WORKING PAPER 16-25

LOCALIZED KNOWLEDGE SPILLOVERS: EVIDENCE FROM THE AGGLOMERATION OF AMERICAN R&D LABS AND PATENT DATA

Kristy Buzard
Syracuse University

Gerald A. Carlino
Research Department, Federal Reserve Bank of Philadelphia

Robert M. Hunt
Payment Cards Center, Federal Reserve Bank of Philadelphia

Jake K. Carr
The Ohio State University

Tony E. Smith
University of Pennsylvania

October 2016

RESEARCH DEPARTMENT, FEDERAL RESERVE BANK OF PHILADELPHIA
Ten Independence Mall, Philadelphia, PA 19106-1574 • www.philadelphiafed.org/research-and-data/

Localized Knowledge Spillovers:
Evidence from the Agglomeration of American R&D Labs and Patent Data*

Kristy Buzard
Syracuse University

Gerald A. Carlino and Robert M. Hunt
Federal Reserve Bank of Philadelphia

Jake K. Carr
The Ohio State University

Tony E. Smith
University of Pennsylvania

September 2016

Abstract

We employ a unique data set to examine the spatial clustering of private R&D labs. Instead of using fixed spatial boundaries, we develop a new procedure for identifying the location and size of specific R&D clusters. Thus, we are better able to identify the spatial locations of clusters at various scales, such as a half mile, 1 mile, 5 miles, and more. Assigning patents and citations to these clusters, we capture the geographic extent of knowledge spillovers within them. Our tests show that the localization of knowledge spillovers, as measured via patent citations, is strongest at small spatial scales and diminishes rapidly with distance.

Keywords: spatial clustering, geographic concentration, R&D labs, localized knowledge spillovers, patent citations

JEL Codes: O31, R12

* We thank Kristian Behrens, Jim Bessen, Satyajit Chatterjee, Gilles Duranton, Vernon Henderson, Andy Haughwout, Jim Hirabayashi, Tom Holmes, Mark Schweitzer, Will Strange, Isabel Tecu, and Elisabet Viladecans-Marsal for comments and suggestions. This paper has benefited from the contribution of outstanding research assistance by Cristine McCollum, Adam Scavette, Elif Sen, and Annette Swahala. The views expressed here are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at www.philadelphiafed.org/research-and-data/publications/working-papers/.

1. INTRODUCTION

Popular accounts suggest that research and development (R&D) facilities are highly spatially concentrated into comparatively few geographic locations such as Silicon Valley and the Route 128 Corridor outside Boston. That R&D labs are geographically concentrated is immediately evident from examining a national map of the locations of private R&D establishments (Figure 1). What is not immediately clear from the map is whether the spatial concentration of R&D is significantly greater than economic activity in general. The primary purpose of the research addressed in this paper is whether the spatial pattern of R&D laboratories observed in Figure 1 is somehow unusual; that is, is it different from what we would expect based on the spatial concentration of the economic activity? We answer this question by using a new location-based data set of private R&D labs to document and analyze patterns in the geographic concentration of U.S. R&D labs.

Rather than using fixed geographic units, such as counties or metropolitan areas, we use continuous measures to identify the spatial structure of the concentrations of R&D labs. Specifically, we use point pattern methods to analyze locational patterns over a range of selected spatial scales (within a half mile, 1 mile, 5 miles, etc.). This approach allows us to consider the spatial extent of the agglomeration of R&D labs and to measure any attenuation of clustering with distance more accurately.[1]

Following Duranton and Overman (2005) — hereafter DO — we look for geographic clusters of labs that represent statistically significant departures from spatial randomness using simulation

[1] Other studies that have used continuous measures of concentration include Marcon and Puech (2003) for French manufacturing firms; Arbia, Espa, and Quah (2008) for patents in Italy; and Kerr and Kominers (2015) in a more general model, one application of which uses data on patent citations. See Carlino and Kerr (2015) for a recent review of this literature.

techniques. We do not assume that "randomness" implies a uniform distribution of R&D activity. Rather, we focus on statistically significant departures of R&D labs at each spatial scale from the distribution of an appropriately defined measure of economic activity at that scale. This is important because studies have shown that manufacturing activity is agglomerated at various spatial scales (e.g., Ellison and Glaeser (1997); Rosenthal and Strange, 2001; and Ellison, Glaeser, and Kerr, 2010) and the large majority of R&D activity is performed by manufacturing firms. Our main results take manufacturing employment as the benchmark, but our findings are robust to alternative benchmarks such as manufacturing establishments and science, technology, engineering, and math (STEM) workers.

While this multiple-scale approach is similar in spirit to that of DO, our test statistics are based on Ripley's *K*-function rather than the "*K*-density" approach of DO.[2] A significant advantage of *K*-functions of which we take advantage is that they can easily be disaggregated to yield information about the *spatial locations* of clusters of R&D labs at various scales.

We begin the analysis by using global *K*-function statistics to test for the presence of significant clustering over a range of spatial scales. We find strong evidence of spatial clustering at even very small spatial scales — distances as small as one-half mile. Clustering exists at these and much larger spatial scales.

Next, we focus on the question of *where* clustering occurs using a more refined procedure based on local *K*-functions. We introduce a novel procedure called the multiscale core-cluster approach to identify the location of clusters and the number of labs in these clusters. Core clusters at each scale are identified in terms of those points with the most significant local clustering at that scale.

[2] The simulation procedure we use to construct the distribution of counterfactual *K*-functions takes edge effects into account since the same edge effects are present in all counterfactuals.

By construction, core clusters at smaller scales tend to be nested in those at larger scales. Such core clusters generate a hierarchy that reveals the relative concentrations of R&D labs over a range of spatial scales. In particular, at scales of 5 and 10 miles, these core clusters reveal the presence of the major agglomerations visible on any map.

A secondary purpose of this article is to show that the local R&D clusters we identify are economically meaningful. In this part of our analysis, we document that patent citations are more highly geographically localized within the clusters of R&D labs we identify than outside them. To do this, we construct treatment versus control tests for the localization of patent citations in the spirit of those found in Jaffe, Trajtenberg, and Henderson (1993), hereafter, JTH. For labs in the Northeast Corridor, we find that citations are on average about three to six times more likely to come from the same cluster as earlier patents than one would predict using a (control) sample of otherwise similar patents. For California, citations are on average roughly three to five times more likely to come from the same cluster as earlier patents than one would predict using the control sample. Our results are robust to drawing the controls more narrowly from patents that share the same patent class and subclass as the citing patents.[3] Finally, we show that our results persist when we use an alternative method to select the controls (Coarsened, Exact Matching) although the tests for the localization of patent citations are at the lower end of our findings, particularly in California.

