$$

where it follows from (16) that for all  $r = 1, \dots, R - 1$ ,

$$\nabla_r D(p_i|p_0) = \ln \left( \frac{p_{ir}}{1 - \sum_{s=1}^{R-1} p_{is}} \right) - \ln \left( \frac{p_{0r}}{1 - \sum_{s=1}^{R-1} p_{0s}} \right). \quad (25)$$

Hence it follows in particular that when  $p_i \neq p_0$ ,<sup>23</sup>

$$\sqrt{n_i} [D(\hat{p}_i|p_0) - D(p_i|p_0)] \xrightarrow{L} N[0, \sigma^2(p_i|p_0)] \quad (26)$$

with variance given in terms of (21), (24), and (25) by,

$$\sigma^2(p_i|p_0) = \nabla D(p_i|p_0)' \Sigma(p_i) \nabla D(p_i|p_0). \quad (27)$$

To gain further insight into the nature of this limiting distribution, consider the behavior of variance as  $p_i \rightarrow p_0$ . In the limit, when  $p_i = p_0$ , it is clear from (25) that the gradient reduces to the zero vector, and hence from (27) that

$$\sigma^2(p_0|p_0) = 0. \quad (28)$$

Thus even for large  $n_i$  this distribution tends to be degenerate for values of  $p_i$  close to  $p_0$ . The consequences of this will be made clear in Section 3.2 below.

We next obtain a more useful form of this result for purposes of empirical analysis. To do so, observe first that approximate normality of  $\sqrt{n_i} [D(\hat{p}_i|p_0) - D(p_i|p_0)]$  with zero mean and variance,  $\sigma^2(p_i|p_0)$ , implies (by a simple change of variables) that  $D(\hat{p}_i|p_0)$  is also approximately normally distributed with mean  $D(p_i|p_0)$  and variance,  $\sigma^2(p_i|p_0)/n_i$ . Since  $p_i$  is not known, it is usually necessary to estimate this variance by  $\sigma^2(\hat{p}_i|p_0)/n_i$ , i.e., by

$$\text{var} [D(\hat{p}_i|p_0)] = \frac{1}{n_i} \nabla D(\hat{p}_i|p_0)' \Sigma(\hat{p}_i) \nabla D(\hat{p}_i|p_0) \quad (29)$$

when testing hypotheses about the mean,  $D(p_i|p_0)$ . However, in view of the exponential rate of convergence of  $\hat{p}_i$  to  $p_i$  (mentioned above), this estimate turns out to yield a good approximation even for rather small sample sizes.

Finally, it is important to note that a problem arises if  $\hat{p}_{ir} = 0$  for one or more components  $r = 1, \dots, R - 1$ . This is seen already in the asymptotic distribution of expression (20) where the estimated covariance matrix,  $\Sigma(\hat{p}_i)$ , becomes singular. In fact, it turns out that each component  $r$  of  $\hat{p}_i$  with  $\hat{p}_{ir} = 0$  can simply be dropped from the estimation of variance in (29). If we now let

$$\hat{p}_i^+ = \{r=1, \dots, R-1: \hat{p}_{ir} > 0\}, \quad (30)$$

then we obtain the following result [see Appendix B for proof]:

**Theorem 1 (Normal approximation with zero values)** *If the true distribution  $p_i$  of industry  $i$  is not completely dispersed (i.e., if  $p_i \neq p_0$ ) and if the size,  $n_i$ , of this industry is sufficiently large, then the  $D$ -index,  $D(\hat{p}_i|p_0)$ , is approximately normally distributed with mean,  $D(p_i|p_0)$ , and variance,*

$$\text{var}[D(\hat{p}_i|p_0)] = \frac{1}{n_i} \sum_{r,s \in \hat{p}_i^+} \nabla_r D(\hat{p}_i|p_0) \Sigma_{rs}(\hat{p}_i) \nabla_s D(\hat{p}_i|p_0) \quad (31)$$

where  $\nabla_r D(\hat{p}_i|p_0)$  is given by (25), and where  $\Sigma_{rs}(\hat{p}_i) = -\hat{p}_{ir}\hat{p}_{is}$  for all distinct regional pairs,  $r$  and  $s$ , with  $\Sigma_{rr}(\hat{p}_i) = \hat{p}_{ir}(1 - \hat{p}_{ir})$  for the  $r^{\text{th}}$  diagonal term.

It is important to notice here that the true distribution,  $p_i$ , need not have zero components. If all are positive, then for large  $n_i$  one will have  $\hat{p}_i^+ = \{1, \dots, R-1\}$  with probability approaching one.

### 3.2 Confidence intervals for the true $D$ -index

Given the general result in the previous section, observe first that if  $\sigma(\hat{p}_i|p_0) = \sqrt{\sigma^2(\hat{p}_i|p_0)}$ , then for any  $\alpha \in (0, 1)$ , the (large sample)  $100(1 - \alpha)\%$  confidence interval for  $D(p_i|p_0)$  is given by

$$D(\hat{p}_i|p_0) \pm \frac{1}{\sqrt{n_i}} z_\alpha \sigma(\hat{p}_i|p_0) \quad (32)$$

where  $z_\alpha$  is the critical  $\alpha$ -value for  $N(0, 1)$ . Notice in particular, that *for any confidence level, the associated confidence intervals become tighter as  $n_i$  increases. Hence the effect of increasing the number of establishments is to sharpen confidence about the true value of localization for industry  $i$ .* Notice also from (28) above, that for observed values of  $\hat{p}_i$  close to  $p_0$ , the standard deviation  $\sigma(\hat{p}_i|p_0)$  in (32) will be very small, thus yielding very tight confidence bounds. So even for small sectors, it may be possible to detect slight deviations from complete spatial dispersion.

To illustrate the accuracy of this approximation, the histogram in Figure 1 shows 1000 simulated draws from the sampling distribution of  $D(\hat{p}_i|p_0)$  for a simple example with  $n_i = 100$ ,  $R = 4$ ,  $p_0 = (0.3, 0.1, 0.2, 0.4)$ , and  $p_i = (0.1, 0.2, 0.4, 0.3)$ .

Hence even for a relatively small sector of 100 establishments, the normality of this sampling distribution is evident. The middle  $\blacktriangle$  shows the location of the true value of  $D(p_i|p_0)$  ( $= 0.2197$ ) for this example. Here we focus on the standard case of  $\alpha = 0.05$  with associated 95% confidence interval given by

$$D(\hat{p}_i|p_0) \pm \frac{1.96}{\sqrt{n_i}} \sigma(\hat{p}_i|p_0). \quad (33)$$

The  $\blacktriangle$ 's to either side of  $D(p_i|p_0)$  in the figure show the mean values (0.0980, 0.3414) of the end points of the confidence interval in (33). The simulated confidence value for this 95% interval [i.e., the percent of

simulated confidence interval estimates containing the true value of  $D(p_i|p_0)$  was 95.3%, indicating that (under the assumption of random-sampling) this asymptotic approximation is quite good.<sup>24</sup>

### 3.3 Hypothesis testing

Turning next to hypothesis testing, observe first that the results above do *not* allow a direct test that industry  $i$  is completely dispersed. In fact the asymptotic normality property of  $D(\hat{p}_i|p_0)$  [in (26) and (27)] fails to hold under the null hypothesis that  $p_i = p_0$ . There are of course a host of other tests (such as the likelihood-ratio test or chi-square goodness-of-fit test) that could easily be applied here (refer to Section 2.3). But for our present purposes, the null hypothesis of “complete spatial dispersion” in (2) is not of much interest by itself.<sup>25</sup> Rather, it is meant to serve as a benchmark against which the *relative dispersion* (or *relative localization*) between industries can be compared. In particular, if for a given pair of industries,  $i$  and  $j$ , it is observed that  $D(\hat{p}_i|p_0) > D(\hat{p}_j|p_0)$ , then (as mentioned above) the question of the most interest is whether location in industry  $i$  is *significantly* more localized than industry  $j$ . Here it is appropriate to test the null hypothesis,  $D(p_i|p_0) - D(p_j|p_0) = 0$ , against the one-sided alternative,  $D(p_i|p_0) - D(p_j|p_0) > 0$ . As an extension of our random-sampling assumption for each industry, we now *assume that  $D(\hat{p}_i|p_0)$  and  $D(\hat{p}_j|p_0)$  are independently distributed*, and hence that their difference,  $D(\hat{p}_i|p_0) - D(\hat{p}_j|p_0)$ , is also asymptotically normally distributed with mean,  $D(p_i|p_0) - D(p_j|p_0)$ , and variance

$$\text{var} [D(\hat{p}_i|p_0) - D(\hat{p}_j|p_0)] = \text{var} [D(\hat{p}_i|p_0)] + \text{var} [D(\hat{p}_j|p_0)] = \frac{1}{n_i} \sigma^2(\hat{p}_i|p_0) + \frac{1}{n_j} \sigma^2(\hat{p}_j|p_0). \quad (34)$$

The  $P$ -value for this one-sided test is then given by

$$P_{ij} = 1 - \Phi \left[ \frac{D(\hat{p}_i|p_0) - D(\hat{p}_j|p_0)}{\sqrt{\frac{1}{n_i} \sigma^2(\hat{p}_i|p_0) + \frac{1}{n_j} \sigma^2(\hat{p}_j|p_0)}} \right] \quad (35)$$

where  $\Phi$  represents the cumulative distribution of a standard normal random variable. As an illustration of this test, consider a second example with the same four regions as in the first example, so that again  $p_0 = (0.3, 0.1, 0.2, 0.4)$ . In addition, suppose industry  $j$  has the same distribution as in the first example,  $p_j = (0.1, 0.2, 0.4, 0.3)$ , and that industry  $i$  has establishment distribution,  $p_i = (0.1, 0.4, 0.2, 0.3)$ . Here industry  $i$  has a larger fraction of establishments concentrated in the smallest region,  $r = 2$ , so that the  $D$ -index now increases from  $D(p_j|p_0) = 0.2197$  to  $D(p_i|p_0) = 0.3584$ . Finally suppose the numbers of establishments are given respectively by  $n_i = 1000$  and  $n_j = 500$ , so that both industries are still relatively small.<sup>26</sup> Here for a test of size  $\alpha = 0.05$ , the results of 1000 simulated samples produced an estimated power of 0.946 (i.e., the null hypothesis was correctly rejected 94.6 % of the time). In addition,

the estimated size of this test under the null hypothesis,  $p_i = p_j$ , was  $\hat{\alpha} = 0.055$ .<sup>27</sup> So again, the normal approximation continues to be working well here.

