Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach

Glenn Ellison

Massachusetts Institute of Technology and National Bureau of Economic Research

Edward L. Glaeser

Harvard University and National Bureau of Economic Research

This paper discusses the prevalence of Silicon Valley–style localizations of individual manufacturing industries in the United States. A model in which localized industry-specific spillovers, natural advantages, and pure random chance all contribute to geographic concentration is used to develop a test for whether observed levels of concentration are greater than would be expected to arise randomly and to motivate new indices of geographic concentration and of coagglomeration. The proposed indices control for differences in the size distribution of plants and for differences in the size of the geographic areas for which data are available. As a consequence, comparisons of the degree of geographic concentration across industries can be made with more confidence. Our empirical results provide a strong reaffirmation of the previous wisdom in that we find almost all industries to be somewhat localized. In many industries, however, the degree of localization is slight. We explore the nature of agglomerative forces in describing patterns of concentration, the geographic scope of localization, and the coagglomeration of related industries and of industries with strong upstream-downstream ties.

We would like to thank Richard Caves, Sara Fisher Ellison, Larry Katz, Bill Miracky, Wally Mullin, Peter Reiss, José Scheinkman, Sherwin Rosen, and two anonymous referees for helpful comments and Matt Botein, David Hwang, Rajesh James, Bruce Sacerdote, and Jacob Vigdor for research assistance. We both thank the National Science Foundation (SBR-9310009 and SBR-9515076, and SBR-9309808) and the Harvard Clark Fund for financial support.

[*Journal of Political Economy*, 1997, vol. 105, no. 5]
 0022-3808/97/0505-0006$02.50

I. Introduction

The concentrations of high-tech industries in Silicon Valley and the auto industry in Detroit are two of the more famous examples of the geographic agglomeration of firms in a single industry. These examples for years have fascinated both practically minded urban planners and economic geographers who are interested in accounting for a striking feature of the economic landscape. More recently, it has been suggested by Krugman (1991*a*) and others that Silicon Valley–style agglomerations may be more the rule than the exception and that from them one may learn about the sources of increasing returns that have appeared in the literature following Marshall (1920). Given the central role increasing returns play in the new theories of growth and international trade, these suggestions have led to a surge of new work. Researchers who are primarily interested in international trade, growth, industrial organization, and business strategy have joined geographers and urban economists in investigating why agglomerations exist.[1] In this paper we step back a bit from this work and reexamine both how industry concentration over and above the general concentration of manufacturing (and industry group coagglomeration) can be measured and what the facts are to be explained.

We begin by proposing a "model-based" index of geographic concentration that has several useful properties. First, the index is scaled so that it takes on a value of zero not if employment is uniformly spread across space, but instead if employment is only as concentrated as it would be expected to be had the plants in the industry chosen locations by throwing darts at a map. Because production in many industries occurs mainly in a few large plants, accounting for lumpiness can be substantial. For example, in the U.S. vacuum cleaner industry (Standard Industrial Classification [SIC] 3635), about 75 percent of the employees work in one of the four largest plants. Thus we would not want to regard it as being concentrated simply because 75 percent of its employment is contained in only four states. Second, the index is designed to facilitate comparisons across industries, across countries, or over time. When plants' location decisions are made as in the model, differences in the size of the industry, the size distribution of plants, or the fineness of the geographic data that are available should not affect the index. Thus one may compare with more confidence, for example, the concentration of American and European industries, the concentration of

[1] For samples of work in these fields, see Creamer (1943), Florence (1948), Hoover (1948), Fuchs (1962), Carlton (1983), Henderson (1988), Enright (1990), Porter (1990), Krugman (1991*a*), and Jaffe, Trajtenberg, and Henderson (1993).

high- and low-tech industries, and the changes in levels of concentration over time.[2]

In our model of location choice, plants sequentially choose locations to maximize profits. We allow for two types of agglomerative forces, which we refer to as spillovers and natural advantage. By locational spillovers we mean both physical spillovers (as in Krugman [1991*b*], where the presence of one firm lowers transportation costs for a second) and intellectual spillovers (as in Glaeser et al. [1992]). Natural advantage includes the forces that lead the wine industry to concentrate in California and large shipyards to locate on bodies of water. When neither of these forces is present, the model reduces to one in which plants choose locations by throwing darts at an appropriately scaled map.

The first result of our theory section is an observational equivalence theorem that shows that the relationship between mean measured levels of concentration and industry characteristics is the same regardless of whether concentration is the result of spillovers, natural advantage, or a combination of the two. One may interpret this result as a warning that geographic concentration by itself does not imply the existence of spillovers; natural advantages have similar effects and may be important empirically. For our purposes, however, the result has a positive message: one can design an index that controls for differences in industry characteristics, regardless of the cause of concentration. The second part of the theory sections analyzes a similar multiple industry model to motivate an index of coagglomeration that may be useful in studies of cross-industry spillovers and shared natural advantages.

The largest portion of the paper uses our indices to describe concentration in U.S. manufacturing. Our first surprising result is that despite the fact that our index imposes more stringent standards for calling an industry concentrated, virtually every industry displays excess concentration (446 of 459 four-digit SIC industries). This does not mean, however, that we take our results as support for the view that Silicon Valleys are ubiquitous. While there are a number of industries that look like Silicon Valley or the auto industry, it is much more common for industries to be only very slightly concentrated. Our measurements suggest that explanations for concentration vary by industry and that natural advantage may often play a role. We also look at concentration at the county, state, and regional

[2] See Krugman (1991*a*) for a discussion of the first two questions and Fuchs (1962) for an analysis of the third. Florence's (1948) observation that industries with larger plants are more concentrated is a particularly clear example of the difficulties that can arise in interpreting results with other indices.

levels and at the coagglomeration of industries with related SIC codes. Here, we find evidence suggesting that spillover benefits are restricted neither to the county level nor to the most narrowly defined industries. Industries also appear to coagglomerate both with important upstream suppliers and with important downstream customers.

II. A Model of Location Choice

In this section, we develop a simple model in which the geographic concentration of an industry is one result of a sequence of profit-maximizing location decisions made by individual plants. Natural advantages of some locations and industry-specific spillovers lead plants to cluster together, and idiosyncratic plant-specific considerations provide the counterbalance that keeps the entire industry from concentrating at a single point.

A. *Natural Advantage, Spillovers, and Localization*

Suppose that an industry consists of N business units (best thought of as manufacturing plants) that choose sequentially to locate in one of the M geographic subunits of a larger entity (e.g., in one of the states of the United States). We assume that the kth business unit chooses its location v_k to maximize its profits given that it will receive profits π_{ki} from locating in area i. To make the model tractable, we assume that these profits are given by

$$\log \pi_{ki} = \log \bar{\pi}_i + g_i(v_1, \ldots, v_{k-1}) + \epsilon_{ki}, \tag{1}$$

where $\bar{\pi}_i$ is a random variable reflecting the profitability of locating in area i for a typical firm in the industry (as influenced by observed and unobserved area characteristics), g_i captures the effects of spillovers created by business units that have previously chosen locations, and ϵ_{ki} is an additional random component reflecting factors that are idiosyncratic to plant k.[5]

"Natural advantages" are included in our model to capture the fact that the plants in an industry will be geographically concentrated whenever their location choices have been influenced by common factors that make some locations more desirable than others. While natural advantage reasons for geographic concentration may

[5] Given the way in which we shall specify the model, one will also be able to regard location decisions as a rational expectations equilibrium of a process in which plants receive spillovers also from plants that choose locations later on.

not be exciting intellectually, they are clearly important when accounting for some of the agglomeration we observe. For example, the localization of the wine industry in California is certainly attributable at least in part to California's favorable climate for growing grapes, and some portion of the agglomeration of large shipyards is due to their desire for locations on large bodies of water.

In our model, the effects of natural advantages on profits are captured by the random variables $\{\overline{\pi}_i\}$, which are chosen by nature at the start of the process when it assigns resource endowments to each area that fit well or poorly with the industry's needs. The expectation $\overline{\pi}_i$ then reflects the average profitability of locating in area i, and the variance of the $\{\overline{\pi}_i\}$ reflects how sensitive profits are to a good fit. For example, these variances might be high in the shipbuilding industry because the profitability of a state will depend greatly on whether nature has put that state on the coast.

If we specify that the $\{\epsilon_{ki}\}$ are independent Weibull random variables independent of the $\{\overline{\pi}_i\}$ and there are no spillovers ($g_i \equiv 0$ for all i), then conditional on a realization of $\overline{\pi}_1, \ldots, \overline{\pi}_M$, our model is a standard logit model and the firms' location choices are conditionally independent random variables with

$$\text{prob}\{v_k = i | \overline{\pi}_1, \ldots, \overline{\pi}_M\} = \frac{\overline{\pi}_i}{\sum_j \overline{\pi}_j}.$$

We have therefore chosen to focus on models in which the distributions of the $\{\overline{\pi}_i\}$ satisfy two parametric restrictions.

First, so that on average across industries the model reproduces the overall distribution of manufacturing activity (e.g., puts many more plants in California and New York than in Wyoming), we assume that

$$E_{\overline{\pi}_1,\ldots,\overline{\pi}_M} \frac{\overline{\pi}_i}{\sum_j \overline{\pi}_j} = x_i, \tag{2}$$

where x_i is area i's share of overall manufacturing employment. In practice, one can think of states with more manufacturing as having higher average profit levels for any of several reasons: plants located there may benefit from spillovers of aggregate activity that are not industry-specific, they may have characteristics (such as nice weather allowing lower equilibrium wages) desired by all industries, and they may have more potential locations to choose from, increasing the fit quality of the best location a plant is able to find.

Second, we assume that the joint distribution of natural advantages is such that there is a single parameter $\gamma^{na} \in [0, 1]$ for which

$$\text{var}\left(\frac{\bar{\pi}_i}{\sum_j \bar{\pi}_j}\right) = \gamma^{na} x_i(1 - x_i). \tag{3}$$

We think of the parameter γ^{na} as capturing the importance of natural advantage to the industry. The $\gamma^{na} = 0$ extreme corresponds to a model in which unobserved state characteristics have no effect on profitability. In this case, the plant's location decisions are independent, with each choosing area i with probability x_i. At the other extreme, when $\gamma^{na} = 1$, state characteristics are so important that they completely overwhelm firm-specific idiosyncratic factors, and the one state that has the best set of endowments will attract all the firms. (The largest variance the random variable $\bar{\pi}_i / [\sum_j \bar{\pi}_j]$ can have consistent with its always being between zero and one and having mean x_i is $x_i[1 - x_i]$.)

