NBER WORKING PAPER SERIES

GEOGRAPHIC CONCENTRATION IN U.S. MANUFACTURING INDUSTRIES: A DARTBOARD APPROACH

Glenn Ellison
Edward L. Glaeser

Working Paper No. 4840

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1994

We would like to thank Richard Caves, Sara Fisher Ellison, Bill Miracky, Wally Mullin, Peter Reiss and José Scheinkman for helpful comments, and Matt Botein, Rajesh James, Bruce Sacerdote and Jacob Vigdor for research assistance. Both authors thank the National Science Foundation (SBR-9310009 and SBR-9309808) and the Harvard Clark Fund for financial support. This paper is part of NBER's research program in Growth. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #4840
August 1994

GEOGRAPHIC CONCENTRATION IN U.S. MANUFACTURING INDUSTRIES: A DARTBOARD APPROACH

ABSTRACT

This paper discusses the prevalence of Silicon Valley-style localizations of individual manufacturing industries in the United States. Several models in which firms choose locations by throwing darts at a map are used to test whether the degree of localization is greater than would be expected to arise randomly and to motivate a new index of geographic concentration. The proposed index controls for differences in the size distribution of plants and for differences in the size of the geographic areas for which data is available. As a consequence, comparisons of the degree of geographic concentration across industries can be made with more confidence. We reaffirm previous observations in finding that almost all industries are localized, although the degree of localization appears to be slight in about half of the industries in our sample. We explore the nature of agglomerative forces in describing patterns of concentration, the geographic scope of localization, and the extent to which agglomerations involve plants in similar as opposed to identical industries.

Glenn Ellison
Department of Economics
Massachusetts Institute of Technology
50 Memorial Drive
Cambridge, MA 02139

Edward L. Glaeser
Department of Economics
113 Littauer Center
Harvard University
Cambridge, MA 02138
and NBER

1 Introduction

The concentration of the computer industry in Silicon Valley and of the auto industry in Detroit are two of the more famous examples of the geographic agglomeration of firms in a single industry. The economics literature motivated by these examples is both old and vibrant. Agglomerations have for years drawn the attention both of urban planners with practical concerns and of economists who wish to understand them simply because they are a striking feature of the economic landscape. More recently, they have been regarded also as a potential source of insights into the nature of the increasing returns and external economies which drive the new theories of growth and international trade. As a result, researchers primarily interested in international trade, growth, industrial organization, and business strategy have joined geographers and urban economists in investigating geographic concentration.[1]

This paper is concerned with measurement issues relevant to work in all these fields. Our "dartboard approach" to studying concentration consists essentially of extending the analogy of firms choosing locations by throwing darts at a map into a useful set of models of location choice in the presence of agglomerative forces. In doing so, we have two main goals. First, we wish to look formally at whether most industries are truly localized. Second, and more importantly, we wish to use the models to guide the development of new tools for the measurement of localization. We hope that the index of localization we propose will facilitate future research into a range of topics involving cross-industry comparisons, *e.g.* how patterns of agglomeration compare in different countries, how levels of concentration have evolved over time, and whether cross-industry patterns provide insights into the nature of the forces which cause agglomeration.

That Silicon Valley-style agglomerations may be more the rule than the exception has been noted by a number of authors (see *e.g.* Krugman (1991a)). Our first goal is to provide a careful reexamination of whether this is indeed the case. The defining characteristic of our dartboard approach (and the motivation for a reexamination) is that we wish to reserve the term "localized" for industries exhibiting levels of concentration beyond those which would be observed if firms had chosen the locations of their plants in a completely random manner. In doing so, we take as exogenous the discreteness of plants.[2] For example, in the U.S. vacuum cleaner industry (S.I.C. 3635) about 75% of the employees work in one of the four largest plants. Given this, we do not want to regard the industry as being

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

[2] We do not mean to say that the determinantion of plant sizes is not a topic with interesting implications for understanding increasing returns, just that it is usefully separated from the measurement of interplant agglomeration.

localized simply because 75% of its employment is contained in four states. Also, even if firms did choose locations for their plants by throwing darts at a map, one should recognize that several of the plants by chance might appear to form a cluster.[3] Our use of the term geographic concentration is further restricted in that we will regard an industry as concentrated only if it displays some agglomeration beyond the overall concentration of U.S. manufacturing.[4] For example, we do not want to call the newspaper industry concentrated just because 12% of all employment in the industry is in California and an additional 9% is in New York. Despite this more stringent definition of localization, our results strikingly reaffirm the belief that localization is widespread.

Our primary focus in this paper is on the development of a new index (and other tools) for the measurement of the degree to which industries are geographically concentrated. We believe that a useful index of geographic concentration must have two properties: it must measure something which is interesting to economists and allow one to make comparisons across industries. Such comparisons are not only of descriptive interest, but are the substance of most inquiries into the nature of geographic concentration.[5] Interindustry (or intertemporal) comparisons are problematic with previously defined indices because the comparisons are greatly affected (in ways which are not completely understood) by variations in industry characteristics and data availability.[6]

We motivate our index with an analysis of two models of location choice: one based on the idea that spillovers (*e.g.* localized knowledge spillovers[7]) may lead firms to wish to locate together, and the other based on the idea that firms want to locate wherever some type of natural advantage (*e.g.* access to raw materials) is present. Both models are capable of accounting for geographic concentration, and they are likely important to varying degrees

[3] In fact, one only needs to throw 6 darts at a map of the U.S. before it is most likely that at least two will hit in some state. Such random agglomerations would be less likely to occur if transportation costs or other "centrifugal" forces give firms a desire to locate away from their competition.

[4] We thus use the term as a synonym for what Henderson (1988) and Krugman (1991a) call localization as opposed to what they term urbanization or geographic concentration. We hope that our use of localization and geographic concentration as interchangeable terms does not create confusion. Again, we do not wish to imply that the overall agglomeration of industries is not an interesting topic, just that it is usefully separated from an examination of intraindustry agglomeration.

[5] For example, Krugman (1991a) discusses whether high tech industries are more concentrated than other industries to investigate the importance of knowledge spillovers and compares the U.S. auto industry with its European counterpart to discuss the potential impact of European integration. Earlier comparative works include Florence's (1948) study of U.S. and British industries and Fuchs's (1962) discussion of changes in the U.S. between 1929 and 1954.

[6] A representative set of these indices are those of Creamer (1943), Florence (1948), Enright (1990) and Krugman (1991a). Florence's observation that industries with larger plants are more concentrated is a particularly clear example of the difficulties in interpreting comparisons.

[7] See Krugman (1991b) for a discussion of other spillovers.

in different industries. (Our best example of natural advantage is the wine industry where it is difficult to separate manufacturing from the growing of grapes and 78% of employment is in California. Our best example of spillovers is the fur industry where 334 plants in New York (most in Manhattan) employ 77% of the industry's workforce.[8]) The main point of our analysis is not that these models can both account for geographic concentration, but rather that regardless of which mechanism generates geographic concentration in a particular industry we can control for the number and size distribution of plants and for the set of geographic areas for which data is available in the same way. It is because of this coincidence that we feel somewhat comfortable that our index may control for these factors in the real world as well.

While the paper is concerned largely with methodology, we try also to provide as detailed a description as space allows of geographic concentration in U.S. manufacturing industries.[9] After all, the ultimate test of an index is whether it provides enlightening results. First, we discuss overall levels of concentration, with one observation being that many industries are only slightly concentrated (with a substantial fraction of what others have identified as concentration being attributable to the discreteness of plants.) Next, we discuss briefly which industries are concentrated. Subsequently, we explore the nature of the spillovers (or other forces) causing agglomeration along a number of dimensions: using data on county, state, and regional agglomeration to investigate the geographical scope; looking at whether their influence is felt within narrowly defined industries or whether these spillovers act more broadly; and examining the degree to which the agglomeration of plants occurs internally within firms.

2 Models of Location Choice

In this section we develop several simple models of location choice. These models will be used to construct a test of whether observed levels of geographic concentration are greater than would be expected to occur randomly, and to motivate our subsequent proposal of an index of geographic concentration.

As a practical matter, the data available for measuring geographic concentration typically consists of a breakdown of an industry's total employment by some geographic subunits, *e.g.* we may find state-by-state employments for an industry in the U.S. or country-by-

[8] Fuchs (1957) provides an excellent discussion of the industry.

[9] The raw data for most of our calculations is from the 1987 Census of Manufactures. We have gone to some length to fill in missing state-industry employment data so that we may analyze the complete set of manufacturing industries. We have also estimated the Herfindahl indices of the plant size distributions for each 4-digit industry. We hope that this data may prove useful in future work as well.

country employments in the European Community. We therefore consider an abstract model in which a geographic whole (*e.g.* the U.S.) is divided into M subunits which have shares $x_1, x_2, \ldots, x_M$ of aggregate employment. We assume the shares $s_1, s_2, \ldots, s_M$ of a given industry's employment located in each of these subunits are also available. With such data, a natural measure of the degree to which employment in the industry departs from the overall pattern of employment is

$$g = \sum_{i=1}^{M}(s_i - x_i)^2.$$

We feel that such a measure is of economic interest in that it emphasizes departures which involve significant fractions of an industry's employment. We will focus on modifications of this measure throughout this paper both because the measure is of economic interest, and because it will prove easier to work with than, say, Gini coefficients.[10]

2.1 A Simple Model

We begin with a simple model we will use to ask whether the concentration of employment within industries is greater than would be expected if all plants were located in an independent random manner. We view the "random" choice of the model as reflecting what would be expected in an industry lacking both agglomerative forces (such as spillovers) and centrifugal forces (such as transportation costs with dispersed demand).

Consider an industry consisting of N business units having shares $z_1, z_2, \ldots, z_N$ of the industry's employment. We write H for the industry Herfindahl index[11] defined by $H = \sum_j z_j^2$. Suppose that each business unit chooses a single location for all of its operations within a country which is divided into M geographic areas having shares $x_1, x_2, \ldots, x_M$ of total employment.[12] As a model of random location, we imagine that each business unit chooses a single location for all of its employees by throwing a dart at the map of the country. Formally, we suppose that the geographic areas in which the firms choose to locate are independent identically distributed random variables $v_1, v_2, \ldots, v_N$, each taking on the values $1, 2, \ldots, M$ with probabilities $p_1, p_2, \ldots, p_M$. We can think of the probabilities $p_1, p_2, \ldots, p_M$ as describing the relative sizes of the states on the map. In trying to test whether this model can describe the geographic concentration of U.S. industries, we will

[10] Florence (1948) provides a lengthy argument for a similar measure.

[11] Note that our definition differs from the conventional use of the term both in that we will usually think of plants rather than firms as the business units in question and in that market shares are shares of employment rather than shipments.

[12] We think of the industry as being small relative to the country so that the $\{x_i\}$ can be treated as fixed regardless of the location decisions of the business units in the industry.

usually take $p_i = x_i$ for all i, so that the random location process would on average produce a pattern of employment shares for the industry matching that we have assumed to prevail in the aggregate.[13]

Let us now examine the degree of localization such a model would produce. The fraction of the industry's employment located in geographic unit i is

$$s_i = \sum_{j=1}^{N} z_j u_{ji},$$

where u_{ji} is the Bernoulli random variable equal to one if and only if $v_j = i$. Define a normalized measure, G, we will refer to as the raw geographic concentration of the industry by

$$G = \frac{\sum_i (s_i - x_i)^2}{1 - \sum_i x_i^2}.$$

Proposition 1 characterizes the raw geographic concentration produced by this model. The fact that we get such a simple answer with the expected value of G depending only on H and not on any details of the plant size distribution is not only interesting, but also useful in that detailed data on plant sizes may be hard to come by.

Proposition 1 *In the model above,*

$$E(G) = \frac{1 - \sum_i p_i^2}{1 - \sum_i x_i^2} H + \frac{\sum_i (p_i - x_i)^2}{1 - \sum_i x_i^2}.$$

For $(p_1, \ldots, p_M) = (x_1, \ldots, x_M)$ *this reduces to*

$$E(G) = H.$$

Proof

The result follows from a straightforward calculation using the fact that the expectation of a sum of random variables is the sum of the expectations regardless of whether the random variables are independent.

$$\begin{aligned}(1 - \sum_i x_i^2) E(G) &= E(\sum_i (s_i - x_i)^2) \\ &= \sum_i E((s_i - p_i + p_i - x_i)^2) \\ &= \sum_i \mathrm{Var}(s_i) + \sum_i (p_i - x_i)^2.\end{aligned}$$

[13] We emphasize that by doing so we are taking as given the concentration of aggregate employment, even though this may be thought of as resulting from (nonindustry-specific) interfirm spillovers. We are interested in exploring intra-industry localizations, not in the fact that there is virtually no manufacturing in the state of Wyoming.

Using $s_i = \sum_j z_j u_{ji}$ and that the u_{ji} for $j = 1, 2, \ldots, N$ are independent Bernoulli random variables we have

$$\begin{aligned}(1 - \sum_i x_i^2)E(G) &= \sum_i \sum_j z_j^2 \mathrm{Var}(u_{ji}) + \sum_i (p_i - x_i)^2 \\ &= \sum_i \sum_j z_j^2 p_i (1 - p_i) + \sum_i (p_i - x_i)^2 \\ &= H(1 - \sum_i p_i^2) + \sum_i (p_i - x_i)^2\end{aligned}$$

as desired.

QED.