Finally, it should again be emphasized that the above testing results are based on the strong assumption of independent random sampling. However, it is important to note that in the presence of positively correlated location patterns between industries  $i$  and  $j$ , the present testing procedure errs on the *conservative* side. In particular, if it can be concluded under independence that industry  $i$  is significantly more concentrated than industry  $j$ , then this conclusion will continue to hold in the presence of positive correlation (co-localization tendency of the two industries). To see this, it is enough to observe that if  $D(\hat{p}_i|p_0)$  and  $D(\hat{p}_j|p_0)$  were positively correlated, then a proper estimate of variance in (34) would be obtained by subtracting a positive-covariance term from the right hand side (yielding a smaller estimated variance). This in turn would increase the second term on the right hand side of (35), resulting in an even smaller  $P$ -value. Hence *if differences in  $D$ -indices are significant under independence, they would be even more significant in the presence of positive correlation.*

## 4 Localization of industries in Japan

In this section, we apply our  $D$ -index to private manufacturing industries in Japan at the 3-digit level.<sup>28</sup> In Section 4.1, we present the  $D$ -index based on economic areas<sup>29</sup> evaluated at the county level.<sup>30</sup> This  $D$ -index is then decomposed in Section 4.2 into parts explained by localization at various regional levels (involving bundles of counties). We also examine changes in the localization levels of industries between 1981 and 1999 in Section 4.3. Finally, our  $D$ -index is compared with Ellison and Glaeser's raw index,  $G$ , in Section 4.4.

### 4.1 Localization within the nation

Figure 2 shows the distribution of  $D$ -values for 3-digit manufacturing industries within the nation in 1999. Note that the  $D$ -values for almost *all* individual manufacturing industries (96%) are greater than that for manufacturing as a whole (with  $D$ -value 1.11). Moreover these differences are all quite significant in terms of the  $P$ -values in (35).<sup>31</sup> This is mainly due to the fact that while manufacturing as a whole is quite dispersed, individual industries tend to be concentrated in a small number of regions, reflecting regional specialization of these industries. But as already noted,<sup>32</sup> those industries at the low end of the scale tend to have the largest number of establishments, and hence contribute substantially to aggregate manufacturing as a whole.

Table 1 lists the most localized and ubiquitous 3-digit industries (with more than 100 establishments) at the national level.<sup>33</sup> The small standard errors in the table indicate that (under independence) sample

Figure 1: Sampling distribution for the estimated  $D$ -index

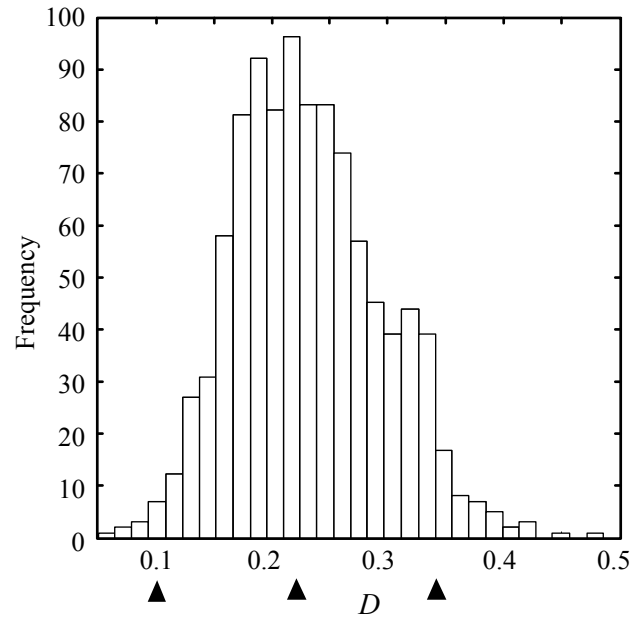
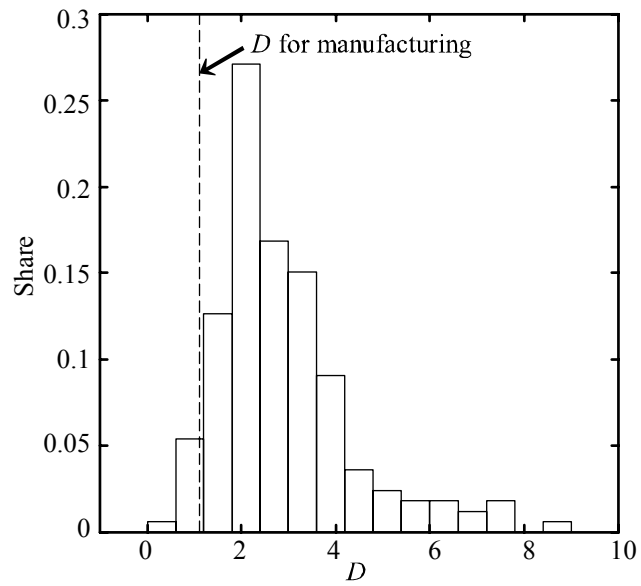


Figure 2: Distribution of the  $D$ -values for the 3-digit manufacturing industries in 1999



sizes are large enough to yield sharp estimates of the true  $D$ -values for individual industries.<sup>34</sup> Notice also that (as already suggested in Section 3.3) the estimates of the  $D$ -indices for ubiquitous industries appear to be quite sharp.

Regional resource endowments seem to play an important role in determining the location of industries. For instance, lacquer ware (JSIC346) is concentrated in regions that are endowed with abundant sumac trees. Similarly, the concentration of pottery (JSIC254) in Toki and Tajimi (Gifu prefecture) and Seto (Aichi prefecture) can be explained by the availability of high quality clay.

In contrast to these industries, the apparel-related industries (JSIC154, 232, 241, 243, 244, 249) and publishing- and printing-related industries (JSIC192, 199, 34C) are highly localized in the largest metro areas, Tokyo, Osaka and Nagoya. Hence agglomeration economies based on demand/production externalities may be more relevant in these cases.<sup>35</sup>

When production externalities are industry specific (such as knowledge spillovers of specially skilled workers), the location of industrial concentrations may be determined almost completely by historical accident. The most localized industry, leather glove and mittens (JSIC245), is mostly concentrated in three small counties, Hikita, Shiratori and Ohuchi (Kagawa prefecture) with a total population of only 38,000. Together these three counties account for more than 90% of all leather glove manufacturing in Japan. Similarly, a small town called Sabae (Fukui prefecture) with only 65,000 inhabitants accounts for more than 90% of all frames for eyeglasses [classified in ophthalmic goods (JSIC326)] manufactured in Japan (and in fact, 20% in the world). But, in both of these cases, there are no strong reasons other than historic why such dramatic industrial concentrations should be found in these locations.

Finally, plant level scale economies may also lead to high localization of industries. This is well illustrated by the case of petroleum refining (JSIC211) along the Pacific coast.

Turning next to ubiquitous industries, transport costs seem to play a major role in the dispersion of industrial activities. Livestock products (JSIC121), bakery and confectionery products (JSIC127) are perishable products subject to high transport costs. Alcoholic beverages (JSIC132) such as beer are typical weight/bulk gaining industries. Thus, their location follows distribution of consumers. The location patterns of industries supplying household products (JSIC308, 347) can be understood similarly.

The ubiquity of cement and its products (JSIC252), may be due to Japan-specific policies promoting government spending on construction of ubiquitous public facilities and roads.<sup>36</sup> Similar arguments can be made for the ubiquity of sawing, planing mills and wood products (JSIC161), sliding doors and screens (JSIC173), fabricated constructional and architectural metal products (JSIC284) .

While the localization tendencies above are all in terms of 3-digit industrial classification, similar properties are also observed at the 2-digit level. As shown in Table 2,<sup>37</sup> all sectors except two exhibit

Table 1: The most localized and ubiquitous 3-digit industries

JSIC	Industry	$D$	Standard error
The most localized industries			
245	Leather gloves and mittens	6.061	0.097
326	Ophthalmic goods, including frames	5.699	0.060
232	Rubber and plastic footwear and its findings	5.615	0.056
243	Cut stock and findings for boots and shoes	5.293	0.069
241	Leather tanning and finishing	5.229	0.062
249	Miscellaneous leather products	4.904	0.077
141	Silk reeling plants	4.569	0.170
199	Service industries related to printing trade	4.333	0.083
192	Publishing industries	4.298	0.036
244	Leather footwear	4.169	0.060
154	Fur apparel and apparel accessories	4.161	0.106
346	Lacquer ware	4.102	0.039
34C	Information recording materials, except newspapers, books, other printed products, etc.	4.084	0.094
254	Pottery and related products	3.910	0.025
211	Petroleum refining	3.889	0.123
The most ubiquitous industries			
252	Cement and its products	0.463	0.008
161	Sawing, planing mills and wood products	0.706	0.010
129	Miscellaneous foods and related products	0.725	0.008
173	Sliding doors and screens	0.792	0.009
121	Livestock products	0.964	0.015
127	Bakery and confectionery products	1.042	0.013
123	Canned and preserved fruit and vegetable product	1.095	0.018
124	Seasonings	1.129	0.017
132	Alcoholic beverages	1.137	0.018
284	Fabricated constructional and architectural metal products	1.186	0.008
308	Electronic parts and devices	1.214	0.011
347	Sundry goods of straw ("tatami") mats, umbrellas and other daily commodities.	1.262	0.017
224	Foamed and reinforced plastic products	1.337	0.018
301	Electrical generating, transmission, distribution and industrial apparatus	1.396	0.012
34D	Manufacturing industries, n.e.c.	1.428	0.02

higher localization than manufacturing ( $D > 1.11$ ).<sup>38</sup> Notice also that at both extremes, i.e., the most localized and the least localized 2-digit sectors, the average  $D$ -values of their 3-digit subsectors are much higher [all are significantly different from those of the corresponding 2-digit sectors according to the  $P$ -values in (35)]. These observations suggest that *localizations of sectors and their subsectors tend to have different spatial extents: each county tends to be specialized in only a few 3-digit subsectors, and those counties with subsector specializations in the same 2-digit sector tend to be close in space, thus forming larger regions specialized in these 2-digit sectors*. For instance, Aichi prefecture has a pair of counties which are among the ten counties with largest concentrations of establishments in structural clay products (JSIC253) and another pair in pottery and related products (JSIC254), both of which belong to the same 2-digit sector of ceramic, stone and clay products (JSIC25).