Thus, using samples of patents 15 to 20 years after those used by JTH — and after the Internet significantly reduced the cost of searching for prior art located anywhere — we confirm their

[3] As a robustness check, we follow Thompson and Fox-Kean (2005) — hereafter TFK — and substitute six-digit technological categories for the three-digit patent class we use to identify controls in our main analysis. The results are found to be highly robust with respect to such controls, suggesting that they are not solely a consequence of technical aggregation.

main result. Moreover, we find that patents inside each cluster receive more citations on average than those outside the cluster in a suitably defined counterfactual area. This suggests that the geography and scale of the clusters we identify is related to the extent of localization of knowledge spillovers, at least as evidenced by patent citations. Moreover, our tests reveal clear evidence of the attenuation of the localization effect as distance increases. In other words, the localization of knowledge spillovers appears strongest at small spatial scales (5 miles or less) and diminishes rapidly with distance.

2. THEORY AND DATA

We introduce a novel data set in this paper, based on the 1998 vintage of the *Directory of American Research and Technology*, which profiles the R&D activities of public and private enterprises in the United States. The directory includes virtually all nongovernment facilities engaged in any commercially applicable basic and applied research. For this paper, our data set contains the R&D establishments ("labs") associated with the top 1,000 publicly traded firms ranked in terms of research and development expenditure in Compustat.[4] These firms represent slightly less than 95 percent of all R&D expenditures reported in the 1999 vintage of Compustat for 1998.[5] Thus, each lab in our data set is associated with its Compustat parent firm and information on its street address and a text description of its research specialization(s) to which

[4] We referenced several additional sources both to cross-check the information provided by this directory and to supplement it when we could not locate an entry for a Compustat listing. Dalton and Serapio (1995) provide a list of locations of U.S. labs of foreign-headquartered firms. In some cases, we found information about the location of a firm's laboratories in the "Research and Development" section of the firm's 10-K filings with the Securities and Exchange Commission. The following company databases were also used to supplement or confirm our main sources: Hoover's Company Records database, Mergent Online, the Harris Selectory Online Database, and the American Business Directory.

[5] Although we cannot know for sure the impact on the analysis of including smaller labs, if these labs tend to cluster near larger labs as is widely believed, then we will underestimate the significance of the labs in our data set. Some clusters that fail our tests of significance may indeed be significantly clustered in that case as well, and some cluster boundaries may be slightly different than what we identify. Our results on patent citation differentials will not be impacted, as these rely on the universe of patents, not only those of the firms who have labs in our data set.

we have assigned the corresponding four-digit Standard Industrial Classification (SIC) codes. Using the address information for each private R&D establishment, we geocoded the locations of more than 3,000 labs (shown in Figure 1).

In this paper, we analyze two major regions of the U.S.: the Northeast Corridor and the state of California. There are 1,035 R&D labs in 10 states comprising the Northeast Corridor of the United States (Connecticut, Delaware, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Virginia, including the District of Columbia — the Washington, D.C., cluster). There are 645 R&D labs in California.

Even at the most aggregate level, it is easy to establish that R&D activity is relatively concentrated in these two regions. For example, in 1998, one-third of private R&D labs (and 29 percent of private R&D expenditures) were located within the Northeast Corridor, compared with 22 percent of total employment (21 percent of manufacturing employment) and 23 percent of the population. California accounted for almost 22 percent of all private R&D labs (and 22 percent of private R&D expenditures) in 1998 compared with 12 percent of total employment (11 percent of manufacturing employment) and 12 percent of the population. Together, these two regions accounted for the majority of all U.S. private R&D labs (and private R&D expenditures) in 1998.[6] This concentration is consistent with Audretsch and Feldman (1996), who report that the top four states in terms of innovation in their data are California, Massachusetts, New Jersey, and New York.

In our formal analysis below, we assess the concentration of R&D establishments relative to a baseline of economic activity as reflected by the amount of manufacturing employment in the zip

[6] Data for private R&D expenditures are from Table A.39 of National Science Foundation (2000).

code. These data were obtained from the 1998 vintage of Zip Code Business Patterns. Given that the vast majority of our R&D labs are owned by manufacturing firms, manufacturing employment represents a good benchmark. It is possible that owners of R&D labs locate these facilities using different factors than they use for locating manufacturing establishments. We address this concern by using total employment data at the census block level for 2002 from the Longitudinal Employer-Household Dynamics (LEHD) survey to identify feasible lab locations within each zip code.[7]

Table 1 presents summary statistics for zip codes in the Northeast Corridor and in California for 1998. The average zip code in the Northeast Corridor had about 29 square miles of land area with a radius of about 2.5 miles in 1998. Since there were approximately 6,044 zip codes in the Northeast Corridor in 1998, there is, on average, one R&D facility for every six zip codes in this part of the country. The average zip code in the Northeast Corridor had about 4,300 jobs in 1998, 13 percent of which were in manufacturing. In California, the average zip code consisted of about 96 square miles of land area with an average radius of slightly less than 4 miles. The average zip code in California had almost 6,000 jobs in 1998, 14 percent of which were in manufacturing. Table 1 also provides descriptive statistics for those zip codes containing one or more R&D labs. These zip codes are physically smaller (with a radius of about 2 miles in each region) and contain three to four times more employment.

2.1 Theory

How do we account for the geographic concentration of R&D activity observed in this paper? Much of the theoretical literature on urban agglomeration economies has focused on externalities

[7] In Appendix A, we report results of our analyses using manufacturing establishments as an alternative benchmark.

in the production of goods and services rather than on invention itself. Nevertheless, the three formal mechanisms primarily explored in the literature — sharing, matching, and knowledge spillovers — are also relevant for innovative activity.

2.1.1 Knowledge Spillovers:

Spatial concentration of economic activity facilitates the spread of tacit knowledge. More than most types of economic activity, R&D depends on knowledge spillovers. A high geographic concentration of R&D labs creates an environment in which ideas move quickly from person to person and from lab to lab. Locations that are dense in R&D activity encourage knowledge spillovers, thus facilitating the exchange of ideas that underlies the creation of new goods and new ways of producing existing goods.

2.1.2 Sharing and Matching

Thick factor markets can arise when innovative activity clusters locally. These clusters allow each of their members to benefit as if they had greater scale through: The development of pools of specialized workers — such as of STEM workers; and greater variety of specialized business services, such as patent attorneys, commercial labs for product testing, and access to venture capital. As Helsley and Strange (2002) have shown, dense networks of input suppliers facilitate innovation by lowering the cost needed to bring new ideas to fruition. Thick labor markets also can improve the quality of matches in local labor markets (Berliant, Reed, and Wang 2006; Hunt 2007). Also, specialized workers can readily find new positions without having to change locations (job hopping).