To help get some intuition for this result, it may help to note why it holds in a couple of limiting cases (assuming that $p_i = x_i$ for all i). First, for any fixed set of geographic areas, the limit as $H \to 0$ describes an industry with an infinite number of small firms. In this case the law of large numbers dictates that a fraction x_i of the industry's employment will be in geographic unit i and G will be zero. Next, for any fixed firm size distribution imagine that the sizes of the geographic areas become arbirarily small, *i.e.* let $M \to \infty$ with $\max_i x_i \to 0$. With only a finite number of firms we can in the limit ignore the probability of two darts hitting any geographic unit. The value of $(s_i - x_i)^2$ will then be approximately z_j^2 if business unit j is located in area i and 0 otherwise. Hence, we can see that the sum of squared deviations will approach the Herfindahl index.

The result also gives us our first intuitive interpretation of a measure of concentration. If an industry has raw concentration G, we can think of the distribution of employment in the industry as being as concentrated as would be expected if $\frac{1}{G}$ randomly selected locations each had a fraction G of the industry's employment.

In testing whether a set of industries exhibits excess geographic concentration, it is useful also to know the variance of G in this model. The expression is not as simple as that for the expectation, and depends also on the fourth moment of the distribution of business unit sizes.

Proposition 2 *For $(p_1, p_2, \ldots, p_M) = (x_1, x_2, \ldots, x_M)$ in the model above*

$$Var(G) = \frac{2}{(1 - \sum x_i^2)^2} \left(H^2 \left(\sum x_i^2 - 2 \sum x_i^3 + (\sum x_i^2)^2 \right) - \sum_j z_j^4 \left(\sum x_i^2 - 4 \sum x_i^3 + 3 (\sum x_i^2)^2 \right.\right.$$

The result follows from a straightforward but tedious calculation, which we omit.

2.2 Two Models of Localization

We now discuss two additional models of the location decision process, each of which is capable of explaining localization in excess of that predicted in the simple model above. The models will thus be useful in developing an index of the extent to which an industry exhibits *excess* geographic concentration.

The models concern location choices which are influenced not only by aggregate employment, but also by the "natural advantages" to locating in certain areas and by localized intraindustry "spillovers." In order to discuss how these factors should be incorporated, it is helpful first to recast the dartboard model of the previous section in more economic terms. Specifically suppose that in an industry like that described above, each business unit locates in whichever state maximizes its profits, and that the profits received by the k^{th} unit when it locates in area i take the form

$$\log \pi_{ki} = \log \overline{\pi_i} + \epsilon_{ki},$$

where $\overline{\pi_i}$ is a measure of the average profitability of area i and ϵ_{ki} is a random variable reflecting idiosyncratic elements of the suitability of the area to the firm in question (because of fixed firm characteristics, preferences of its management, the success of its search for a site, etc.). If we assume that the $\{\epsilon_{ki}\}$ are independent and have the Weibull distribution, then it is a standard result that firm k's location v_k is a random variable with

$$\text{Prob}\{v_k = i\} = \frac{\overline{\pi_i}}{\sum_j \overline{\pi_j}}.[14]$$

Our standard dartboard model can be obtained as a special case by assuming that the states have no distinguishing features which affect their average profitability other than differences in aggregate employment, and that the positive spillover of aggregate employment on profits takes the form $\overline{\pi_i} = x_i$. With this specification

$$\text{Prob}\{v_k = i\} = \frac{x_i}{\sum_j x_j} = x_i.$$

Because this dependence of average profits on aggregate employment leads to location choices which on average recreate aggregate agglomeration given the error structure we have assumed, we shall take it as a starting point for our subsequent models.[15]

[14]See McFadden (1973).

[15]Rather than thinking of this dependence as reflecting aggregate spillovers, it is also possible to obtain such a relation indirectly by assuming that the profitability at each potential site is independent and ex-ante identical, but that larger states have more sites to choose from (proportionally to their aggregate employment) so that the best location in a larger state is on average superior.

2.2.1 A Model of Natural Advantage

Our first model of industry localization is motivated by the observation that the business units in an industry will appear to be clustered whenever their location decisions are influenced by factors which can be regarded as giving a "natural" advantage to certain of the geographic areas. Our prototypical example is the wine industry. Clearly, the localization of the industry in is in large part due to California's climatic natural advantage in growing grapes. Similarly, the concentration of industries which import or export bulky commodities in coastal states reflects a natural advantage in access to transportation.[16] Perhaps because such factors so straightforwardly lead firms to cluster together they have generally received less attention than spillovers in discussions of industry localization. They are, however, an essential component of a complete description of agglomeration.

The simplest way to add natural advantage to the location choice model described above is to assume that firm k's profits when it locates in state i are again of the form

$$\log \pi_{ki} = \log \overline{\pi_i} + \epsilon_{ki},$$

but with the average profitability of state i, $\overline{\pi_i}$, now taken to be a nonnegative random variable reflecting all of the ways in which nature has chosen to make state i unique (which affect profits in the same way for all plants). Conditional on a realization of the $\{\overline{\pi_i}\}$, the probability that each business unit locates in state i is

$$p_i \equiv \frac{\overline{\pi_i}}{\sum_j \overline{\pi_j}}.$$

The larger are the differeces between the p's and the x's, the more we can think of locational patterns as being influenced by natural advantage.

We analyze a specification of this model in which the importance of natural advantage is neatly parameterized by a single constant $\gamma_0 \in [0,1]$, by assuming that the state profit levels $\{\overline{\pi_i}\}$ are independent of the $\{\epsilon_{ki}\}$, and that their distribution is such that $E(p_i) = x_i$ and $\text{Var}(p_i) = \gamma_0 x_i(1-x_i)$.[17] Note that when $\gamma_0 = 0$, there are no common shocks and we obtain our standard dartboard model of random locations. At the other extreme, when $\gamma_0 = 1$ each p_i has the largest possible variance given its mean and support, so that with probability one the differences in state characteristics are so extreme that all business units will cluster in a single state.

To explore the level of raw geographic concentration such a model produces for intermediate γ_0 and how this depends on the structure of the industry, it is helpful to restate

[16] One formal study of such an effect is Carlton (1983), which finds that energy prices are an important determinant of plant location decisions in several industries.

[17] For example, one could assume that $\overline{\pi_i} = x_i + \eta_i$ with the $\{\eta_i\}$ being mean zero random variables with $\sum_i \eta_i = 0$ (with probability one) and $\text{Var}(\eta_i) = \gamma_0 x_i(1-x_i)$. Another example is given later in this section.

the model using a dartboard metaphor. We can think of the business units' choices of location as a two stage process. In the first stage nature chooses (from some set of possible dartboards) a single dartboard on which the geographic areas have sizes $p_1, p_2, \ldots, p_M$, reflecting the importance and allocation of comparative advantage across the areas (the larger areas being those with greater average profits). In the second stage, all business units, being influenced by the same levels of comparative advantage, independently throw darts at this board to choose their locations.

The following proposition shows that the expected raw concentration is linearly increasing in γ_0 and again depends on the distribution of the plant sizes only through H.

Proposition 3 *In the two stage model of comparative advantage described above*

$$E(G) = \gamma_0 + (1 - \gamma_0)H.$$

<u>Proof</u>

Using the result of Proposition 1 we have

$$(1 - \sum_i x_i^2)E(G) \;=\; E_p\left((1 - \sum_i p_i^2)H + \sum_i (p_i - x_i)^2\right).$$

We have assumed that $E(p_i) = x_i$, and $\mathrm{Var}(p_i) = \gamma_0 x_i(1 - x_i)$. Hence,

$$\begin{aligned}(1 - \sum_i x_i^2)E(G) &= H(1 - \sum_i x_i^2 + \gamma_0 x_i(1 - x_i)) + \sum_i \gamma_0 x_i(1 - x_i) \\ &= H(1 - \gamma_0 + (\gamma_0 - 1)\sum_i x_i^2) + \gamma_0(1 - \sum_i x_i^2) \\ &= (1 - \sum_i x_i^2)((1 - \gamma_0)H + \gamma_0)\end{aligned}$$

as desired.

QED.

For concreteness, it may help to note that one specification satisfying the conditions above is obtained by assuming that the $\{\overline{\pi_i}\}$ are independent random variables with $\overline{\pi_i}$ having a chi-square distribution with $2\frac{1-\gamma_0}{\gamma_0}x_i$ degrees of freedom. The induced distribution of $p_i = \frac{\overline{\pi_i}}{\sum_j \overline{\pi_j}}$ is then $\beta(\frac{1-\gamma_0}{\gamma_0}x_i, \frac{1-\gamma_0}{\gamma_0}(1 - x_i))$, and hence has mean x_i and variance $\gamma_0 x_i(1 - x_i)$.[18]

[18] More generally the same distribution for p_i is obtained whenever $\overline{\pi_i} \sim \Gamma(\frac{1-\gamma_0}{\gamma_0}x_i, \lambda)$. The joint distribution of $(p_1, \ldots, p_M)$ is Dirichlet with parameters $(\frac{1-\gamma_0}{\gamma_0}x_1, \ldots, \frac{1-\gamma_0}{\gamma_0}x_M)$. See Johnston and Kotz (1972, p. 231) for a description of this distribution. The density function is $f(y_1, \ldots, y_{M-1}) = \Gamma(\frac{1-\gamma_0}{\gamma_0})\prod_{i=1}^M \Gamma(\frac{1-\gamma_0}{\gamma_0}x_i)^{-1} p_i^{\frac{1-\gamma_0}{\gamma_0}x_i - 1}$.

2.2.2 A Model of Spillovers

Our second model of industry localization is motivated by the idea that externalities or spillovers may lead firms to desire to locate their plants near other plants in the industry. We use the term spillovers quite broadly here to refer to technological spillovers, gains from interfirm trade, the effect of local knowledge on the location of spinoff firms, etc. – essentially any forces which lead firms to choose locations near other firms in the industry.

To model such factors, one might assume that the profit of business unit k if located in area i is of the form

$$\log \pi_{ki} = \log \overline{\pi}(x_i, v_1, \ldots, v_{i-1}, v_{i+1}, v_M) + \epsilon_{ki},$$

where as before v_j is the location of plant j. This formulation allows average profits within a state to be affected generally by both the aggregate employment and the location of the other plants in the industry (but not by state characteristics). To make the analysis tractable and to aid interpretation, we again examine a simple parametric specification of this model. In particular, we consider for $\gamma_0 \in [0,1]$ profit functions of the form

$$\log \pi_{ki} = \log(x_i) + \sum_{\ell \neq k} e_{k\ell}(1 - u_{\ell i})(-\infty) + \epsilon_{ki},$$

where the $\{e_{k\ell}\}$ are Bernoulli random variables equal to one with probability γ_0, $u_{\ell i}$ is an indicator for whether $v_\ell = i$, and the $\{\epsilon_{ki}\}$ are again independent Weibull random variables independent of the $\{e_{k\ell}\}$.

To motivate this formulation, note that the first term $\log(x_i)$ is the same dependence of profits on aggregate employment necessary to reproduce (on average) the pattern of aggregate employment. The second expression involves two main assumptions made largely for tractability. First, we have assumed that the effect of plant ℓ's location on plant k's profit depends only on whether they are in the same area, not on the distance between different areas. Second, rather than assuming a continuous distribution for the magnitude of the spillovers, we take the spillovers to have an extreme two point support – they are either strong enough so that firms k and ℓ will have negative infinity profits if they locate apart, or they are nonexistent so that k's profits are independent of ℓ's location. As the probability that any pair of firms has such a crucial spillover between them, γ_0 clearly indexes the importance of spillovers.[19]

In this spillover model, one needs to be more careful in specifying the decision processes of the business units. We assume the the business units choose locations in some preordained order, and that each firm in turn maximizes its profits conditioning only on the

[19] Note that in a violation of accepted practice are using γ_0 to represent a completely different parameter here than in the previous model. We have made this decision to emphasize that the predicted mean concentration of the two models will turn out to be identical.

location decisions of the firms which have moved previously. We shall assume also that the indicator variables $\{e_{k\ell}\}$ for whether spillovers exist between pairs of firms are symmetric and transitive in the sense that $e_{k\ell} = 1 \Rightarrow e_{\ell k} = 1$ and $e_{k\ell} = 1$ and $e_{\ell m} = 1 \Rightarrow e_{km} = 1$.[20] In this case, the process we have specified is a rational expectations equilibrium in which each firm earns nonnegative profits and the resulting distribution of locations is independent of the order in which the business units make their choices. Note that for $\gamma_0 = 0$ the model is again our standard dartboard model, and for $\gamma_0 = 1$ all firms will cluster in a single area.

To analyze the geographic concentration such a model produces it helps again to think of the firms' location choices in terms of a dart throwing metaphor. Each business unit is represented by a dart which will be thrown to choose a location. In the first stage, nature randomly decides to weld some of the darts together into clusters, with the distribution of her decisions being such that each pair of darts has probability γ_0 of being in the same welded cluster. In the second stage, each cluster of welded darts is thrown independently (with all darts in a cluster hitting a single point).

In this model, the business units' locations $v_1, \ldots, v_N$ are identically distributed random variables, each taking on the value $i \in \{1, 2, \ldots, M\}$ with probability x_i. Note, however, that the $\{v_j\}$ are not independent; instead it is straightforward to show that $\text{Corr}(u_{ki}, u_{\ell i}) = \gamma_0$ for all i and all $\ell \neq k$.[21] Proposition 4 characterizes the raw geographic concentration produced by this model.

Proposition 4 *In the model of spillovers described above*

$$E(G) = \gamma_0 + (1 - \gamma_0)H.$$

Proof

$$\begin{aligned}
(1 - \sum_i x_i^2)E(G) &= \sum_i \text{Var}(\sum_j z_j u_{ji}) \\
&= \sum_i \sum_j z_j^2 \text{Var}(u_{ji}) + \sum_i \sum_{j,k,j\neq k} z_k z_j \text{Cov}(u_{ki}, u_{ji}) \\
&= \sum_i \sum_j z_j^2 x_i(1 - x_i) + \sum_i \sum_{j,k,j\neq k} z_k z_j \gamma_0 x_i(1 - x_i) \\
&= (1 - \sum x_i^2)(\sum_j z_j^2 + \gamma_0 \sum_{j,k,j\neq k} z_k z_j).
\end{aligned}$$

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

[21] The reader may note that this is the only property of the joint distribution which is necessary for the proposition, and hence the proposition applies to any formulation of the interdependent profits which induces this correlation in location choices.