Aside from regional specialization, however, higher  $D$ -values for disaggregate industries may in part be due simply to industrial classifications which sometimes group industries together that are governed by very different locational determinants. For instance, the fifth most concentrated 2-digit sector is precision instruments and machinery (JSIC32) which contains ophthalmic goods (JSIC326) localized in Sabae, as mentioned above. In addition, however, this same sector also contains watches, clocks, clockwork-operated devices and parts (JSIC327) which has no strong economic linkages with ophthalmic goods (JSIC326).

## 4.2 Spatial decomposition of the $D$ -index

In this section, we develop a hierarchical decomposition of the  $D$ -index utilizing eq. (13). Starting with counties as our basic geographic units (3363 in total), we construct a more economically meaningful aggregate unit, designated as a *metro area*. Each metro area consists of a set of counties representing the employment area for a common business core.<sup>39</sup> Based on this definition, we identified 359 metro areas for year 2000.<sup>40</sup> Within each metro area we then identified those constituting the *business area*, and designated all others as the *residential area*.<sup>41</sup> Finally, at a higher level, we defined the set of all counties in metro areas (2485 in number) to be the *urban area* for Japan, and designated the set of all other counties (878 in number) as the *rural area*.

Table 3 summarizes the contribution of each level in this hierarchy to the  $D$ -index for manufacturing industries at the 1-digit, 2-digit, and 3-digit levels (where by definition the 1-digit level corresponds to all manufacturing).

To illustrate how these values are obtained, it is convenient to focus on the higher level decompositions only. For any manufacturing industry  $i$  (at either the 1-, 2-, or 3-digit level) the first decomposition of  $D$

Table 2: The most localized and ubiquitous 2-digit sectors

JSIC	Industry	$D$	Average $D$ of 3-digit subsectors
	The most localized sectors		
24	Leather tanning, leather products and fur skins	2.878	4.827
23	Rubber products	2.711	3.503
14	Textile mill products, except apparel and other finished products made from fabric and similar materials.	2.188	3.510
19	Publishing, printing and allied industries	2.026	3.462
32	Precision instruments and machinery	1.96	3.314
	The most ubiquitous sectors		
12	Food	0.652	1.579
16	Lumber and wood products	0.811	1.500
17	Furniture and fixtures	1.025	1.819
13	Beverages, tobacco and feed	1.128	2.248
30	Electrical machinery, equipment and supplies	1.238	1.971

Table 3: Average share (%) of each regional level in the  $D$ -index (standard errors in parentheses)

	between urban and rural	within urban				within rural
		among metro	within metro			
			between business district and residential area	within business area	within residential area	
(1) manufacturing	11.5	45.5	18.4	14.0	6.1	4.5
(2) 2-digit	8.3 (0.74)	43.9 (1.65)	13.1 (0.82)	16.8 (1.29)	10.3 (0.96)	7.7 (1.71)
(3) 3-digit	6.1 (0.23)	44.0 (0.78)	11.5 (0.28)	17.9 (0.51)	13.2 (0.48)	7.4 (0.63)

with respect to the urban-rural partition can be written as

$$D(\widehat{p}_i|p_0) = D(\widehat{q}_i|q_0) + \{\widehat{q}_{i,urban}D(\widehat{p}_{i|urban}|p_{0|urban}) + \widehat{q}_{i,rural}D(\widehat{p}_{i|rural}|p_{0|rural})\} \quad (36)$$

where  $\widehat{q}_i = (\widehat{q}_{i,urban}, \widehat{q}_{i,rural})$  now denotes the observed urban-rural distribution of industry  $i$ , and where  $\widehat{p}_{i|urban}$  and  $\widehat{p}_{i|rural}$  denote the observed conditional distributions of industry  $i$  across the counties in the urban and rural areas, respectively. This can be extended to a second level of decomposition by considering the metro-area partition defining the urban area. If these metro areas are enumerated as  $m = 1, \dots, M$ , then the term,  $D(\widehat{p}_{i|urban}|p_{0|urban})$ , can be decomposed further as follows:

$$D(\widehat{p}_{i|urban}|p_{0|urban}) = D(\widehat{q}_{i|metro}|q_{0|metro}) + \sum_{m=1}^M \widehat{q}_{im}D(\widehat{p}_{i|m}|p_{0|m}) \quad (37)$$

where  $\widehat{q}_{i|metro} = (\widehat{q}_{im} : m = 1, \dots, M)$  now denotes the observed distribution of industry  $i$  across metro areas, and  $\widehat{p}_{i|m}$  denotes the observed conditional distribution of  $i$  across the counties in metro area  $m$ .

To evaluate this decomposition, it is convenient to focus on all manufacturing (1-digit level) where  $i$  denotes the single ‘‘aggregate’’ manufacturing industry. Here the ‘‘between urban and rural’’ share of the  $D$ -index is given in terms of (36) and (37) by

$$\frac{D(\widehat{q}_i|q_0)}{D(\widehat{p}_i|p_0)} \times 100 = 11.5 \quad (38)$$

and similarly, the ‘‘within rural’’ share is given by

$$\frac{\widehat{q}_{i,rural}D(\widehat{p}_{i|rural}|p_{0|rural})}{D(\widehat{p}_i|p_0)} \times 100 = 4.5. \quad (39)$$

The ‘‘within urban’’ share then consists of the remainder:

$$\frac{\widehat{q}_{i,urban}D(\widehat{p}_{i|urban}|p_{0|urban})}{D(\widehat{p}_i|p_0)} \times 100 = 100 - (11.5 + 4.5) = 84. \quad (40)$$

This latter share can be further decomposed as in (37). Here the ‘‘among metro’’ share is given by

$$\frac{\widehat{q}_{i,urban}D(\widehat{q}_{i|metro}|q_{0|metro})}{D(\widehat{p}_i|p_0)} \times 100 = 45.5 \quad (41)$$

and the ‘‘within metro’’ share is given by

$$\frac{\widehat{q}_{i,urban} \sum_{m=1}^M \widehat{q}_{im}D(\widehat{p}_{i|m}|p_{0|m})}{D(\widehat{p}_i|p_0)} \times 100 = 84 - 45.5 = 38.5. \quad (42)$$

Lower (i.e., finer) level decompositions can be obtained in a similar manner.

At all industry aggregation levels, localization “among metro areas” explains nearly half of the  $D$ -value. Notice however that localization “within metro areas” is also fairly large. In addition, notice that as industry classification becomes finer, the average shares of concentration “within business areas” as well as “within residential areas” increase, while the average shares “between business areas and residential areas” decrease. Since the average  $D$ -value increases as classification becomes finer (1.1, 1.9, and 2.9, respectively, for 1-digit, 2-digit, and 3-digit industries), the decrease in average shares “between business areas and residential areas” shows that this increased localization is not simply attributable to greater concentration in business areas. In fact, the increased shares for both “within business areas” and “within residential areas” show that this distinction itself is less clear for finer classifications of industries.

Now, let us look across 3-digit industries at the relationship between the  $D$ -value and its regional components. Table 4 shows the correlation<sup>42</sup> between the log of the  $D$ -index and its corresponding shares for various decomposition levels. Observe that the share of concentration “among metro areas” and that “within business areas” are both positively correlated with  $D$ . This reflects the fact that *more localized industries are found not only in a fewer number of metro areas, but also in a fewer number of counties within the corresponding business areas.*

This localization pattern can also be seen in relationship between the population size of a metro area and the degree of localization of industries found in that metro area. Figure 3 plots the highest, median, and lowest  $D$ -values for all industries located in each metro area against the population size of that metro area.<sup>43</sup> The correlations of these three values with the population size of a metro area are 0.52, 0.20 and  $-0.31$ , respectively. The plot roughly suggests that *a larger metro area attracts both more localized and ubiquitous industries, while smaller metro areas tend to contain mostly ubiquitous industries.* This, in turn, is roughly consistent with Christaller’s (1933) well-known *Hierarchy Principle* of industrial locations, namely that industries present in a given metro area are also present in all the larger metro areas.<sup>44</sup>

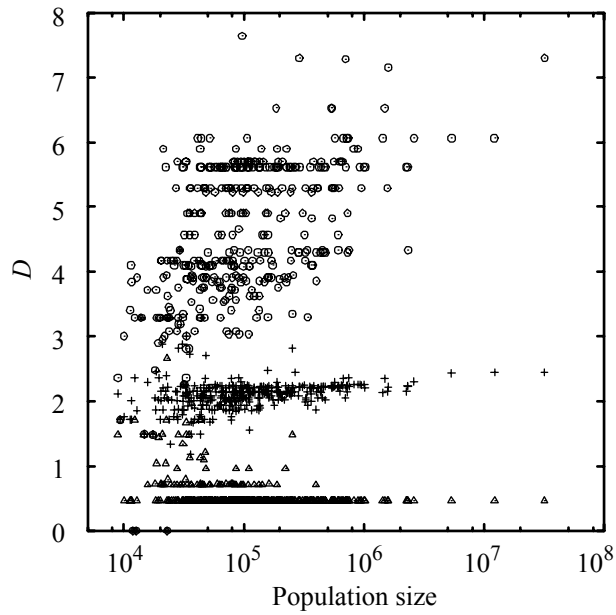
### 4.3 Changes in the industrial localization over time

