

GEOGRAPHIC CONCENTRATION AND ESTABLISHMENT SCALE

Thomas J. Holmes and John J. Stevens*

Abstract—This paper shows that plants located in areas where an industry concentrates are larger, on average, than plants in the same industry outside such areas. In some sectors, such as manufacturing, the differences are substantial. The connection between size and concentration is stronger than what we would expect to find if plants were randomly distributed like darts on a dartboard.

I. Introduction

GEOGRAPHIC concentration of industry is pervasive; this fact is consistent in new and past work. Relatively little is known, however, about the relationship between industry concentration and the scale of plants. Specifically, do plants located in areas where an industry concentrates tend on average to be larger, the same as, or smaller than those located outside such areas? This paper provides a clear answer to this question. Plants located in areas where an industry concentrates are larger, on average, than plants in the same industry outside such areas, and in some sectors, such as manufacturing, the differences are substantial.

An important issue that must be addressed in our analysis is that, from chance alone, we would expect to see a positive correlation across locations between industry concentration and average plant size. The issue is analogous to that considered by Ellison and Glaeser (1997). They observe that even if plants are randomly distributed across locations like darts on a dartboard, by chance pockets of concentration will emerge. A dartboard-like process will also lead to a positive correlation between average plant size and industry concentration. If, by chance, a location gets a very large plant, then both average plant size and industry concentration at the location will tend to be high.

We are able to explore this issue by using plant-level County Business Patterns (CBP) data. Our idea is to consider how the expectation of an individual plant's size varies with local concentration, *excluding* that part of the concentration contributed by that plant. If plants were randomly distributed across locations in a manner independent of size, there would be a zero correlation between a plant's size and its excluded concentration measure. But we find a positive correlation in virtually every major sector of the economy. Our results are particularly striking in manufacturing, where we find that plants in the highest quintile of the own-plant-excluded measure of concentration are on average 64% larger than plants in the lowest quintile.

Received for publication May 5, 2000. Revision accepted for publication July 27, 2001.

* University of Minnesota and Federal Reserve Bank of Minneapolis, and the Board of Governors of the Federal Reserve System, respectively.

We are grateful to the editor and the anonymous referees for helpful comments. Holmes acknowledges support from the NSF through grant SES 9906087. Views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

We initially expected that average plant size might turn out to be *smaller* in areas where an industry concentrates, contrary to what we found. There is a large literature that emphasizes the vertical disintegration that occurs in industrial districts (for example, Piore and Sabel, 1984; Holmes, 1999). The emergence of a vast network of intermediate-good suppliers in such an area would appear to increase the incentive to open a small plant that specializes in a narrow product niche. The classic work of Lichtenberg (1960) documents the importance of this phenomenon for the dress industry. In New York City, where the industry is highly concentrated, he found, using data from the 1950s, that plants were small and specialized compared with plants outside New York. It turns out that the dress industry is still heavily concentrated in New York and that plants there continue to be small. However, the dress industry is the rare exception to the general rule that plants tend to be large in areas where an industry concentrates.

One possible explanation for our result is that the four-digit Standard Industrial Classification (SIC) codes (Census Bureau, 1987) we use to define industries may be aggregating fundamentally different subindustries. Within a given four-digit manufacturing industry, we may have one subindustry that includes plants that provide custom-made products and that make minor alterations to goods manufactured elsewhere. The activities of these plants are similar to those of retail and service establishments. These plants may tend to be small and geographically diffuse for the same reason that retail and service establishments are small and geographically diffuse. Within that same four-digit manufacturing industry, there may also be a subindustry of plants that uses mass-production techniques to produce mass-marketed products. These plants may tend to be large and geographically concentrated. Consider the cigarette industry. In our data there are fifteen cigarette plants; eight of them had over 1,000 employees and are located in and around North Carolina. There is a single cigarette manufacturing plant in Kansas, and this plant has less than five employees. We don't know what this Kansas plant did, but we are confident it did not make Marlboro or other mass-marketed brands of cigarettes.

An important implication of this explanation is that when we use SIC codes to define industries, as is the standard practice, we may get a distorted picture of the true level of geographic concentration in an industry. We may very well be interested in the geographic concentration of hard-core manufacturing plants in an industry, and standard measures may tend to understate this concentration. To determine the potential quantitative significance of this issue, we calculated the change in the average Ellison-Glaeser (EG) index of concentration when small plants are excluded. It turns out

that eliminating small plants has a significant effect: an increase in the average EG measure of concentration.

An alternative explanation for our result is that plants located in areas where an industry concentrates may enjoy productivity advantages over plants in other areas. Plants in concentrated areas may grow large to exploit their areas' productivity advantages, advantages that may arise either from natural factors at the locations or from some kind of agglomeration benefits. Thus, our work is related to the large literature that compares the productivity of plants in concentrated areas with the productivity of plants outside these areas (for example, Henderson, 1986; Ciccone and Hall, 1996). Although a plant's productivity is likely to be of more inherent interest than its size, size is likely to be correlated with productivity and is significantly easier to measure. By using an employment-based measure of size, we are able to conduct an analysis of the entire private sector of the economy. No comparable analysis exists that uses productivity instead of size and that looks at the entire private sector.

This paper is related to Florence (1948) and Kim (1995), who both examine, among other things, the relationship *across* industries between average plant size in the industry and the geographic concentration of the industry. The conceptual experiment is different in our analysis, as we are looking *within* industries and comparing the size of plants across locations. Our use of CBP data to make comparisons across locations within an industry is analogous to Glaeser et al. (1992). The difference is that they focus on the growth of industry employment at locations, whereas we focus on plant size.