The desired result now follows from the substitution $\sum_{j\neq k} z_j z_k = (\sum_j z_j)^2 - \sum_j z_j^2 = 1 - H$. **QED.**

The most notable feature of the result in Proposition 4 is that our model of spillovers and our model of natural advantage yield identical functional forms for the relationship between the expected level of geographic concentration and the other industry characteristics (the plant size distribution and the sizes areas for which employment breakdowns are available). This coincidence motivates the index of concentration proposed below. The coincidence of the results of Propositions 3 and 4 also tells us, in some sense, that we cannot distinguish comparative advantage from spillover theories based only on the mean levels of geographic concentration.[22]

Real world location decisions are likely to be affected by both natural advantage and by spillovers, so it is probably worth noting that a combination of the two factors also produces a level of raw concentration which is related to the industry characteristics in the same way. Specifically, consider a three stage model (we give only the dartboard version) where in the first stage Nature chooses $(p_1, p_2, \ldots, p_M)$ from a distribution with $E(p_i) = x_i$ and $\mathrm{Var}(p_i) = \gamma_1 x_i(1 - x_i)$; in the second Nature randomly welds each pair of darts with probability γ_2; and in the third the welded clusters are independently thrown at a dartboard in which the states have sizes $(p_1, p_2, \ldots, p_M)$.

Proposition 5 *In the three stage model above*

$$E(G) = \gamma_0 + (1 - \gamma_0)H,$$

with $\gamma_0 = \gamma_1 + \gamma_2 - \gamma_1\gamma_2$.

The proof is similar to those of the previous propositions and is therefore omitted.

3 An Index of Geographic Concentration

Suppose 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.

[22]The theories will differ in their predictions for higher moments of G. Recall, however, that we have not fully specified either model (leaving out the higher moments of the $\{p_i\}$ in the natural advantage model and the full joint distribution of the $\{e_{kl}\}$ in the spillover model). Varying these elements, each model can produce a range of predictions for $\mathrm{Var}(G)$. For this reason, we do not feel that attempts to distinguish the theories on such grounds will be fruitful.

As a convenient index of the degree to which an industry is geographically concentrated we propose the use of a measure γ defined by

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

We believe that this index has a number of attractive features. It reflects a property which is naturally meaningful in emphasizing large deviations from the distribution of aggregate employment. Because $E(\gamma) = 0$ when data is generated by our standard dartboard model, it is clearly interpretable as measuring *excess* concentration beyond that which would be expected to occur randomly. Finally and most importantly the index allows us to easily perform meaningful comparisons of the degrees of concentration in different industries, *e.g.* comparing a U.S. industry with its counterpart in another country, or comparing concentration using 3- and 4-digit industry definitions.

To justify such comparisons, we note simply that if the location decision process of plants is accurately described by either or both of the natural advantage and spillover models of the previous section then

$$E(\gamma) = \frac{E(G) - H}{1 - H} = \gamma_0,$$

i.e. the index is an unbiased estimator of the fundamental parameter which describes the strength of natural advantage or spillovers. That the index controls for the number and size distribution of plants and (subject to the caveat below) for the sizes of the geographic subunits for which data is available in both of these modes gives us some hope that it will allow us to compare the strength of these forces in real world industries as well.

In making the transition from the models to the real world one caveat is necessary with regard to comparisons based on differing geographic subunits. Each of our models takes an extreme view of the geographic scope of the forces which produce localization. For example, when spillovers are important they are assumed to accrue only if the firms locate in the identical geographic subunit. In practice, spillovers would likely have an effect which declines more smoothly and provides some benefit to locating in nearby areas as well (more so when the areas in question are smaller). If we estimate γ using county level data, it will reflect only the added probability of pairs of plants locating in the same county, while a γ estimated from state level data will reflect the typically larger increase in the probability of the pair locating in the same state.[23]

[23] The location decision process of the natural advantage model depends heavily on the definition of the subareas, and thus the conditions under which we can compare estimates based on different geographic subunits are less obvious. One case in which such comparisons are completely justified is that of the gamma-

The fact that we can regard γ as a parameter estimate clearly suggests that it should be interpreted in light of its standard error. What this standard error is, however, can not be determined given the assumptions we have made so far. In particular, (and we consider it a feature of our paper that this is true) the results on mean concentration above have been derived without ever specifying the higher moments of the $\{\overline{\pi_i}\}$ in the natural advantage model or joint distribution of the indicators $\{e_{k\ell}\}$ in the spillover model. A straightforward calculation of the standard errors gives

$$\mathrm{Var}(\gamma - \gamma_0) = \frac{1}{(1-H^2)(1-\sum x_i^2)^2} \sum_{i,r=1}^{M} \sum_{j,k,\ell,m=1}^{N} z_j z_k z_\ell z_m \mathrm{Cov}(\mu_{ji}\mu_{ki}, \mu_{\ell r}\mu_{mr}).$$

The covariance terms will depend on the unspecified elements of the models. To give a feel for the magnitude of one of the sources of measurement error in our calculations of γ's, we will present later standard errors (obtained from simulations) from one complete specification – a natural advantage model where the distribution of the $(p_1, p_2, \ldots, p_M)$ is assumed to be Dirichlet with parameters $(\frac{1-\gamma_0}{\gamma_0}x_1, \frac{1-\gamma_0}{\gamma_0}x_2, \ldots, \frac{1-\gamma_0}{\gamma_0}x_M)$.[24]

4 Data

By design, the data requirements for this paper are fairly simple. We require the distribution of employment across a set of geographic areas for a set of industries and the Herfindahl indices of firm and plant employment shares for those industries. Our definition of industries is the finest one possible given data availability constraints – the 459 manufacturing industries defined by the 4-digit classifications of the Census Bureau's 1987 S.I.C. system. Given this decision, we settled on the finest geographic areas for which we felt we could obtain reliable employment breakdowns – the 50 states plus the District of Columbia. Our

distributed $\{\overline{\pi_i}\}$ described at the end of Section 2.2.1. In this model, we can regard natures choice of a dartboard as resulting from an assignment of a probability (or more precisely of an independent gamma-distributed $\overline{\pi_i}$) to each square inch of the country, with the probability of a dart hitting each state (or other subunit) being obtained by summing the probabilities of its hitting each of the square inch plots within the state. More formally, suppose that geographic area 1 is divided into subareas $11, 12, \ldots, 1r$ with shares $x_{11}, \ldots, x_{1r}$ of total employment (with $x_{11} + \cdots + x_{1r} = x_1$). When $(p_{11}, \ldots, p_{1r}, p_2, \ldots, p_M)$ is Dirichlet with parameters $(\frac{1-\gamma_0}{\gamma_0}x_{11}, \ldots, \frac{1-\gamma_0}{\gamma_0}x_{1r}, \frac{1-\gamma_0}{\gamma_0}x_2, \ldots, \frac{1-\gamma_0}{\gamma_0}x_M)$, $(p_{11} + \cdots + p_{1r}, p_2, \ldots, p_M)$ is Dirichlet with parameters $(\frac{1-\gamma_0}{\gamma_0}x_1, \frac{1-\gamma_0}{\gamma_0}x_2, \ldots, \frac{1-\gamma_0}{\gamma_0}x_M)$. Note that this specification is extreme as well in that natural advantages of nearby square inches are likely to be spatially correlated, so again we would expect to find larger γ's when using coarser subdivisions.

[24] One question of interpretation arises in defining the standard errors. Do we treat γ as an estimate of the variance of the *ex ante* distribution from which the $\{p_i\}$ are drawn or as an estimate of the *ex post* realization of $\frac{\sum(p_i - x_i)^2}{1-\sum x_i^2}$. We take this latter interpretation, drawing $\{p_i\}$ for which this expression is equal to γ from the conditional distribution induced by the Dirichlet.

source for all of this data is the 1987 Census of Manufactures, although despite our simple data requirements the process of constructing the data is quite involved. While the data we have filled in is by necessity speculative, we hope that our data may be helpful to others in the future.

Our construction of state-industry employments relies on the Census of Manufactures' listings of state-industry employments and on the reported totals for manufacturing employment in each state and in each 2-, 3-, and 4-digit industry. A very substantial data filling procedure was necessitated by two limitations of the raw data. First, employment in any state-industry with fewer than 150 employees is omitted from the raw data (presumably to save space). Because states with less than 150 employees contain a nontrivial fraction of employment in some industries, it is desirable to fill in these numbers. Second, and more importantly, to protect the confidentiality of individual responses the Census often reports state-industry employment only as falling into one of five categories corresponding to employments of 100-249, 250-499, 500-999, 1000-2500, and the unfortunately low topcode of 2500+.[25] To give some idea of the magnitude of these restrictions, simply setting employment in each of these cells to its lower bound unequivocally identifies the location of 90% of employment in the median industry and 80% on average.

To complete our data set, we have filled in data for all 2-, 3-, and 4- digit state-employments using the census data and the adding up constraints across states within industries and across subindustries within states. Our algorithm is based on the idea of imposing the upper and lower bounds of the reported ranges, and adjusting employments within the ranges to try to satisfy the adding up constraints in both directions. Data for state-industries with less than 150 employees is filled by a similar procedure which uses also the number of "missing" establishments created by the nonreporting of these cells. More details are given in Appendix A.[26]

Throughout most of the paper we will think of plants as the business units of the model, and thus rely on a Herfindahl index H_p of the employment shares of plants to control for the size distribution of business units. While the Census does not publish this information it does make available for each 4-digit industry the total employment and the total number of

[25] Fortunately, employment in many topcoded cells is relatively small, because the largest state employments tend to occur in states with several firms so that withholding restrictions do not apply.

[26] An alternate approach used by several past authors, *e.g.* Enright (1990), is to reduce the number of topcodes by obtaining data from the County Business Patterns (which does not have an identical sample, but which has a much higher topcode) and use means of ranges, dropping industries where topcodes can not be avoided or where the ranges are too large. Drawbacks of such an approach are that the important information in the adding up constraints (and often the existence of small state-industries) is being ignored, and that industries with interesting agglomerations often end up being dropped.

plants in each of ten employment size ranges.[27] Subject to disclosure rules, the total number of employees within the plants of a size category is also usually reported. The withholding is somewhat more of a problem here than in the construction of the state employments because it is primarily the shares of the largest plants which are obscured by the nondisclosure rules. For each industry we used the data to estimate a sum of squared plant shares using a two step procedure: employees were first allocated between the classes where data was withheld, 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. To get a rough idea of the magnitude of the resulting measurement errors we conducted tests of our algorithm on simulated data. The details of the data construction and of the simulations are reported in Appendix B. The simulations suggest that measurement errors are not likely to substantially bias our results. A complete list of our estimated plant Herfindahls is given in Appendix C.

A firm level Herfindahl index, H_f, was taken directly from the Census of Manufactures' *Concentration Ratios in Manufacturing*. The data are based on shares of shipments of the top 50 firms rather than on all firms' employment, and are available for 444 of the 459 industries, with the values for the remaining fifteen (highly concentrated) industries withheld because of disclosure rules. We drop these industries whenever our analysis uses H_f. In the restricted sample of 444 industries, the mean of H_f is 0.068 and that of H_p is 0.025. It is interesting to note as an aside that following Scherer (1975) one may use the ratio of these two concentration measures for an industry as an estimate of the effective number of plants operated by the large firms.[28] In doing so, we find a mean across industries of 3.8 plants/firm and a median of 2.7, figures roughly comparable to those reported by Scherer twenty years ago. The ratio takes on its largest value, 30.1, in S.I.C. 2813, industrial gases.

Finally, for our analysis of the geographic scope of concentration we obtained a dataset of 1987 county level employments for 3-digit industries. The dataset had been constructed by filling in County Business Patterns data using an algorithm which consists largely of using mean plant sizes for all nondisclosed employments.[29] Some comparisons of this data with our primary dataset are given at the end of Appendix A.

[27] The ranges are those determined by the lower bounds of 1, 5, 10, 20, 50, 100, 250, 500, 1000, and 2500 employees.

[28] The estimate is literally valid only if each firm's activities are divided evenly between the same number of plants.

[29] See Gardocki and Baj (1985)

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

5.1 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 that which would be expected to arise randomly.

We begin with what is clearly the most compelling application of our simple dartboard model, assuming that the *plants* in an industry are the business units choosing their locations in an independent random manner. The prediction of the model is that $E(G) = H_p$, with the difference between G and H_p being heteroskedastic with a variance given in Proposition 2. For the full sample of 459 industries we find that the means of G and H_p are 0.77 and 0.28, respectively. The simple dartboard model predicts that the sample average of the G's should have a mean of 0.28 and a standard deviation of 0.0005, so this difference is highly significant indicating that there is excess localization relative to random location choice.

Looking at the industry-by-industry estimates the prevalence of excess localization which previous authors have noted is strikingly confirmed. The level of raw concentration G exceeds that which would be expected to arise randomly in 446 of the 459 industries.[30] In fact, 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.

Before discussing patterns of geographic concentration in more detail, we would like to comment briefly on an alternate application of the dartboard model which might potentially account for higher levels of concentration. Specifically, one could apply the model by assuming instead that the *firms* in an industry are the business units, with each choosing a common location for all of its plants. While this extreme is plainly counterfactual, it may provide a more reasonable test for the hypothesis that locations are random in some instances. For example, for a number of years Maytag had exactly two plants in which it manufactured washing machines, and both of these were located in Newton, Iowa. The

[30] The difference between G and H_p is larger than twice its standard deviation in 369 of the 446 industries in which the difference is positive, and none of the 13 industries in which the difference is negative.

two location decisions were not independent, and given that the entire industry (SIC 3633) consists of only 18 plants, treating them as independent observations might lead to a misleading conclusion that locations are unlike those expected from independent random dart throws.