One concrete specification of the distribution of the $\{\bar{\pi}_i\}$ consistent with these requirements is to assume that the $\{\bar{\pi}_i\}$ are independent random variables that are scaled so that $2[(1 - \gamma^{na})/\gamma^{na}]\bar{\pi}_i$ has a χ^2 distribution with $2[(1 - \gamma^{na})/\gamma^{na}]x_i$ degrees of freedom. In this case, we have $E(\bar{\pi}_i) = x_i$ and $\text{var}(\bar{\pi}_i) = [\gamma^{na}/(1 - \gamma^{na})]x_i$, so it is easy to see that unobserved state characteristics have a negligible effect on average profitability levels when γ^{na} is close to zero and that profits vary greatly with the realized suitability of state characteristics when γ^{na} is close to one.

The second class of explanations for agglomeration we examine are what we call "spillover" theories. We use the term broadly to refer to technological spillovers, gains from sharing labor markets, gains from interfirm trade, the effect of local knowledge on the location of spin-off firms, and any other forces that might provide increased profits to firms locating near other firms in the same industry. While it might be descriptively more accurate to suppose that a plant receives more benefits from locating near some plants than others and that the fraction of the potential benefits that are realized varies smoothly with proximity, we consider instead (to make the model tractable) spillovers of an "all or nothing" variety. For each pair of plants, either the plants receive no benefits from colocation or the spillovers between them have infinite magnitude and are extremely localized geographically, so the plants receive the full potential benefits if they choose identical locations and no benefits at all if they locate in separate areas (regardless of proximity).

Formally, we incorporate spillovers whose importance is indexed by a parameter $\gamma^s \in [0, 1]$ by assuming that

$$\log \pi_{ki} = \log(\bar{\pi}_i) + \sum_{l \neq k} e_{kl}(1 - u_{li})(-\infty) + \epsilon_{ki}, \qquad (4)$$

where the $\{e_{kl}\}$ are Bernoulli random variables equal to one with probability γ' that indicate whether a potentially valuable spillover exists between each pair of plants, and u_{li} is again an indicator for whether plant l is located in area i. We assume also that the existence of spillovers between plants is a symmetric, transitive relationship in the sense that $e_{kl} = 1 \Rightarrow e_{lk} = 1$ and $e_{kl} = 1$ and $e_{lm} = 1 \Rightarrow e_{km} = 1$.[4] This assumption is also motivated by the properties of the location decision process it induces: the process in which the kth plant chooses its location taking into account only the locations of the first $k - 1$ plants is also a rational expectations equilibrium of a model in which plants are forward looking, and the resulting distribution of locations is independent of the order in which the plants make their choices.

In describing this specification of spillovers, we sometimes extend the dartboard metaphor and imagine a two-stage process in which nature first randomly chooses to weld some of the darts into clusters (representing groups of plants that are sufficiently interdependent that they will always locate together) and then each cluster is thrown randomly at the dartboard to choose a location. The importance of spillovers is captured by the parameter γ', which indicates the fraction of pairs of firms between which a spillover exists.

Writing s_i for the share of the industry's employment in area i and x_i for the share of aggregate manufacturing employment in area i, one can construct a measure of an industry's geographic concentration by setting $G \equiv \sum_i (s_i - x_i)^2$. In the model we have described, the $\{x_i\}$ are taken as exogenous and the $\{s_i\}$ are determined endogenously by $s_i = \sum_k z_k u_{ki}$, where z_k is the kth plant's (exogenously fixed) share of the industry's employment and u_{ki} is an indicator variable equal to one if plant k chooses to locate in state i. The principal result of this section is a characterization of how the expected value of G is related to the parameters characterizing the strength of natural advantages and spillovers, the industry's plant size distribution, and the sizes of the areas for which employment breakdowns are available when location decisions are made in accordance with the model described above.

PROPOSITION 1. In any specification of the location choice model in which plants $1, 2, \ldots, N$ sequentially choose locations to maximize

[4] Note that we have *not* fully specified the joint distribution of the $\{e_{kl}\}$. The proposition below will apply to all distributions with these properties. To see that at least one such joint distribution exists, consider the case in which the $\{e_{kl}\}$ are perfectly correlated, so that with probability γ_0 all the firms are completely interdependent and with probability $1 - \gamma_0$ all their profits are independent.

profit functions that satisfy equations (2), (3), and (4),

$$E(G) = (1 - \sum x_i^2)[\gamma + (1 - \gamma)H],$$

where $H \equiv \sum_k z_k^2$ is the Herfindahl index of the industry's plant size distribution and $\gamma = \gamma^{na} + \gamma^s - \gamma^{na}\gamma^s$.

Proof. Writing p_i for $\bar{\pi}_i / (\sum_j \bar{\pi}_j)$ and p for $p_1, \ldots, p_M$, we can expand $E(G)$ using the law of iterated expectations as

$$E(G) = \sum_i E_p E[(s_i - x_i)^2 | p]$$

$$= \sum_i E_p \operatorname{var}(s_i | p) + E_p(s_i - x_i | p)^2.$$

Using the identity $s_i = \sum_j z_j u_{ji}$ and expanding the variance terms gives

$$E(G) = \sum_i E_p \Bigg[\sum_j z_j^2 \operatorname{var}(u_{ji} | p) + \sum_{j,k,j \neq k} z_j z_k \operatorname{cov}(u_{ji} u_{ki} | p) + E(s_i - x_i | p)^2 \Bigg]$$

$$= \sum_i E_p \Bigg\{ \sum_j z_j^2 p_i (1 - p_i) + \sum_{j,k,j \neq k} z_j z_k [\gamma^s p_i + (1 - \gamma^s) p_i^2 - p_i^2] + (p_i - x_i)^2 \Bigg\}.$$

Our specification of natural advantage ([2] and [3]) assumed $E(p_i) = x_i$ and $E[(p_i - x_i)^2] = \gamma^{na}(x_i - x_i^2)$, which together imply $E(p_i - p_i^2) = (1 - \gamma^{na})(x_i - x_i^2)$. Also, from our definition $H = \sum_j z_j^2$ we have $\sum_{j,k,j \neq k} z_j z_k = (\sum_j z_j)^2 - \sum z_j^2 = 1 - H$. Substituting each of these into the equation above gives

$$E(G) = \sum_i [H(1 - \gamma^{na})(x_i - x_i^2) + (1 - H)\gamma^s(1 - \gamma^{na})(x_i - x_i^2) + \gamma^{na}(x_i - x_i^2)]$$

$$= \left(1 - \sum_i x_i^2\right)[\gamma^{na} + \gamma^s - \gamma^{na}\gamma^s + (1 - \gamma^{na} - \gamma^s + \gamma^{na}\gamma^s)H].$$

Q.E.D.

The most interesting aspect of proposition 1 is that it establishes something of an observational equivalence result between the effects of natural advantages and spillovers on expected concentration lev-

els. An analysis of the mean concentration of industries will allow one only to estimate $\gamma = \gamma^s + \gamma^{na} - \gamma^s\gamma^{na}$, and any estimated $\gamma \in [0, 1]$ is compatible with a pure natural advantage model, a pure spillover model, or models with various combinations of the two factors.[5]

The conclusion of proposition 1 is not a pessimistic statement. It is helpful in that it indicates that it will be possible to construct an index of concentration that "controls" for differences in industry and data characteristics without knowing what combination of natural advantages or spillovers is responsible for the agglomeration of each industry.

B. Coagglomeration

To discuss the degree to which pairs or groups of industries appear to be coagglomerated, we consider now a model in which N plants, each belonging to one of r industries in an industry group, choose locations. We use N_j, w_j, and H_j, respectively, for the number of plants in the jth industry, the jth industry's share of the total employment in those r industries, and the plant Herfindahl of the jth industry, and H for the plant Herfindahl of the group.

To produce a model in which these industries will exhibit some degree of coagglomeration, one could modify the discrete choice model above to allow for natural advantages that are correlated across industries or for spillovers that are not purely industry-specific. For example, plants in the cane sugar refining and shipbuilding industries might be coagglomerated because coastal locations provide higher profits both for shipyards and for importers of bulky commodities. On the other hand, the coagglomeration of various textile industries might be attributable to the presence of spillovers between plants in similar but not identical lines of business. Formally, this would involve making the average profits $\bar{\pi}_i^j$ and $\bar{\pi}_i^k$ that plants in the jth and kth industries receive when locating in area i correlated random variables and allowing the probability that a crucial spillover exists between two plants to depend on whether or not they belong to the same industry.

While such a model is not difficult to create, analyzing it is tedious and not particularly enlightening. Therefore, rather than character-

[5] While the result is limited in that only the effects on first moments of measured concentration are considered, we believe that attempts to distinguish natural advantage from spillover theories will not be fruitful because the higher moments of G will depend on a number of additional assumptions (e.g., on higher moments of the distribution of the area-specific average profit levels and on the full joint distribution of the indicator variables for whether spillovers exist between pairs of firms). Hence, pure natural advantage and pure spillover theories are each compatible with a range of findings for the higher moments.

ize expected concentration as a function of moments and so forth, we have chosen instead to give a more reduced-form theorem that relates the concentration of the group to the correlations in location choices induced by natural advantages and spillovers.

PROPOSITION 2. In an r industry location choice model, suppose that the distributions of average profit levels and spillovers are such that the indicator variables $\{u_{ki}\}$ for whether the kth plant locates in area i satisfy $E(u_{ki}) = x_i$ and

$$\operatorname{corr}(u_{ki}, u_{li}) = \begin{cases} \gamma_j & \text{if plants } k \text{ and } l \text{ both belong to industry } j \\ \gamma_0 & \text{otherwise.} \end{cases}$$

Let $G = \sum_i (s_i - x_i)^2$, where s_i is area i's share of the aggregate employment in the r industries, and $H = \sum_j w_j^2 H_j$ be the plant Herfindahl of the aggregate of the r industries. Then

$$E(G) = \left(1 - \sum_i x_i^2\right)\left[H + \gamma_0\left(1 - \sum_{j=1}^{r} w_j^2\right) + \sum_{j=1}^{r} \gamma_j w_j^2 (1 - H_j)\right].$$

Proof. Write $z_{j1}, \ldots, z_{jn_j}$ for the sizes of plants in the jth industry. Our assumptions on correlations then give

$$\begin{aligned} E(G) &= \sum_i \operatorname{var}(s_i) \\ &= \sum_i \left[\sum_{j,l} z_{jl}^2 \operatorname{var}(u_{jli}) + \sum_{j,l,l',l\neq l'} z_{jl} z_{jl'} \operatorname{cov}(u_{jli}, u_{jl'i}) \right. \\ &\quad \left. + \sum_{j,j',l,l',j\neq j'} z_{jl} z_{j'l'} \operatorname{cov}(u_{jli}, u_{j'l'i})\right] \\ &= \left[\sum_i x_i(1 - x_i)\right]\left(\sum_{j,l} z_{jl}^2 + \sum_{j,l,l',l\neq l'} z_{jl} z_{jl'} \gamma_j + \sum_{j,j',l,l',j\neq j'} z_{jl} z_{j'l'} \gamma_0\right) \\ &= \left(1 - \sum_i x_i^2\right)\left[H + \sum_j \gamma_j \left(w_j^2 - \sum_l z_{jl}^2\right) + \gamma_0\left(1 - \sum_j w_j^2\right)\right] \\ &= \left(1 - \sum_i x_i^2\right)\left[H + \gamma_0\left(1 - \sum_j w_j^2\right) + \sum_j \gamma_j w_j^2 (1 - H^j)\right]. \end{aligned}$$

Q.E.D.