One recent paper that looks at plant-level size data is Bernard and Jensen (1999). There is nothing about geographic concentration in their paper, but they report an interesting result that is related to the main result here. They found that exporting plants tend to be larger than plants in the same industry that do not export. This finding is consistent with our result, because we may presume that plants in locations where an industry is concentrated are likely to be exporters (at least in the sense of exporting to other regions within the United States). This conclusion follows because these locations will need to trade the products they specialize in for products they do not make.

Another closely related paper is that of Dumais, Ellison, and Glaeser (1997). They use plant-level data from manufacturing to link the dynamics of geographic concentration with plant turnover. Their key finding is that, on the one hand, the birth of new plants tends to disperse employment away from areas where it is already concentrated. On the other hand, the closure of existing plants tends to further concentrate the industry in areas where it is already concentrated. It is well documented that larger plants tend to have lower turnover rates; see for example, Dunne, Roberts, and Samuelson (1989). Thus, Dumais et al.'s findings of reduced turnover in industry agglomerations are consistent

with our findings of larger plant size. They also present evidence that even though the aggregate levels of concentration have remained constant, there is some shift over time in the centers of concentration. The small young plants with high turnover that exist outside centers of concentration may in part serve the function of experimental draws in a search for new centers of concentration.

The rest of the paper is organized as follows. Section II presents the basic framework relating plant scale and industry concentration, and it discusses how we deal with the Ellison-Glaeser "dartboard" issue. Section III describes the CBP data. Sections IV and V present our results.

II. Framework

An often-used measure of industry specialization in a given location is the employment location quotient. This measure compares the relative concentration of industry employment in a given location with the relative concentration of industry employment in the nation. Let $x_{i,l}$ and $n_{i,l}$ denote the employment and number of plants in industry i at location l . Let x_i , n_i , x_l , and n_l denote the corresponding measures aggregated to the industry and location levels, respectively. The standard employment location quotient, $Q_{i,l}^x$, can then be written as the product of two other quotients, a plant quotient and a size quotient:

$$Q_{i,l}^x \equiv Q_{i,l}^n \times Q_{i,l}^s, \quad (1)$$

where

$$Q_{i,l}^x = \frac{x_{i,l}/x_l}{x_i/x},$$

$$Q_{i,l}^n = \frac{n_{i,l}/n_l}{n_i/n},$$

$$Q_{i,l}^s = \frac{x_{i,l}/n_{i,l}}{x_i/n_i}.$$

This decomposition of the standard employment location quotient captures the two possible sources of industry specialization in location l : differences in the number of industry plants per capita from that at the national level ($Q_{i,l}^n$) and differences in average size from the national average ($Q_{i,l}^s$). Taking logs in equation (1) and letting lowercase q 's represent the natural logs of their uppercase counterparts, we have

$$q_{i,l}^x \equiv q_{i,l}^n + q_{i,l}^s. \quad (2)$$

We are interested in the relationship between $q_{i,l}^s$ and $q_{i,l}^x$, that is to say, the relationship between scale and concentration. We can summarize this relationship by looking at

$$\beta^s = \frac{\text{cov}(q_{i,l}^s, q_{i,l}^x)}{\text{var}(q_{i,l}^x)},$$

which is of course the slope coefficient in a regression of q^s on q^x . We will refer to this as a *location-level* regression, as it treats each industry-location pair as an observation. An analogous slope coefficient is

$$\beta^n = \frac{\text{cov}(q_{i,l}^n, q_{i,l}^x)}{\text{var}(q_{i,l}^x)}$$

Given equation (2), we know that $\beta^s + \beta^n = 1$. The magnitude of β^s is of some interest. In a world where the size distribution of plants is independent of concentration, $\beta^s = 0$ and $\beta^n = 1$. All the variation in concentration is accounted for by variation in the number of plants. In another extreme case, all variation is accounted for by variation in size. The measure β^s is an indication of how far we are between these cases. Note that in the Lichtenberg story discussed in the introduction, we expect $\beta^s < 0$ and thus $\beta^n > 1$.

A second method is to treat each establishment as an observation and to ask how an individual plant size varies with concentration. In this case, for plant e we let

$$q_e^x \equiv q^x$$

denote the plant-level employment location quotient (that is, each plant is assigned the employment location quotient corresponding to its location and industry), and we let

$$q_e^s \equiv \ln \left(\frac{x_e}{x_i/n_i} \right)$$

be the corresponding size quotient. Similar to the location-level analysis, the regression coefficient is given by

$$\beta_e^s = \frac{\text{cov}(q_e^s, q_e^x)}{\text{var}(q_e^x)}$$

We will refer to this as a *plant-level* regression. Note that this approach differs from the location-level regression, which is concerned with average plant size at a location rather than individual plant size (compare q^s and q_e^s).