Looking first at the overall level of concentration, we note that for the 444 industries for which H_f was available the means of G and H_f are 0.074 and 0.068, respectively. While the overall level of concentration appears to be approximately that predicted by the theory, there is a difference between correctly predicting the overall level of concentration and predicting the pattern of concentration across industries. A more demanding test of the firm-random location theory is obtained by estimating the parameters in the regression $G_i = \alpha_0 + \alpha_1 H_{fi} + \epsilon_i$ and testing whether $\alpha_0 = 0$ and $\alpha_1 = 1$. OLS estimates of this equation are given in Table 1. Each of the equalities is rejected individually with a t-statistic of at least 10, and an F-test of the joint hypothesis yields an $F_{2,442}$ statistic of 177.8, also rejecting strongly. The model thus fails to account for pattern of concentration across industries, which should not be surprising given that we know that multiplant firms do often choose multiple locations. The comparison does provide some intuition for the degree to which manufacturing industries are localized: they are approximately as localized as would be expected if units as large as firms' operations in each industry chose locations at random.

Table 1: Test of the Firm-Random Location Theory

Equation: $G_i = \alpha_0 + \alpha_1 H_{fi} + \epsilon_i$		
Parameter	Coeff. Estimate	Std. Error
Constant	0.047*	0.005
H_f	0.394*	0.057
$R^2 = 0.09$		

* indicates significance at the 1% level.

5.2 Levels of Geographic Concentration

From the previous section we know that the degree of localization in U.S. manufacturing industries is not zero. In this section we try to use our models to get a feel for how much concentration there is. It seems likely that the agglomerative forces reflected in our models will vary greatly from industry to industry. We therefore begin by imposing no structure across industries and simply computing the index γ defined by (2) for each of the 459 4-

digit industries in our sample.[31] Recall that γ can be interpreted either as the probability with which any pair of plants choose their locations jointly or as a measurement of the importance of natural advantage in location choice. A complete list of the γ's we find is contained in Appendix C.

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 tallest bar is that corresponding to values of γ between 0 and 0.01. The distribution appears to be quite skewed, with a mean of 0.051 and a median of 0.026. Approximately 43% of the industries have $\gamma < 0.02$, while 26% have $\gamma > 0.05$.

How large are these values? Recall that if an industry has many equal sized plants (so that $H \approx 0$), the natural advantage and spillover models both predict that $E(G) = \gamma$. A similar level of concentration would result from completely independent random location decisions by $1/\gamma$ equal sized plants. Hence, for the 118 industries with $\gamma < 0.01$ we can think of agglomerative forces as being sufficiently weak so that if not for the fewness of plants, production would be no more concentrated than if it were scattered at 100 equal sized sites. While there is no justification for any definition of the phrase "not very localized," we feel that it would be an appropriate description of such a pattern, and we apply it both to these and to the other 88 industries with $\gamma < 0.02$. At the other extreme, we shall refer to industries with $\gamma > 0.05$ as "very localized". This category contains 119 industries.

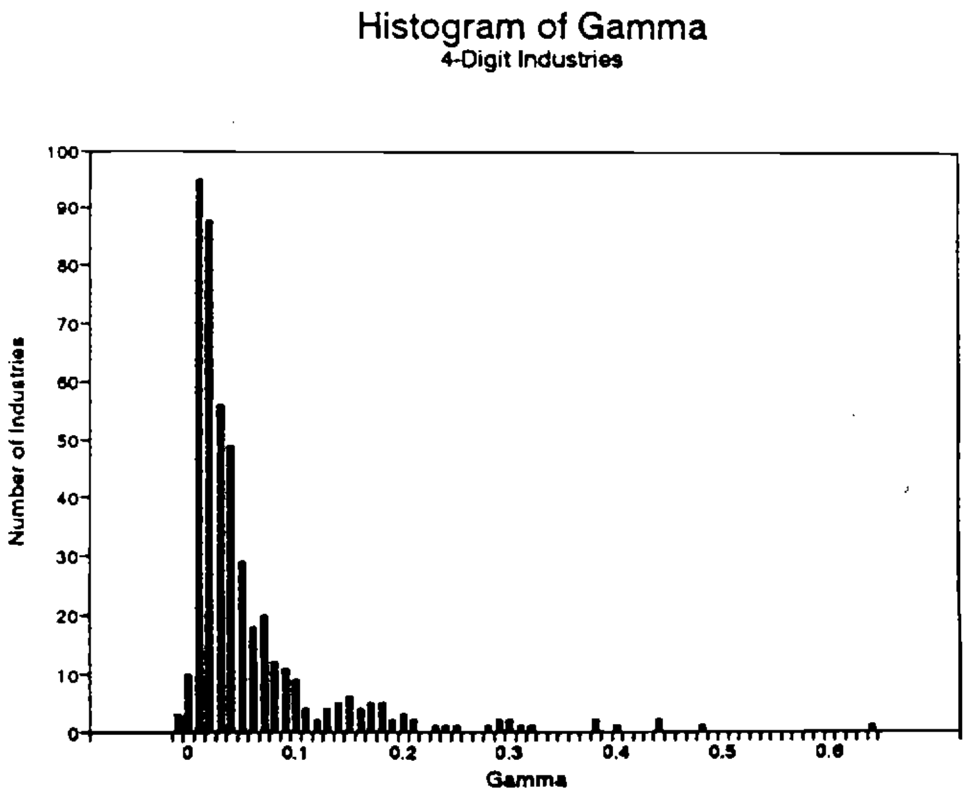
The reader should keep in mind that if one views locations as being generated by a random process, an individual industry's γ is a parameter estimate with a nontrivial standard deviation. To get a feel for the size of this uncertainty in our measurements, we computed standard errors by simulating a special case of our natural advantage model – that of Dirichlet-distributed state sizes.[32] Among industries with $H_p < 0.02$ the mean of the estimated standard errors is 0.02. The means for industries with H_p 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.

To provide some intuition for the importance of accounting properly for random agglomeration when constructing an index of geographic concentration, Table 2 lists the frequency with which γ/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 which 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

[31] Note that we assume here and throughout the rest of the paper that the plants are the business units choosing locations.

[32] The computation requires also a more complete specification of the plant size distribution. For this purpose, we took the plant sizes to be those used as an intermediate step to the Herfindahl calculation.

Figure 1: Histogram of γ



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 do actual agglomerations of plants. Our index may then give a somewhat different picture of geographic concentration than would a discussion of raw concentrations.

Table 2: Raw Concentration Attributable to Spillovers/Comparative Advantage

Fraction of Industries with γ/G in Range.		
Range	All Industries	High G Industries
Less than 0	0.03	0.03
0.00 – 0.25	0.09	0.10
0.25 – 0.50	0.22	0.16
0.50 – 0.75	0.32	0.19
0.75 – 1.00	0.33	0.53

5.3 Patterns of Concentration

An attempt to explore formally the industry characteristics which tend to be associated with localization is well beyond the scope of this paper.[33] We would, however, like to present a few more tables concerning the geographic concentration in different industies.

In Table 3 we summarize the levels of geographic concentration of the 4-digit subindustries of each 2-digit manufacturing industry. For each 2-digit industry, the table lists the fraction of subindustries which 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.

In hopes that the forces affecting geographic concentration might be clearest at the extremes, Table 4 lists the 15 most and the 15 least localized industries in terms of the index γ. As Krugman (1991a) 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. The next most concentrated, wine, may be largely attributable to the natural

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

Table 3: Concentration by 2-digit Category

2-digit industry	# of 4-digit subindustries	Percent of 4-digit industries with $\gamma < 0.02$	$\gamma \in [0.02, 0.05]$	$\gamma > 0.05$
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 misc. 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 equip.	37	41	46	14
37. Transportation equipment	18	28	33	39
38. Instruments and related products	17	47	41	11
39. Miscellaneous manufacturing ind.	18	44	22	33

advantage of California in growing grapes. The concentration of oilfield machinery (in the Houston/Galveston area) may be partially attributable to the location of oil production.

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, *e.g.* vacuum cleaners and small arms ammunition, where raw concentration is substantial, but employment turns out to be concentrated in a few very large (randomly scattered) plants.[34]

6 Scope of Geographic Concentration

In this section we examine two different aspects of the scope of geographic concentration. First, we discuss the scope in the sense of industrial definition, *i.e.* whether concentration is principally a phenomenon which exists at the level of individual industries or whether it is characteristic of broad industry classes as well. Next, we discuss the geographic scope of concentration, comparing data at the county, state, and regional levels.

6.1 Industry Definition

Table 5 provides a simple look at the concentration of 2-, 3-, and 4-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 2-digit to the 3-digit level. This naturally raises two questions of scope. Is there any correlation in the location decisions of firms which share only a two digit industry class, or is the concentration of 2-digit industries entirely a consequence of the localization of its 3-digit subindustries? Are location decisions influenced as strongly by the locations of plants belonging to different 4-digit industries within the same 3-digit class as they are by the locations of plants belonging to their own 4-digit industry? In this section, we develop a framework for addressing such questions and apply it to our data.

Consider an industry with r subindustries having shares $w_1, w_2, \ldots, w_r$ of the overall industry employment. Write H^j for the plant Herfindahl of the j^{th} subindustry, and $H = \sum_{j=1}^{r} w_j^2 H^j$ for the plant Herfindahl of the broader industry. Suppose that plants choose their locations in a manner which is nearly identical to that of our spillover model, but that the probability of any pair of darts being welded together (*i.e.* that a crucial spillover exists between the firms they represent) being γ_j if both darts correspond to plants within the

[34] 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_p is larger, so the list may contain many indnstries with large H_p simply because this is where we have made the largest errors in measurement.

Table 4: Most and Least Localized Industries

15 Most Localized Industries

4-digit industry	H_p	G	γ
2371. Fur Goods	0.007	0.63	0.63
2084. Wines, Brandy, Brandy Spirits	0.041	0.50	0.48
2252. Hosiery, n.e.c.	0.008	0.44	0.44
3533. Oil and Gas Field Machinery	0.015	0.44	0.43
2251. Women's Hosiery	0.028	0.42	0.40
2273. Carpets and Rugs	0.013	0.39	0.38
2429. Special Product Sawmills, n.e.c.	0.009	0.38	0.37
3961. Costume Jewelry	0.017	0.33	0.32
2895. Carbon Black	0.054	0.34	0.30
3915. Jewelers' Materials, Lapidary	0.025	0.32	0.30
2874. Phosphatic Fertilizers	0.066	0.34	0.29
2061. Raw Cane Sugar	0.038	0.32	0.29
2281. Yarn Mills, Except Wool	0.005	0.29	0.28
2034. Dehydrated Fruits, Veg's, Soups	0.030	0.30	0.28
3761. Guided Missiles, Space Vehicles	0.046	0.28	0.25

15 Least Localized Industries

4-digit industry	H_p	G	γ
3021. Rubber and Plastics Footwear	0.06	0.05	-0.013
2032. Canned Specialties	0.03	0.02	-0.012
2082. Malt Beverages	0.04	0.03	-0.010
3635. Household Vacuum Cleaners	0.18	0.18	-0.009
3652. Prerecorded Records and Tapes	0.04	0.03	-0.008
3482. Small Arms Ammunition	0.18	0.18	-0.004
3324. Steel Investment Foundries	0.04	0.04	-0.003
3534. Elevators and Moving Stairways	0.03	0.03	-0.001
2052. Cookies and Crackers	0.03	0.03	-0.0009
2098. Macaroni and Spaghetti	0.03	0.03	-0.0008
3262. Vitreous China Table, Kitchenware	0.13	0.13	-0.0006
2035. Pickles, Sauces, Salad Dressings	0.01	0.01	-0.0003
3821. Laboratory Apparatus and Furniture	0.02	0.02	-0.0002
2062. Cane Sugar Refining	0.11	0.11	0.0002
3433. Heating Equipment except Electric	0.008	0.009	0.0002

Table 5: Concentration and Industry Definition

Industry Definition	Industry means H_p	G	γ
2-digit	0.007	0.032	0.026
3-digit	0.014	0.058	0.045
4 digit	0.028	0.077	0.051

j^{th} subindustry and γ_0 otherwise. We will assume that $\gamma_0 \leq \text{Min}_{j=1\ldots r}\gamma_j$, so that spillovers are always more powerful (in expectation) between plants which are more similar.[35]

Again writing G for the raw geographic concentration of the broader industry we have

Proposition 6 *In the model above,*

$$E(G) = H + \gamma_0(1 - \sum_{j=1}^{r} w_j^2) + \sum_{j=1}^{r} \gamma_j w_j^2 (1 - H^j).$$

Proof

Write n_j for the number of business units in the j^{th} subindustry and $z_{j1}, \ldots, z_{jn_j}$ for the sizes of plants in that subindustry. Writing $u_{j\ell i}$ for the Bernoulli random variable indicating whether the ℓ^{th} plant in subindustry j locates in area i, the assumption on welding implies

$$\text{Corr}(u_{j\ell i}, u_{j'\ell' i}) = \begin{cases} \gamma_j & \text{if } j = j' \text{ and } \ell \neq \ell' \\ \gamma_0 & \text{if } j \neq j'. \end{cases}$$

We then have

$$\begin{aligned}
(1 - \sum_i x_i^2)E(G) &= \sum_i \text{Var}(s_i) \\
&= \sum_i \Bigg[\sum_{j,\ell} z_{j\ell}^2 \text{Var}(u_{j\ell i}) + \sum_{j,\ell,\ell',\ell\neq\ell'} z_{j\ell} z_{j\ell'} \text{Cov}(u_{j\ell i}, u_{j\ell' i}) \\
&\quad + \sum_{j,j',\ell,\ell',j\neq j'} z_{j\ell} z_{j'\ell'} \text{Cov}(u_{j\ell i}, u_{j'\ell' i}) \Bigg] \\
&= (\sum_i x_i(1 - x_i)) \Bigg[\sum_{j,\ell} z_{j\ell}^2 + \sum_{j,\ell,\ell',\ell\neq\ell'} z_{j\ell} z_{j\ell'} \gamma_j + \sum_{j,j',\ell,\ell',j\neq j'} z_{j\ell} z_{j'\ell'} \gamma_0 \Bigg]
\end{aligned}$$

[35] We require also that the welding relation is again symmetric and transitive. The assumption that $\gamma_0 \leq \text{Min}\gamma_j$ ensures that such a joint distribution on the welding probabilities exists so that the model is well defined.

$$E(G) = H + \sum_j \gamma_j (w_j^2 - \sum_\ell z_{j\ell}^2) + \gamma_0 (1 - \sum_j w_j^2)$$
$$= H + \gamma_0 (1 - \sum_j w_j^2) + \sum_j \gamma_j w_j^2 (1 - H^j).$$

QED.