The proposition characterizes the expected concentration of the aggregate employment in an industry group in terms of two factors. The first is simply the tendency of plants in each individual industry to agglomerate as captured by the single parameter γ_j (for the jth

industry), which reflects the influence of natural advantage and spillovers as in proposition 1. The second, γ_0, captures the tendency for plants in one industry to locate near plants in the others. The $\gamma_0 = 0$ extreme corresponds with the case in which there are spillovers or shared natural advantages across industries within the group (beyond the spillovers from aggregate activity). At the other extreme, when $\gamma_0 = \gamma_1 = \cdots = \gamma_n$, average profit levels are perfectly correlated across industries and spillovers are group-specific rather than industry-specific. For example, a pure spillover model satisfying the conditions of the theorem would be one in which the probability that a pair of plants had a crucial spillover between them was γ_j if each belonged to industry j and γ_0 if they belonged to different industries.[6]

III. Indexes of Geographic Concentration

In this section, we propose indices that may be used to measure the geographic concentration of an industry and the coagglomeration of groups of industries, and we discuss the properties of these indices.

A. *An Index of Industry Concentration*

Beginning with the single-industry problem, suppose that we are given data containing the shares $s_1, s_2, \ldots, s_M$ of an industry's employment in each of M geographic areas, the shares $x_1, x_2, \ldots, x_M$ of total employment in each of those areas, and the Herfindahl index $H = \sum_{j=1}^{N} z_j^2$ of the industry plant size distribution. As an index of the degree to which an industry is geographically concentrated, we propose the use of a measure γ defined by

$$\gamma \equiv \frac{G - \left(1 - \sum_i x_i^2\right)H}{\left(1 - \sum_i x_i^2\right)(1 - H)} \equiv \frac{\sum_{i=1}^{M}(s_i - x_i)^2 - \left(1 - \sum_{i=1}^{M} x_i^2\right)^2 \sum_{j=1}^{N} z_j^2}{\left(1 - \sum_{i=1}^{M} x_i^2\right)\left(1 - \sum_{j=1}^{N} z_j^2\right)}. \tag{5}$$

[6] Provided that $\gamma_0 \leq \min_j \gamma_j$, it is always possible to define such a joint distribution.

Note that if the plants' location decisions are made in accordance with the model of the previous section, then proposition 1 implies that the index γ is an unbiased estimate of the quantity $\gamma^{na} + \gamma^{s} - \gamma^{na}\gamma^{s}$ that captures the strength of the agglomerative forces in the model. For this reason the index has a number of desirable properties.

1. The index is easy to compute given the available data. In practice, the best available data on concentration are often a breakdown of total employment by some geographic subunits, for example, state-by-state employments for an industry in the United States or country-by-country employments for the European Community, and very little information is available on plant size distributions (on which our index requires only one moment).

2. The scale of the index allows one to make comparisons with a no-agglomeration benchmark in that $E(\gamma) = 0$ if the data are generated by the simple dartboard model of random location choices with no natural advantages or industry-specific spillovers.

3. The index is comparable across industries in which the size distribution of firms differs. Specifically, if each plant's location decision is made as in the model above, then the expected value of the concentration index is independent both of the number of plants and of their distribution.

4. The index is also comparable across industries regardless of differences in the level of geographic aggregation at which employment data are available in the different industries. While the geographic areas are built directly into the model specification, one can formalize this statement by supposing that the model describes how firms choose from a large set of M geographic areas (e.g., one for each square mile of the United States) and asking that the expected value of the index be unchanged no matter how the employment data are combined into M' larger aggregates before the index is computed.

To see that our index has this property given one specification of our location decision process, consider the example mentioned earlier in which $2[(1 - \gamma^{na})/\gamma^{na}]\bar{\pi}_i$ has a χ^2 distribution with $2[(1 - \gamma^{na})/\gamma^{na}]x_i$ degrees of freedom. The location process is then equivalent to drawing $(p_1, p_2, \ldots, p_M)$ from a Dirichlet distribution with parameters

$$\left(\frac{1 - \gamma^{na}}{\gamma^{na}}x_1, \frac{1 - \gamma^{na}}{\gamma^{na}}x_2, \ldots, \frac{1 - \gamma^{na}}{\gamma^{na}}x_M\right)$$

and then having each spillover-tied cluster of plants choose its location independently, with the probability of choosing area i being p_i.

When $(p_1, \ldots, p_M)$ has a Dirichlet distribution with parameters

$$\left(\frac{1-\gamma^{na}}{\gamma^{na}}x_1, \ldots, \frac{1-\gamma^{na}}{\gamma^{na}}x_M\right),$$

the distribution of $(p_1 + p_2, p_3, \ldots, p_M)$ is Dirichlet with parameters

$$\left(\frac{1-\gamma^{na}}{\gamma^{na}}(x_1 + x_2), \frac{1-\gamma^{na}}{\gamma^{na}}x_3, \ldots, \frac{1-\gamma^{na}}{\gamma^{na}}x_M\right).$$

Hence, data generated by aggregating areas 1 and 2 in an M-location model with parameters $\gamma^{na}, \gamma^s, x_1, \ldots, x_M$ will have the same distribution as data generated from an $M-1$ location model with parameters $\gamma^{na}, \gamma^s, x_1 + x_2, x_3, \ldots, x_M$. Repeating this argument as multiple areas are combined, we conclude that regardless of how areas are combined together, the expected value of an index computed from aggregated data remains $\gamma^{na} + \gamma^s - \gamma^{na}\gamma^s$.

When one makes the transition from the models to the real world, a caveat is necessary with regard to comparisons based on data at different levels of geographic aggregation. Our model imposes an extreme limitation on the geographic scope of forces that produce localization in two ways. First, when potential spillovers exist, they are realized only if firms choose to locate in the same geographic area. Second (at least in the χ^2 specification), natural advantages are drawn independently for each geographic area. In practice, we would expect that spillovers might provide some benefit also to plants locating in nearby areas. In this case, an estimate of γ that is computed from county-level data (and hence reflects only the added probability with which pairs of plants locate in the same county) would be expected to be smaller than an estimate that is computed from state-level data and reflects the additional colocations due to spillovers felt at some distance and to correlated natural advantages.

While the properties above can be taken as formalizing our motivation that an index should allow for meaningful comparisons across industries and with the null of no concentration, they are not axioms that determine our index uniquely. For example, any other unbiased estimator of $\gamma^{na} + \gamma^s - \gamma^{na}\gamma^s$ that could be computed from available data would also satisfy those properties. Our particular choice is to some degree arbitrary, although it does reflect a concern that the index reflects economically significant localizations. On these grounds, we would, for example, be uncomfortable with indexes based on plant count data (because such data tend to be dominated by very small plants that account for only a small portion of employment).

B. Interpreting the Scale of the Index

While the scale of γ is such that one can interpret a value of zero as indicating a complete lack of agglomerative forces, we would also like to be able to talk about whether particular positive values of the index are "large" or "small." We discuss here several ways in which one might try to get a feel for the scale of γ.

First and most informally, we find it useful in trying to interpret values of γ to keep a mental list of the γ's of things with which we feel somewhat familiar so that they can be used for comparisons. Appendix C of the working paper version of this paper (Ellison and Glaeser 1994) contains a complete list of the γ's of each four-digit manufacturing industry. Looking at industries that have previously attracted attention for their concentration, one can find there, for example, that the measured γ's for the U.S. automobile and automobile parts industries (SICs 3711 and 3714) are 0.127 and 0.089. The photographic equipment industry (SIC 3861) has a γ of 0.174. The carpet industry's (SIC 2273) γ is 0.378. The computer industry is a bit harder to find in the SIC codes, but the γ's for SICs 3571 (electronic computers), 3572 (computer storage devices), and 3674 (semiconductors and related devices) are 0.059, 0.142, and 0.064, respectively. As a reference point at the other extreme, one can look up industries that one could not imagine to be concentrated and find that the γ's of the bottled and canned soft drink (SIC 2086), manufactured ice (SIC 2096), newspaper (SIC 2711), and miscellaneous concrete products (SIC 3272) industries are 0.005, 0.012, 0.002, and 0.012, respectively.

Another source of reference points is the agglomeration of aggregate manufacturing activity. The model and index of this paper can also be applied to measure the concentration of overall U.S. manufacturing activity relative to the land area of the states. We typically think of manufacturing concentration as substantial. Computing our index using state manufacturing employment for s_i and land area for x_i, we find a γ of 0.055. On the other hand, if we restrict our attention to the states east of the Mississippi, manufacturing employment is much closer to being proportional to land area: the largest states—Georgia, Michigan, and Illinois—have far more manufacturing than the smallest—the District of Columbia, Rhode Island, and Delaware—and the raw correlation between manufacturing employment and land area is .50. The measured γ of manufacturing employment shares relative to land area in this subset is 0.019.

While the comparisons above may help build intuition, they do not provide an estimated dollar magnitude for the impact of natural

advantages/spillovers. One way to do this is to note that γ reflects the effect of natural advantages/spillovers on each state's share of manufacturing in an industry and that many previous studies have estimated the elasticity of plant locations with respect to cost differences. In the model (with a large number of plants), the effect of nature's allocation of natural advantage is to make the share p_i of plants that will locate in state i a random variable with mean x_i and variance $\gamma x_i(1 - x_i)$. For a state with $x_i = 0.02$ this means, for example, that when $\gamma = 0.01$, the standard deviation of p_i is $0.7E(p_i)$. A wide range of estimated new plant share–cost elasticities can be found in the literature.[7] If we assume the elasticity to be 25, the magnitude of the effects of natural advantages/spillovers on location decisions when $\gamma = 0.01$ is then similar to the effect of a cost shock whose standard deviation is 3 percent of total costs. The effect of natural advantages/spillovers with a γ magnitude of 0.10 would be similar to the effects of a cost shock with a standard deviation of 9 percent.[8] To put such differences in perspective, after one controls for education, tenure, and so forth in a log wage regression, the standard deviation of wage rates across states is about 8–10 percent of the level of wages. We would therefore regard a γ of 0.01 as indicating that cost differences are fairly small and a γ of 0.10 as indicating that cost differences are substantial.