Even if there is some sense that in an underlying distribution we have $\beta^s = 0$ (hence $\beta_e^s = 0$), with a finite number of plants we would expect to find $\beta^s > 0$ and $\beta_e^s > 0$. To see why this assertion is so, suppose a location happens to receive a large plant. This plant will drive up both the average size and the industry concentration at l , with the result that a regression of $q_{i,l}^s$ on $q_{i,l}^x$ will have an estimate of β^s that is greater than zero. Although the OLS estimator is nonetheless consistent, the presence of a bias problem when working with a finite number of plants motivates our use of a location quotient in which the current plant is excluded from the calculation.¹ We call this an

excluded LQ. The intuition for using such a measure is that if a region tends to get large firms in a particular industry, then the location should still tend to have large firms if we ignore the current observation. Therefore, we define \tilde{q}_e^x for plant e to be the log of the plant-level location quotient with plant e excluded from the calculation, and likewise for \tilde{q}_e^s .² Specifically,

$$\tilde{q}_e^x = \ln \left(\frac{x_{i,l} - x_e}{x_l - x_e} \frac{x_i - x_e}{x - x_e} \right)$$

and

$$\tilde{q}_e^s \equiv \ln \left(\frac{x_e}{x_i - x_e} \frac{n_i - 1}{n_i} \right),$$

where it should be clear how plant e 's employment is excluded from all calculations. Let $\tilde{\beta}_e^s$ denote the slope coefficient in a regression of \tilde{q}_e^s on \tilde{q}_e^x . It is worthwhile to note that for a given industry and location, there will be a number of different $(\tilde{q}_e^x, \tilde{q}_e^s)$ pairs—one corresponding to each plant size. This difference is an immediate consequence of excluding individual plants.

We can illustrate the usefulness for the excluded LQ measure with a simple example. Suppose there are N plants in each of I industries, where I is an even number. Let $i \leq I/2$ denote the set of first-half industries, and $i > I/2$ the set of second-half industries. Suppose there are two types of locations, type A and type B , and there are 50 locations of each type. First-half industries tend to be located in type A locations, and second-half industries in type B locations. Specifically, assume that for $i \leq I/2$, the N plants are randomly distributed across locations in an i.i.d. manner, with a probability $2/150$ of landing in any given type A location and a probability $1/150$ of landing in any given type B . For $i > I/2$, the plants are also distributed i.i.d., but the probabilities are reversed. Thus, the expected total number of plants across all industries is the same at each location. But the type A locations have, in expectation, twice as many plants in first-half industries as do type B locations. And type B locations have twice as many plants in second-half industries.

Plants come in two employment sizes, low or high. We assume the high-employment plants have twice the employ-

both augment the vectors and force the recalculation of each element of the vector. Thus, each element of the design matrix is asymptotically approaching its true value as the matrix is being augmented.

² This quotient is not defined when a plant is a singleton, that is, it is the only one in a given industry and location. This factor turns out not to be an issue in our analysis using Census divisions, because there are only 136 singleton plants out of the 5,545,540 in our data, and we eliminate them from the analysis. By comparison, if we use counties instead of Census divisions we have to eliminate 26,563 singleton plants.

¹ In typical asymptotics arguments used to demonstrate consistency, additional observations simply augment the dependent and independent vectors of observations. In this model, however, additional observations

TABLE 1.—RESULTS FOR A SIMULATED MODEL WITH NO FUNDAMENTAL CONNECTION BETWEEN SIZE AND CONCENTRATION

N	Location Level	Plant Level	
	β^s	β_e^s	$\tilde{\beta}_e^s$
100	0.2389	0.2122	0.0022
500	0.0951	0.0823	-0.0021
1,000	0.0611	0.0555	-0.0018
10,000	0.0071	0.0068	-0.0134

ment of low-employment plants. Each plant has a 50% probability of being high employment. Importantly, we assume the size draw is independent of the location draw. Thus, there is no fundamental connection between specialization and plant size in this economy.

We simulated this example, varying the number of plants, N , between 100 and 10,000 and setting the number of industries I to 100. We report the coefficients in table 1. (Note the coefficients change only negligibly if we take different sets of random draws.) We have also experimented with different values for the model's parameters, and the results are qualitatively the same.

The second column of table 1 presents—for each N the estimate of β^s , the coefficient of the location-level regression of $q_{i,l}^s$ on $q_{i,l}^x$. In this example, by construction, the draw of size is independent of location. Yet for $N = 100$ we get a value for β^s of 0.24, an indication of a high degree of positive correlation between average size at a location and concentration at a location. As N is increased, β^s declines to zero, because the sample distribution gets close to the “true” distribution, where the average size is the same at each location.

The third column presents β_e^s , the coefficient of the plant-level regression of q_e^s on q_e^x , where the employment LQ is constructed to *include* the plant's own employment. For $N = 100$, we get an average value of 0.21, which is strictly positive, as in the location-level case. This value is positive for the same reason as in the location-level case. When N is small, a large plant will on average tend to be located at a location with a high LQ, because the large plant will itself be raising the LQ of its location. As N gets large, this effect disappears, and the slope coefficient goes to zero.

Note that the plant-level coefficient β_e^s is strictly less than the location-level coefficient for each N . Since the empirical results below have this same feature, it is useful to explain the discrepancy. If in the location-level regression we were to use the *average* of the *log* size as the left-hand size variable, the coefficient estimate would be identical to β_e^s .³ But we instead use the *log* of the *average* size. Since the log function is concave, the estimated coefficients are different. We have experimented with Monte Carlo simulations of

grouped-data regression models and have found that when the left-hand side is the log of the average and when the error term is positively correlated with the right-side variable (which are both true here), the coefficient estimate from the grouped-data regression tends to be larger than the coefficient from the individual-data regression.

The fourth column of table 1 reports the coefficient $\tilde{\beta}_e^s$ from the excluded regression. Unlike the other coefficients, this coefficient is essentially zero even when N is small. In this model, the plants are distributed i.i.d. across locations independent of size. When a plant's own employment is excluded from the LQ measure, the plant's size is uncorrelated with the measure.