Given data on state-industry employments for an industry and for each of its subindustries, raw geographic concentrations may be computed both for the larger industry and for the subindustries. Write G_j for the raw concentration in the j^{th} subindustry and $\hat{\gamma}_j$ for $\frac{G_j - H^j}{1 - H^j}$. An unbiased estimate of the degree of intersubindustry spillovers may then be obtained from the raw concentrations by setting

$$\hat{\gamma_0} = \frac{G - H - \sum_{j=1}^r \hat{\gamma}_j w_j^2 (1 - H^j)}{1 - \sum_{j=1}^r w_j^2}.$$

To discuss the degree to which spillovers are general, we define a new measure λ by

$$\lambda = \frac{\hat{\gamma_0}}{\sum_j w_j \hat{\gamma}_j}.$$

Note that this measure should be zero if there are *no* spillovers between plants in different subindustries and one if spillovers are equally strong regardless of the subindustry to which plants in the same broad industry belong.[36]

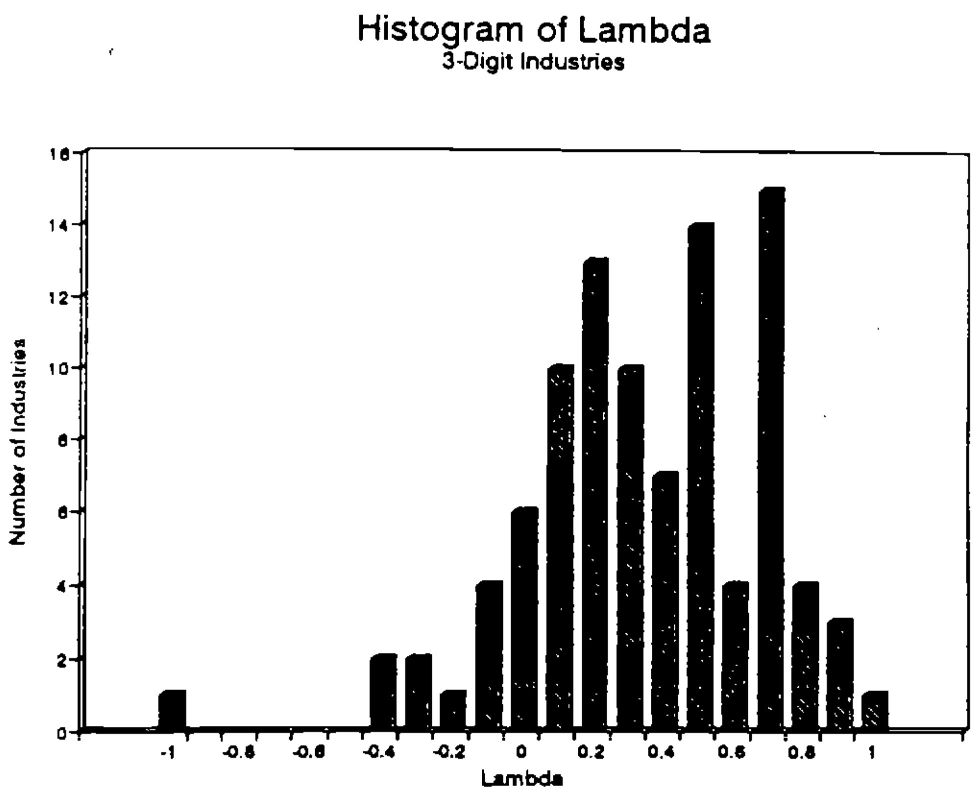
There are 97 3-digit industires with more than one 4-digit subindustry. A histogram for λ on these three-digit industries is given in Figure 2. The values of λ are fairly evenly spread between 0 and 0.8, indicating that there certainly is some clustering of similar 4-digit industries. In answer to our introductory question, however, it appears that spillovers are nearly as strong across 4-digit industries in the same 3-digit industry as within the 4-digit industries themselves only in about 20% of the cases.

Moving on to yet broader industry classes, Table 6 reports the results of the identical calculation using the 3-digit subindustries of each 2-digit industry. The mean value of λ across 2-digit industries is 0.29. There is great variation across industries. In four cases (furniture, industrial machinery, electronic and electric equipment, and transportation equipment) the data indicate that there is no concentration at all at the 2-digit level. On the other hand, there is substantial concentration of the 3-digit industries within the 2-digit tobacco, textile, and lumber industries.

We should note while we have tended to use the word "spillovers" in this section, several factors may explain the results. Technology or knowledge spillovers are one possibility, but

[36] Of course, λ is a random variable so these statements apply literally only to $\frac{\gamma_0}{\sum_j w_j \gamma_j}$. Note also that λ is not an unbiased estimate of this expression. An earlier version of this paper defined λ by a linear interpolation which in practice yields values almost identical to those we report here.

Figure 2: Extent of Spillovers between 4-digit Industries



spatially correlated natural advantages or other "spillovers" like minimizing transportation costs given intersubindustry trade could also account for positive values of λ. While a more detailed analysis is clearly called for, the list of 2-digit industries where λ is largest suggests that we may be detecting in large part that the "natural advantages" which are important to the subindustries are similar.

Table 6: Extent of Spillovers between 3-digit Industries

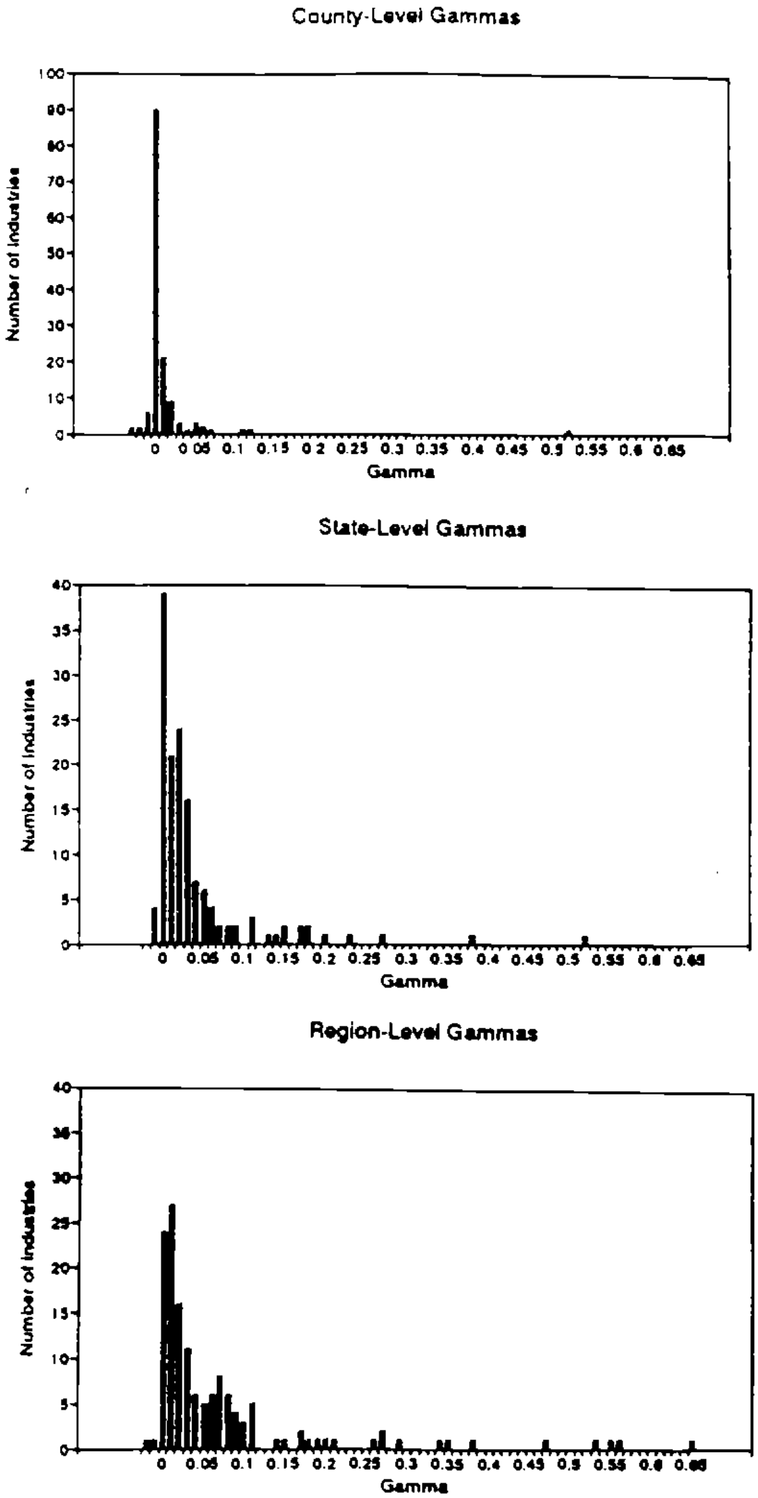
2-digit industry	$\hat{\gamma}_0$	λ	2-digit industry	$\hat{\gamma}_0$	λ
Food and kindred products	0.002	0.14	Rubber and misc. plastics	0.003	0.38
Tobacco products	0.151	0.88	Leather and leather products	0.017	0.31
Textile mill products	0.115	0.61	Stone, clay, and glass products	0.002	0.20
Apparel and other textiles	0.010	0.29	Primary metal industries	0.012	0.41
Lumber and wood products	0.016	0.63	Fabricated metal products	0.003	0.22
Furniture and fixtures	0.001	0.02	Industrial machinery and equip.	0.000	0.00
Paper and allied products	0.005	0.31	Electronic & other electric equip.	0.000	0.02
Printing and publishing	0.005	0.48	Transportation equipment	-0.001	-0.08
Chemicals and allied products	0.007	0.25	Instruments and related products	0.013	0.36
Petroleum and coal products	0.007	0.12	Miscellaneous manufacturing	0.011	0.34

6.2 Geographic Scope of Concentration

In Section 3, we noted that the γ's estimated from county-, state-, or region-level data should be identical (in expectation) provided the scope of spillovers is such that advantages are gained only if firms choose identical locations. 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/3-digit dataset using counties, states, and the nine census regions as the units of observation.

Figure 3 presents histograms of the γ's estimated from the three levels of data. Comparing first the county- and state-level estimates note that substantially more concentration is apparent at the state level. The median γ's at the two levels are 0.005 and 0.023, with the median of the ratio between them being 0.25, so that typically the effect of spillovers is such that about one fourth of the excess tendency of plants to locate in the same state involves

Figure 3: Concentration at the County, State, and Regional Level



plants locating in the same county. We draw two conclusions. First, given that states have many more than 4 counties, spillovers appear to have a stronger effect at very small distances. Second, spillovers are still quite substantial at a range beyond that of counties. In only a few cases do spillovers appear both to be substantial and limited in scope to the county level.[37] The rubber and plastics footwear industry seems to be the unique example where concentration is substantially greater at the county level than at the state level, *i.e.* where tightly grouped clusters of plants are spread (excessively) evenly across the states as if to minimize transportation costs.

Measured levels of state and regional concentration are more similar, although the regional data shows a much thicker tail of very concentrated industries. (The mean γ's are 0.044 and 0.078.) The general pattern of slightly more than half of the tendency of firms to locate in the same region being accounted for by the tendency to locate in the same state appears to hold equally well for industries which are very unconcentrated and very concentrated at the state level, although there is considerable variation about this norm.[38]

7 Geographic Concentration within the Firm

In this section we investigate the tendency to locate together of plants belonging to the same firm. The issue is interesting not only in its own right, but also in that such a tendency might 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. Suppose that firm j consists of n_j plants having shares $z_{j1}, \ldots, z_{jn_j}$ of the industry's employment. Assume that the location choices of the plants are again made as in our spillover model with the probability of a pair of darts being welded being equal to γ_0 if they correspond to plants in different firms and $\gamma_1 > \gamma_0$ if they belong to the same firm. The model is thus a special case of the model of Section 6.1 with the firms analogous to subindustries and the expected degree of spillovers within each firm assumed to be identical. A direct corollary of Proposition 6 is

Proposition 7 *In the model above,*

$$E(G) = H_p + \gamma_0(1 - H_f) + \gamma_1(H_f - H_p).$$

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

[38]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.

In trying 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 OLS regression

$$G_i - H_{pi} = \alpha_0(1 - H_{pi}) + \alpha_1(H_{fi} - H_{pi}) + \epsilon_i$$

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

We estimated the regression above for our sample of 444 4-digit industries. The parameter estimates are $\hat{\alpha_0}$ and $\hat{\alpha_1}$ are 0.046 (s.e. 0.005) and 0.068 (s.e. 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 can not 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.