2.1.3 Connection Between Theory and Evidence

We impose statistical requirements on our tests for localization to determine whether R&D labs are clustered. This approach is based on a test of a simple location model (i.e., R&D locations are more clustered than would be expected from random draws from the distribution of overall manufacturing employment).

In Section 6, we provide evidence that the clustering of R&D labs is related to knowledge spillovers by studying the relative geographic concentration of citations to patents originating in the clusters we identify. It's possible that technologically related activities may cluster to benefit from agglomeration forces other than knowledge spillovers (such as sharing and better matching of workers and firms). These other sources of agglomeration could potentially explain some of the geographic concentration of technologically related research activity. To address this issue, our basic approach (JTH's approach) is to construct a control sample of patents that have the same technological and temporal distribution as the citations to account for these other agglomeration forces. Our test for knowledge spillovers is whether the citation matching frequency is significantly greater than the control matching frequency. Put differently, our test is whether citations are more localized relative to what would be expected given the existing distribution of technological related activity.[8]

3. GLOBAL CLUSTER ANALYSIS

A key question is whether the overall patterns of R&D locations in the two regions we examine exhibit more clustering than would be expected from the spatial concentration of manufacturing in those regions. However, since we are interested in possible clustering of R&D labs at scales

[8] In Section 6.5.3, we develop an alternative benchmark or backcloth for analyzing R&D clustering with respect to STEM workers to address the concern that we may be mingling knowledge spillovers with labor market pooling. As we will see, our main findings are highly robust to the use of alternative backcloths.

below the average sizes of zip codes, it is necessary to refine this hypothesis. To address this question statistically, we start with the null hypothesis that R&D locations are mainly determined by the distribution of manufacturing employment.

We obtained total employment data at the census block level for 2002 from the LEHD survey[9] and used this to identify feasible lab locations within each zip code area.[10] Blocks with zero employment are clearly infeasible (such as public areas and residential zones), and blocks with higher levels of total employment are hypothesized to offer more location opportunities. It is also implicitly hypothesized that accessibility to manufacturing within a given zip code area is essentially the same at all locations within that zip code. So, even in blocks where there is no manufacturing, locations are regarded as feasible as long as there is some type of employment present.[11]

In summary, our basic null hypothesis, H_0, is that *lab locations are influenced by the distribution of manufacturing employment at the zip code level and by the distribution of total employment within each zip code area.*

Locations consistent with H_0 are then generated by a three-stage Monte Carlo procedure in which (i) zip code locations are randomly selected in proportion to manufacturing employment levels, (ii) census block locations within these zip codes are selected in proportion to total

[9] More specifically, the LEHD offers publicly available Workplace Area Characteristic (WAC) data at the census block level as part of the larger LEHD Origin-Destination Employment Statistics (LODES) database.

[10] There are two exceptions that need to be mentioned. First, the state of Massachusetts currently provides no data to LEHD. So, here we substituted 2011 ArcGIS Business Analyst Data for Massachusetts, which provides both geocoded locations and employment levels for more than 260,000 establishments in Massachusetts. These samples were aggregated to the census block level and used to approximate the LEHD data. While the time lag between 1998 and 2011 is considerable, we believe that the zoning of commercial activities is reasonably stable over time. Similar problems arose with the District of Columbia, where only 2010 WAC data were available.

[11] An additional advantage of using total employment levels at scales as small as census blocks is that they are less subject to censoring than finer employment classifications.

employment levels, and (iii) point locations within blocks are selected randomly. It should be mentioned that actual locations are almost always along streets and cannot, of course, be random within blocks. But, as discussed in Section 3.2 below, blocks themselves are sufficiently small to allow such random effects to be safely ignored at the scales of most relevance for our purposes.

By repeating this procedure separately for the Northeast Corridor (with a set of $n = 1,035$ location choices) and for California (with n = 645 location choices), one generates a pattern, $X = (x_i = (r_i, s_i) : i = 1,..,n)$, of potential R&D locations that is consistent with H_0, where (r_i, s_i) represents the latitude and longitude coordinates (in decimal degrees) at point i. This process is repeated many times for each R&D location in the data set. In this way, we can test whether the *observed point pattern*, $X^0 = (x_i^0 = (r_i^0, s_i^0) : i = 1,..,n)$, of R&D locations is "more clustered" than would be expected if the pattern were generated randomly (i.e., randomly drawn from the manufacturing employment distribution).

3.1 *K*-Functions

The most popular measure of clustering for point processes is Ripley's (1976) *K*-function, $K(d)$, which (for any given mean density of points) is essentially the expected number of additional points within distance d of any given point.[12] In particular, if $K(d)$ is higher than would be expected under H_0, then this may be taken to imply *clustering* of R&D locations relative to manufacturing at a spatial scale *d.* For testing purposes, it is sufficient to consider sample estimates of $K(d)$. If for any given point i in pattern $X = (x_i : i = 1,..,n)$, we denote the number

[12] The term "function" emphasizes the fact that values of $K(d)$ depend on distance, *d.*

(count) of additional points in *X* within distance *d* of *i* by $C_i(d)$, then the desired *sample estimate*, $\hat{K}(d)$, is given simply by the average of these point counts (i.e., by [13])

$$\hat{K}(d) = \frac{1}{n}\sum_{i=1}^{n} C_i(d). \qquad (1)$$

As described in Section 3, we draw a set of *N* point patterns, $X^s = (x_i^s : i = 1,..,n)$, $s = 1,..,N$, for each of a selection of radial distances, $D = (d_1,..,d_k)$, and calculate the resulting sample *K*-functions, $\{\hat{K}^s(d) : d \in D\}$, $s = 1,..,N$. For each spatial scale, $d \in D$, these values yield an approximate sampling distribution of $K(d)$ under our null hypothesis, H_0.

Hence, if the corresponding value, $\hat{K}^0(d)$, for the observed point pattern, X^0, of R&D locations is sufficiently large relative to this distribution, then this can be taken to imply significant clustering relative to manufacturing. More precisely, if the value $\hat{K}^0(d)$ is treated as one additional sample under H_0, and if the number of these $N+1$ sample values at least as large as $\hat{K}^0(d)$ is denoted by $N^0(d)$, then the fraction

$$p(d) = \frac{N^0(d)}{N+1} \qquad (2)$$

is a (maximum likelihood) estimate of the *p-value* for a one-sided test of hypothesis H_0.