Finally, we present a few magnitude calculations derived strictly from our model. In the model, the portion of γ due to spillovers is readily interpretable as an added colocation probability. For example, in an industry with 20 large plants, the expected number of large plants with which a given plant will colocate (on top of random colocations) is approximately 0.2 for $\gamma^s = 0.01$ and one for $\gamma^s = 0.05$. The magnitudes of natural advantages in the model are defined only in relation to the assumed magnitudes of the non-industry-specific advantages of the large states and the firm-specific idiosyncratic factors. To try to derive intuition from such a definition, table 1 records for several values of γ^{na} how likely it is that natural advantage will make Iowa a better location (for a firm with no idiosyncratic prefer-

[7] Given that energy costs are as small as 0.5 percent of total costs in some of the industries considered, the substantial energy price elasticities in Carlton's (1983) classic study imply elasticities of new plant shares to total costs in the 100–500 range. Carlton, however, finds much lower elasticities to wage differentials, and others (e.g., Bartik 1985) have failed to find significant responses to energy and other cost differences in other industries. In the literature on local tax rates and location decisions, Bartik's estimates imply elasticities with respect to total costs of around 50, whereas others (e.g., Schmenner, Huber, and Cook 1987) find very small elasticities. Crihfield (1990) finds the effects of taxes on growth rates to be small.

[8] These differences would be scaled up (down) linearly for smaller (larger) elasticities.

TABLE 1

EFFECT OF γ NATURAL ADVANTAGE RELATIVE TO STATE SIZE

γ^{na}	prob$\{\bar{\pi}_{IA} > \bar{\pi}_{GA}\}$	prob$\{\bar{\pi}_{IA} > \bar{\pi}_{MI}\}$	prob$\{\bar{\pi}_{IA} > \bar{\pi}_{CA}\}$
.005	.07	.006	.00
.01	.14	.03	.00
.02	.20	.08	.006
.05	.25	.14	.04
.10	.26	.15	.07
1.00	.27	.17	.09

ences) than Georgia, Michigan, and California.[9] For $\gamma^{na} = 0.01$, natural advantages are at times sufficiently powerful to make Iowa as attractive as Georgia, but they are rarely enough to overcome the nonindustry-specific advantages of the larger states. Natural advantage becomes sufficiently important to make Iowa as good as Michigan with a reasonable probability when γ^{na} is between 0.02 and 0.05, and Iowa starts to be comparable to California at times when γ^{na} is between 0.05 and 0.10.

In describing our results, we shall generally adopt the convention of referring to those industries with γ's above 0.05 as being highly concentrated and to those with γ's below 0.02 as being not very concentrated.

C. *Measurements of Coagglomeration*

Suppose now that we are given area industry employment and plant size data for each of r industries belonging to some group. As in Section II*B*, use G^j, H_j, and w_j for the raw concentration, the plant Herfindahl index, and the employment share of the jth industry. Let $\hat{\gamma}_j$ be the value of our index of concentration as computed from the data on the jth industry. Write G for the raw concentration of employment in the group as a whole and $H = \sum_j w_j^2 H_j$ for the group's plant Herfindahl index. As an index of the degree to which the industries in the group are coagglomerated, we propose the use of a measure γ^c defined by

$$\gamma^c \equiv \frac{\left[G \Big/ \left(1 - \sum_i x_i^2\right)\right] - H - \sum_{j=1}^{r} \hat{\gamma}_j w_j^2 (1 - H_j)}{1 - \sum_{j=1}^{r} w_j^2}. \tag{6}$$

[9] The figures pertain to the χ^2 specification of average profits. Iowa has approximately 1 percent of manufacturing employment, Georgia 3 percent, Michigan 5 percent, and California 11 percent.

Note that proposition 2 implies that γ^c is an unbiased estimate of the parameter γ_0 in the model of Section II*B*, and as such, it has the same robustness properties as γ does with regard to changes in the firm size distribution and in the level of data aggregation. As a measure of the importance of group-specific natural advantages and spillovers, magnitudes have the same meaning as they do for γ. An estimate of $\gamma^c = 0$ may be interpreted as indicating that there is no more agglomeration of plants in the group than that attributable to the tendencies of plants to locate near other plants in the same industry and where aggregate manufacturing employment is high.

In discussing the scope of spillovers/natural advantages, we find it useful also to rescale this measure, defining an index, λ, of the degree to which spillovers are general by

$$\lambda \equiv \frac{\gamma^c}{\sum_j w_j \hat{\gamma}_j}. \tag{7}$$

We interpret a value of $\lambda = 0$ as indicating that any spillovers/natural advantages found within the industry group are completely industry-specific. We interpret a value of $\lambda = 1$ as indicating that they are perfectly general in the sense that any spillovers benefit firms in all industries equally and natural advantages are perfectly correlated.[10]

IV. Data

Our index requires the distribution of employment across a set of geographic areas for a set of industries and the Herfindahl index of plant employment shares for those industries. There is a trade-off between locational fineness and industrial fineness in the available data, and for this paper, we have chosen to focus on the most narrowly defined industries possible: the 459 manufacturing industries defined by the four-digit classifications of the Census Bureau's 1987 SIC system. Given this decision, we settled on the 50 states plus the District of Columbia as our geographic division, and even at this level of disaggregation, a complicated and somewhat speculative data construction process was necessary.

Our construction of state-industry employments relies on data from the *Census of Manufactures*. These data are incomplete in that some state-industry employments are categorized or top-coded to protect confidentiality. Moreover, employment data are not re-

[10] Note that because of the parameter estimates in the denominator, λ is not an unbiased estimator of $\gamma_0/(\sum_j w_j \gamma_j)$ in our model.

ported for state-industries with fewer than 150 employees.[11] To complete our state-industry data set, we used a fairly elaborate computer program that tried to exploit the information contained in the across-state and across-industry adding-up constraints and in the state-industry plant count data when estimating employments in categorized and unreported state-industries. Some details on the process are provided in Appendix A.

Given our interpretation of the model as describing the location decision process of manufacturing plants, we must also construct an estimate of the Herfindahl index of plant employment shares in each four-digit industry. For this purpose, the relevant and available (subject to disclosure restrictions) census data consist of the number of plants and the total employment within plants in each of 10 employment size ranges.[12] We estimate Herfindahl indices from these data by a two-step procedure: employees were first allocated between the classes, and a Herfindahl index was then estimated by a procedure similar to that recommended by Schmalensee (1977), but taking into account the additional information available here in the form of the category divisions. The details of the data construction algorithm and a simulation analysis of the measurement errors it may create are discussed in Appendix B.

While we cannot be sure of the accuracy of our data-filling procedure, we do feel that it is an improvement over those that are typically used. The changes are likely to be particularly important in very small industries and highly geographically concentrated industries. The state employment data and the plant Herfindahl indices are available from the authors on request (the latter are listed also in app. C of Ellison and Glaeser [1994]).

Finally, to allow for a more thorough analysis of the geographic scope of concentration, we obtained a data set of 1987 county-level employments for three-digit industries. The data set had been constructed by filling in County Business Patterns data using an algorithm that consists largely of using mean plant sizes for nondisclosed employments (see Gardocki and Baj 1985). Some comparisons of these data with our main data set are also given in Appendix A.

V. Basic Results on Geographic Concentration

In this section we describe the patterns of geographic concentration in U.S. manufacturing industries. We begin at the broadest level with

[11] To give some idea of the magnitude of these restrictions, simply setting employment in each cell to its lower bound unequivocally identifies the location of 90 percent of employment in the median industry and 80 percent on average.

[12] The nondisclosures here are somewhat more problematic because they tend to obscure primarily the shares of the largest plants.

a discussion of whether any geographic concentration exists before moving on to discuss a few aspects in a little more detail.

A. *Are Industries Geographically Concentrated?*

The single most crucial question one must ask before further studying the geographic concentration of industries is whether geographic concentration really exists. While a number of previous writers have noted that localization appears to be widespread, we present here for the first time formal tests of the more stringent hypothesis that the extent of localization is greater than what would be expected to arise randomly.

In the simplest dartboard model in which the plants in an industry choose their locations in an independent random manner and there are no industry-specific spillovers or natural advantages, the result of proposition 1 is that $E(G) = (1 - \sum_i x_i^2)H$. The mean values of G and $(1 - \sum_i x_i^2)H$ across the 459 manufacturing industries in our data set are 0.74 and 0.27, respectively, and the difference between these two numbers is highly significant.[13]

When we look more closely at the industry-by-industry numbers, we find a prevalence of localization that we think is striking, even in light of the comments on the ubiquity of concentration found in Krugman (1991*a*) and so forth. The level of raw concentration G exceeds what would be expected to arise randomly in 446 of the 459 industries.[14] The flip side of this result—that in only 13 industries are plants more evenly distributed than would be expected at random—is interesting in that it indicates that the need to be near final consumers is rarely an overwhelming force in location decisions.

Because one might worry that manufacturing employment is not a good measure of the final demand in consumer goods industries, we performed this calculation also with population rather than manufacturing employment as the measure of state size. Such a calculation identifies 14 industries as being more evenly distributed than random (six of the 13 above and eight others), with the overall correlation between the two measures being .993.

[13] Under the null of $\gamma' = \gamma^m = 0$, a lengthy calculation shows that

$$\text{var}(G) = 2\left\{H^2[\Sigma x_i^2 - 2\Sigma x_i^3 + (\Sigma x_i^2)^2] - \sum_j z_j^4[\Sigma x_i^2 - 4\Sigma x_i^3 + 3(\Sigma x_i^2)^2]\right\}.$$

Using this formula, we estimate the standard deviation of the sample mean under the null to be 0.0005.

[14] The difference between G and $(1 - \sum_i x_i^2)H$ is larger than twice its standard deviation in 369 of the 446 industries in which the difference is positive and in none of the 13 industries in which the difference is negative.

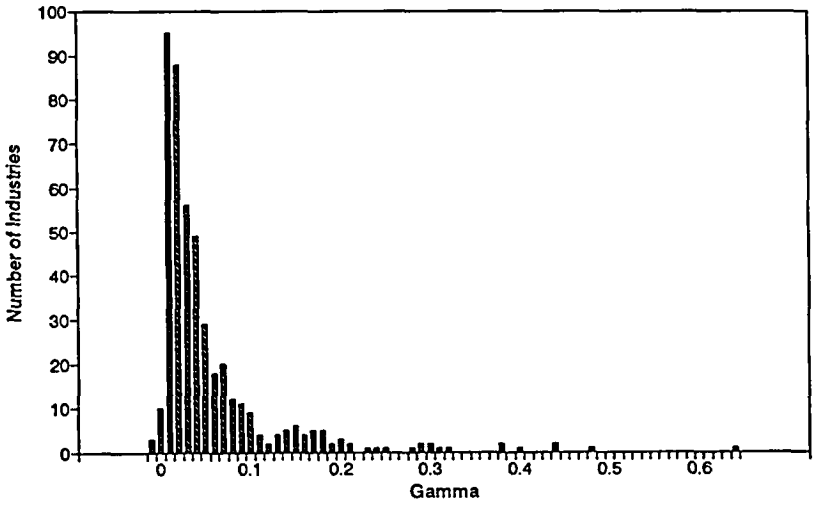


FIG. 1.—Histogram of γ (four-digit industries)

B. How Concentrated Are They?

In this subsection, we try to use our models to get a feel for how much concentration there is. We begin by imposing no structure across industries and simply computing the index γ defined by (5) for each of the 459 four-digit industries in our sample. A complete list of the γ's we find can be found in appendix C of Ellison and Glaeser (1994) and is also available from the authors on request.[15]

A histogram illustrating the frequency distribution of these γ's is presented in figure 1. In the figure, each bar represents the number of industries for which γ lies in an interval of width 0.01. The distribution in the figure appears to be quite skewed, with the mean being 0.051 and the median being 0.026. The most striking feature of the figure is the large number of industries falling into the range we described as not very concentrated ($\gamma < 0.02$). The tallest bar is the one corresponding to values of γ between zero and 0.01, and 43 percent of the industries have $\gamma < 0.02$. On the other side, the figure displays a thick right tail, with slightly more than a quarter of the

[15] If one interprets γ's as estimates of $\gamma^{na} + \gamma^{a} - \gamma^{na}\gamma^{a}$ (as opposed to estimates of the realized sum of squared differences between the p's and the x's), these γ's are measured with substantial errors. To get a feel for the magnitudes, we computed standard errors by simulating a special case of our natural advantage model: that of Dirichlet-distributed state sizes. Among industries with $H < 0.02$, the mean of the estimated standard errors is 0.02. The means for industries with H in the ranges 0.02–0.05, 0.05–0.10, and 0.10–1.0 are 0.024, 0.041, and 0.072, respectively.