Now consider an alternative version of this dartboard model where plants are not distributed i.i.d. Suppose instead some complementarity exists between large and small firms, so that they are found in fixed proportions across locations; for every large firm there is a corresponding small firm. In this case, we would actually find a negative coefficient in the excluded regression. The neighbors of a large plant would on average be smaller than the neighbors of a small plant.

Below we find a positive coefficient on the excluded regression. This result is inconsistent with the both the i.i.d. case and the fixed-proportions case just mentioned. But it is consistent with a model where a plant's size draw is positively correlated with its relative probability of landing in a location.

III. Data

The CBP data set is unique in its coverage of all private sectors of the economy and its link to location. For each of the six million plants in the United States, we have the primary four-digit industry code, location (state and county), and a size measure (employment). The employment variable is in size categories, 1–4, 5–9, and so on. We handle the employment data by using the mean employment size in each cell category to estimate the employment of each plant. For example, the average employment of plants in the 5–9 category is 6.6, so we treat all plants in this category as though they had 6.6 employees.⁴ At high levels of geographic aggregation the actual data on employment are available for many industries, but we found that for the location-level regressions the results were virtually the same whether or not we used the actual data or the estimated data. Since the employment size classes are narrowly defined, the

³ The location-level case is an example of a grouped-data problem (see Greene, 1997). In general, the coefficient estimate from a grouped-data regression does not equal the coefficient from the individual data regression. But it would in this case, because all the observations in a given group (that is, location) have the same right-side variable.

⁴ In the U.S. totals, plants with over 1,000 employees are aggregated into one size grouping, but this category is broken into four groups in the county-level data. For these four groups, we assumed a log normal distribution and used GMM to estimate the following cell averages: 1,208 (for 1,000–1,499), 1,891 (for 1,500–2,499), 3,367 (for 2,500–4,999), and 9,370 (for over 5,000).

measurement error is relatively small. The year of the data we used is 1992.⁵

We consider two extreme cases for the level of geographic aggregation in our analysis. At one extreme, we look at county-level data, the finest level of geographic detail possible with our data. At the other extreme, we look at region-level data, where the regions combine several states. To define our regions, we use the nine divisions defined by Census.⁶ The use of the Census divisions provides a systematic way of aggregating the counties, and the choice has precedent in Kim (1995). Moreover, the population is nicely distributed across the Census divisions, with the largest division possessing slightly less than three and one-half times the population of the smallest region. For the remainder of the paper, it will be convenient to refer to these geographic units as *regions* rather than use the Census term *divisions*.

We define industries at the finest level of detail that is possible given the CBP data. We used four-digit industries in most cases and three-digit in other cases.⁷ But before presenting our formal analysis of the detailed industry data, it is useful to consider a preliminary analysis in which industries are defined at the major sector level.

Table 2 presents this preliminary analysis. It reports Q^x , Q^n , and Q^s for the major industry sectors and selected regions. For each major sector, except manufacturing, the displayed regions are those for which the sector is most and least concentrated. For manufacturing we report the quotients for each region. For all sectors besides wholesale, retail, and services, there is substantial variation in Q^x across regions. Consider manufacturing. In the East South Central region, $Q^x = 1.40$, meaning manufacturing employment is 40% higher there than it would be if manufacturing were evenly distributed across regions on the basis of total regional employment. In the Mountain region, Q^x is only 0.70. This degree of variation is surprising in that the data in this table were aggregated at both the industry level (to major sectors) and the geographic level (to Census regions). With high industry aggregation we are less likely to find concentration, as diverse industries are treated alike. For example, treating manufacturing alike hides the fact that Georgia specializes in carpets and Washington specializes in airplanes. Geographic aggregation also masks specialization within the region. For example, combining Washington and California into the Pacific division hides the fact that Cali-

⁵ After the project was completed, we obtained the 1997 CBP, and the results are virtually the same with these more recent data. The 1977 data also yield very similar results.

⁶ The nine Census divisions are New England (ME, NH, VT, MA, RI, CT), Middle Atlantic (NJ, NY, PA), East North Central (OH, IN, IL, MI, WI), West North Central (MN, IA, MO, ND, SD, NE, KS), South Atlantic (DE, MD, DC, VA, WV, NC, SC, GA, FL), East South Central (KY, TN, AL, MS), West South Central (AR, LA, OK, TX), Mountain (MT, ID, WY, CO, NM, AZ, UT, NV), Pacific (WA, OR, CA, AK, HI).

⁷ There are 772,150 plants for which we have aggregated SIC code information rather than detailed SIC information. We exclude these, which leaves us with 5,545,540 plants.

TABLE 2.—LOCATION QUOTIENTS BY MAJOR SECTOR

Sector	Region	$Q^x_{i,l}$	$Q^n_{i,l}$	$Q^s_{i,l}$
Agriculture	Pacific	1.37	1.12	1.23
	East South Central	0.77	0.77	0.99
Mining	West South Central	3.74	4.19	0.89
	New England	0.13	0.57	0.22
Construction	Mountain	1.25	1.23	1.02
	New England	0.72	1.03	0.70
Manufacturing	East South Central	1.40	1.02	1.37
	East North Central	1.25	1.07	1.17
	New England	1.08	1.13	0.96
	West North Central	1.01	0.95	1.07
	South Atlantic	0.95	0.85	1.12
	Pacific	0.91	1.17	0.78
	West South Central	0.88	0.90	0.98
	Middle Atlantic	0.88	0.99	0.89
Transportation	Mountain	1.11	1.13	0.98
	New England	0.78	0.86	0.91
Wholesale	West North Central	1.10	1.16	0.94
	New England	0.85	0.83	1.02
Retail	Mountain	1.10	1.07	1.03
	Middle Atlantic	0.88	0.99	0.89
Fire	Middle Atlantic	1.29	1.09	1.18
	East South Central	0.71	0.86	0.82
Services	Middle Atlantic	1.10	0.99	1.11
	East South Central	0.86	0.90	0.95

fornia specializes in food processing, while Washington specializes in airplanes.