As a simple specification test for this model, we estimated also the unconstrained regression

$$G_i = \beta_0 + \beta_1 H_{fi} + \beta_2 H_{pi} + \epsilon_i$$

and performed a Wald test of the restriction $\beta_0 + \beta_1 + \beta_2 = 1$. The test does reject the specification at the 5% level, although we note that the test would no longer reject if we made the minor bias correction suggested by Appendix B. While we believe the results of this section are of interest, we thus admit that they clearly should be interpreted with some caution.

8 Conclusion

In this paper we have developed a new framework for the analysis of geographic concentration based on a dart throwing metaphor. Using a series of very simple models, we obtain characterizations of both "random" agglomeration and of agglomeration caused by spillovers and natural advantage. Our most important theoretical result is that it is possible to control for industry characteristics in a fairly robust manner when measuring geographic concentration. This leads us to propose two new "natural" indices, γ and λ, to measure the localization of industries and the relative strength of cross-industry agglomerations.

The empirical work of this paper is largely descriptive. Besides reaffiming that localization is ubiquitous, we have tried to develop rough sketches of the variation of localization across industries, the geographic and industry levels at which it is most prominent, and its relationship to the structure of multiplant firms.

The existence of geographic concentration has attracted the attention of researchers in many fields and many potential explanations have been proposed. In the future, we hope that our measurement techniques will prove useful both in descriptive work on the nature of geographic concentration and in attempts to use the facts uncovered to assess the relative importance of various agglomerative forces.

References

Carlton, Dennis (1983): "The Location and Employment Decisions of New Firms: An Econometric Model with Discrete and Continuous Exogenous Variables," *Review of Economics and Statistics*, 65, 440–449.

Creamer, Daniel (1943): "Shifts of Manufacturing Industries," in *Industrial Location and Natural Resources*. Washington, DC: US Government Printing Office.

Enright, Michael (1990): *Geographic Concentration and Industrial Organization*. Ph.D. thesis, Harvard University.

Florence, P. Sargant (1948): *Investment, Location and Size of Plant*. London: Cambridge University Press.

Fuchs, Victor (1957): *The Economics of the Fur Industry*. New York: Columbia University Press.

Fuchs, Victor (1962): *Changes in the Location of Manufacturing in the US Since 1929*. New Haven: Yale University Press.

Gardocki, Bernard C., Jr. and Baj, John (1985): "Methodology for Estimating Nondisclosure in County Business Patterns Data," Mimeo.

Glaeser, Edward L., Kallal, Hedi D., Scheinkman, Jose, and Shleifer, Andrei (1992): "Growth in Cities," *Journal of Political Economy*, 100, 1126–1152.

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

Hoover, E.M. (1948): *The Location of Economic Activity*. New York: McGraw Hill.

Jaffe, Adam, Trajtenberg, Manuel and Henderson, Rebecca (1993): "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108, 577–598.

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

Krugman, Paul (1991b): "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, 483–499.

McFadden, Daniel (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Fronteirs in Eoconometrics*. New Yourk: Academic Press, 1974, pp. 105–42.

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

Scherer, F. Michael. et al. (1975): *The Economics of Multi-plant Operation: an International Comparisons Study*. Cambridge, Mass: Harvard University Press.

Schmalensee, Richard (1977): "Using the H-index of Concentration with Published Data," *Review of Economics and Statistics*, 59, 186–193.

Appendix A

This appendix describes the process by which state-industry employment figures were constructed. The 1987 Census of Manufactures reports the employment or a range of employments for all state-industries with at least 150 employees. Table 7 indicates the number of these state-industries for which data is categorized, the number of these which are topcoded at 2500 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.

Table 7: Extent of Withheld Data

	Industry Definition		
	2-digit	3-digit	4-digit
# Industries	21	141	460
# Cells with Ranges	153	1776	5700
# Topcodes	46	268	487
Avg. Employment Fraction	0.02	0.11	0.20

Before beginning to fill in the data, we first adjust the upper or lower bounds on any 2- or 3-digit state industry for which a sharper bound can be obtained by summing the upper or lower bounds of the subindustries which comprise it. This reduces the number of 2- and 3-digit state industries without upper bounds to 13 and 157, respectively. In addition, a total of 82 and 680 bounds are tightened on cells where a non-topcoded range had been given.

The filling process begins with the 21×51 matrix of 2-digit data. First, a rough estimate of the total employment in cells which 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 employess 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 less 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 and 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. The manner in which this is done usually results in topcodes not being filled in until virtually all active rows and columns contain a topcode, and rows/columns with multiple topcodes not being filled until there are no rows/columns with a single topcode 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 filling the 2-digit data, the process is repeated on the 3- and 4- digit data, the only difference being 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 3-digit subindustries of a 2-digit industry adding up to the employment in that state in the 2-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 4-digit data, these rounded-to-zero employments are occasionally a nontrivial fraction of the total employment in an idustry.

While there is no way to tell that this algorithm is doing well, it is at least possible to tell that it is doing badly to the extent that the algorithm is unable to make the state or industry totals add up (although due to rounding errors totals are off by up to 400 employees in industries where no data is withheld). Of the 21 2-digit industries, the maximum error in the adding up constraints is 508 employees with all other industries within 400. In the 3-digit industries and 4-digit industries there are two and six industries where the error is greater than 400, with two 4-digit industries having errors greater than 1000 employees, the maximum being 2010 (although these two are very big industries). The average error in the state adding up constraints are 31, 177, and 558 at the 2-, 3- and 4-digit level. In all but one of the 2-digit industries and in all but 6 of the 3-digit industries it was never necessary to fill in multiple topcodes 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 sufficienty transparent that we felt confident that we could reasonably simulate it. Absent that, we present here a small test of the accuracy of our procedure based on data obtained separately from the County Business Patterns for the area where our procedure is most suspect, filling in topcodes in the 4-digit data.

Data was available from County Business Patterns on state-industry employment for 171 of the 487 4-digit state-industries where employment was topcoded at 2500 or more. The CBP's sample differs somewhat from the Census of Manufactures, and as a result the CBP reported employment is below 2500 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 it is 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 and in each case our data fit extremely well, giving our estimates a misleadingly high 0.98 correlation with the CBP data. Across the remaining 137 state-industries the mean and standard deviation of employment are virtually identical in our data and in the CBP data and the correlation between the two is 0.74. (The means are 5329 and 5304, the standard deviations 3451 and 3306.) For comparison, if the Census of Manufactures had reported ranges for this data using the CBP ranges (2500-4999, 5000-10000, and 10000-20000) and we had constructed estimates simply by filling in the mean of the appropriate range, the correlation coefficient would be higher (0.93), but the sample means and variance would be much further from those of the CBP data. (The mean would be 5939, and the standard deviation 4314).

While the results above suggest that our procedure has some accuracy in filling, the most important question is clearly what implications errors in assigning state-employments have on the computation of G. Even a procedure which 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 of that our filling in of topcodes 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 topcoded cells in the 61 industries where the CBP data allowed all topcodes to be filled in (and where there was at least one topcode). For this purpose we took the CBP data to report an employment of 2500 whenever it actually reported a smaller number. Comparing our previously estimated G with the value G_{cbp}, we find that the means are 0.054 and 0.050, with correlation of 0.96. The absolute value of the difference between the two has a median of 0.0015, with the value being larger than

0.005 in 11 of the 61 industries. While this suggests that our filling of topcodes 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 topcodes to fill because they tended to have fewer topcodes than the average industry with at least one topcode (1.5 vs. 2.5). On the other hand, the majority of 4-digit industries have no topcoded cells to begin with. Also, while we the filled topcodes would appear to be the greatest potential problem with our algorithm, this test says nothing about biases due to the filling of nontopcoded ranges and of state-industry employments of less 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 a state/3-digit industry calculations with our standard dataset and with state totals from our county-level dataset. (Recall that this latter dataset had been constructed entirely from CBP data using mean establishment sizes to fill in missing values.) Because the latter dataset 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 14 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 datasets as giving roughly similar measurements. The differences are both larger than 0.015 and larger than 20% of the larger G for only eight SIC codes: 213, 281, 302, 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 the variable H_p 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_p by a procedure which is much more akin to filling in data than to imposing any distributional assumptions and estimating parameters, and which therefore will admittedly be ad hoc. The algorithm has two main steps: the first consisting of allocating employees across size classes to obtain a regular data structure, and the second of computing an expected sum of squares for the plants within each class using a rule of thumb recommeded by Schmalensee (1977).

In 316 of the 459 industries 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, we first for each size class used the sample of industrues 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 be of the form $a_0 + a_1 log(1 + n)$ and the variance of the form $b_0 + b_1(1/n)$, with the parameters estimated by OLS regressions.) Employment in each of the combined classes were 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 the mean and with its boundary at the closer of the two endpoints of the size range. H_p is estimated simply by taking the sum of the squares of the plant shares for this particular allocation of employees across plants. Schmalensee (1977) 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 which approximates the Herfindahl index given the available data. To assess the accuracy of this procedure, we constructed a simulated dataset of 5000 industrues. The simulated industries 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 (1106), 143 (286), and 287 (2101), respectively. These parameters were obtained from sample statistics (and the estimated H_p) of our 459 industry sample. The data produced by the simulations bears a superficial resemblance to the actual data, although it tends to contain far more extreme outliers (*e.g.* industries with over 95% of employment in a single plant.) We created a simulated dataset 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 3200 of the 5000 industries.

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

$$E(\gamma - \gamma_0 | \hat{H}_p) = (1 - \gamma_0)E(\frac{H_p - \hat{H}_p}{1 - \hat{H}_p} | \hat{H}_p).$$

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

One cannot estimate $E(H_p | \hat{H}_p)$ without making assumptions about the distribution of H_p. While our simulated H_p's do not match the observed distribution of plant Herfindahls, we hope that they will at least provide results which are indicative of the magnitude of the bias our procedure produces. Over our 5000 industry sample, a OLS regression of H_p on $\hat{H}_p$ yields an estimated constant of 0.0003 (t-stat: 1.3), with the estimated coefficient on $\hat{H}_p$ being 1.04 (t-stat: 228.9). Restricting the regression to the observations with $\hat{H}_p < 0.3$ to eliminate the effect of unreasonable industries gives estimates of 0.0001 (t-stat: 0.5) and 1.05 (t-stat: 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, $\hat{H}_p$, and $\hat{H}_p^2$ to get an idea of the magnitude of the measurement error in a typical industry gives the estimate $\hat{\sigma}^2 = 0.00003 + 0.003\hat{H}_p + 0.007\hat{H}_p^2$.

If we believe these results, then for a typical industry with γ_0 small we will underestimate γ_0 by about $0.05H_p$. Given that the mean of H_p is less than 0.03, this bias is fairly small. To correct this bias, one could simply multiply all of our previous estimates of H_p 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.

Appendix C

Industry	Employment	Plant Herfindahl	Gamma
2011 Meat packing plants	113.9	0.008	0.042
2013 Sausages and other prepared meats	78.7	0.004	0.006
2015 Poultry slaughtering and processing	147.9	0.005	0.055
2021 Creamery butter	1.7	0.045	0.147
2022 Cheese, natural and processed	33.0	0.009	0.131
2023 Dry, condensed, and evaporated dairy products	14.1	0.056	0.015
2024 Ice cream and frozen desserts	20.3	0.008	0.001
2026 Fluid milk	72.4	0.002	0.002
2032 Canned specialties	24.5	0.032	-0.012
2033 Canned fruits and vegetables	65.1	0.006	0.044
2034 Dehydrated fruits, vegetables, and soups	10.1	0.030	0.280
2035 Pickles, sauces, and salad dressings	21.4	0.013	0.000
2037 Frozen fruits and vegetables	49.8	0.011	0.079
2038 Frozen specialties, n.e.c.	37.5	0.015	0.002
2041 Flour and other grain mill products	13.3	0.009	0.019
2043 Cereal breakfast foods	16.0	0.064	0.018
2044 Rice milling	4.5	0.053	0.136
2045 Prepared flour mixes and doughs	12.1	0.020	0.015
2046 Wet corn milling	8.6	0.060	0.137
2047 Dog and cat food	13.4	0.018	0.011
2048 Prepared feeds, n.e.c.	34.5	0.002	0.019
2051 Bread, cake, and related products	161.9	0.003	0.001
2052 Cookies and crackers	45.3	0.028	-0.001
2053 Frozen bakery products, except bread	9.9	0.035	0.013
2061 Raw cane sugar	6.2	0.038	0.289
2062 Cane sugar refining	5.5	0.107	0.000
2063 Beet sugar	7.9	0.031	0.074
2064 Candy and other confectionery products	45.8	0.012	0.046
2066 Chocolate and cocoa products	11.0	0.107	0.038
2067 Chewing gum	5.2	0.157	0.072
2068 Salted and roasted nuts and seeds	8.8	0.079	0.025
2074 Cottonseed oil mills	2.8	0.032	0.168
2075 Soybean oil mills	7.0	0.020	0.070
2076 Vegetable oil mills, n.e.c.	0.9	0.084	0.049
2077 Animal and marine fats and oils	9.8	0.009	0.010
2079 Edible fats and oils, n.e.c.	9.3	0.021	0.031
2082 Malt beverages	31.9	0.042	-0.010
2083 Malt	1.4	0.072	0.238
2084 Wines, brandy, and brandy spirits	13.9	0.041	0.480
2085 Distilled and blended liquors	9.0	0.035	0.079
2086 Bottled and canned soft drinks	95.6	0.002	0.005
2087 Flavoring extracts and syrups, n.e.c.	9.1	0.018	0.025
2091 Canned and cured fish and seafoods	6.7	0.020	0.061
2092 Fresh or frozen prepared fish	38.2	0.007	0.059
2095 Roasted coffee	10.7	0.026	0.032
2096 Potato chips and similar snacks	33.1	0.011	0.008
2097 Manufactured ice	4.7	0.005	0.012
2098 Macaroni and spaghetti	6.6	0.028	-0.001
2099 Food preparations, n.e.c.	58.0	0.003	0.013
2111 Cigarettes	32.0	0.223	0.169
2121 Cigars	2.5	0.107	0.158
2131 Chewing and smoking tobacco	3.3	0.083	0.200