TABLE 2

RAW CONCENTRATION ATTRIBUTABLE TO SPILLOVERS/COMPARATIVE ADVANTAGE: FRACTION OF INDUSTRIES WITH $(1 - \sum x_i^2)\gamma / G$ IN RANGE

Range	All Industries	High-G Industries
<0	.03	.03
.00–.25	.09	.10
.25–.50	.22	.16
.50–.75	.32	.19
.75–1.00	.33	.53

industries having a γ of at least 0.05 and 59 having a γ of at least 0.10. While the automobile, computer, carpet, and other industries people have used as examples of concentration are far from typical, there are a substantial number of industries that have received less attention and are similarly concentrated. We would thus like to amend our earlier conclusion that concentration is remarkably widespread to read that slight concentration is remarkably widespread, with the more extreme concentration that has attracted attention existing in a smaller subset of industries.

To provide a rough idea of how important it is to account properly for random agglomeration when constructing an index of geographic concentration, table 2 lists the frequency with which the ratio $(1 - \sum_i x_i^2)\gamma / G$ falls into a number of intervals, both for all industries and for the subsample of those in the upper quartile of raw geographic concentration. We can think of the fraction as a rough measure of the portion of raw concentration that is legitimately attributable to some form of spillovers/natural advantage rather than to randomness. The table indicates that the two components are comparable in magnitude and that there is a great variation in the mix between them. In roughly one-third of the industries (both overall and among the industries with high raw concentration), the fact that plants are discrete units and that some clusters appear at random accounts for at least as large a part of measured raw concentration as actual agglomerations of plants do. It is, therefore, not surprising that our index gives a somewhat different picture of geographic concentration than previous discussions of raw concentrations have.

C. Patterns of Concentration

While an attempt to explore formally the industry characteristics that tend to be associated with localization is well beyond the scope

of this paper, we felt that a couple of simple tables would be of interest.[16]

Table 3 summarizes the levels of geographic concentration of the four-digit subindustries of each two-digit manufacturing industry. For each two-digit industry, the table lists the fraction of subindustries that fall in the not very localized ($\gamma < 0.02$), intermediate, and very localized ($\gamma > 0.05$) ranges. High levels of geographic concentration are most prevalent in the tobacco, textile, and leather industries and most rare in the paper, rubber and plastics, and fabricated metal products industries.

Table 4 lists the 15 most and the 15 least localized industries in terms of the index γ. As Krugman (1991*a*) has previously noted, there is no obvious single factor accounting for extreme concentration. The most concentrated industry, furs, is probably explained both by the local transfer of knowledge from one generation to the next and as a response to buyers' search costs. Furs also have an unusually high ratio of value to weight that may make physical transportation costs less important. The next most concentrated industry, wine, may be largely attributable to the natural advantage of California in growing grapes. Natural advantage may also be important in the carbon black, raw cane sugar, and phosphatic fertilizer industries (and perhaps very indirectly in the oil field machinery industry). While a single spillover-based explanation may account for the concentration of the various textile industries in the Southeast, the remaining industries seem quite disparate.

The list of the 15 least concentrated industries is also something of a mixed bag. The industries certainly do not stand out as being those in which spreading out to be close to final consumers is important, and the list contains several industries, for example, vacuum cleaners and small-arms ammunition, in which raw concentration is substantial, but employment turns out to be concentrated in a few very large (randomly scattered) plants.[17]

D. The Geographic Scope of Concentration

In Section III, we noted that the γ's estimated from county-, state-, or region-level data should be identical (in expectation) provided that the scope of spillovers is such that advantages are gained only

[16] For interesting work on this topic, see Henderson (1988) and Enright (1990).

[17] In interpreting these latter cases, the reader should keep in mind that the errors in measuring γ include both the inherent uncertainty of analyzing random dart throws and errors in filling in census nondisclosures. Each of these components is larger when H is larger, so the list may contain many industries with a large H simply because this is where we have made the largest errors in measurement.

TABLE 3

CONCENTRATION BY TWO-DIGIT CATEGORY

TWO-DIGIT INDUSTRY	NUMBER OF FOUR-DIGIT SUBINDUSTRIES	PERCENTAGE OF FOUR-DIGIT INDUSTRIES WITH		
		$\gamma < .02$	$\gamma \in [.02, .05]$	$\gamma > .05$
20 Food and kindred products	49	47	18	35
21 Tobacco products	4	0	0	100
22 Textile mill products	23	9	13	78
23 Apparel and other textile products	31	13	42	45
24 Lumber and wood products	17	29	47	24
25 Furniture and fixtures	13	69	8	23
26 Paper and allied products	17	53	47	0
27 Printing and publishing	14	71	14	14
28 Chemicals and allied products	31	38	24	38
29 Petroleum and coal products	5	60	0	40
30 Rubber and miscellaneous plastics	15	73	27	0
31 Leather and leather products	11	0	36	64
32 Stone, clay, and glass products	26	58	27	15
33 Primary metal industries	26	39	35	27
34 Fabricated metal products	38	61	32	8
35 Industrial machinery and equipment	51	49	26	26
36 Electronic and other electric equipment	37	41	46	14
37 Transportation equipment	18	28	33	39
38 Instruments and related products	17	47	41	11
39 Miscellaneous manufacturing industries	18	44	22	33

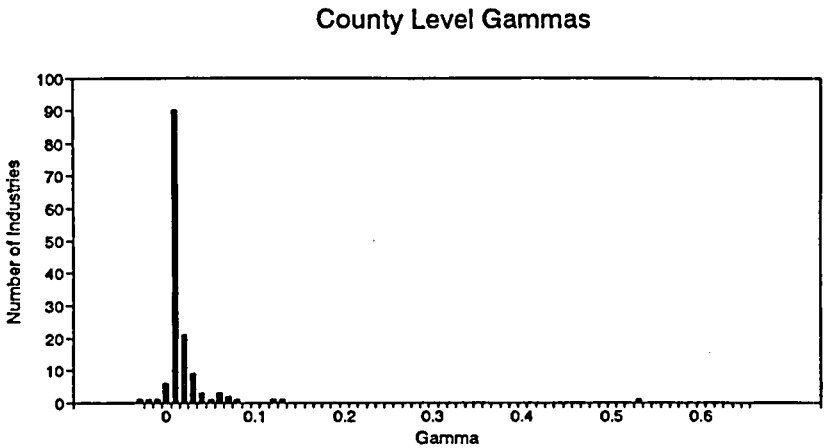
TABLE 4

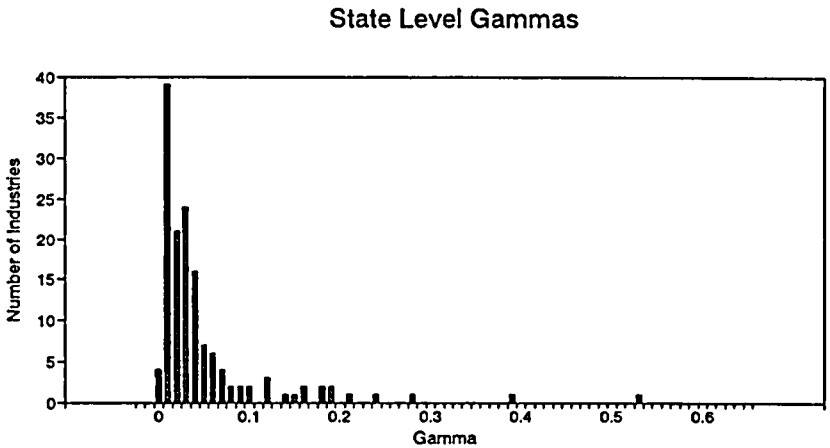
MOST AND LEAST LOCALIZED INDUSTRIES

Four-Digit Industry	H	G	γ
	15 Most Localized Industries		
2371 Fur goods	.007	.60	.63
2084 Wines, brandy, brandy spirits	.041	.48	.48
2252 Hosiery not elsewhere classified	.008	.42	.44
3533 Oil and gas field machinery	.015	.42	.43
2251 Women's hosiery	.028	.40	.40
2273 Carpets and rugs	.013	.37	.38
2429 Special product sawmills not elsewhere classified	.009	.36	.37
3961 Costume jewelry	.017	.32	.32
2895 Carbon black	.054	.32	.30
3915 Jewelers' materials, lapidary	.025	.30	.30
2874 Phosphatic fertilizers	.066	.32	.29
2061 Raw cane sugar	.038	.30	.29
2281 Yarn mills, except wool	.005	.27	.28
2034 Dehydrated fruits, vegetables, soups	.030	.29	.28
3761 Guided missiles, space vehicles	.046	.27	.25
	15 Least Localized Industries		
3021 Rubber and plastics footwear	.06	.05	−.013
2032 Canned specialties	.03	.02	−.012
2082 Malt beverages	.04	.03	−.010
3635 Household vacuum cleaners	.18	.17	−.009
3652 Prerecorded records and tapes	.04	.03	−.008
3482 Small-arms ammunition	.18	.17	−.004
3324 Steel investment foundries	.04	.04	−.003
3534 Elevators and moving stairways	.03	.03	−.001
2052 Cookies and crackers	.03	.03	−.0009
2098 Macaroni and spaghetti	.03	.03	−.0008
3262 Vitreous china table, kitchenware	.13	.12	−.0006
2035 Pickles, sauces, salad dressings	.01	.01	−.0003
3821 Laboratory apparatus and furniture	.02	.02	−.0002
2062 Cane sugar refining	.11	.10	.0002
3433 Heating equipment except electric	.01	.01	.0002

if firms choose identical locations, with natural advantages being independent across geographic areas. If, on the other hand, the effect of spillovers (or the spatial correlation of natural advantage) is smoothly declining with distance, then those γ's will reflect the excess probability with which pairs of firms tend to locate in the same county, state, and region, respectively. To investigate the geographic scope of spillovers, we estimated γ's from our county/three-digit data set using counties, states, and the nine census regions as the units of observation.