For example, if $N = 999$ and $N^0(d) = 10$ so that $P(d) = 0.01$, then under H_0, there is estimated to be only a one-in-a-hundred chance of observing a value as large as $\hat{K}^0(d)$. Thus, at spatial

[13] These average counts are usually normalized by the estimated mean density of points. But since this estimate is constant for all point patterns considered, it has no effect on testing results.

scale *d*, there is significant clustering of R&D locations at the 0.01 level of statistical significance.

3.2 Test Results for Global Clustering

Our Monte Carlo test for clustering was carried out with $N = 999$ simulations at radial distances, $d \in D = \{0.25, 0.5, 0.75, 1, 2, ..., 99, 100\}$, (i.e., at quarter-mile increments up to a mile and at one-mile increments from 1 to 100 miles). Before discussing these results, it should be noted that quarter-mile distances are approximately the smallest scale at which meaningful clustering can be detected within our present spatial framework. Recall that since locations consistent with the null hypothesis are distributed randomly within each census block, they cannot reflect any locational constraints inside such blocks. For example, if all observed lab locations are street addresses, then, at scales *smaller* than typical block sizes, these locations will tend to exhibit some degree of spurious clustering relative to random locations. If relevant block sizes are taken to be approximated by their associated (circle-equivalent) radii, then since the average radius of the LEHD blocks with positive employment is 0.15 miles in the Northeast Corridor (ignoring Massachusetts) and 0.13 miles in California, this suggests that 0.25 miles is a reasonable lower bound for tests of clustering. In fact, the smallest radius used in most of our subsequent analyses is 0.5 miles.[14]

Given this range of possible spatial scales, our results show that clustering in the Northeast Corridor is so strong (relative to manufacturing employment) that the estimated *p*-values are 0.001 for all scales considered. The results are the same for California up to about 60 miles, and

[14] Since mean values can sometimes be misleading, it is also worth noting that only 6.2 percent of all the LEHD block radii exceed 0.5 miles in the Northeast. This percentage is about 4 percent for California.

they remain below 0.05 up to about 90 miles. Thus, our conjecture that private R&D activities exhibit significant agglomeration is well supported by this data.[15]

3.3 Variations in Global Clustering by Spatial Scale

Further analysis of these sampling distributions (both in terms of Shapiro-Wilk (1965) tests and normal quintile plots (not shown)) showed that they are well approximated by normal distributions for all the spatial scales tested. So, to obtain a sharper discrimination between results at different spatial scales, we calculated the *z*-scores for each observed estimate, $\hat{K}^0(d)$, as given by

$$z(d) = \frac{\hat{K}^0(d) - \bar{K}_d}{s_d} \quad , \quad d = \{0.25, 0.5, 0.75, 1, 2, ..., 99, 100\} \qquad (3)$$

where $\bar{K}_d$ and s_d are the corresponding sample means and standard deviations for the $N+1$ sample *K*-values.

The *z*-scores for the Northeast Corridor are depicted in Figure 2a, and those for California are shown in Figure 2b. Significance levels decrease nearly monotonically for California, while in the Northeast, we see a hump-shaped pattern. The high *z*-scores are consistent with the significance of the Monte Carlo results noted previously but add more detailed information about the patterns of significance.[16] Observe that in both figures, clustering is most significant at

---

[15] In addition, it should be noted that since 0.001 is the smallest possible *p*-value obtainable in our simulations (i.e., $1/(N+1)$ with $N = 999$), these results actually underestimate statistical significance in many cases. While *N* could, of course, be increased, this sample size appears to be sufficiently large to obtain reliable estimates of sampling distributions under H_0.

[16] The benchmark value of $z = 1.65$, shown as a dashed line in both Figures 2a and 2b, corresponds to a *p*-value of 0.05 for the one-sided tests of H_0 in expression (2) above.

smaller scales but exhibits rapid attenuation as scales increase. This pattern is consistent with empirical research on human capital spillovers and agglomeration economies mentioned in the Introduction.[17]

3.4 Relative Clustering of R&D Labs by Industry

We believe that the distribution of manufacturing employment provides a reasonably objective basis for assessing patterns of clustering by private R&D facilities. Nevertheless, the reasons for establishing an R&D lab in a particular location may differ from those that determine the location of manufacturing establishments. For example, R&D labs may be drawn to areas with a more highly educated labor force than would be typical for most manufacturing establishments. Some R&D labs may co-locate not because of the presence of spillovers but rather because of subsidies provided by state and local governments (as, for example, when technology parks are partially subsidized).

To explore such differences, we begin by grouping all labs in terms of their primary industrial research areas at the two-digit SIC level.[18] With respect to this grouping, our null hypothesis is simply that there are no relevant differences between the spatial patterns of labs in each group (i.e., the spatial distribution of labs in any given industry is statistically indistinguishable from the distribution of all labs). The simplest formalization of this hypothesis is to treat each group of labs as a typical random sample from the distribution of all labs. More precisely, if *n* is the total number of labs (where $n = 1035$ for the Northeast and $n = 645$ for California) and if n_j denotes

[17] See Carlino and Kerr (2015) for a review of the literature on the localization of knowledge spillovers.

[18] We assign labs to an industry based on information contained in the *Directory of American Research and Technology*. In the Northeast Corridor, there are 19 industrial groupings corresponding to SICs 10, 13, 20-23, 26-30, 32-39, and 73. In California, there are 16 industrial groupings corresponding to SICs 13, 16, 20, 26, 28-30, 32-39, and 73. The industry names of these SICs are included in Tables 2a and 2b.

the number of these labs associated with industry j, then our null hypothesis, H_0^j, for industry j is that *the spatial distribution of R&D labs in industry j is not statistically distinguishable from that of a random sample of size n_j from all n labs.* Such random samples are easily constructed by randomly permuting (reordering) the lab indices $1,..,n$ and choosing the first n_j of these (as is also done in DO). With respect to clustering, one can then compare $\hat{K}(d)$ values for the observed pattern of labs in industry j with those for set of *N* such randomly sampled patterns and derive both p-values, $P_j(d)$ and *z*-scores, $z_j(d)$ comparable with those in expressions (2) and (3), respectively. If $P_j(d)$ is sufficiently low [or $z_j(d)$ is sufficiently high], then it can be concluded that there is significantly more clustering at scale *d* for labs in industry j than would be expected under hypothesis H_0^j.

This has two benefits. First, it sets a much higher bar in our tests of spatial concentration. Second, we can implement these tests with very high precision at even the smallest of spatial scales. Using this counterfactual method, we find the strongest evidence for the spatial concentration of R&D labs occurring at very small spatial scales (such as within a two- to three-block area). Before reporting the results of these (random permutation) tests, it must be stressed that such results are only meaningful *relative* to the population of all R&D labs, and, in particular, allow us to say nothing about clustering of R&D labs in general. But the benefits of this approach are two-fold. First, since the pattern of all R&D labs has already been shown to exhibit significant clustering relative to manufacturing employment (at all scales tested), the present results help to sharpen these general findings. Moreover, while this sharpening could in principle be accomplished by simply repeating the global tests above for each industry, the present approach avoids all issues of location feasibility at small scales. In particular, since the

exact locations of all labs are known, we can use this information to compare relative clustering among industries at all scales.