Table 2 is a preview of the next section's main result. The table reveals a positive relationship between plant scale and industry concentration. Consider manufacturing first. For the East South Central region, the plant quotient is 1.02 and the size quotient is 1.37. Thus, the extra 40% in total employment in manufacturing in the East South Central region is due to its having plants that are on average 37% larger than the U.S. average, not by having more plants than a typical region. Analogously, the Mountain region has a relatively small share of manufacturing employment, not because the region has relatively few plants, but because the plants it has tend to be small. For the other sectors, the size difference between the leading region and the lagging region is not as sharp as it is with manufacturing. Nevertheless, with the exception of wholesale, plants in the leading region are bigger on average than plants in the lagging region. That this comparison holds in eight out of nine cases is striking.

IV. Analysis

To explore this relationship further, we compute the slope coefficients corresponding to the framework described in section II. Table 3 presents the results of both the location- and plant-level regressions using four-digit industries and Census regions. The first row shows the results for all industries, and the remaining rows show the results using

TABLE 3.—REGRESSION ESTIMATES FOR CENSUS REGIONS

Sector	Location Level	Plant Level	
	β^s	β_e^s	$\tilde{\beta}_e^s$
All	0.322 (0.00576)	0.183 (0.00153)	0.161 (0.00152)
Agriculture	0.313 (0.0903)	0.141 (0.0119)	0.128 (0.0119)
Mining	0.193 (0.0162)	0.0304 (0.00613)	0.0240 (0.00608)
Construction	0.410 (0.0429)	0.225 (0.00668)	0.213 (0.00668)
Manufacturing	0.436 (0.00674)	0.334 (0.00377)	0.274 (0.00370)
Transportation	0.316 (0.0314)	0.183 (0.00792)	0.151 (0.00785)
Wholesale	0.135 (0.0154)	0.0907 (0.00478)	0.0747 (0.00477)
Retail	0.224 (0.0245)	0.228 (0.00446)	0.217 (0.00446)
FIRE	0.398 (0.0308)	0.265 (0.00488)	0.253 (0.00487)
Services	0.359 (0.0153)	0.137 (0.00312)	0.121 (0.00311)

Standard errors shown in parentheses.

only the four-digit industries within each of the major sectors. The first column is an estimate of β^s for the location-level data when weighting by the number of plants in each industry. The estimates of the location-level coefficient β^s are large across all the sectors. The second is the coefficient from the plant-level regression that does not exclude a plant's own employment in the calculation of the employment LQ. These estimates are less than the location-level estimates, as happened in the simulation model in table 1 and for the same reasons.

Notice that the estimated coefficients in the location-level regressions are quite precise. For the case of all industries (row 1), there are 7,875 observations (nine regions times 875 industries); other major industry groups will have less observations. The establishment-level regressions have far more observations, as each establishment is treated as a single observation. In the nonexcluded plant regression the right-side variable (relative concentration) is the same for all plants in the same region. Although the left-side values will still differ due to different draws of the error term, there is a sense in which the common right-side values inflate precision (the sense in which each industry-region pair is really a single observation).⁸ This is not a problem for our results, because precision in the plant-level regressions is not an important issue, given the precision of our location-level results, which do treat each industry-region pair as a single observation.

The third column is the plant-level coefficient using the excluded LQ. As we predicted, the excluded coefficients are all lower than the nonexcluded coefficients. For example, in

⁸ This downward bias to the standard errors is a well-known problem associated with all regressions in which right-side variables are common within groups. See Moulton (1986) for a discussion of this issue.

TABLE 4.—REGRESSION ESTIMATES FOR COUNTIES

Sector	Location Level	Plant Level	
	β^s	β_e^s	$\tilde{\beta}_e^s$
All	0.365 (0.000816)	0.190 (0.000603)	0.073 (0.000572)
Agriculture	0.212 (0.00593)	0.083 (0.00367)	0.006 (0.00360)
Mining	0.232 (0.00474)	0.141 (0.00407)	0.087 (0.00397)
Construction	0.402 (0.00394)	0.221 (0.00235)	0.063 (0.00222)
Manufacturing	0.415 (0.00145)	0.274 (0.00165)	0.126 (0.00159)
Transportation	0.373 (0.00406)	0.208 (0.00281)	0.070 (0.00265)
Wholesale	0.338 (0.00196)	0.182 (0.00177)	0.048 (0.00167)
Retail	0.262 (0.00221)	0.200 (0.00169)	0.049 (0.00161)
FIRE	0.518 (0.00314)	0.240 (0.00191)	0.152 (0.00180)
Services	0.463 (0.00156)	0.212 (0.00115)	0.071 (0.00107)

Standard errors shown in parentheses.

manufacturing, the coefficient falls from 0.334 to 0.274. Nevertheless, the coefficients with the excluded variable do not fall that much; they retain at least 80% of the original coefficient with the nonexcluded variable. This result is very different than what happens with the numerical example in table 1, where the coefficients fell to zero. We conclude that only a small part of the connection between concentration and plant size in the location-level analysis in column 2 is being driven by dartboard factors.