Figure 2 presents histograms of the γ's estimated from the three





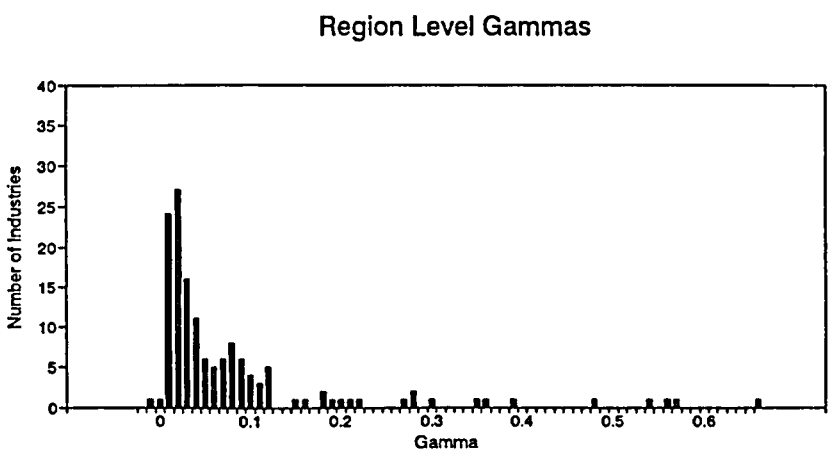


FIG. 2.—Concentration at the county, state, and regional levels

levels of data. Comparing the county- and state-level estimates, we find substantially more concentration at the state level. The median γ at the county level is 0.005, and the median γ at the state level is 0.023. The median of the ratio between them is 0.25, so typically the effect of spillovers is such that about one-fourth of the excess tendency of plants to locate in the same state involves plants' locating in the same county. We draw two conclusions. First, because one-fourth of all excess colocations do involve plants' locating in the same county (while states have many more than four counties), within-county spillovers are stronger than nearby-county spillovers. Second, "localized" spillovers are still quite substantial at a range beyond that of counties. In only a few cases do spillovers appear to be both substantial and limited in scope to the county level.[18] The rubber and plastics footwear industry seems to be the unique example in which concentration is substantially greater at the county level than at the state level, that is, where tightly grouped clusters of plants are spread (excessively) evenly across the states as though to minimize transportation costs.

Measured levels of state and regional concentration are more similar, although the regional data show a much thicker tail of very concentrated industries. (The mean γ's are 0.044 and 0.078.) The general pattern that slightly more than half of the tendency of firms to locate in the same region is accounted for by the tendency to locate in the same state appears to hold equally well for industries that are very unconcentrated and very concentrated at the state level, although there is considerable variation about this norm.[19]

VI. Evidence on Coagglomeration

In this section, we present some descriptive evidence on the coagglomeration of industries. First, we examine the extent to which geographic concentration tends to be a characteristic of broadly or narrowly defined industries by discussing the coagglomeration of SIC-similar industries. Next, to explore the importance of transportation costs or information flows between buyers and sellers, we look at the coagglomeration of pairs of industries with strong upstream-downstream relationships.

[18] The most notable cases are fur goods, building paper and board mills, and periodicals.

[19] Industries notable for unusually high (relative) regional concentration include ordnance and accessories, nonferrous foundries, and cigarettes. Industries in which state-level clusters are unusually dispersed include photographic equipment and supplies, radio and television receiving equipment, and periodicals.

TABLE 5

CONCENTRATION AND INDUSTRY DEFINITION

	INDUSTRY MEANS		
INDUSTRY DEFINITION	H	G	γ
Two-digit	.007	.031	.026
Three-digit	.014	.056	.045
Four-digit	.028	.074	.051

A. *Industry Definition*

Table 5 provides a simple look at the concentration of two-, three-, and four-digit industries. While raw geographic concentration increases steadily as we move to finer industry definitions, the increase in γ appears to come more abruptly as we move from the two-digit to the three-digit level. This naturally raises two questions of scope. Is there any correlation in the location decisions of firms that share only a two-digit industry class, or is the concentration of two-digit industries entirely a consequence of the localization of their three-digit subindustries? Are location decisions influenced as strongly by the locations of plants belonging to different four-digit industries within the same three-digit class as they are by the locations of plants belonging to their own four-digit industry?

To address the latter question, we calculated for each of the 97 three-digit industries with more than one four-digit subindustry our measures γ^c and λ of the degree to which the four-digit subindustries are coagglomerated. Recall that the scale of γ^c is the same as that of γ, whereas λ measures the strength of coagglomerative forces relative to agglomerative forces. A value of $\lambda = 0$ would indicate that the subindustries exhibit no coagglomeration at all, and a value of $\lambda = 1$ would indicate that the natural advantages and spillovers that exist are (three-digit) group-specific rather than (four-digit) industry-specific. Figure 3 contains a histogram of the values of λ we estimate, which are fairly evenly spread between zero and 0.8. From this we conclude that there is some coagglomeration of four-digit industries, but it is rare for spillovers to be almost completely general to three-digit classes. Perhaps most interesting, the histogram suggests that there is considerable heterogeneity across industries in the specificity of spillovers.

Let us move on to yet broader industry classes. Table 6 reports the values of the γ^c and λ obtained from a similar calculation using the three-digit subindustries of each two-digit industry. The mean value of λ across two-digit industries is 0.29. There is again a great

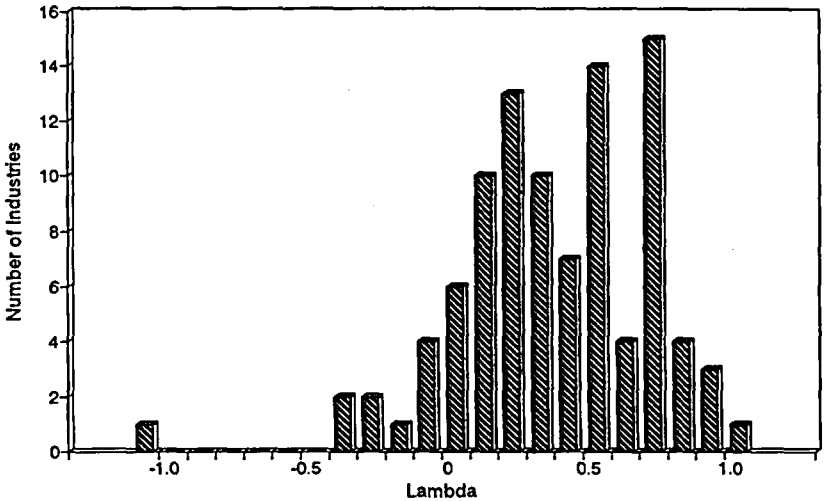


FIG. 3.—Histogram of λ: extent of spillovers between four-digit subindustries of three-digit industries.

TABLE 6

EXTENT OF SPILLOVERS BETWEEN THREE-DIGIT INDUSTRIES

Two-Digit Industry	γ^c	λ
Food and kindred products	.002	.14
Tobacco products	.151	.88
Textile mill products	.115	.61
Apparel and other textiles	.010	.29
Lumber and wood products	.016	.63
Furniture and fixtures	.001	.02
Paper and allied products	.005	.31
Printing and publishing	.005	.48
Chemicals and allied products	.007	.25
Petroleum and coal products	.007	.12
Rubber and miscellaneous plastics	.003	.38
Leather and leather products	.017	.31
Stone, clay, and glass products	.002	.20
Primary metal industries	.012	.41
Fabricated metal products	.003	.22
Industrial machinery and equipment	.000	.00
Electronic and other electric equipment	.000	.02
Transportation equipment	−.001	−.08
Instruments and related products	.013	.36
Miscellaneous manufacturing	.011	.34

variation across industries. In four cases (furniture, industrial machinery, electronic and electric equipment, and transportation equipment), the data indicate that there is no coagglomeration at all at the two-digit level. On the other hand, there is substantial coagglomeration of the three-digit subindustries within the two-digit tobacco, textile, and lumber industries.

B. *Coagglomeration and Upstream-Downstream Relationships*

We examine here the coagglomeration of industries with strong upstream-downstream ties in hopes that it may provide some suggestive evidence on economizing on transportation costs as a motivation for agglomeration. Our analysis focuses on two lists of 100 industry pairs that we constructed using data from the Census Bureau's six-digit commodity-by-industry direct requirements table: one consisting of the 100 (downstream) industries that receive the largest value of inputs per dollar value of output from a single upstream industry (paired with that supplier) and the other consisting of the 100 (upstream) industries that sell the largest portion of their output to one downstream industry. For example, the first list contains the ice cream and frozen dessert industry paired with the milk industry (from which it purchases a large amount of inputs per dollar of output), and the second contains the engine electrical equipment industry paired with the motor vehicle industry (to which it sells a large portion of its output).

In the set of 100 industry pairs in which the downstream industry is heavily dependent on an upstream input, we find clearly significant evidence that there is a tendency to coagglomerate: for 77 of the pairs, γ^c is positive. The mean level of γ^c for these pairs is 0.018, which we would not regard as being particularly large, although nine of the pairs have coagglomeration γ^c's in the range we would call very concentrated (above 0.05). Of the 100 pairs in which the upstream industry has an important customer, 68 exhibit some coagglomeration, with the mean of γ^c being 0.015 and 10 of the pairs having γ^c above 0.05. Table 7 lists the top 15 pairs of each type (ranked on input/output dependency) along with the γ^c and the λ of the pair.