In this section, we look at changes in the degree of localization of 3-digit manufacturing industries in Japan between 1981 and 1999. Since this industrial classification has been disaggregated for most sectors between these two periods, we have attempted to reconcile the two classifications by aggregating the 1999 classification to that of 1981, resulting in 148 industries. The  $D$ -index of manufacturing as a whole decreased by more than 10% during this period. Among individual industries,  $D$ -values decreased for 63% of these industries (refer to Figure 4).<sup>45</sup>

Table 4: Correlation between  $D$  and the share of  $D$  at each region level (3-digit)

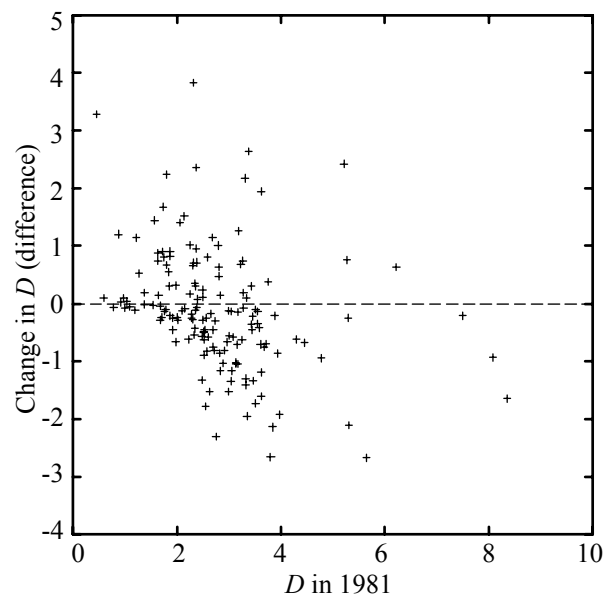
urban-rural	among metro	between business area and residential area	within business area	within residential area	within rural area
-0.46	0.45	-0.25	0.34	-0.43	-0.22

Figure 3: Population size and industrial structure of metro areas



Note: Markers  $\circ$ ,  $+$  and  $\triangle$  indicate respectively the highest, median and lowest  $D$ -values of industries located in each metro area.

Figure 4: Change in the  $D$ -value between 1981 and 1999



As depicted in Figure 5, an increase in  $D$  appears to be associated with an increase in the share of concentration among metro areas (with correlation of 0.39, significant at 1% level). No other changes in shares were found to be significantly related to changes in  $D$ . This result suggests that changes in the localization of industries take place mainly at the level of metro areas.

#### 4.4 Comparison with Ellison and Glaeser's $G$

Finally, in this section, we compare our  $D$ -index with the widely used “ $G$ ” index of localization proposed by Ellison and Glaeser (1997). In particular if we now let  $M_{ir}$  denote the *employment level* of industry  $i$  in region  $r$  (rather than the number of establishments), and let the *total employment level* of industry  $i$  be denoted by  $M_i = \sum_r M_{ir}$ , then the share of region  $r$  in the employment of industry  $i$  is given by  $s_{ir} = M_{ir}/M_i$ . In this case, an alternative reference measure is given by the share of total industrial employment in each region  $r$ , i.e., by  $x_r = \sum_j M_{jr}/\sum_j M_j$ . If sum-of-squared errors is adopted as the relevant measure of deviation, then the  $G$ -index of localization for industry  $i$  is given by:

$$G(s_i, x) \equiv \sum_r (s_{ir} - x_r)^2 \quad (43)$$

where  $s_i = (s_{ir} : r \in \mathbf{R})$  and  $x = (x_r : r \in \mathbf{R})$ .

It should also be noted here that the popular “ $\gamma$ ” index proposed by Ellison and Glaeser (1997) is based on  $G$ . More precisely,  $\gamma$  extends  $G$  to include establishment-size effects within each industry. Unfortunately, since establishment-size data is not publicly available in Japan (as well as in many other countries), we can only make empirical comparisons of  $D$  with  $G$ , and not  $\gamma$ . However, it should be emphasized that the basic measure of localization in  $\gamma$  is still given by  $G$ , and in fact that there is often little practical difference between the two. To see this, let  $S = \sum_r x_r^2$  denote the regional “employment-diversity” index, and for each industry  $i$  with establishment sizes  $(M_{ij} : j = 1, \dots, m_i)$ , let  $H_i = \sum_{j=1}^{m_i} (M_{ij}/M_i)^2$  denote the “establishment-size-diversity” index for  $i$ . Then the  $\gamma$ -index for industry  $i$  is defined by<sup>46</sup>

$$\gamma_i = \frac{G_i - (1 - S)H_i}{(1 - S)(1 - H_i)} = a_i G_i - b_i \quad (44)$$

where  $a_i = 1/[(1 - S)(1 - H_i)]$ ,  $b_i = H_i/(1 - H_i)$ , and where  $G_i$  denotes the  $G$ -index for industry  $i$ . Hence for any fixed  $S$  and  $H_i$ , this index is seen to be a simple positive affine transformation of  $G_i$ . In particular this implies that if each industry were to have the same number of establishments, all of equal size within each industry, then this affine function would be the same for all industries. Under these conditions, the orderings of  $\gamma_i$  and  $G_i$  values would be *identical*. More generally, if each industry has many establishments, all of which are roughly comparable in size, then values of  $H_i$  are not only similar, they must all be very close to zero.<sup>47</sup> Hence one can expect that  $a_i \approx 1/(1 - S)$  and  $b_i \ll G_i/(1 - S)$ , so

that again the orderings of  $\gamma_i$  and  $G_i$  will be almost the same.<sup>48</sup>

Figure 6 depicts the relationship between values of the  $G$  and our  $D$  for all 3-digit manufacturing industries in 1999. It is not surprising that these indices exhibit substantial disagreement on a case-by-case basis, given the fundamental difference between their implicit notions of “localization.” While they do have a fairly high positive correlation [0.5 between  $\log(G)$  and  $D$ ], it should be emphasized that this does not imply similar behavior of the two indices (see Appendix A for details).

One reason for this positive correlation in the Japanese case may be that more localized industries (in terms of  $D$ ) tend to have smaller employment shares, as seen in Figure 7.<sup>49</sup> Also recall (Section 2.2, see also Appendix A) that *for a given regional employment distribution within an industry, smaller total employment levels yield higher degrees of localization according to the  $G$ -index*. It is this relation between employment shares and degrees of localization for the  $G$ -index that appears to be creating this positive correlation. However, while the  $D$ -value is independent of the size variation of industries, the ordering of  $G$ -values for industries can change drastically as relative sizes change (see Appendix A for further details).

Finally it should be emphasized that the above comparison of the  $G$ -index with our  $D$ -index is based directly on Ellison and Glaeser’s definition of  $G$ . While this is the simplest and most obvious comparison to make, it can be argued that such comparisons implicitly involve a number of different dimensions:

- (1) What “industry-size distributions” are used?
- (2) What regional “reference distributions” are used?
- (3) What “functional forms” are used to compare these distributions?

Here it turns out that  $D$  and  $G$  differ in all three dimensions: not only are their functional forms different, but also their industry-size distributions ( $i$ -establishment shares,  $\hat{p}_{ir}$ , for  $D$  versus  $i$ -employment shares,  $s_{ir}$ , for  $G$ ), and their regional reference distributions (economic area shares,  $p_{0r}$ , for  $D$  versus total employment shares,  $x_r$ , for  $G$ ). One referee described this as comparing “apples with oranges.” So why not construct a range of “intermediate” indices by modifying either  $D$  or  $G$  (or both) to yield pairs of indices that are more comparable? For example, to achieve greater comparability with respect to dimensions (1) and (2) above, one might modify  $G$  to involve comparisons of  $\hat{p}_{ir}$  with  $p_{0r}$  [i.e.,  $G' = \sum_r (\hat{p}_{ir} - p_{0r})^2$ ], or modify  $D$  to involve comparisons of  $s_{ir}$  and  $x_r$  [i.e.,  $D' = \sum_r s_{ir} \ln(s_{ir}/x_r)$ ]. Such modifications are certainly of interest, and may ultimately help to clarify the relationships between these different dimensions. But for our present purposes, we choose not to attempt such an ambitious program.