Turning now to the test results, the *p*-values for each of the 19 two-digit SIC industries in the Northeast Corridor are reported in Table 2a for selected distances. As stated previously, we are able to analyze relative clustering at all scales, regardless of how small. In particular, at the quarter-mile scale, we find that seven of these 19 industries (37 percent) are significantly more localized (at the 0.05 percent level) than are R&D labs in general.[19] Moreover, none are significantly more dispersed.[20] Table 2b reports the *p*-values for each of the 16 two-digit SIC industries in California for selected distances. We find that, at a distance of a quarter-mile, eight of these 16 industries (50 percent) are significantly more localized (at the 0.05 percent level) than are R&D labs in general.[21] Again, none are significantly more dispersed.

A graphical representation of these results is presented in Figure 3, where the *z*-scores for each of the seven industries in the Northeast with most significant clustering is shown in Figure 3a, and those for the seven (of eight) most significant California industries are shown in Figure 3b.[22] Because we are especially interested in the attenuation of *z*-scores at small scales, these *z*-scores are calculated in increments of 0.25 miles up to five miles. For all but one of these industries in the Northeast, the clustering of R&D labs is by far most significant at very small spatial scales —

[19] The seven industries are Textile Mill Products; Stone, Clay and Glass; Fabricated Metals; Chemicals and Allied Products (this category includes drugs); Instruments and Related Products; Miscellaneous Manufacturing Industries; and Business Services.

[20] With respect to dispersion, two of the 19 industries are found to be significantly more dispersed starting at a distance of five miles, and a third industry exhibits some degree of relative dispersion at 50 miles.

[21] The eight industries are Chemicals and Allied Products; Rubber Products; Primary Metal Products; Industrial and Commercial Machinery; Electronics; Transportation Equipment; Measuring, Analyzing, and Controlling Equipment; and Business Services.

[22] To conserve on space, the graph of the *z*-scores for the Chemicals and Allied Products is not shown in Figure 3b since the labs doing R&D in this industry accounted for less than 1 percent of all labs in California.

a quarter mile or less. The lone exception is Miscellaneous Manufacturing Industries (SIC 39), where the highest *z*-score occurs at a distance of just under two miles. In California, the clustering of R&D labs is most significant at very small spatial scales for only four of the seven industries shown in Table 3b. Two of the other industries, Electronics and Business Services have local peaks at one-half mile and at one mile, respectively.

In addition, Figure 3a shows rapid attenuation of *z*-scores at small scales for all seven industries in the Northeast. Moreover, for most of these industries, there is essentially a monotonic decline in *z*-scores at all scales shown. While degrees of significance at larger scales vary among industries, the relative clustering of labs in both the Chemicals and Business Services industries continues to be significant at all scales shown. (For Business Services in particular, all but one these labs are associated with firms engaged in the computer programming or data processing subcategories.) Turning to California, Figure 3b shows rapid attenuation of *z*-scores at small scales for four of these seven industries. The other three industries, Industrial and Commercial Machinery, Electronics, and Business Services (mostly in the subcategory, Computers and Data Processing) exhibit an opposite trend in which relative clusters becomes more significant at larger scales.

Finally, it is of interest to note that three industries are among the most significantly clustered industries in both the Northeast and California, namely Chemicals, Business Services, and the Manufacturing, Analyzing, and Controlling Equipment industry. Here, the Chemical industry (SIC 28) merits some special attention, if for no other reason than this category includes labs engaged in pharmaceutical R&D, a very important segment of the U.S. economy. In our data, this category of labs accounts for about 40 percent of all labs in the Northeast, a share more than twice as large as any other two-digit SIC industry. In California, the Chemicals industry accounts

for about 16 percent of the labs we study. Thus, at least within the geographic area we study, this industry is seen to be a major contributor to the overall clustering pattern of R&D shown in Figures 2a and 2b. But it should be equally clear from Figures 3a and 3b that significant clustering occurs in many other industries as well. So, clustering of R&D labs is by no means specific to drugs and chemicals.

4. LOCAL CLUSTER ANALYSIS

While the above global analysis can identify spatial *scales* at which clustering is most significant, it does not tell us *where* clustering occurs. In this section, we use a variation of our techniques to identify clustering in the neighborhood of specific R&D labs. The main tool for accomplishing this is the *local* version of sample *K*-functions for individual pattern points (first introduced by Getis, 1984).[23] This local version at each point i in the observed pattern is simply the count of all additional pattern points within distance *d* of i. In terms of the notation in expression (1) above, the *local K-function,* $\hat{K}_i$, at point i is given for each distance, *d*, by

$$\hat{K}_i(d) = C_i(d) .^{24} \qquad (4)$$

Hence, the global *K*-function, $\hat{K}$, in expression (1) is simply the average of these local functions.

4.1 Local Testing Procedure

[23] The interpretation of the population *local K-function,* $K_i(d)$, for any given point *i* is simply the expected number of additional pattern points within distance *d* of point *i.* Hence, $\hat{K}_i(d)$ is basically a single-sample (maximum likelihood) estimate of $K_i(d)$. For a range of alternative measures of local spatial association, see Anselin (1995).

[24] It should be noted that the original form proposed by Getis (1984) involves both an "edge correction" based on Ripley (1976) and a normalization based on stationarity assumptions for the underlying point process. However, in the present Monte Carlo framework, these refinements have little effect on tests for clustering. Hence, we choose to focus on the simpler and more easily interpreted "point count" version above.

For the local testing procedure, we use the same null hypothesis employed in Section 3: R&D labs are distributed in a manner proportional to manufacturing employment at the zip code level and proportional to total employment at the block level.[25] The only substantive difference from the procedure used in that section is that the location, x_i, of point i is held fixed. The appropriate simulated values, $\hat{K}_i^s(d),\ s=1,..,N$, under H_0 are obtained by generating point patterns, $X^s=(x_j^s: j=1,..,n-1)$, $s=1,..,N$, representing all $n-1$ points other than i. The resulting *p*-values for a one-sided test of H_0 with respect to point i then take the form,

$$P_i(d) \ = \ \frac{N_i^0(d)}{N+1}\ ,\ i=1,..,n, \qquad (5)$$

where $N_i^0(d)$ is again the number of these $N+1$ draws that produce values at least as large as $\hat{K}_i^0(d)$.