Industry	Employment	Plant Herfindahl	Gamma
2141 Tobacco stemming and redrying	6.9	0.045	0.178
2211 Broadwoven fabric mills, cotton	72.3	0.025	0.170
2221 Broadwoven fabric mills, manmade fiber and silk	88.3	0.007	0.228
2231 Broadwoven fabric mills, wool	14.0	0.042	0.087
2241 Narrow fabric mills	18.5	0.011	0.074
2251 Women's hosiery, except socks	29.3	0.028	0.398
2252 Hosiery, n.e.c.	36.5	0.008	0.437
2253 Knit outerwear mills	59.0	0.012	0.065
2254 Knit underwear mills	19.3	0.082	0.020
2257 Weft knit fabric mills	34.9	0.019	0.191
2258 Lace and warp knit fabric mills	20.5	0.014	0.116
2259 Knitting mills, n.e.c.	3.8	0.071	0.094
2261 Finishing plants, cotton	16.5	0.019	0.123
2262 Finishing plants, manmade	27.9	0.022	0.188
2269 Finishing plants, n.e.c.	11.7	0.020	0.098
2273 Carpets and rugs	53.3	0.013	0.378
2281 Yarn spinning mills	89.0	0.005	0.284
2282 Throwing and winding mills	18.3	0.025	0.206
2284 Thread mills	6.5	0.051	0.207
2295 Coated fabrics, not rubberized	10.3	0.020	0.001
2296 Tire cord and fabrics	5.1	0.121	0.178
2297 Nonwoven fabrics	13.8	0.023	0.038
2298 Cordage and twine	6.9	0.017	0.034
2299 Textile goods, n.e.c.	16.4	0.009	0.021
2311 Men's and boys' suits and coats	55.2	0.010	0.043
2321 Men's and boys' shirts	76.7	0.004	0.062
2322 Men's and boys' underwear and nightwear	17.2	0.032	0.097
2323 Men's and boys' neckwear	7.4	0.018	0.106
2325 Men's and boys' trousers and slacks	93.3	0.004	0.064
2326 Men's and boys' work clothing	33.1	0.009	0.090
2329 Men's and boys' clothing, n.e.c.	52.2	0.006	0.025
2331 Women's, misses', and juniors' blouses and shirts	73.4	0.002	0.038
2335 Women's, misses', and juniors' dresses	112.7	0.001	0.098
2337 Women's, misses', and juniors' suits and coats	55.2	0.003	0.034
2339 Women's, misses', and juniors' outerwear, n.e.c.	107.3	0.002	0.028
2341 Women's and children's underwear	53.7	0.006	0.053
2342 Brassieres, girdles, and allied garments	13.8	0.024	0.019
2353 Hats, caps, and millinery	17.2	0.013	0.044
2361 Girls' and children's dresses and blouses	30.9	0.007	0.030
2369 Girls' and children's outerwear, n.e.c.	40.8	0.008	0.046
2371 Fur goods	2.2	0.007	0.630
2381 Fabric dress and work gloves	4.8	0.027	0.102
2384 Robes and dressing gowns	8.7	0.029	0.024
2385 Waterproof outerwear	6.4	0.057	0.075
2386 Leather and sheep-lined clothing	2.1	0.034	0.100
2387 Apparel belts	10.5	0.013	0.167
2389 Apparel and accessories, n.e.c.	8.3	0.015	0.020
2391 Curtains and draperies	27.1	0.008	0.025
2392 Housefurnishings, n.e.c.	50.5	0.006	0.036
2393 Textile bags	8.8	0.011	0.005
2394 Canvas and related products	16.7	0.005	0.010
2395 Pleating and stitching	14.1	0.009	0.026

Industry	Employment	Plant Herfindahl	Gamma
2396 Automotive and apparel trimmings	44.2	0.016	0.074
2397 Schiffli machine embroideries	5.9	0.025	0.153
2399 Fabricated textile products, n.e.c.	30.5	0.008	0.005
2411 Logging	85.8	0.001	0.062
2421 Sawmills and planing mills, general	148.3	0.001	0.039
2426 Hardwood dimension and flooring mills	29.7	0.005	0.062
2429 Special product sawmills, n.e.c.	2.2	0.009	0.374
2431 Millwork	89.0	0.005	0.013
2434 Wood kitchen cabinets	67.0	0.002	0.011
2435 Hardwood veneer and plywood	20.5	0.008	0.050
2436 Softwood veneer and plywood	38.9	0.008	0.187
2439 Structural wood memebers, n.e.c.	24.6	0.003	0.026
2441 Nailed wood boxes and shook	5.9	0.009	0.018
2448 Wood pallets and skids	25.7	0.001	0.006
2449 Wood containers, n.e.c.	5.4	0.023	0.026
2451 Mobile homes	39.9	0.005	0.037
2452 Prefabricated wood buildings	25.4	0.006	0.025
2491 Wood preserving	11.8	0.005	0.029
2493 Reconstituted wood products	22.0	0.011	0.029
2499 Wood products, n.e.c.	56.3	0.002	0.006
2511 Wood household furniture	135.9	0.003	0.077
2512 Upholstered household furniture	82.1	0.004	0.130
2514 Metal household furniture	30.1	0.010	0.013
2515 Mattresses and bedsprings	24.4	0.004	0.007
2517 Wood television and radio cabinets	5.9	0.072	0.010
2519 Household furniture, n.e.c.	5.9	0.050	0.004
2521 Wood office furniture	31.0	0.009	0.045
2522 Office furniture, except wood	49.7	0.036	0.050
2531 Public building and related furniture	21.8	0.012	0.008
2541 Wood partitions and fixtures	40.6	0.002	0.003
2542 Partitions and fixtures, except wood	33.5	0.007	0.011
2591 Drapery hardware and blinds and shades	20.6	0.018	0.006
2599 Furniture and fixtures, n.e.c.	29.3	0.005	0.007
2611 Pulp mills	14.2	0.051	0.047
2621 Paper mills	129.1	0.008	0.039
2631 Paperboard mills	52.3	0.011	0.024
2652 Setup paperboard boxes	8.7	0.011	0.037
2653 Corrugated and solid fiber boxes	105.7	0.001	0.001
2655 Fiber cans, drums, and similar products	12.5	0.009	0.006
2656 Sanitary food containers	15.8	0.047	0.028
2657 Folding paperboard boxes	50.7	0.004	0.002
2671 Paper coated and laminated packaging	15.0	0.018	0.018
2672 Paper coated and laminated, n.e.c.	30.9	0.017	0.010
2673 Bags: plastics, laminated, and coated	36.6	0.009	0.011
2674 Bags: uncoated paper and multiwall	17.1	0.013	0.025
2675 Die-cut paper and board	15.7	0.011	0.011
2676 Sanitary paper products	38.4	0.020	0.033
2677 Envelopes	27.6	0.007	0.008
2678 Stationery products	11.2	0.021	0.025
2679 Converted paper products, n.e.c.	29.6	0.009	0.012
2711 Newspapers	434.4	0.002	0.002
2721 Periodicals	110.0	0.005	0.067

Industry	Employment	Plant Herfindahl	Gamma
2731 Book publishing	70.1	0.008	0.062
2732 Book printing	43.5	0.012	0.011
2741 Miscellaneous publishing	69.5	0.005	0.008
2752 Commercial printing, lithographic	403.9	0.000	0.004
2754 Commercial printing, gravure	23.8	0.032	0.016
2759 Commercial printing, n.e.c.	125.8	0.001	0.004
2761 Manifold business forms	53.3	0.003	0.003
2771 Greeting cards	21.5	0.091	0.037
2782 Blankbooks and looseleaf binders	39.1	0.007	0.007
2789 Bookbinding and related work	29.7	0.005	0.020
2791 Typesetting	37.6	0.002	0.014
2796 Platemaking services	31.8	0.002	0.010
2812 Alkalies and chlorine	5.0	0.061	0.058
2813 Industrial gases	8.1	0.005	0.011
2816 Inorganic pigments	8.3	0.041	0.031
2819 Industrial inorganic chemicals, n.e.c.	72.2	0.053	0.017
2821 Plastics materials and resins	56.3	0.012	0.029
2822 Synthetic rubber	10.4	0.063	0.165
2823 Cellulosic manmade fibers	10.5	0.224	0.159
2824 Organic fibers, noncellulosic	45.4	0.043	0.140
2833 Medicinals and botanicals	11.6	0.042	0.089
2834 Pharmaceutical preparations	131.6	0.015	0.023
2835 Diagnostic substances	15.4	0.033	0.059
2836 Biological products, except diagnostic	13.3	0.023	0.010
2841 Soap and other detergents	31.7	0.016	0.004
2842 Polishes and sanitation goods	20.6	0.010	0.018
2843 Surface active agents	9.1	0.017	0.040
2844 Toilet preparations	57.9	0.011	0.055
2851 Paints and allied products	55.2	0.003	0.007
2861 Gum and wood chemicals	2.6	0.041	0.061
2865 Cyclic crudes and intermediates	22.8	0.019	0.010
2869 Industrial organic chemicals, n.e.c.	100.3	0.012	0.069
2873 Nitrogenous fertilizers	7.4	0.025	0.031
2874 Phosphatic fertilizers	9.4	0.066	0.291
2875 Fertilizers, mixing only	7.5	0.006	0.020
2879 Agricultural chemicals, n.e.c.	16.1	0.038	0.031
2891 Adhesives and sealants	20.9	0.005	0.012
2892 Explosives	13.8	0.113	0.003
2893 Printing ink	11.1	0.005	0.015
2895 Carbon black	1.8	0.054	0.300
2899 Chemical preparations, n.e.c.	37.9	0.006	0.006
2911 Petroleum refining	74.6	0.011	0.088
2951 Asphalt paving mixtures and blocks	14.6	0.003	0.009
2952 Asphalt felts and coatings	13.5	0.009	0.010
2992 Lubricating oils and greases	11.2	0.007	0.013
2999 Petroleum and coal products, n.e.c.	1.9	0.027	0.061
3011 Tires and inner tubes	65.4	0.025	0.038
3021 Rubber and plastics footwear	10.9	0.060	-0.013
3052 Rubber and plastics hose and belting	23.2	0.026	0.038
3053 Gaskets, packing, and sealing devices	28.4	0.011	0.016
3061 Mechanical rubber goods	49.8	0.008	0.047
3069 Fabricated rubber products, n.e.c.	54.3	0.006	0.022

Industry	Employment	Plant Herfindahl	Gamma
3081 Unsupported plastics film and sheet	48.4	0.006	0.006
3082 Unsupported plastics profile shapes	25.2	0.007	0.005
3083 Laminated plastics plate, sheet, and profile shapes	17.3	0.025	0.005
3084 Plastics pipe	12.5	0.008	0.010
3085 Plastics bottles	25.1	0.007	0.012
3086 Plastics foam products	61.3	0.004	0.004
3087 Custom compounding of purchased plastics resins	17.3	0.008	0.012
3088 Plastics plumbing fixtures	7.5	0.023	0.014
3089 Plastics products, n.e.c.	384.9	0.001	0.005
3111 Leather tanning and finishing	14.6	0.013	0.025
3131 Footwear cut stock	5.0	0.032	0.142
3142 House slippers	3.7	0.104	0.066
3143 Men's footwear, except athletic	31.6	0.018	0.073
3144 Women's footwear, except athletic	26.6	0.012	0.055
3149 Footwear, except rubber, n.e.c.	9.2	0.025	0.088
3151 Leather gloves and mittens	3.1	0.028	0.035
3161 Luggage	11.4	0.027	0.041
3171 Women's handbags and purses	9.5	0.021	0.144
3172 Personal leather goods, n.e.c.	7.2	0.024	0.059
3199 Leather goods, n.e.c.	7.1	0.011	0.023
3211 Flat glass	14.8	0.055	0.019
3221 Glass containers	41.1	0.013	0.011
3229 Pressed and blown glass, n.e.c.	36.3	0.020	0.038
3231 Products of purchased glass	51.1	0.005	0.002
3241 Cement, hydraulic	19.1	0.009	0.010
3251 Brick and structural clay tile	16.6	0.007	0.036
3253 Ceramic wall and floor tile	9.5	0.039	0.023
3255 Clay refractories	6.4	0.027	0.078
3259 Structural clay products, n.e.c.	2.1	0.048	0.160
3261 Vitreous plumbing fixtures	9.7	0.041	0.014
3262 Vitreous china table and kitchenware	5.4	0.126	-0.001
3263 Semivitreous table and kitchenware	1.8	0.109	0.088
3264 Porcelain electrical supplies	10.7	0.030	0.044
3269 Pottery products, n.e.c.	10.5	0.016	0.012
3271 Concrete block and brick	18.6	0.002	0.004
3272 Concrete products, n.e.c.	70.0	0.001	0.012
3273 Ready-mixed concrete	96.8	0.001	0.010
3274 Lime	5.7	0.033	0.063
3275 Gypsum products	12.1	0.013	0.013
3281 Cut stone and stone products	12.5	0.011	0.036
3291 Abrasive products	23.4	0.038	0.028
3292 Asbestos products	4.0	0.107	0.009
3295 Minerals, ground or treated	8.8	0.011	0.005
3296 Mineral wool	21.5	0.020	0.015
3297 Nonclay refractories	7.7	0.020	0.042
3299 Nonmetallic mineral products, n.e.c.	7.6	0.009	0.004
3312 Blast furnaces and steel mills	188.1	0.018	0.067
3313 Electrometallurgial products	3.9	0.072	0.148
3315 Steel wire and related products	24.7	0.012	0.013
3316 Cold finishing of steel shapes	16.4	0.027	0.032
3317 Steel pipe and tubes	19.6	0.010	0.038
3321 Gray and ductile iron foundries	82.4	0.011	0.029