Figure 5: Change in  $D$  and its inter-metro area share

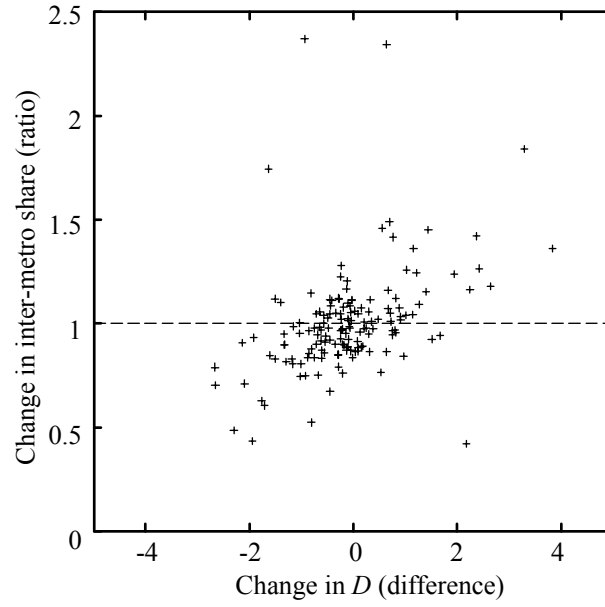
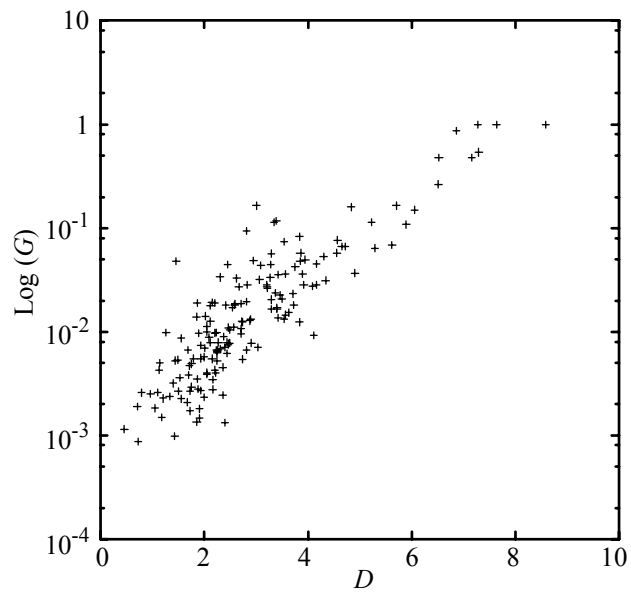


Figure 6: Relationship between  $D$  and Ellison and Glaeser's  $G$



## 5 Concluding remarks

In this paper we have developed a  $D$ -index of industrial localization based on Kullback-Leibler divergence. This measure not only has a natural interpretation in terms of relative likelihoods, but also has a simple asymptotic normal distribution theory that allows the construction of both confidence bounds on the degree of localization for individual industries and tests of the relative degree of localization between industries. As shown in Section 4.3, this testing procedure can also be applied to detect temporal changes in the degree of localization for individual industries.

Given these positive features of the  $D$ -index, it is important to emphasize that it does have certain limitations. First, with respect to the random sampling framework used to motivate this index, it should be clear that actual industrial location behavior is in fact a dynamical process in which the location of any new establishment tends to depend quite heavily on the locations of existing ones.<sup>50</sup> Moreover, given the prevalence of multi-plant firms, our independence assumption for plant locations is at best a convenient fiction. In particular, this implies that our proofs of asymptotic normality for  $D$  [see eqs.(4, 5)] may not hold. However, it should be emphasized that asymptotic normality itself is much more robust, and hence that our hypothesis tests may still be reasonable as long as dependencies between establishment are not “too strong.”<sup>51</sup>

Finally, since no explicit geographic relationships among regional units are embodied in  $D$ , this index can at best give only a limited indication of the actual spatial extent of localization (as illustrated in Section 4.2 above). For example, concentrations of establishments in contiguous counties may yield the same  $D$ -values as concentrations in widely separated counties. Hence, while the former suggests the possibility of a larger geographic concentration of establishments, this can only be captured by appropriate spatial decompositions of  $D$ . Such issues are more appropriately addressed by models at the establishment level, where modifiable areal unit problems do not arise [as for example in Duranton and Overman (2005) and Marcon and Puech (2003)].

# Appendix.

## A Choice of reference distribution

In essence, our  $D$  index amounts to a measure of deviation from a given reference distribution representing an implicit null hypothesis about industrial location patterns. However, there are other choices for both the reference distribution used and the measure of deviation from that distribution. Hence it is appropriate in this appendix to compare our choice with one of the most popular measures of this type, the  $G$ -index of Ellison and Glaeser (1997), introduced in Section 4.4, where an alternative reference measure is given by the share of each region in all industrial employment.<sup>52</sup>

An obvious motivation for measuring localization in relative terms is to remove the effect of regional population size. For if there is significant size variation among regions, then it is natural to expect to find larger employment levels for industries in larger regions. But unless all the regions have identical shares of each industry, this normalization can distort the index in certain systematic ways. For under this type of normalization, those industries with a larger share of total national employment will appear to be more ubiquitous. This is mainly due to the fact that the employment share distribution,  $s_{ir}$ , for large industries  $i$  will then be very similar to the total regional employment shares,  $x_r$ . At the other extreme, small industries are likely to appear to be very localized – even if their employment distribution is in fact quite uniform across regions. These observations are well illustrated by the following two simple examples.

### Regional diversity effect

Difference in industrial diversity among regions can lead to ambiguities in the interpretation of the  $G$ -index.<sup>53</sup> In particular, while large metro areas, like Tokyo, tend to have greater employment in each industry than do small metro areas, the employment shares of individual industries tend to be smaller in large metro areas simply because of the greater industrial diversity in those areas. Hence, indices based on relative concentration, such as  $G$ , tend to undervalue the localization of industries in large metro areas, even when their absolute levels of employment are quite high.

To see this, consider the following *core-periphery system of two regions*, where the core region has both manufacturing and agriculture, while the periphery has only agriculture (as illustrated in Figure 8). Here, manufacturing is a “localized” sector (of size  $1 - q$ , where  $0 < q < 1$ ) completely concentrated in region  $B$ , while agricultural is a “ubiquitous” sector (of size  $q$ ) uniformly distributed over the two regions. But under index  $G$  we have  $G_{localized} = \frac{1}{2}(1 - q)^2$  and  $G_{ubiquitous} = \frac{1}{2}q^2$ , where the subscripts denote the “localized” manufacturing sector and “ubiquitous” agricultural sector respectively. Hence

Figure 7: Relationship between  $D$  and employment share of an industry

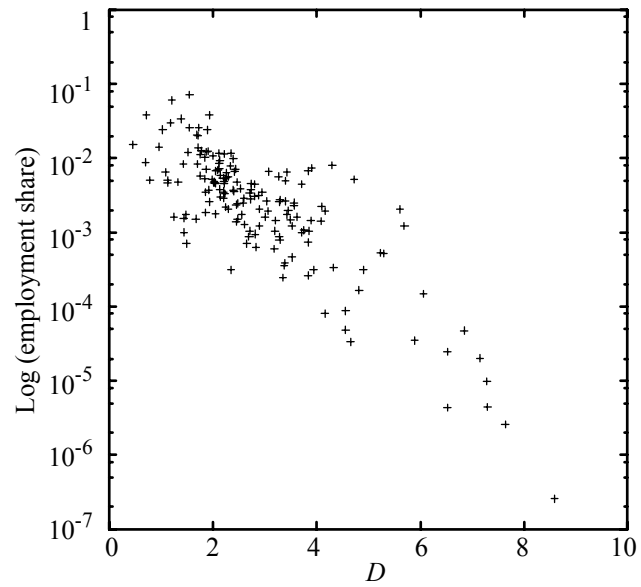
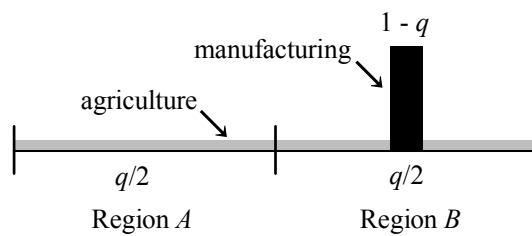


Figure 8: Core-periphery regional system



$G_{ubiquitous} \gtrless G_{localized}$  if  $q \lesseqgtr 1/2$ , and it follows that when the “ubiquitous” sector is relatively small, the  $G$ -index *always* evaluates this sector as more localized than the “localized” sector. Thus, by completely ignoring the spatial aspect of the employment distribution, and by defining “localization” solely in terms of the underlying population (or employment) distribution, a spatially concentrated sector may appear to be *more ubiquitous* than a spatially dispersed sector.

It should also be noted that this same example can be applied to the  $\gamma$ -index of Ellison and Glaeser. For if we assume (not unreasonably) that there are a large number of small “ubiquitous” farms and a large number of small “clustered” manufacturing establishments in Region B, then (as argued in Section 4.4 above) the ordering of  $\gamma$  and  $G$  should be the same in this case.

Note finally for this example that if the reference distribution in our  $D$ -index is replaced by regional employment shares [i.e., if  $D(p_i|p_0)$  is replaced by  $D(s_i|x)$ ] then it can be easily verified that  $D_{localized} = 0$  and  $D_{ubiquitous} = \log(2)$  for *any* choice of  $q$ . So the same problem arises, and it is clear in this case that it is the choice of *reference distributions* (and not the measure of deviation) that is creating this anomaly.

### Industry size effect

Next, we examine the effect of industry size on alternative indices of localization. Again consider an economy with two regions and two industries. Suppose that each industry is relatively concentrated in a different region, and that the interregional distribution of employment is symmetric between the two, i.e., that each region has the same share of total employment in its specializing industry. Specifically, let the total employment for industry 1 [resp., 2] be given by  $q$  [resp.,  $1 - q$ ], where  $0 \leq q \leq 1$ . In addition, let the share of region A [resp., B] in the employment for industry 1 [resp. 2] be given by  $p$ , where  $0 \leq p \leq 1$  (as summarized in Table 5 below). Note that the regional shares of employment for the two industries are symmetric (i.e., region A has a share,  $p$ , of industry 1 and region B has the same share,  $p$ , of industry 2).

Table 5: Employment concentration of industries with different size

	industry 1	industry 2
region A	$pq$	$(1 - p)(1 - q)$
region B	$(1 - p)q$	$p(1 - q)$

The values of  $G$  for industries 1 and 2 are given respectively by  $G_1 = 2(2p - 1)^2(1 - q)^2$  and  $G_2 = 2(2p - 1)^2q^2$ . Here it can readily be verified that  $G_1 > G_2$  if  $p \neq 1/2$  and  $q < 1/2$ . Hence, unless employment for each industry is evenly distributed across regions, *the smaller industry is always evaluated as more concentrated*. As discussed above, this example gives a dramatic illustration in which the total employment distribution is always “closer” to that of the larger industry, thus making that industry appear more ubiquitous.<sup>54</sup> Note finally that the argument for  $\gamma$  made in the last example again

shows that if both industries have a large number of small firms, then the ordering of  $\gamma$  and  $G$  should be the same, so that  $\gamma$  tends to exhibit the same problem.