For comparison, tables 4 and 5 show the same results using counties and metropolitan statistical areas (MSAs). Comparing the county and Census-region estimates, we observe that the location-level estimates and most of the

TABLE 5.—REGRESSION ESTIMATES FOR THE TEN LARGEST MSAs

Sector	Location Level	Plant Level	
	β^s	β_e^s	$\tilde{\beta}_e^s$
All	0.402 (0.00580)	0.197 (0.00229)	0.149 (0.00225)
Agriculture	0.833 (0.102)	0.457 (0.0241)	0.394 (0.0234)
Mining	0.261 (0.0198)	0.0427 (0.0151)	0.0205 (0.0148)
Construction	0.542 (0.0428)	0.304 (0.0104)	0.271 (0.0104)
Manufacturing	0.443 (0.00604)	0.288 (0.00539)	0.181 (0.00513)
Transportation	0.452 (0.0282)	0.217 (0.0130)	0.136 (0.0126)
Wholesale	0.273 (0.0156)	0.0900 (0.00719)	0.0491 (0.00715)
Retail	0.479 (0.0279)	0.447 (0.00765)	0.409 (0.00762)
FIRE	0.365 (0.0378)	0.172 (0.00690)	0.144 (0.00685)
Services	0.409 (0.0128)	0.155 (0.00398)	0.127 (0.00397)

Standard errors shown in parentheses.

non-excluded plant-level estimates are larger using the county data than they are with the Census-region data. This difference is not surprising, as greater geographic disaggregation implies that a single large establishment will have a greater relative impact on measures of size and concentration. This hypothesis is confirmed by the county-level excluded estimates which, with the exception of mining, are uniformly lower than their region counterparts. Since counties and Census divisions are somewhat arbitrary geographic units, it might be expected that the results would differ when functional geographic units like MSAs are used. This is not what we find. Using only the ten most populated MSAs, we observe the same patterns as before.⁹ The location-level estimates are the largest, followed by the nonexcluded and excluded plant-level estimates. The excluded plant-level estimates remain large, even though they are smaller than their nonexcluded counterparts. Thus, our results appear to be quite robust to the way we define geographic units.

Of the major sectors, manufacturing tends to exhibit the strongest positive relationship between size and specialization. Table 6 explores this further by looking at the raw data for a random set of industries. We choose the four-digit industry in each two-digit manufacturing industry with the lowest SIC code. For each of the 20 displayed industries, the employment and size quotients are reported for three Census regions. The three regions are those with the largest and smallest values for Q^x as well as the region with Q^x closest to one. Notice the pattern between the employment and size quotients. In all but four of the displayed industries, the size ordering is the same as the employment ordering. The difference between the middle and bottom rows is of a much larger magnitude than the difference between the top and middle rows. The size quotient in the bottom row of the least concentrated regions is remarkably small; it is less than 0.20 in more than half the cases. These tiny manufacturing plants in these least concentrated regions could very well be performing retail and service functions, as suggested in the introduction.

As an aside, we have also looked at the raw data for SIC 2335, "Women's and Misses' Dresses." This is the industry discussed in the classic Lichtenberg (1960) study. The industry continues to be concentrated in the New York City area, just as it was in the earlier study (the metro area contains 30% of all U.S. employment). The concentration index for the Middle Atlantic region (which contains New York) was $Q_x = 2.70$, the highest concentration index of all the regions. The size index for this region is only 0.90, so plants in this region tend to be smaller than the average U.S. plant (and if we look just at the New York MSA, it is even

TABLE 6.—SELECTED MANUFACTURING LOCATION QUOTIENTS

Four-Digit Industry		Region	$Q_{i,l}^x$	$Q_{i,l}^s$
2011	Meat-packing plants	West North Central	5.73	2.28
		East South Central	1.01	0.78
		New England	0.04	0.14
2111	Cigarettes	South Atlantic	4.90	1.17
		Middle Atlantic	0.01	0.02
		West North Central	0.001	0.001
2211	Broadwoven fabric mills, cotton	South Atlantic	4.47	1.96
		West South Central	0.41	0.45
		East North Central	0.02	0.05
2311	Men's and boys' suits and coats	East South Central	2.56	1.55
		West South Central	0.85	1.06
		Mountain	0.11	0.31
2411	Logging	East South Central	2.64	0.97
		New England	0.99	1.02
		Middle Atlantic	0.17	0.64
2511	Wood household furniture	South Atlantic	2.74	2.10
		East North Central	0.68	0.96
		West South Central	0.34	0.58
2611	Pulp mills	East South Central	4.55	1.07
		New England	0.90	1.17
		West North Central	0.01	0.04
2711	Newspapers	New England	1.24	1.27
		South Atlantic	1.00	1.20
		East South Central	0.78	0.70
2812	Alkalis and chlorine	West South Central	3.61	1.60
		Pacific	0.76	0.78
		West North Central	0.10	0.14
2911	Petroleum refining	West South Central	4.49	1.36
		East North Central	0.88	1.22
		New England	0.005	0.03
3011	Tires and inner tubes	East South Central	5.42	1.91
		West North Central	0.93	0.79
		Mountain	0.005	0.01
3111	Leather tanning and finishing	New England	2.28	0.81
		South Atlantic	0.92	1.83
		Mountain	0.04	0.05
3211	Flat glass	East South Central	4.08	2.28
		South Atlantic	0.98	0.87
		New England	0.06	0.20
3312	Blast furnaces and steel mills	East North Central	2.71	1.74
		East South Central	1.07	0.76
		New England	0.04	0.11
3411	Metal cans	East North Central	1.43	1.17
		South Atlantic	1.00	1.10
		New England	0.18	0.39
3511	Turbines and turbine generator sets	New England	3.40	0.92
		Pacific	0.88	1.19
		Mountain	0.001	0.005
3612	Transformers, except electronic	East South Central	2.70	2.17
		West North Central	0.90	0.87
		New England	0.20	0.27
3711	Motor vehicles and car bodies	East North Central	3.39	1.84
		West North Central	1.06	0.93
		New England	0.01	0.02
3812	Search and navigation equipment	Pacific	1.89	1.11
		West North Central	1.01	0.60
		East South Central	0.15	0.45
3911	Jewelry, precious metal	New England	3.32	1.47
		West North Central	0.76	1.67
		East South Central	0.10	0.58