Industry	Employment	Plant Herfindahl	Gamma
3322 Malleable iron foundries	4.2	0.197	0.072
3324 Steel investment foundries	20.3	0.040	-0.003
3325 Steel foundries, n.e.c.	22.9	0.012	0.040
3331 Primary copper	3.3	0.135	0.194
3334 Primary aluminum	17.3	0.050	0.053
3339 Primary nonferrous metals, n.e.c.	11.0	0.044	0.004
3341 Secondary nonferrous metals	12.5	0.008	0.016
3351 Copper rolling and drawing	22.6	0.029	0.018
3353 Aluminum sheet, plate, and foil	26.1	0.063	0.009
3354 Aluminum extruded products	30.7	0.013	0.001
3355 Aluminum rolling and drawing, n.e.c.	0.9	0.084	0.031
3356 Nonferrous rolling and drawing, n.e.c.	17.9	0.031	0.016
3357 Nonferrous wiredrawing and insulating	64.9	0.008	0.018
3363 Aluminum die-castings	28.1	0.010	0.021
3364 Nonferrous die-casting, except aluminum	12.9	0.010	0.036
3365 Aluminum foundries	26.3	0.008	0.021
3366 Copper foundries	8.2	0.007	0.012
3369 Nonferrous foundries, n.e.c.	4.0	0.117	0.103
3398 Metal heat treating	18.0	0.004	0.026
3399 Primary metal products, n.e.c.	13.8	0.105	0.059
3411 Metal cans	39.4	0.006	0.009
3412 Metal barrels, drums, and pails	8.7	0.014	0.042
3421 Cutlery	10.5	0.039	0.056
3423 Hand and edge tools, n.e.c.	41.9	0.008	0.008
3425 Saw blades and handsaws	7.7	0.039	0.039
3429 Hardware, n.e.c.	85.2	0.007	0.008
3431 Metal sanitary ware	8.0	0.064	0.030
3432 Plumbing fixture fittings and trim	17.1	0.023	0.003
3433 Heating equipment, except electric	20.5	0.008	0.000
3441 Fabricated structural metal	80.9	0.006	0.004
3442 Metal doors, sash, and trim	74.7	0.003	0.003
3443 Fabricated plate work (boiler shops)	74.7	0.004	0.010
3444 Sheet metal work	100.2	0.001	0.003
3446 Architectural metal work	28.0	0.004	0.004
3448 Prefabricated metal buildings	25.8	0.009	0.006
3449 Miscellaneous metal work	22.9	0.006	0.014
3451 Screw machine products	42.7	0.002	0.027
3452 Bolts, nuts, rivets, and washers	52.0	0.006	0.029
3462 Iron and steel forgings	26.8	0.017	0.024
3463 Nonferrous forgings	7.3	0.082	0.022
3465 Automotive stampings	119.8	0.013	0.177
3466 Crowns and closures	8.1	0.056	0.039
3469 Metal stampings, n.e.c.	95.5	0.002	0.018
3471 Plating and polishing	71.1	0.001	0.012
3479 Metal coating and allied services	41.5	0.002	0.014
3482 Small arms ammunition	9.0	0.184	-0.004
3483 Ammunition, except for small arms, n.e.c.	41.5	0.041	0.003
3484 Small arms	13.3	0.067	0.080
3489 Ordnance and accessories, n.e.c.	23.9	0.166	0.004
3491 Industrial valves	45.9	0.009	0.006
3492 Fluid power valves and hose fittings	27.9	0.010	0.037
3493 Steel springs, except wire	5.0	0.024	0.048

Industry	Employment	Plant Herfindahl	Gamma
3494 Valves and pipe fittings, n.e.c.	25.1	0.010	0.017
3495 Wire springs	19.7	0.009	0.014
3496 Miscellaneous fabricated wire products	35.1	0.003	0.004
3497 Metal foil and leaf	10.4	0.033	0.033
3498 Fabricated pipe and fittings	20.0	0.004	0.020
3499 Fabricated metal products, n.e.c.	72.5	0.002	0.006
3511 Turbines and turbine generator sets	22.9	0.091	0.023
3519 Internal combustion engines, n.e.c.	64.0	0.034	0.070
3523 Farm machinery and equipment	57.0	0.013	0.064
3524 Lawn and garden equipment	24.9	0.043	0.014
3531 Construction machinery	81.1	0.016	0.061
3532 Mining machinery	13.6	0.016	0.057
3533 Oil and gas field machinery	24.8	0.015	0.433
3534 Elevators and moving stairways	10.2	0.028	-0.001
3535 Conveyors and conveying equipment	31.5	0.005	0.018
3536 Hoists, cranes, and monorails	7.0	0.020	0.015
3537 Industrial trucks and tractors	20.1	0.016	0.004
3541 Machine tools, metal cutting types	31.7	0.019	0.035
3542 Machine tools, metal forming types	13.8	0.018	0.071
3543 Industrial patterns	8.6	0.006	0.051
3544 Special dies, tools, jigs, and fixtures	114.4	0.001	0.053
3545 Machine tool accessories	48.5	0.003	0.037
3546 Power-driven handtools	16.8	0.037	0.045
3547 Rolling mill machinery	3.9	0.067	0.084
3548 Welding apparatus	18.7	0.028	0.040
3549 Metalworking machinery, n.e.c.	11.3	0.011	0.041
3552 Textile machinery	15.6	0.012	0.165
3553 Woodworking machinery	8.9	0.016	0.033
3554 Paper industries machinery	17.1	0.022	0.096
3555 Printing trades machinery	25.0	0.032	0.016
3556 Food products machinery	19.2	0.008	0.014
3559 Special industry machinery, n.e.c.	83.3	0.003	0.007
3561 Pumps and pumping equipment	35.2	0.010	0.008
3562 Ball and roller bearings	36.9	0.021	0.043
3563 Air and gas compressors	23.8	0.021	0.020
3564 Blowers and fans	24.8	0.008	0.003
3565 Packaging machinery	22.6	0.010	0.018
3566 Speed changers, drives, and gears	17.9	0.019	0.019
3567 Industrial furnaces and ovens	16.6	0.010	0.006
3568 Power transmission equipment, n.e.c.	22.0	0.014	0.014
3569 General industrial machinery, n.e.c.	40.8	0.004	0.004
3571 Electronic computers	151.9	0.019	0.059
3572 Computer storage devices	43.3	0.113	0.142
3575 Computer terminals	15.0	0.046	0.005
3577 Computer peripheral equipment, n.e.c.	76.2	0.030	0.031
3578 Calculating and accounting equipment	12.8	0.060	0.008
3579 Office machines, n.e.c.	28.5	0.053	0.015
3581 Automatic vending machines	7.9	0.062	0.005
3582 Commercial laundry equipment	4.6	0.054	0.019
3585 Refrigeration and heating equipment	133.3	0.008	0.011
3586 Measuring and dispensing pumps	9.4	0.083	0.002
3589 Service industry machinery, n.e.c.	35.2	0.005	0.014

Industry	Employment	Plant Herfindahl	Gamma
3592 Carburetors, pistons, rings, and valves	21.7	0.038	0.042
3593 Fluid power cylinders and actuators	20.2	0.052	0.025
3594 Fluid power pumps and motors	14.8	0.034	0.003
3596 Scales and balances, except laboratory	6.7	0.027	0.023
3599 Industrial machinery, n.e.c.	228.5	0.000	0.005
3612 Transformers, except electronic	32.2	0.018	0.021
3613 Switchgear and switchboard apparatus	44.8	0.010	0.008
3621 Motors and generators	74.6	0.008	0.021
3624 Carbon and graphite products	9.8	0.033	0.042
3625 Relays and industrial controls	66.6	0.010	0.008
3629 Electrical industrial apparatus, n.e.c.	14.5	0.017	0.010
3631 Household cooking equipment	21.9	0.050	0.030
3632 Household refrigerators and freezers	25.7	0.107	0.035
3633 Household laundry equipment	16.7	0.128	0.124
3634 Electric housewares and fans	25.1	0.019	0.107
3635 Household vacuum cleaners	11.3	0.182	-0.009
3639 Household appliances, n.e.c.	16.0	0.061	0.030
3641 Electric lamp bulbs and tubes	22.2	0.027	0.033
3643 Current-carrying wiring devices	47.9	0.017	0.009
3644 Noncurrent-carrying wiring devices	21.5	0.023	0.012
3645 Residential lighting fixtures	22.5	0.009	0.027
3646 Commercial lighting fixtures	22.7	0.022	0.018
3647 Vehicular lighting equipment	15.5	0.139	0.022
3648 Lighting equipment, n.e.c.	14.4	0.017	0.010
3651 Household audio and video equipment	30.9	0.035	0.016
3652 Prerecorded records and tapes	13.3	0.039	-0.008
3661 Telephone and telegraph apparatus	112.3	0.021	0.009
3663 Radio and television communications equipment	126.0	0.015	0.021
3669 Communications equipment, n.e.c.	21.9	0.017	0.030
3671 Electron tubes	28.4	0.057	0.043
3672 Printed circuit boards	66.6	0.005	0.041
3674 Semiconductors and related devices	184.6	0.014	0.064
3675 Electronic capacitors	21.7	0.023	0.029
3676 Electronic resistors	15.7	0.022	0.016
3677 Electronic coils and transformers	23.9	0.009	0.018
3678 Electronic connectors	42.8	0.017	0.036
3679 Electronic components, n.e.c.	162.6	0.008	0.022
3691 Storage batteries	24.2	0.017	0.010
3692 Primary batteries, dry and wet	10.7	0.045	0.049
3694 Engine electrical equipment	67.3	0.045	0.054
3695 Magnetic and optical recording media	25.6	0.028	0.065
3699 Electrical equipment and supplies, n.e.c.	60.3	0.008	0.015
3711 Motor vehicles and car bodies	281.3	0.016	0.127
3713 Truck and bus bodies	37.8	0.009	0.008
3714 Motor vehicle parts and accessories	389.6	0.006	0.089
3715 Truck trailers	27.5	0.013	0.014
3716 Motor homes	15.1	0.055	0.150
3721 Aircraft	268.2	0.053	0.023
3724 Aircraft engines and engine parts	139.6	0.042	0.047
3728 Aircraft parts and equipment, n.e.c.	188.2	0.029	0.032
3731 Ship building and repairing	120.2	0.080	0.014
3732 Boat building and repairing	57.2	0.005	0.046

Industry	Employment	Plant Herfindahl	Gamma
3743 Railroad equipment	22.1	0.085	0.123
3751 Motorcycles, bicycles, and parts	7.4	0.077	0.010
3761 Guided missiles and space vehicles	166.7	0.046	0.249
3764 Space propulsion units and parts	31.8	0.145	0.112
3769 Space vehicle equipment, n.e.c.	15.1	0.157	0.005
3792 Travel trailers and campers	17.2	0.011	0.087
3795 Tanks and tank components	16.7	0.157	0.023
3799 Transportation equipment, n.e.c.	15.4	0.015	0.021
3812 Search and navigation equipment	369.4	0.011	0.039
3821 Laboratory apparatus and furniture	17.1	0.020	0.000
3822 Environmental controls	26.5	0.035	0.011
3823 Process control instruments	53.3	0.010	0.017
3824 Fluid meters and counting devices	10.1	0.032	0.022
3825 Instruments to measure electricity	85.2	0.014	0.031
3826 Analytical instruments	31.2	0.014	0.039
3827 Optical instruments and lenses	20.1	0.027	0.061
329 Measuring and contolling devices, n.e.c.	41.0	0.015	0.004
41 Surgical and medical instruments	73.1	0.007	0.011
2 Surgical appliances and supplies	78.5	0.005	0.005
3 Denta equipment and supplies	14.6	0.017	0.022
4 X-ray apparatus and tubes	8.7	0.049	0.017
45 Electromedical equipment	29.2	0.021	0.025
3851 Opthalmic goods	24.2	0.020	0.027
3861 Photographic equipment and supplies	88.0	0.067	0.174
3873 Watches, clocks, watchcases, and parts	11.8	0.031	0.005
3911 Jewelry, precious metal	35.5	0.005	0.095
3914 Silverware and plated ware	6.9	0.065	0.049
3915 Jewelers' materials and lapidary work	7.1	0.025	0.298
3931 Musical instruments	12.2	0.017	0.014
3942 Dolls and stuffed toys	4.4	0.027	0.086
3944 Games, toys, and children's vehicles	30.9	0.017	0.011
9 Sporting and athletic goods, n.e.c.	53.6	0.005	0.003
Pens and mechanical pencils	8.4	0.048	0.030
Lead pencils and art goods	5.6	0.045	0.030
Marking devices	7.5	0.007	0.005
5 Carbon paper and inked ribbons	7.3	0.035	0.008
1 Costume jewelry	22.2	0.017	0.320
3965 Fasteners, buttons, needles, and pins	9.8	0.018	0.042
3991 Brooms and brushes	12.3	0.014	0.007
3993 Signs and advertising specialties	66.3	0.001	0.006
3995 Burial caskets	8.7	0.026	0.050
3996 Hard surface floor coverings, n.e.c.	7.6	0.139	0.097
3999 Manufacturing industries, n.e.c.	68.3	0.003	0.008