More generally it should be emphasized that size variations among industries may be due to fundamental structural differences between these industries, such as production technology or market structures. Hence it is our view that indices of industrial localization should not depend on such industry-specific characteristics. In particular, our  $D$ -index is independent of sector size (and in the example above we have  $D_1 = D_2$  for any  $p$  and  $q$ ). However, if the reference distribution for  $D$  is set to the distribution of aggregate industry [i.e., if  $D(p_i|p_0)$  is replaced by  $D(s_i|x)$ ], then it can be verified that  $D_1 \gtrless D_2 \Leftrightarrow p \gtrless 1/2$  when  $q < 1/2$ . Hence *relative specialization influences the degree of localization*. More specifically, even if interregional distribution is symmetric for industries  $i$  and  $j$  (as in this example), if the region where industry  $i$  is localized is relatively more specialized in  $i$  than the other region is in  $j$ , then industry  $i$  is evaluated to be more localized than  $j$ . Hence for the  $D$ -index, this choice of reference distributions can in some cases create a confusion between the degree of specialization for industries and their degree of geographic concentration.

## B Proof of Theorem 1

From the derivation in Section 3.1, we know that for large  $n_i$  the  $D$ -index,  $D(\hat{p}_i|p_0)$ , is approximately normally distributed with mean,  $D(p_i|p_0)$ , and variance,  $\sigma^2(\hat{p}_i|p_0)/n_i$  (eq., (29)) if  $\hat{p}_{ir} > 0$  for all  $i$  and  $r$ . Thus, we are left to show that each component  $r$  of  $\hat{p}_i$  with  $\hat{p}_{ir} = 0$  can simply be dropped from the estimation of variance (29).

In the present case this is less of a problem since we are only considering a one-dimensional function of  $\hat{p}_i$ . Here it suffices to establish a limiting form for  $\text{var}[D(\hat{p}_i|p_0)]$  when one or more components of  $\hat{p}_i$  approaches zero. To do so, let us simplify the notation to

$$g(x) = \nabla D(x|p_0) \tag{45}$$

and consider the limit of the quadratic form,  $\nabla D(x|p_0)' \Sigma(x) \nabla D(x|p_0) = \sum_r \sum_s g(x_r) \Sigma_{rs}(x) g(x_s)$  as  $x_r \rightarrow 0$ . First observe from (21) and (25) that if for each  $r = 1, \dots, R-1$  we now let

$$\theta_r(x) = \ln \left( 1 - \sum_{s=1}^{R-1} x_s \right) + \ln \left( \frac{p_{0r}}{1 - \sum_{s=1}^{R-1} p_{0s}} \right) \tag{46}$$

then for all distinct regional pairs,  $rs$ ,

$$\begin{aligned} g(x_r)_{\Sigma_{rs}}(x)g(x_s) &= [\ln(x_r) - \theta_r(x)] (-x_r x_s) [\ln(x_s) - \theta_s(x)] \\ &= - [x_r \ln(x_r) - x_r \theta_r(x)] [x_s \ln(x_s) - x_s \theta_s(x)]. \end{aligned} \quad (47)$$

But since  $\lim_{x_r \rightarrow 0} x_r \ln(x_r) = 0$ , and since

$$\lim_{x_r \rightarrow 0} \theta_r(x) = \ln \left( 1 - \sum_{s \neq r} x_s \right) + \ln \left( \frac{p_{0r}}{1 - \sum_{s=1}^{R-1} p_{0s}} \right) \quad (48)$$

is bounded, it follows from (45) that

$$\lim_{x_r \rightarrow 0} g(x_r)_{\Sigma_{rs}}(x)g(x_s) = 0 \quad (49)$$

Next consider the  $r^{\text{th}}$  diagonal term

$$\begin{aligned} g(x_r)^2_{\Sigma_{rr}}(x) &= [\ln(x_r) - \theta_r(x)]^2 x_r (1 - x_r) \\ &= x_r [\ln(x_r)]^2 - 2 [x_r \ln(x_r)] \theta_r(x) + x_r \theta_r(x)^2 - [x_r \ln(x_r) - x_r \theta_r(x)]^2 \end{aligned} \quad (50)$$

and observe from the arguments above that the last three terms go to zero as  $x_r \rightarrow 0$ , so that

$$\lim_{x_r \rightarrow 0} g(x_r)^2_{\Sigma_{rr}}(x) = \lim_{x_r \rightarrow 0} x_r [\ln(x_r)]^2. \quad (51)$$

To establish the limit of the right hand side of (51), let  $f(x_r) = [\ln(x_r)]^2$  and  $h(x_r) = 1/x_r$  so that

$$x_r [\ln(x_r)]^2 = \frac{f(x_r)}{h(x_r)}. \quad (52)$$

Observe in addition that

$$\frac{f'(x_r)}{h'(x_r)} = \frac{2 [\ln(x_r)] / x_r}{-x_r^{-2}} = -2x_r \ln(x_r) \quad (53)$$

implies  $\lim_{x_r \rightarrow 0} [f'(x_r)/h'(x_r)] = 0$ , and moreover that  $\lim_{x_r \rightarrow 0} f(x_r) = \infty = \lim_{x_r \rightarrow 0} h(x_r)$ . By L'Hospital's Rule<sup>55</sup> it then follows that  $\lim_{x_r \rightarrow 0} [f(x_r)/h(x_r)] = 0$  and hence that

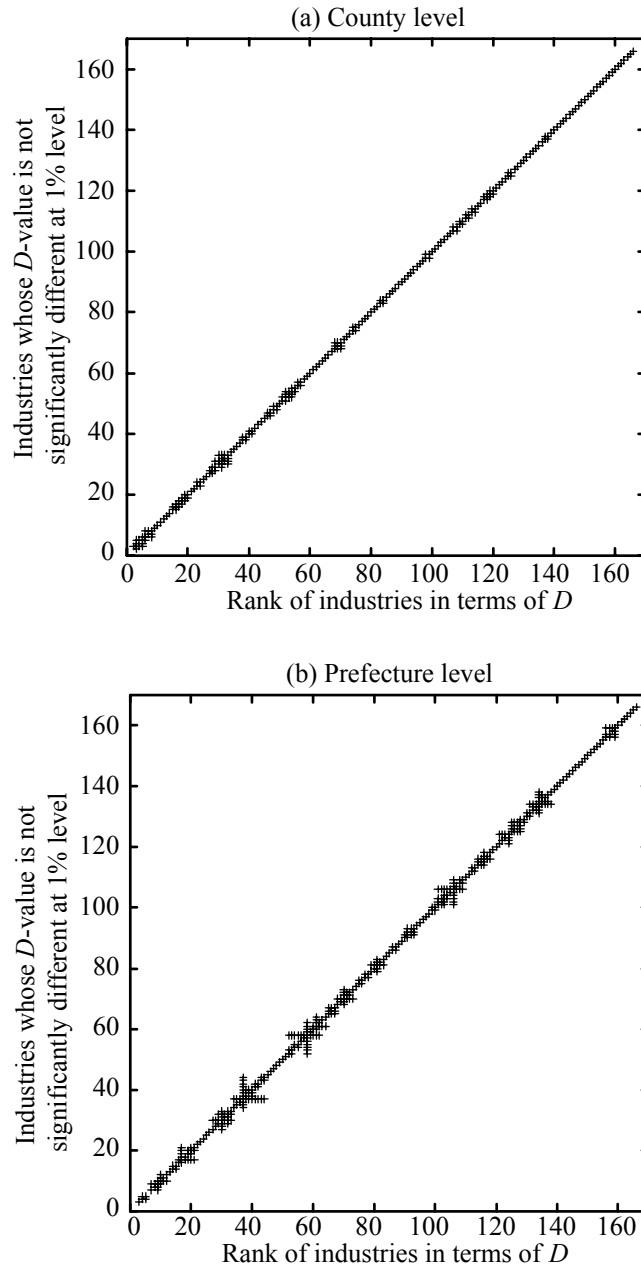
$$\lim_{x_r \rightarrow 0} g(x_r)^2_{\Sigma_{rr}}(x) = 0 \quad (54)$$

Thus each component  $r$  of  $\widehat{p}_i$  with  $\widehat{p}_{ir} = 0$  can simply be dropped from the estimation of variance in (29).<sup>56</sup> If we now let  $\widehat{p}_i^+ = \{r=1, \dots, R-1: \widehat{p}_{ir} > 0\}$ , then we obtain Theorem 1 ■

## C Inter-industry comparison of localization degrees

Figure 9 indicates the industries (vertical axis) whose  $D$ -values are not significantly different (at 1% level) from that of a given industry (horizontal axis), based on 1999 data. The regional units are counties in Diagram (a) and prefectures in Diagram (b).<sup>57</sup> Industries are ranked in terms of their  $D$ -values (i.e., the industry with the largest  $D$  is rank 1). Notice that at both of these regional levels, most industries are quite distinguishable in terms of their degree of localization. This is mainly due to the large sample sizes (industry sizes), which yield sharp confidence bounds on individual  $D$ -values.