An attractive feature of these local tests is that the resulting *p*-values for each point i in the observed pattern can be *mapped*. This allows one to check visually for *regions* of significant clustering. In particular, groupings of very low *p*-values serve to indicate not only the location but also the *approximate size* of possible clusters. Such groupings based on *p*-values necessarily suffer from "multiple testing" problems, which we address in later sections and more systematically in Appendix B.

4.2 Test Results for Local Clustering

[25] Later in the paper, we replace manufacturing employment with manufacturing establishments and STEM workers as robustness checks.

For our local cluster analyses, simulations were again performed using $N = 999$ test patterns of size $n-1$ for each of the *n* (=1,035 in the Northeast Corridor and 645 in California) R&D locations in the observed pattern, X^0. The set of radial distances (in miles) used for the local tests was $D = \{0.25, 0.5, 0.75, 1, 2, 5, 10, 11, 12.., 100\}$. But, unlike the global analyses previously in which clustering was significant at all scales, there is considerable variation in significance levels across labs located at different points in space. For example, it is not surprising to find that many isolated R&D locations exhibit no local clustering whatsoever. Moreover, there is also considerable variation in significance at different spatial scales. At very large scales (perhaps, 50 miles), one tends to find a few large clusters associated with those mega regions containing most of the labs (within the Washington–Boston corridor or the San Francisco Bay Area). At very small scales (say 0.25 miles), one tends to find a wide scattering of small clusters, mostly associated with locations containing multiple labs (such as industrial parks). In our present setting, the most meaningful patterns of clustering appear to be associated with intermediate scales between these two extremes.

A visual inspection of the *p*-value maps generated by our test results showed that the clearest patterns of distinct clustering can be captured by the three representative distances, $D = \{1, 5, 10\}$. Of these three, the single best distance for revealing the overall clustering pattern in the entire data set appears to be five miles, as illustrated for the Northeast Corridor and California in Figures 4a and 4b, respectively.[26] As seen in the legend, those R&D locations, i, exhibiting maximally significant clustering [$P_i(5) = 0.001$] are shown in black, and those with *p*-values not exceeding 0.005 are shown as dark gray. Here, it is evident that essentially all of the most significant locations occur in four distinct groups in the Northeast Corridor, which can be

[26] In the Appendix B, we report results for *all* distances in *D* as a robustness check.

roughly described (from north to south) as the "Boston," "New York City," "Philadelphia," and "Washington, D.C.," agglomerations.[27] In California, there are again three distinct groups, roughly described (from north to south) as "San Francisco Bay Area," "Los Angeles area (mainly Irvine)," and "San Diego." While these patterns are visually compelling, it is important to establish such results more formally.

5. IDENTIFYING SPATIAL CLUSTERS

The global cluster analysis in Section 3 identified the *scales* at which clustering is most significant (relative to manufacturing employment). The local cluster analysis in Section 4.1 provided information about *where* clustering is most significant at each spatial scale. But neither of these methods formally identifies or defines specific "clusters" of labs. In this section, we apply some additional techniques to identify clusters, which we call the *multiscale core-cluster* approach.

As discussed in Appendix B, a number of cluster-identification techniques have been developed to identify sequences of clusters that are individually "most significant" in an appropriate sense.[28] The present approach is based more directly on the *K*-function methods previously, and in particular, focuses on the *multiscale* nature of local *K*-functions. More specifically, this clustering procedure starts with the local point-wise clustering results in Section 4.1 and seeks to identify subsets of points that can serve as "core" cluster points at a given selection of relevant scales, *d*. Here, we again focus on the three scales, $D = \{1,5,10\}$, used in Section 4.1. At each scale, $d \in D$, we define a *core point* to be a maximally significant R&D lab, i.e., with a local *K*-

[27] Two exceptions are the small but significant agglomerations identified in the analysis — one in Pittsburgh and one in Buffalo.

[28] This sequential approach is designed specifically to overcome the problem of "multiple testing," as discussed further in Appendix B.

function *p*-value of 0.001 (using the 999 simulations of *K* at distance *d* in Section 5.1). In order to exclude "isolated" points that simply happen to be in areas with little or no manufacturing, we also require that there be at least *four* other R&D labs within this *d*-mile radius. Finally, to identify distinct clusters of such points, we create a *d*-mile-radius buffer around each core point (in ArcMap). We designate the set of points (labs) in each connected component of these buffer zones as a *core cluster* of points at scale *d*. Hence, each such cluster contains a given set of "connected" core points along with all other points that contributed to their maximal statistical significance at scale *d*. These concepts are best illustrated by examples.

We begin with the single most striking example of multiscale clustering in our data set, namely the San Francisco Bay Area in California shown in Figure 5. Starting at the 10-mile level, we see one large cluster (represented by dashed gray curve), that essentially covers the entire Bay Area. At the five-mile level (represented by solid gray curves), the dominant core cluster is seen to be perfectly nested in its 10-mile counterpart, corresponding almost exactly to what is typically regarded as Silicon Valley. The smaller secondary cluster of labs is approximately centered around the Lawrence Livermore National Laboratory complex. Finally, at the one-mile level (represented by black curves), the heaviest concentration of core clusters essentially defines the traditional "heart" of Silicon Valley, stretching south from the Stanford Research Park area to San Jose. In short, this statistical hierarchy of clusters is in strong agreement with the most well-known R&D concentrations in the San Francisco Bay Area.

A second example, from the Northeast Corridor, is provided by the hierarchical complex of R&D clusters in the Boston area, shown in Figure 6a. Here again, the entire Boston area is itself a single 10-mile cluster. Moreover, within this area, there is again a dominant five-mile core cluster containing the five major one-mile clusters in the Boston area. The largest of these is

concentrated around the university complex in Cambridge, while the others are centered at points along Route 128 surrounding Boston. This is seen more clearly in Figure 6b,[29] which also shows that most R&D labs in the Boston area are located in close proximity to major transportation routes, including Interstate Routes 90, 93, 95, and 495.

Note, finally, that while the clusters in both Figures 5 and 6a tend to be nested by scale, this is not always the case. For example, the five-mile "Livermore Lab" cluster in Figure 5 is seen to be mostly outside the major 10-mile cluster. Here, there is a concentration of six R&D labs within two miles of each other, although Livermore is relatively far from the Bay Area. So, while this concentration is picked up at the five-mile scale, it is too small by itself to be picked up at the 10-mile scale.