⁹ The 10 most populous MSAs in 1992 were New York–Northern New Jersey–Long Island, Los Angeles–Anaheim–Riverside, Chicago–Gary–Lake County, San Francisco–Oakland–San Jose, Philadelphia–Wilmington–Trenton, Detroit–Ann Arbor, Dallas–Fort Worth, Washington, Houston–Galveston–Brazoria, and Boston–Lawrence–Salem.

TABLE 7.—RELATIVE SIZE BY DEGREE OF CONCENTRATION

Sector	Average \bar{Q}_e^s for \bar{Q}_e^x in:						
	Quintile 1	2	3	4	5	Bottom 5%	Top 5%
Census Regions							
All	0.91	0.99	1.00	1.02	1.09	0.83	1.22
Agriculture	0.94	1.03	1.02	0.93	1.14	1.12	0.59
Mining	0.76	1.04	1.07	1.38	0.81	0.62	0.66
Construction	0.89	0.98	1.02	1.10	1.01	0.84	1.19
Manufacturing	0.76	0.91	1.01	1.11	1.25	0.67	1.35
Transportation	0.93	1.10	0.97	1.11	0.91	0.93	1.22
Wholesale	0.95	0.97	1.04	1.00	1.04	0.96	1.06
Retail	0.93	1.00	1.03	1.02	1.03	0.91	1.10
FIRE	0.86	1.00	0.99	1.03	1.13	0.81	1.29
Services	0.95	0.96	0.99	1.01	1.10	0.85	1.20
Counties							
All	0.86	1.00	1.04	1.03	1.12	0.77	1.30
Agriculture	0.91	0.99	0.98	0.99	1.16	0.81	1.39
Mining	0.71	0.83	0.73	0.87	2.10	0.62	3.34
Construction	0.90	1.01	1.03	1.04	1.05	0.85	1.11
Manufacturing	0.83	0.87	0.96	1.04	1.31	0.78	1.52
Transportation	0.83	1.06	1.02	1.10	1.14	0.81	1.20
Wholesale	0.88	1.00	1.06	1.06	1.09	0.83	1.08
Retail	0.93	1.03	1.07	1.04	0.96	0.84	0.95
FIRE	0.68	0.91	1.09	1.11	1.29	0.53	1.54
Services	0.86	0.99	1.02	1.03	1.15	0.77	1.39

smaller, at 0.77). Thus, the dress industry is an exception to the general rule that plants are larger in areas where an industry concentrates.

Since the relationship between size and concentration in table 6 seems stronger when we go from the bottom range of concentration to the middle than it is when we go from the middle range to the top, it is useful to look at the data in a way that does not impose a linear relationship. Table 7 presents a cross-tabulation of the mean \bar{Q}_e^s across plants within each quintile of \bar{Q}_e^x and for the top and bottom 5% of \bar{Q}_e^x . (Recall that \bar{Q}_e^x is the excluded LQ employment measure for plant e and \bar{Q}_e^s is plant size divided by the industry's mean plant size excluding the current plant.) With few exceptions, these numbers indicate a strong positive relationship between concentration and relative size. This relationship is especially pronounced for manufacturing and services, where relative size steadily in-

creases with concentration. If we look at all industries together, we see that the average difference in size between the top 5% and the bottom 5% is 47% for regions (1.22 versus 0.83) and 67% for counties (1.30 versus 0.77). If we look only at manufacturing, the difference between these groups is a factor of two for both regions and counties.

V. Implications for Measures of Concentration

Measuring the extent of geographic concentration is of fundamental interest and has received substantial attention lately (Ellison and Glaeser, 1997). One explanation for our findings is that our SIC industry definitions are poorly measuring what plants are doing. A four-digit industry might include small, geographically diffuse plants that per-

TABLE 8.—ELLISON-GLAESER INDICES FOR FOUR-DIGIT MANUFACTURING

Areas	Statistic	EG Index by Plant Employment				
		0+	20+	50+	100+	250+
Census regions	Mean	0.087	0.091	0.094	0.102	0.120
	25%	0.013	0.014	0.014	0.014	0.012
	Median	0.043	0.044	0.048	0.053	0.066
	75%	0.100	0.108	0.124	0.134	0.174
States	Mean	0.045	0.047	0.048	0.051	0.058
	25%	0.008	0.008	0.009	0.009	0.005
	Median	0.022	0.025	0.025	0.027	0.031
	75%	0.053	0.056	0.061	0.066	0.085
Counties	Mean	0.011	0.011	0.011	0.012	0.012
	25%	0.002	0.002	0.002	0.002	0.001
	Median	0.005	0.005	0.005	0.005	0.004
	75%	0.010	0.010	0.011	0.012	0.012
No. of Industries		459	459	458	457	428
Share of emp. (%)		100	93.4	84.8	74.9	57.4

form a retail or service function, as well as large, geographically concentrated plants that do hard-core manufacturing.