Figure 9: Test of the difference in localization degrees between industries



## Reference

- Bartle, Robert G., *The Elements of Real Analysis, Second Edition* (New York: John Wiley & Sons, Inc, 1976).
- Bourguignon, Francois, “Decomposable income inequality measures,” *Econometrica* 47 (1979), 901-920.
- Brühlhart, Marius, and Rolf Traeger, “An account of geographic concentration patterns in Europe,” forthcoming in *Regional Science and Urban Economics* (2005).
- Christaller, Walter., *Die Zentralen Orte in Süddeutschland* (Gustav Fischer, 1933), English translation by Carlisle W. Baskin, *Central Places in Southern Germany* (Englewood Cliffs, NJ: Prentice Hall, 1966).
- Cover, Thomas M., and Joy A. Thomas, *Elements of information theory* (New York: John Wiley & Sons, Inc., 1991).
- Dembo, Amir, and Ofer Zeitouni, *Large Deviation Techniques* (Boston: Jones and Bartlett Publishers, 1993).
- Doukhan, Paul, *Mixing: Properties and Examples* (New York: Springer-Verlag, 1994).
- Duranton, Gilles, and Henry G. Overman, “Testing for localization using micro-geographic data,” forthcoming in *Review of Economic Studies* (2005).
- , and Diego Puga, “Micro-foundation of urban agglomeration economies,” in J. Vernon Henderson and Jacques-François Thisse (eds.), *Handbook of Urban and Regional Economics, vol.4* (Amsterdam: North-Holland, 2004), 2063-2117.
- Ellison, Glenn, and Edward L. Glaeser, “Geographic concentration in U.S. manufacturing industries: a dartboard approach,” *Journal of Political Economy* 105 (1997), 889-927.
- Gibbs, Allison L., and Francis Edward Su, “On Choosing and Bounding Probability Metrics,” *International Statistical Review* 70, 419-435 (2002).
- Japan Statistics Bureau, *Establishment and Enterprise Census* (1981,1999).
- , *Population Census* (2000).
- Krugman, Paul, *Geography and Trade* (Cambridge, MA: MIT Press, 1991), Ch.2.
- Kullback, Solomon, *Information Theory and Statistics* (New York: John Wiley & Sons, Inc., 1959).

———, and Richard A. Leibler, “On information and sufficiency,” *Annals of Mathematics and Statistics* 22 (1951) 79-86.

Kanemoto, Yoshitsugu, and Kazuyuki Tokuoka, “The proposal for the standard definition of the metro area in Japan,” *Journal of Applied Regional Science* 7 (2002), 1-15.

Marcon, Eric, and Florence Puech, “Evaluating the geographic concentration of industries using distance-based methods,” *Journal of Economic Geography* 3 (2003), 409-428.

Mori, Tomoya, Nishikimi, Koji, and Tony E. Smith, “Some empirical regularities in spatial economies: a relationship between industrial location and city size,” Discussion paper No.551, Institute of Economic Research, Kyoto University (2002).

Ripley, Brian D., “Modelling spatial patterns,” *Journal of the Royal Statistical Society B*, 39 (1977), 172-192.

Rao, C. Radhakrishna, *Linear Statistical Inference and Its Applications, Second Edition* (New York: John Wiley & Sons, Inc., 1973).

Rosenthal, Stuart S., and William C. Strange, “Evidence on the Nature and sources of agglomeration economies,” in J. Vernon Henderson and Jacques-François Thisse (eds.), *Handbook of Urban and Regional Economics, vol.4* (Amsterdam: North-Holland, 2004), 2119-2171.

Salas, Rafael, “Multilevel interterritorial convergence and additive multidimensional inequality decomposition,” *Social Choice and Welfare* 19 (2002), 207-218.

Shorrocks, Anthony F., “The class of additively decomposable inequality measures,” *Econometrica* 48 (1980), 613-625.

———, “Inequality decomposition by factor components,” *Econometrica* 50 (1982), 193-211.

———, “Inequality decomposition by population subgroups,” *Econometrica* 52 (1984), 1369-1385.

Smith, Tony E., “A Central Limit Theorem for Spatial Samples”, *Geographical Analysis* 12 (1980), 299-324.

Statistical Information for Consulting Analysis, *Tokei de miru Shi-Ku-Cho-Son no Sugata*, in Japanese (1998).

Theil, Henri, *Economics and Information Theory* (Amsterdam: North-Holland, 1967).

Thistle, Paul D., "Large sample properties of two inequality indices," *Econometrica* 58 (1990), 725–728.

Wilks, Samuel Stanley, *Mathematical Statistics* (New York: John Wiley & Sons, Inc., 1962).

## Notes

<sup>1</sup>See, Rosenthal and Strange (2004) for a survey of this literature.

<sup>2</sup>A notable exception is the  $K$ -density approach of Duranton and Overman (2005) and Marcon and Puech (2003) [based on Ripley (1977)]. But while this method provides a powerful framework for statistical analyses of industrial localization (and in particular is free of the border biases arising from internal regional subdivisions), it requires location data at the level of *individual establishments* within each industry, and such data is often not available.

<sup>3</sup>It should also be noted that unlike other existing indices, our index requires only regional-level data that is widely available. In particular, it does not require information about either the locations or sizes of specific establishments [as for example in Duranton and Overman (2005), Marcon and Puech (2003), Ellison and Glaeser (1997)]. Hence this index should provide a handy tool for a wide range of researchers interested in studying spatial localization.

<sup>4</sup>As shown in Section 4.3, it is also useful for testing inter-temporal changes in the degree of localization within a given industry.

<sup>5</sup>A notable exception is the measure of “topographic concentration” by Brülhart and Traeger (2005), discussed in footnote 8 below.

<sup>6</sup>The implications of alternative choices of reference distributions are discussed in more detail in Section 4.4 and Appendix A.

<sup>7</sup>The related literature for such decompositions is found in footnote 17.

<sup>8</sup>At this point it should be noted that an approach similar to the present one has been developed independently by Brülhart and Traeger (2005) for the analysis of inter-temporal changes in spatial concentration. Following the works of Bourguignon (1979) and Shorrocks (1980, 1982, 1984) these authors propose a class of “generalized entropy” measures (GE) based on their desirable decomposability properties. Among these is the present  $D$ -index [designated as GE(1)], where the appropriate reference distribution is implicitly taken to be defined by the choice of units for analysis. Our hypothesis of *complete spatial dispersion* corresponds to the choice of areal units of analysis, designated by Brülhart and Traeger as *topographic concentration*. However, in the present approach we take this reference distribution to represent an explicit null hypothesis for testing differences in spatial concentration between industries. Under the assumption of independent random sampling (discussed in Section 2.1 below), each null hypothesis leads to an explicit asymptotic normal distribution for differences in spatial concentra-

tion (as developed in Section 3.1 below). In contrast, the above authors adopt a nonparametric “block bootstrap” testing procedure for identifying significant changes in spatial concentration over time. While this procedure does allow for possible spatial dependencies over time, it is not specific to any particular index of spatial concentration.

<sup>9</sup>It is of interest to note, however, that this approximation at the establishment level may be more reasonable than at the level of individual workers. In particular, one could equally well calculate the degree of localization for distributions of employment rather than establishments. But since the locations of workers are tied to those of the establishments in which they are employed, statistical independence at the worker level seems even less plausible.

It should also be noted that patterns of establishment location may differ depending on their employment size [as reported by Duranton and Overman (2005) for the case of the UK]. However, the application of our  $D$ -index (introduced below) to Japanese manufacturing industries suggests that the degrees of localization based on employment-size data and establishment data are in close agreement (Spearman’s rank correlation is greater than 0.9). That is, the establishment size variation among industries as well as the size-dependence of locational patterns appears to have only minor influence on the degree of localization measured by our index.

This result still leaves open the possibility that correlation of location behavior occurs at more aggregated level, i.e., the location of multiple establishments may be correlated to some extent given the prevalence of multi-unit firms. However, the present limitations of data availability make it difficult to address these questions.

<sup>10</sup>A recent survey of such measures can be found in Gibbs and Su (2002).

<sup>11</sup>Kullback-Leibler divergence can also be motivated from an information-theoretic viewpoint [e.g., Cover and Thomas (1991)].

<sup>12</sup>While the reference distribution,  $p_{0r}$ , is always positive for all  $r$  for our purposes, the  $i$ -establishment distributions,  $p_{ir}$ , need not be positive. Hence the definition of  $D(p_i|p_0)$  implicitly includes the convention that  $0 \ln(0) = 0$ .

<sup>13</sup>The function  $D(\cdot|p_0)$  is strictly convex on the interior of the probability simplex in  $\mathcal{R}^{R-1}$ , and hence achieves its maximum values at the vertices of this simplex. For the vertex with  $p_{ir} = 1$ , the local maximum value of  $D$  is easily seen to be  $-\ln(p_{0r}) > 0$ , so that the global maximum is achieved when  $p_{0r}$  achieves its minimum value over  $r = 1, \dots, R$ . In other words,  $D$  is as large as possible when all establishments are concentrated in the region of smallest (economic) size.

<sup>14</sup>See, e.g., Sanov's Theorem in Dembo and Zeitouni (1993).

<sup>15</sup>It is well known that, the statistic,  $-2\ln \lambda$ , is asymptotically *chi-square distributed with  $R - 1$  degrees of freedom*, and hence, provides a well-defined statistical test of the null hypothesis,  $p_i = p_0$ .

<sup>16</sup>For the Japanese case in particular, there is a strong negative correlation (Spearman's rank correlation of -.66) between the number of establishments and the  $D$ -index for 3-digit manufacturing industries. So more ubiquitous industries do indeed tend to have larger numbers of establishments.

<sup>17</sup>Studies of the decomposability properties of entropy measures in information theory date back to the seminal work of Kullback (1959) [as summarized in Cover and Thomas (1991)], and were first introduced into economics by Theil (1967). It was later shown by Bourguignon (1979) and Shorrocks (1980, 1982, 1984) that these decomposability properties essentially characterize entropy (i.e., are uniquely exhibited by a somewhat more general class of entropies). A systematic survey of this work can be found in Brülhart and Traeger (2005). See also Salas (2002) for a recent survey.

<sup>18</sup>A notable exception here is the work of Brülhart and Traeger (footnote 8 above) who have independently used the same decomposition techniques to compare industrial concentrations both within and between 16 countries of Western Europe.

<sup>19</sup>It should be noted that Ellison and Glaeser (1997) do calculate the variance of  $G$  under the null hypothesis of no regional-endowment effects or spillover effects ( $\gamma^{na} = 0 = \gamma^s$ ), and make an implicit appeal to asymptotic normality of  $G$  in order to test this null hypothesis. However, no attempt is made to derive general confidence intervals for either  $E(G)$  or  $E(\gamma)$ . With respect to divergence indices, it should also be noted that Brülhart and Traeger (2005) apply bootstrap methods to estimate confidence intervals for an index identical to our  $D$ .