VII. Geographic Concentration within the Firm

In this section, we investigate the tendency of plants belonging to the same firm to locate together. While our data set does not allow us to provide detailed descriptive evidence on the topic, we felt that some treatment of the issue was necessary to see whether such a

TABLE 7

COAGGLOMERATION OF UPSTREAM-DOWNSTREAM INDUSTRY PAIRS

Upstream Industry	Downstream Industry	γ^c	λ
Pairs with Downstream Industry Relying on Upstream Input			
2026 Fluid milk	2021 Ice cream and frozen desserts	.005	1.05
2824 Organic fibers, noncellulosic	2296 Tire cord and fabrics	.078	.55
2011 Meat packing	2013 Sausages and prepared meats	.014	.50
3312 Blast furnaces and steel mills	3316 Cold finishing of steel shapes	.048	.74
3312 Blast furnaces and steel mills	3449 Miscellaneous metal work	.017	.28
2421 Sawmills and planing mills	2439 Structural wood members	.016	.44
3339 Primary nonferrous metals	3356 Nonferrous rolling and drawing	.014	1.20
3312 Blast furnaces and steel mills	3315 Steel wire and related	.019	.31
3312 Blast furnaces and steel mills	3412 Metal barrels, drums, pails	.015	.23
3312 Blast furnaces and steel mills	3465 Automotive stampings	.052	.48
3714 Motor vehicle parts, accessories	3711 Motor vehicles, car bodies	.107	1.02
3312 Blast furnaces and steel mills	3441 Fabricated structural metal	.004	.09
2075 Soybean oil mills	2079 Edible fats and oils not elsewhere classified	.033	.69
2421 Sawmills and planing mills	2941 Wood preserving	.027	.70
3312 Blast furnaces and steel mills	3448 Prefabricated metal buildings	−.006	−.11

Pairs with Upstream Industry Having Large Downstream Buyer				
3313	Electrometallurgical products	3312	Blast furnaces and steel mills	.059	.85
2083	Malt	2082	Malt beverages	.032	87.19
3493	Steel springs except wire	3711	Motor vehicles, car bodies	.006	.05
3714	Motor vehicle parts, accessories	3711	Motor vehicles, car bodies	.107	1.02
2087	Flavoring extracts and syrups	2086	Bottled and canned soft drinks	.001	.19
3465	Automotive stampings	3711	Motor vehicles, car bodies	.149	1.05
2395	Pleating and stitching	3711	Motor vehicles, car bodies	−.028	−.23
3694	Engine electrical equipment	3711	Motor vehicles, car bodies	.011	.09
3292	Asbestos products	3714	Motor vehicle parts, accessories	−.019	−.22
3255	Clay refractories	3312	Blast furnaces and steel mills	.044	.64
2732	Book printing	2731	Book publishing	.000	.00
2076	Vegetable oil mills not elsewhere classified	2079	Edible fats and oils not elsewhere classified	.003	.10
2074	Cottonseed oil mills	2048	Prepared feeds not elsewhere classified	.020	.67
2399	Fabricated textile not elsewhere classified	3711	Motor vehicles, car bodies	−.017	−.15
3331	Primary copper	3351	Copper rolling and drawing	.000	−.01

tendency could account for a significant portion of the localization we have identified.

To analyze the potential for measuring agglomeration within the firm, we consider an industry consisting of r firms with shares $w_1, w_2, \ldots, w_r$ of the industry's employment. Let $H_f = \sum_j w_j^2$ be the Herfindahl index of the firms' employment shares. To avoid confusion, we shall use H_p in this section for the Herfindahl index of the plants' employment shares. Suppose that firm j consists of n_j plants having shares $z_{j1}, \ldots, z_{jn_j}$ of the industry's employment. Suppose that the location choices of the plants are made analogously to those of our multi-industry model (with the firms analogous to subindustries), with the correlation of the location choice indicator variables u_{ki} and u_{li} being γ_0 if plants k and l belong to different firms and $\gamma_1 > \gamma_0$ if they belong to the same firm. A direct corollary of proposition 2 is proposition 3.

PROPOSITION 3. In the model above,

$$E(G) = \left(1 - \sum_i x_i^2\right)[H_p + \gamma_0(1 - H_f) + \gamma_1(H_f - H_p)].$$

When one tries to apply the prediction of this model to recover γ_1, a great obstacle arises: state-firm employments are much harder to find than state-industry employments. As a result, we cannot separately estimate γ_0 and γ_1 for a single industry. What we try to do instead is to identify average values of γ_0 and γ_1 using cross-industry variation. Specifically, we note that if one makes the heroic assumption that the parameters γ_{0i} and γ_{1i} for industry i are random variables whose conditional means are independent of H_{pi} and H_{fi}, then the coefficients $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from the ordinary least squares (OLS) regression

$$\frac{G_i}{1 - \sum_j x_j^2} - H_{pi} = \alpha_0(1 - H_{fi}) + \alpha_1(H_{fi} - H_{pi}) + \epsilon_i$$

are consistent for $E(\gamma_0)$ and $E(\gamma_1)$.

We estimated the regression above for our sample of 444 four-digit industries. The parameter estimates for $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are 0.046 (standard error 0.005) and 0.068 (standard error 0.067), respectively. While the first coefficient estimate is highly significant, the second is quite imprecise. Hence, while the point estimate is that plants belonging to the same firm are slightly more agglomerated than other plants in the same industry, we cannot rule out a substantially higher level of intrafirm agglomeration. Given that the mean

of $H_f - H_p$ is only 0.04, we can say fairly confidently that only a very small portion of total geographic concentration is attributable to intrafirm agglomerations.

VIII. Conclusion

In this paper, we have developed a model for the analysis of geographic concentration that captures both the "random" agglomeration a dart-throwing model would produce and additional agglomeration caused by localized industry-specific spillovers and natural advantages (which we feel have received less attention than they merit given their empirical importance). The model suggests that it is possible to control for industry characteristics in a fairly robust manner when measuring geographic concentration, and we have proposed new indices for the measurement of the localization of industries and the relative strength of cross-industry agglomerations.

While reaffirming that geographic concentration is ubiquitous and that there are many highly concentrated industries, the results are clearly not as proconcentration as some previous statements. Many industries are only slightly concentrated, and some of the most extreme cases of concentration are likely due to natural advantages. Clearly, though, there remains significant concentration to be explained. We have tried also to provide a quick summary of some of the patterns that exist in the coagglomeration of related industries, in the geographic scope of agglomeration, and so forth, and there remains in each case a great deal of heterogeneity to be explored.

Appendix A

This Appendix describes the process by which state-industry employment figures were constructed. For each state-industry with at least 150 employees, the 1987 *Census of Manufactures* reports employment rounded to the nearest multiple of 100 or categorizes employment as belonging to one of five ranges: 100–249, 250–499, 500–999, and 1,000–2,500. Table A1 indi-

TABLE A1

EXTENT OF WITHHELD DATA

	INDUSTRY DEFINITION		
	Two-Digit	Three-Digit	Four-Digit
Industries	21	141	460
Cells with ranges	153	1,776	5,700
Top codes	46	268	487
Average employment fraction	.02	.11	.20

cates the number of these state-industries for which data are categorized, the number of those that are top-coded at 2,500 or more employees, and the average across industries of the fraction of employees whose state cannot be determined simply by assigning each state its minimum possible employment.

Before beginning to fill in the data, we first adjust the upper or lower bounds on any two- or three-digit state-industry for which a sharper bound can be obtained by summing the upper or lower bounds of the subindustries that constitute it. This reduces the number of two- and three-digit state-industries without upper bounds to 13 and 157, respectively. In addition, a total of 82 and 680 bounds are tightened on cells in which a non-top-coded range had been given.

The filling process begins with the 21 × 51 matrix of two-digit data. First, a rough estimate of the total employment in cells that are reported as zero is made for each state and for each industry. The estimate is simply 35 times the number of missing firms with 20 or more employees plus 6 times the number of missing firms with fewer than 20 employees, provided that this total is less than 150 times the number of empty cells in the appropriate row or column. (Each of these estimates is fewer than 600 employees.)

The main part of the algorithm assigns values within the given range to each cell, trying to do so in a manner that makes the sums of the rows and columns as close as possible to those indicated by the reported totals for employment in each industry and manufacturing employment in each state. While this could be treated as a large optimization problem with a number of variables equal to the number of categorized state-industry employments, this approach was deemed intractable. Instead an admittedly ad hoc procedure was used to sequentially fill in cells. Essentially, the procedure repeatedly looks at the matrix of data, identifies the categorized cells for which there is the least uncertainty as to employment, fills in employment of those cells, and again looks at the matrix in which the filled-in numbers are accepted as fact.

The process of identifying which cells to fill in follows a set of priorities. First, if there are any rows or columns for which all categorized cells must be set to the minimum or maximum to satisfy adding-up constraints, those cells are chosen. Next, the algorithm looks for rows or columns in which only a single element is unknown. If all rows and columns have multiple unknown cells, the algorithm selects the row or column in which there is the least variance possible within the unknown ranges. As a result of the manner in which this is done, usually top codes are not filled in until virtually all active rows and columns contain a top code, and rows/columns with multiple top codes are not filled until there are no rows/columns with a single top code remaining. When filling cells in a row with multiple unknown elements, the algorithm looks at the departures from expected employment in the row and column of each unknown cell and adjusts the cells in a direction calculated loosely on the analogy of calculating conditional means of normal random variables. The amount by which a cell is adjusted is limited by the constraint that its row/column must be able to sum as well.

After the two-digit data are filled in, the process is repeated on the three- and four-digit data. The only difference is that instead of using the constraint that the state-industry employments should add up to the state total manufacturing employment, we use the set of constraints dictated, for example, by employment within each state in the three-digit subindustries of a two-digit industry adding up to the employment in that state in the two-digit industry.

In addition, the previously estimated state and industry total employments in states whose employments are reported as zero are allocated across state-industries by an algorithm identical to that described above. In the four-digit data, these rounded-to-zero employments are occasionally a nontrivial fraction of the total employment in an industry.

While there is no way to tell whether this algorithm is doing well, it is at least possible to tell whether it is doing badly to the extent that the algorithm is unable to make the state or industry totals add up (although because of rounding errors, totals are off by up to 400 employees in industries in which no data are withheld). Of the 21 two-digit industries, the maximum error in the adding-up constraints is 508 employees, with all other industries within 400. In the three-digit industries and four-digit industries, there are two and six industries in which the error is greater than 400; two four-digit industries have errors greater than 1,000 employees, the maximum being 2,010 (although these two are very big industries). The average errors in the state adding-up constraints are 31, 177, and 558 at the two-, three-, and four-digit levels. In all but one of the two-digit industries and in all but six of the three-digit industries, it was never necessary to fill in multiple top codes at the same time.

We would have liked to simulate a data-withholding process to provide rough estimates of the bias and variance of measurement error on the raw geographic concentration measure G induced by our data filling. However, the census's withholding process is not sufficiently transparent that we felt confident that we could reasonably simulate it. Without that, we present here a small test of the accuracy of our procedure based on data obtained separately from County Business Patterns (CBP) for the area in which our procedure is most suspect, filling in top codes in the four-digit data.[20]

Data were available from CBP on state-industry employment for 171 of the 487 four-digit state-industries in which employment was top-coded at 2,500 or more. The CBP's sample differs somewhat from the *Census of Manufactures*, and as a result the CBP reported that employment is below 2,500 in 30 of these state-industries. We dropped these state-industries from our test. (We chose not to use CBP data as an input to our algorithm precisely because they are often incompatible with range and adding-up constraints in the *Census of Manufactures* data.) Of the remaining 141 state-industries, four have very large employments; in each case our data fit extremely well, giving our estimates a misleadingly high .98 correlation with the CBP data. After we delete these four state-industries, the mean and standard deviation

[20] These data have previously been used by Enright (1990), among others, to fill in some of the top-coded *Census of Manufactures* data.

of employment in the remaining 137 state-industries are virtually identical in our data and in the CBP data; the correlation between the two is .74. (The means are 5,329 and 5,304; the standard deviations are 3,451 and 3,306.) For comparison, if the *Census of Manufactures* had reported ranges for these data using the CBP ranges (2,500–4,999, 5,000–10,000, and 10,000–20,000) and we had constructed estimates simply by filling in the mean of the appropriate range, the correlation coefficient would be higher (.93), but the sample means and variance would be much farther from those of the CBP data. (The mean would be 5,939 and the standard deviation 4,314.)