To evaluate the potential quantitative significance of this possibility, we calculate in table 8 the EG index for manufacturing, using only plants that meet various size criteria. The first column, labeled "0+," includes all manufacturing plants; the second column, labeled "20+," includes only manufacturing plants with 20 or more employees; and so on.

The table reports the index calculated for regions, states, and counties. We include states here for comparison with Ellison and Glaeser's original analysis, since their paper primarily focused on state-level geographic definitions. The "0+" column (all plants) uses the same underlying universe of plants as the EG analysis. Our estimates are similar to the original EG estimates, despite the fact that our data come from a different year (1992 instead of 1987), we use different data (County Business Patterns versus the Census of Manufactures), and we employ a different strategy for dealing with Census disclosure problems (we use plant-level cell counts in the CBP; they use state-level data and a complicated imputation procedure). With the state-level data, we find a mean and a median EG index of 0.045 and 0.022, which are slightly less than the corresponding estimates of 0.051 and 0.026 reported by EG. Just as they do, we find that concentration is very minimal at the county level. The median for the county data is only 0.005, which is the same number that Ellison and Glaeser get at the county level with three-digit industries.

Table 8 shows a clear pattern for the region-level and state-level cases. When small plants are thrown out of the data set, there is a substantial increase in the EG index.¹⁰ If we looked instead at a raw measure of concentration rather than the EG index, then of course this increase would happen, because dartboard issues become important when there are relatively few plants. But the EG methodology explicitly takes into account the dartboard issues; there is no a priori reason to expect that their index should increase when small establishments are excluded.

Note that the biggest changes are at the top of the distribution. The 25th percentile changes hardly at all; in the region case it goes from 0.013 with all plants to 0.012 when plants with less than 250 are excluded. But the 75th percentile increases dramatically, going from 0.100 to 0.174 over the same range. We find these results to be very interesting, because they suggest that the distribution of hard-core manufacturing is more geographically concentrated than previously thought.

¹⁰ As the cutoff is raised, some industries start to drop out of the sample (an industry needs two plants above the size threshold to be in). With the 250+ cutoff, 428 out of 459 industries are still in. Recalculating table 8 using only those 428 industries does not alter the observed pattern.

It is also noteworthy that the pattern continues to hold when plants in the 100–249 category are removed. Presumably, plants in this size range do more than provide retail and service functions. So our first proposed explanation for our results, which argues that plants with different functions are being combined in the same industries, is not likely to be the whole story. Our second explanation, that plants in concentrated areas expand to exploit productivity advantages, may also be a factor.

For the county-level data, there is nothing going on as we move across a row and remove small plants. But there is little going on at the starting point where all the small plants are included. The EG index is extremely small here, just as Ellison and Glaeser found in their original paper. Evidently, geographic connections in manufacturing extend beyond the county level.

REFERENCES

- Bernard, Andrew B., and J. Bradford Jensen, "Exporters, Jobs, and Wages in U.S. Manufacturing, 1976–1987," *Brookings Papers on Economic Activity, Microeconomics*, Washington DC (1995).
- "Exceptional Exporter Performance: Cause, Effect, or Both?" *Journal of International Economics* 47:1 (1999), 1–25.
- Ciccone, Antonio, and Robert E. Hall, "Productivity and the Density of Economic Activity," *American Economic Review* 86:1 (1996), 54–70.
- Dumais, Guy, Glenn Ellison, and Edward L. Glaeser, "Geographic Concentration as a Dynamic Process," NBER working paper no. 6270 (1997).
- Dunne, Timothy, Mark J. Roberts, and Larry Samuelson, "The Growth and Failure of U.S. Manufacturing Plants," *Quarterly Journal of Economics* 104:4 (1989), 671–698.
- Ellison, Glenn, and Edward Glaeser, "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy* 105:5 (1997), 889–927.
- Florence, P. Sargant, *Investment, Location, and Size of Plant: A Realistic Inquiry into the Structure of British and American Industries* (Cambridge, U.K.: University Press, 1948).
- Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman, and Andrei Shleifer, "Growth in Cities," *Journal of Political Economy* 100:6 (1992), 1126–1152.
- Greene, William H., *Econometric Analysis*, 3d ed. (New York: Macmillan Publishing Company, 1997).
- Henderson, J. Vernon, "Efficiency of Resource Use and City Size," *Journal of Urban Economics* 19:1 (1986), 47–70.
- Holmes, Thomas, "Localization of Industry and Vertical Disintegration," *The Review of Economics and Statistics* 81:2 (1999), 314–325.
- Kim, Sukkoo, "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U.S. Regional Manufacturing Structure, 1860–1987," *Quarterly Journal of Economics* 110:4 (1995), 881–908.
- Lichtenberg, Robert M., *One-Tenth of a Nation: National Forces in the Economic Growth of the New York Region* (Cambridge, MA: Harvard University Press, 1960).
- Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32:3 (1986), 385–397.
- Piore, Michael J., and Charles F. Sabel, *The Second Industrial Divide* (New York: Basic Books, Inc., 1984).
- U.S. Bureau of Census, *Standard Industrial Classification Manual* (Washington DC: Office of Management and Budget, 1987).