These examples illustrate the attractive features of the multiscale core-cluster approach. First and foremost, this approach adds a scale dimension not present in other clustering methods. In essence, it extends the multiscale feature of local *K*-functions from individual points to clusters of points. Moreover, this approach helps to overcome the particular limitations of significance-maximizing approaches mentioned previously. First, the shapes of individual core clusters are seen to be more sensitive to the actual configuration of points than those found in significance-maximizing methods.[30] In addition, since all core clusters are determined simultaneously, the path-dependency problem of sequential methods does not arise.

In summary, an overall depiction of core clusters for both the Northeast Corridor and California (at scales, $d = 5,\ 10$) is shown in Figures 7a and 7b, respectively. Figure 7a shows the four major clusters identified for the Northeast Corridor (one each in Boston, New York/Northern New

[29] For visual clarity, only core cluster points (and not their associated buffers) are shown in Figure 6b.

[30] This point is demonstrated in Appendix B.

Jersey, Philadelphia/Wilmington, and Washington, D.C.), while Figure 7b shows the three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

Finally, it should be stressed that this multiscale approach is not a substitute for more standard approaches such as significance-maximizing. While it does yield a meaningful hierarchy of statistically significant clusters, it provides no explicit method for rank ordering clusters in terms of statistical significance. In particular, this approach by itself cannot be used to gauge the relative statistical significance of clusters (such as determining whether clustering in Boston is more significant than in New York). Moreover, such representational schemes presently offer no formal criteria for choosing the key parameter values by which they are defined (the *d*-scales to be represented, the *p*-value thresholds and d -neighbor thresholds for core points, and even the connected-buffer approach to identifying distinct clusters).[31] Thus, the primary objective of this more heuristic procedure is to produce explicit representations of clusters that capture both their relative shapes and concentrations in a natural way. The ultimate value of such clusters for our purposes can only be determined by testing their economic significance — to which we now turn.

6. CLUSTERING OF R&D LABS AND CLUSTERING OF PATENT CITATIONS

So far, we have established a body of evidence demonstrating that R&D labs are indeed clustered, and we have posited a method for identifying specific clusters in space. In this section, we test whether these clusters are related to knowledge spillovers that are potentially attenuated by distance. To do this, we study the relative geographic concentration of citations to patents

[31] It should be noted that certain, more systematic procedures may be possible. For example, the selection of "best representative" *d*-scales could be in principle accomplished by versions of *k*-means procedures in which the within-group versus between-group variations in patterns are minimized.

originating in our clusters. These citations are a concrete indication of the transmission of information from one inventor to another.

We follow the general approach developed in JTH, but it is modified to reflect the geographic clustering of R&D labs we identify in this paper. As described earlier, JTH test for the "localization" of knowledge spillovers by constructing measures of geographic concentration of citations contained in two groups of patents — a treatment group and a control group. The treatment group represents a set of patents that cite a specific, earlier patent obtained by an inventor living in a particular geographic area (in the JTH study either a state or a metropolitan area). For each treatment patent, JTH use a process to select a potential control patent that is similar to the treatment patent but does not cite the earlier patent. For patents in the treatment and control groups, JTH calculate the proportion of those patents obtained by an inventor living in the same geographic area as the inventor of the earlier patent. The difference of these two proportions is a test statistic for the localization of knowledge spillovers. In their study, JTH found that, relative to the pattern reflected in the sample of control patents, patent citations were two times more likely to come from the same state and about two to six times more likely to come from the same metropolitan area.

We construct a comparable test statistic, with several refinements, and we substitute the R&D clusters identified in Section 5 for the state and metropolitan area geography used by JTH. This provides us with an alternative way to test for possible localized knowledge spillovers at much smaller spatial scales than are found in much of the preceding literature. Recall that the boundaries of our clusters are determined by interrelationships among the R&D labs in our sample and, therefore, should more accurately reflect the appropriate boundaries in which knowledge spillovers are most likely to be at work. In that sense, the geography of our clusters

should be better suited for studying knowledge spillovers than states, metropolitan areas, or other political or administrative boundaries.

6.2 Construction of the Citations Data Set

For this analysis, we use data obtained from the NBER Patent Data Project.[32] The data span the years 1996–2006. We identify the inventors on a patent using data on inventor codes found in the Patent Network Dataverse (Lai, D'Amour, and Fleming, 2009). Patents are assigned to locations based on the zip code associated with the *residential* address of the first inventor on the patent.[33] We do not use the address of the assignee (typically the company that first owned the patent) because this may not reflect the location where the research was conducted (e.g., it may be the address of the corporate headquarters and not the R&D facility). While it's possible that an inventor's home lies outside of a cluster while his professional work takes place inside a cluster, this type of measurement error would bias our results against finding significant location differentials. As a robustness check, we repeated our main analysis using the zip code of the second inventor on the patent. While the sample size is smaller because not all patents list two or more inventors, the results were virtually the same as we report below. [34]

For our tests, we rely primarily on the boundaries identified by our five-mile and 10-mile core clusters located in the Northeast Corridor and in California.[35] For each core cluster at a given scale, we assemble four sets of patents. The first set, which we call *originating patents*, represent

[32] See https://sites.google.com/site/patentdataproject/. We use the files pat76_06_assg.dta and cite_7606.dta.

[33] We used the location information contained in the file inventors5s_9608.tab downloaded from http://dvn.iq.harvard.edu/dvn/dv/patent. Note that this approach implies that our inventors are located at the centroid of the zip code where they live. We have zip codes information for almost 99 percent of the patents with a first inventor residing in the United States.

[34] Results are available from the authors upon request.

[35] In Section 6.4.1 that follows, we report comparable tests for larger and smaller clusters.

those patents granted in the years 1996–1997 by an inventor living in the cluster. We call the second set of patents *citing patents*. These consist of all subsequent patents, including patents where the residential address of the first inventor is located outside the U.S., that cite one or more of the originating patents, after excluding patents with the same inventor or that were initially assigned to the same company as the originating patent. We exclude these self-citations because these are unlikely to represent the knowledge spillovers we seek to identify.[36]

For every citing patent, we attempt to match it to an appropriate control patent. When we are successful, we include the citing patent in a set we call *treatment patents* and the matched patent in a set we call *control patents*. We select control patents using the following approach. For a given citing patent, the set of potential control patents must have an application date after the grant date of the originating patent that is cited. Potential control patents also cannot cite the originating patent. The application date of potential control patents must be within one year (six months on either side) of the application date of the treatment patent. Finally, as was done by JTH, potential control patents must have the same three-digit primary patent class as the treatment patent.[37] In this way, potential controls are drawn from patents in the same technological field.