While the results above suggest that our procedure has some accuracy in filling in, the most important question is clearly what implications errors in assigning state employments have on the computation of G. Even a procedure that is quite inaccurate might yield reasonable estimates of G if it simply assigns clusters of employment to the wrong states. As a rough estimate of the effect that our filling in of top codes has on the computation of G, we constructed a measure of G_{CBP} by substituting the CBP employment totals for our filled-in employment totals for all top-coded cells in the 61 industries in which the CBP data allowed all top codes to be filled in (and where there was at least one top code). For this purpose we took the CBP data to report employment of 2,500 whenever it actually reported a smaller number. Comparing our previously estimated G with the value G_{CBP}, we find that the means are 0.052 and 0.048, with a correlation of .96. The absolute value of the difference between the two has a median of 0.0014, with the value being larger than 0.005 in 11 of the 61 industries. While this suggests that our filling in of top codes does not induce significant bias or large measurement errors, we should point out that the industries in which this test was performed may have been among the easier industries with top codes to fill in because they tended to have fewer top codes than the average industry with at least one top code (1.5 vs. 2.5). On the other hand, the majority of four-digit industries have no top-coded cells to begin with. Also, while the filled-in top codes would appear to be the greatest potential problem with our algorithm, this test says nothing about biases due to the filling in of non-top-coded ranges and of state-industry employments of fewer than 150.

For another look at the sensitivity of measured levels of concentration to the way in which we filled in the data, we compared the values of G obtained from state/three-digit industry calculations with our standard data set and with state totals from our county-level data set. (Recall that this latter data set had been constructed entirely from CPB data using mean establishment sizes to fill in missing values.) Because the latter data set is not based on the 1987 SIC revision, the comparisons below involve only the 96 SIC codes whose definitions were unchanged. The values of G from the two data sources differ (in absolute value) by less than 0.005 in 59 of the 96 industries. The difference is between 0.01 and 0.02 in 13 industries, and greater than 0.02 in eight. In several of these cases, however, the values of G are quite large, so that we may regard the two data sets as giving roughly similar measurements. The differences are both larger than 0.015 and larger than 20 percent of the larger G for only six SIC codes: 213, 315, 321,

375, 386, and 387. The data for these industries should perhaps be treated with some caution.

Appendix B

This Appendix discusses the manner in which an estimated plant Herfindahl index, H, was constructed from the census data and the potential implications for our measurements of geographic concentration. Given that a significant amount of information about the distribution of plant shares within each industry is available, we have chosen to construct H by a procedure that is much more akin to filling in data than to imposing any distributional assumptions and estimating parameters, and therefore will admittedly be ad hoc. The algorithm has two main steps: the first consists of allocating employees across size classes to obtain a regular data structure, and the second consists of computing an expected sum of squares for the plants within each class using a rule of thumb recommended by Schmalensee (1977).

When nondisclosure constraints do not bind, the *Census of Manufactures* reports for each industry the number of plants and the total employment in plants belonging to each of 10 employment size categories: 1–4, 5–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1,000–2,499, and 2,500 or more. In 316 of the 459 industries, however, the Census Bureau has withheld data on the total employment within a size class (typically one with three or fewer plants). In this case, the census data instead contain the combined employment in this class and another indicated class. To perform a rough separation of the employment in combined classes, for each size class we first used the sample of industries for which the total employment is reported to estimate the mean and variance of employment/plant as a function of the number of plants in the class. (The mean was assumed to have the form $a_0 + a_1 \log[1 + n]$ and the variance the form $b_0 + b_1[1/n]$, with the parameters estimated by OLS regressions.) Employment in each of the combined classes was then set so that departures from the predicted means were inversely proportional to the predicted variances, provided that this did not violate the upper and lower bounds on plant size.

The second step procedure essentially consists of assuming that the sizes of the plants within each class are discretely uniformly spread on a range centered on the mean, with its boundary at the closer of the two end points of the size range. The index H is estimated simply by taking the sum of the squares of the plant shares for this particular allocation of employees across plants. Schmalensee reports that this assumption of linear shares within a class seems to give the best estimates of the Herfindahl index in a similar problem.

We do not regard this procedure as an attempt to assign employments to plants, but just as a complicated function that approximates the Herfindahl index given the available data. To assess the accuracy of this procedure, we constructed a simulated data set of 5,000 industries. The simulated indus-

tries were created by assuming that the plant sizes in industry i consist of n_i draws from a lognormal distribution with mean μ_i and standard deviation σ_i. The parameters n_i, μ_i, and σ_i were themselves realizations of independent lognormal random variables with means (standard deviations) 527 (1,106), 143 (286), and 287 (2,101), respectively. These parameters were obtained from sample statistics (and the estimated H) of our 459-industry sample. The data produced by the simulations bear a superficial resemblance to the actual data, although they tend to contain far more extreme outliers (e.g., industries with over 95 percent of employment in a single plant). We created a simulated data set modified to preserve confidentiality by combining employment in any size class with two or fewer plants with the employment in the next lower nonempty size class. This modification involved withholding data in 3,200 of the 5,000 simulated industries.

We applied our algorithm to this data set to produce estimated plant Herfindahls, $\hat{H}$, and compared them to the true H. On average, the estimated Herfindahls were slightly smaller than the true values, the ratio of the means being 1.05. We principally use estimates of H in the paper as a part of the computation of γ for each industry. Note that if we set $\gamma = [G/(1 - \sum_i x_i^2) - \hat{H}]/(1 - \hat{H})$, where $G = (1 - \sum_i x_i^2)[\gamma_0 + (1 - \gamma_0)H + \epsilon]$ with $E(\epsilon|H, \hat{H}) = 0$, then

$$E(\gamma - \gamma_0|\hat{H}) = (1 - \gamma_0)E\left(\frac{H - \hat{H}}{1 - \hat{H}}\middle|\hat{H}\right).$$

Hence, if $E(H|\hat{H}) = \hat{H}$, then our estimates of γ_0 will be unbiased.

One cannot estimate $E(H|\hat{H})$ without making assumptions about the distribution of H. While our simulated H's do not match the observed distribution of plant Herfindahls, we hope that they will at least provide results that are indicative of the magnitude of the bias our procedure produces. Over our 5,000-industry sample, an OLS regression of H on $\hat{H}$ yields an estimated constant of 0.0003 (t-statistic 1.3), with the estimated coefficient on $\hat{H}$ being 1.04 (t-statistic 228.9). Restricting the regression to the observations with $\hat{H} < 0.3$ to eliminate the effect of unreasonable industries gives estimates of 0.0001 (t-statistic 0.5) and 1.05 (t-statistic 173.8). Adding a quadratic term to this regression, we find the coefficient to be insignificant, suggesting that nonlinearity is not a problem. Regressing the squared error from the linear regression on a constant, $\hat{H}$, and $\hat{H}^2$ to get an idea of the magnitude of the measurement error in a typical industry gives the estimate $\hat{\sigma}^2 = 0.00003 + 0.003\hat{H} + 0.007\hat{H}^2$.

If we believe these results, then for a typical industry in which the true value of γ is small, we shall underestimate γ by about $0.05H$. Given that the mean of H is less than 0.03, this bias is fairly small. To correct this bias, one could simply multiply all our previous estimates of H by 1.05. The correction is not large, however, and given that we have limited confidence in the simulations, we decided not to impose it.

References

Bartik, Timothy J. "Business Location Decisions in the United States: Estimates of the Effects of Unionization, Taxes, and Other Characteristics of States." *J. Bus. and Econ. Statis.* 1 (January 1985): 14–22.

Carlton, Dennis W. "The Location and Employment Choices of New Firms: An Econometric Model with Discrete and Continuous Endogenous Variables." *Rev. Econ. and Statis.* 65 (August 1983): 440–49.

Creamer, Daniel. "Shifts of Manufacturing Industries." In *Industrial Location and National Resources.* Washington: Government Printing Office (for Nat. Resources Planning Board), 1943.

Crihfield, John B. "Manufacturing Supply: A Long-Run, Metropolitan View." *Regional Sci. and Urban Econ.* 20 (November 1990): 327–49.

Ellison, Glenn, and Glaeser, Edward L. "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." Working Paper no. 4840. Cambridge, Mass.: NBER, August 1994.

Enright, Michael. "Geographic Concentration and Industrial Organization." Ph.D. dissertation, Harvard Univ., 1990.

Florence, P. Sargant. *Investment, Location and Size of Plant: A Realistic Inquiry into the Structure of British and American Industries.* Cambridge: Cambridge Univ. Press, 1948.

Fuchs, Victor. *Changes in the Location of Manufacturing in the United States since 1929.* New Haven, Conn.: Yale Univ. Press, 1962.

Gardocki, Bernard C., Jr., and Baj, John. "Methodology for Estimating Nondisclosure in County Business Patterns Data." Manuscript. De Kalb: Northern Illinois Univ., Center Governmental Studies, 1985.

Glaeser, Edward L.; Kallal, Hedi D.; Scheinkman, José A.; and Shleifer, Andrei. "Growth in Cities." *J.P.E.* 100 (December 1992): 1126–52.

Henderson, J. Vernon. *Urban Development: Theory, Fact, and Illusion.* New York: Oxford Univ. Press, 1988.

Hoover, Edgar M. *The Location of Economic Activity.* New York: McGraw-Hill, 1948.

Jaffe, Adam B.; Trajtenberg, Manuel; and Henderson, Rebecca. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Q.J.E.* 108 (August 1993): 577–98.

Krugman, Paul. *Geography and Trade.* Cambridge, Mass.: MIT Press, 1991. (*a*)

———. "Increasing Returns and Economic Geography." *J.P.E.* 99 (June 1991): 483–99. (*b*)

McFadden, Daniel L. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by Paul Zarembka. New York: Academic Press, 1974.

Marshall, Alfred. *Principles of Economics: An Introductory Volume.* 8th ed. London: Macmillan, 1920.

Porter, Michael E. *The Competitive Advantage of Nations.* New York: Free Press, 1990.

Schmalensee, Richard. "Using the *H*-Index of Concentration with Published Data." *Rev. Econ. and Statis.* 59 (May 1977): 186–93.

Schmenner, Roger W.; Huber, Joel C.; and Cook, Randall L. "Geographic Differences and the Location of New Manufacturing Facilities." *J. Urban Econ.* 21 (January 1987): 83–104.