<sup>20</sup>For convenience we use the same regional notation,  $r$ , for the numerical region index in this section.

<sup>21</sup>This implies in particular that for all  $r, s = 1, \dots, R - 1$ ,  $\text{var}(N_{ir}) = n_i p_{ir}(1 - p_{ir})$ , and  $\text{cov}(N_{ir}, N_{is}) = -n_i p_{ir} p_{is}$ . In addition, the variance of  $N_{iR}$  and the covariance between  $N_{iR}$  and  $N_{ir}$  for  $r = 1, 2, \dots, R - 1$  are given by

$$\begin{aligned} \text{var}(N_{iR}) &= n_i \left( 1 - \sum_{s=1}^{R-1} p_{is} \right) \sum_{s=1}^{R-1} p_{is} \\ &= n_i z' [\text{diag}(p_i) - p_i p_i'] z, \end{aligned}$$

$$\begin{aligned}\text{cov}(N_{iR}, N_{ir}) &= -n_i p_{ir} \left( 1 - \sum_{s=1}^{R-1} p_{is} \right) \\ &= -n_i [\text{diag}(p_i) - p_i p_i'] z,\end{aligned}$$

where  $z$  is a  $(R - 1)$ -vector with all elements equal to one.

<sup>22</sup>If one or more components of  $p_i$  are zero, then special care must be taken in defining the appropriate region of differentiation (to be explained below).

<sup>23</sup>This asymptotic normality result is in fact an instance of more general results obtained by Thistle (1990) (i.e., the limiting case of his Theorem 3 as  $\beta \rightarrow 0$ ).

<sup>24</sup>A number of examples of this type were simulated, and all produced similar results.

<sup>25</sup>For the Japanese manufacturing to which we apply our  $D$ -index in Section 4, there is no industry  $i$  for which the null hypothesis  $p_i = p_0$  cannot be rejected under the likelihood-ratio test (refer to footnote 15).

<sup>26</sup>For the case of Japan, the average and median number of establishments for 3-digit industries are 4692 and 1941, respectively. The industry at the 10 percentile point has 105 establishments. It is to be noted, however, that those industries with small numbers of establishments are rather specialized. The smallest 10% (17 industries) include the tobacco industry, eight arms related industries, and two heavily natural-resource oriented industries (coke and briquettes manufacturing). Thus, most market oriented private industries have fairly large numbers of establishments at the 3-digit level.

<sup>27</sup>To estimate the true size of this test, we set  $p_i = p_j = (.1, .4, .2, .3)$ , and sampled 1000 values of both  $\hat{p}_i$  and  $\hat{p}_j$ , from this common distribution [yielding 1000 samples of the test statistic,  $D(\hat{p}_i|p_0) - D(\hat{p}_j|p_0)$ ]. For the rejection region of size  $\alpha = .05$  based on normal theory, our estimate  $\hat{\alpha}$  was simply the fraction of times this test statistic fell in the rejection region, i.e., the fraction of time that the true null hypothesis,  $p_i = p_j$ , was rejected.

<sup>28</sup>The establishments and employment data used in this section are classified according to the Japanese Standard Industry Classification (JSIC) taken from the Establishment and Enterprise Census of Japan [Japan Statistics Bureau (1981, 1999)]. In addition to 3-digit classification, we also consider 2-digit classification later in this section.

<sup>29</sup>Economic area is obtained by subtracting the forest, undeveloped area, lakes and marshes from the total area of a county [Data source: Statistical Information Institute for Consulting and Analysis (1998)].

The total economic area is  $120,205km^2$ , equivalent to 31.8% of total area in Japan. The size of economic area in a county varies from  $0.47km^2$  to  $720.38km^2$  with average  $37.21km^2$ .

As pointed out by one of the referees, economic area is to some extent an endogenous variable. For instance, forest may be converted to industrial area. But 70% of Japanese land was covered by forest in 1998, and for our purposes it seems reasonable to assume that the forest continues to cover large portion of Japanese land at the present time.

It should also be noted that several alternative definitions of economic area might be appropriate, depending on the objectives of a given study. But the validity of our  $D$ -index itself is independent of any particular definition of economic area.

<sup>30</sup>County here is equivalent to *shi-ku-cho-son* in the Japanese Census. County boundaries are as of October 1, 2001. The number of counties in Japan is 3363.

<sup>31</sup>This is most evident in Figure 9 of Appendix C below, where it is seen that all but the smallest differences in  $D$ -value rankings are significant at the 1% level.

<sup>32</sup>Refer to Section 2.3 and footnote 16.

<sup>33</sup>The industries that are excluded from Table 1 but have higher  $D$ -values than those listed are all the eight arms-related industries (JSIC331-339,  $D = 6.52 \sim 8.59$ ), fur skin manufacturing (JSIC248,  $D = 5.89$ ), Coke (JSIC213,  $D = 4.82$ ), Briquette and briquette balls manufacturing (JSIC214,  $D = 4.66$ ), Mechanical leather products, excluding gloves and mittens (JSIC242,  $D = 4.56$ ).

<sup>34</sup>See Appendix C for more detail on the comparison of  $D$ -values among industries.

<sup>35</sup>See, e.g., Duranton and Puga (2004) for a survey of the micro foundations of agglomeration economies.

<sup>36</sup>However, the ubiquity of cement-related industries may also be partly explained by the ubiquity of lime in Japan.

<sup>37</sup>Again, ordnance and accessories (JSIC33) is excluded in the list since it has only 29 establishments. The  $D$ -value for this industry is 5.43.

<sup>38</sup>The size of the 99% confidence band for each industry is smaller than 0.001.

<sup>39</sup>Derivation of metro areas is based on *Urban Employment Area (UEA)* developed by Kanemoto and Tokuoka (2002) which aggregates counties based on commuting patterns, and is comparable to the Core Based Statistical Area (CBSA) of the U.S. While their definition of the UEA requires that the central county of a metro area should have the so-called *densely inhabited district* (defined in the

Population Census of Japan [Japan Statistics Bureau (2000)] with more than 10,000 population, the threshold population size in our definition is 5,000. But, the basic result is not affected by the choice of the threshold.

<sup>40</sup>It should be noted however that our  $D$ -index values for these metro areas are based on 1999 employment data.

<sup>41</sup>Business areas were constructed by first ranking counties by their numbers of establishments, and then choosing the highest ranked counties constituting 70% of all establishments.

<sup>42</sup>These product-moment correlations are all significant at 1% level.

<sup>43</sup>Population data are taken from the Population Census of Japan [Japan Statistics Bureau (2000)].

<sup>44</sup>See Mori, Nishikimi and Smith (2002) for a formal test of this principle.

<sup>45</sup>All changes in the  $D$ -index are significant at 1% level. Note that  $D$ -index values are likely to be positively correlated between time period, so that (as mentioned at the end of section 3.3) this level of significance may even be conservative.

<sup>46</sup>See Ellison and Glaeser (1997).

<sup>47</sup>This is seen most easily when all firms in industry  $i$  are of exactly the same size, so that  $H_i = m_i(1/m_i)^2 = 1/m_i \approx 0$ . More generally, simple continuity considerations suggest that this will continue to hold so long as  $M_i$  is very large compared to any single  $M_{ij}$ .

<sup>48</sup>Empirical evidence for this can be found in the Ellison-Glaeser (1997) itself. For example, the orderings of  $G$  and  $\gamma$  for the 15 most localized industries in their Table 4 are identical. Moreover, while this is less evident for the 15 most ubiquitous industries shown in Table 4, these values of  $\gamma$  are all so close to zero that their actual ordering can be expected to be unstable.

<sup>49</sup>The (product-moment) correlation between  $D$  and the log of employment share is -0.75.

<sup>50</sup>This is one of the issues first dealt with by Ellison and Glaeser (1997).

<sup>51</sup>For example, the case of “within multi-establishment” dependencies can be handled within the current asymptotic framework so long as there is a bound,  $m$ , on the number of establishments within each multi-establishment firms. For then, by ordering establishments by multi-plant firm, the resulting sequence can be regarded as  $m$ -dependent, so that the estimates  $\hat{p}_i = (\hat{p}_{i,r})$  are still asymptotically multi-normally distributed [for a spatial version of this result see, for example, Smith (1980)]. More generally, asymptotic normality continues to hold as long as dependencies between establishments are not “too strong” [see,

for example, Doukhan (1994)].

<sup>52</sup>Another popular measure is the *locational Gini coefficient* [e.g., Krugman (1991)] which is basically the sum of the difference between the cumulative share of industry-specific employment  $\sum_{v \leq r} s_{iv}$  and that of total employment  $\sum_{v \leq r} x_v$  for each region  $r$ , where regions are ordered by the ratio,  $s_{ir}/x_r$ .

<sup>53</sup>In 1999, the most diverse metro area in Japan was Tokyo with positive employment in 157 3-digit manufacturing industries out of the total 166, while the least diverse was Kamifurano with only 10 industries.

<sup>54</sup>While this problem with the  $G$ -index is most evident when one industry is large relative to total manufacturing, the same problem persists for finer industrial classifications as well. One way to see this statistically is to imagine cases in which the employment of each industry is drawn from the same regional “population distribution,” so that industries differ statistically only in terms of their sample size. Then the Law of Large Numbers tells us that larger industries (samples) can be expected to resemble the population distribution more closely than smaller industries (samples).

<sup>55</sup>For details of L’Hospital’s Rule, see for example Bartle (1976).

<sup>56</sup>It should be noted that since the limit of  $\theta_r(x)$  is well defined and bounded for any subset of  $x$  approaching zero, the above argument is directly extendable to the case of multiple zero components.

<sup>57</sup>There are 47 prefectures in Japan.