The set of potential control patents for a given treatment patent may overlap with the set of potential controls for other treatment patents. To rule out any possibility that this overlap may affect our tests, we randomized the order in which treatment patents were matched to control patents, and we randomized the selection of a specific control patent when there was more than

[36] We do this using the pdpass variable in the data set pat76_06_assg and the Invnum in the Consolidated Inventor Dataset. For details, see Lai, D'Amour, and Fleming (2009).

[37] We match on the variable class in the data set pat76_06_assg. This is the original primary classification of the patent. We feel it is important to use a "real time" classification because these are what other researchers might rely upon around the time a patent was issued.

one potential control patent from which to choose.[38] The main results reported below allow for the selection of control patents with replacement. In other words, a given control patent may be matched to more than one citing patent. As a robustness check (not shown), we repeat the analysis by sampling potential controls *without* replacement.[39] In this case, a potential control patent can be matched with one citing patent at most. While this reduces the rate at which we can match control patents to citing patents, it does not materially affect the test statistics.[40]

6.3 The Test Statistics

For any given cluster scale, d $(=5,\ 10)$, let η_o denote the number of *originating patents* indexed $\{o_i : i = 1, \cdots, \eta_0\}$ that were granted to inventors living in one of the core clusters at scale *d* in the years 1996–1997.[41] Let η_i denote the number of subsequent citations $\{c_{ij} : j = 1, \cdots, \eta_i\}$ to o_i (after removing self-citations) over the years 1996–2006. For each of these citing patents, c_{ij}, designated as *treatment patents*, we attempted to identify a unique *control patent*, $\tilde{c}_{ij}$, with the same three-digit patent class and with an application date within one year of the treatment patent

[38] Two random numbers are assigned to each citing patent. The first is used to set the order in which citing patents are matched. The second is used, in conjunction with a random number assigned to every potential control patent, to select a patent associated with the minimum absolute difference between the two random numbers. In JTH, when multiple potential control patents exist, they select the one with a grant date that is nearest to the grant date of the treatment patent as the control the patent.

[39] Randomization of the order of matching control patents to citing patents should rule out any bias resulting from an unknown systematic pattern in the timing of patents being issued for specific technology fields. One concern is that our sampling procedure could violate the independence of the control group and the treatment (citing) group. This is possible if a control patent also appears in the set of treatment patents — if the control patent for one treatment patent is a citing patent for a different originating patent. We find that these two groups are independent since there is absolutely no overlap between the citing patents and control patents either in the Northeast Corridor or the California samples.

[40] These results are available from the authors upon request.

[41] The following formulation of the proportions used for testing purposes is based largely on Murata et al. (2015).

(see previous description). We are not always successful in doing so. Let $\tilde{\eta}_i (\leq \eta_i)$ denote the number of treatment patents, c_{ij}, for which a control, $\tilde{c}_{ij}$, was found.

Among these $\tilde{\eta}_i$ treatment patents, we count the number of patents, m_i, for which the residential address of the first inventor on the citing patent is located in the *same* core cluster as the originating patent it cites. The fraction of all such patents at scale *d*, i.e., the *treatment proportion*, is given by[42]

$$p = \frac{\sum_{i=1}^{\eta_o} m_i}{\sum_{i=1}^{\eta_o} \tilde{\eta}_i} = \frac{1}{\tilde{\eta}} \sum_{i=1}^{\eta_o} m_i \,. \qquad (6)$$

Similarly, let $\tilde{m}_i$ denote the number of matched control patents, $\tilde{c}_{ij}$, in which the residential address of the first inventor is located in the same cluster as the originating patent cited by the treatment patent. The *control proportion* is then given by

$$\tilde{p} = \frac{\sum_{i=1}^{\eta_o} \tilde{m}_i}{\sum_{i=1}^{\eta_o} \tilde{\eta}_i} = \frac{1}{\tilde{\eta}} \sum_{i=1}^{\eta_o} \tilde{m}_i \,. \qquad (7)$$

The resulting test statistic is simply the difference between these proportions, i.e., $p - \tilde{p}$. Under the null hypothesis of "no localization of knowledge spillovers," this difference of independent proportions is well known to be asymptotically normal with mean zero and thus provides a well-defined test statistic.[43]

[42] The dependency of fraction, *p* (and all other quantities in (6)) is taken to be implicit.

[43] In JTH, the standardized test statistic, $(p - \tilde{p}) / \sqrt{[p(1-p) + \tilde{p}(1-\tilde{p})] / n}$, is asserted to be *t* distributed. In fact, the *t* distribution is not strictly accurate. However, for the present large sample size, $n > 50,000$, this is of little consequence since the *t* and standard normal distributions are virtually identical.

6.4 Main Results

Table 3a presents the results of our localization or matching rate tests among five-mile clusters in the Northeast Corridor, while Table 3b shows the results for the 10-mile clusters. As the last row of Table 3a shows, inventors living in the five-mile clusters obtained 8,526 patents in 1996–1997 (column A). Those patents subsequently received 76,730 citations from other patents during the sample period (column B). Our matching algorithm, with replacement, was able to match 85 percent of the citing patents with an appropriate control patent (column H). Among the treatment patents, 3.69 percent (column G) had a first inventor living in the same cluster as the patent it cited; this is the treatment proportion. Among the control patents, only 0.62 percent (column J) had a first inventor living in the same cluster as the patent cited by the treatment patent; this is the control proportion. As shown in the next to the last column of the table, on average, a given patent citing an earlier patent in a five-mile cluster is a little more than six times as likely to have a first inventor living in that cluster than would be expected by chance alone. This value is on the higher side of the range reported by JTH for their test of localization at the metropolitan area level. As the last row of the table 3a shows, the difference between the treatment and control proportions is highly statistically significant (column L). In addition, the location differential — defined as the ratio of treatment and control proportions — is at least around 3.0.

Table 3b presents the results of our localization tests among 10-mile clusters in the Northeast Corridor. At a somewhat larger spatial scale, we find there are more originating patents, more citing patents, and, thus, more treatment and control patents. Both the treatment and control proportions (columns G and J) are higher than was found among the five-mile clusters. The *t* statistic associated with the difference in these proportions is even higher than was found for the

smaller clusters. At the same time, the location differential is somewhat smaller. On average, a given patent citing an earlier patent in a 10-mile cluster is 3.6 times as likely to have a first inventor living in that cluster than would be expected by chance alone. This value is on the lower side of the range reported by JTH for their test of localization at the metropolitan area level. There are a number of specific clusters where this differential is substantially higher. For example, the location differential is more than twice the four cluster average in the Washington, D.C., and Philadelphia clusters, and a little more than one-third higher in the Boston cluster.

Tables 4a and 4b present the results of our localization tests among five- and 10-mile clusters, respectively, in California. Compared with the Northeast Corridor, we find many more originating patents, citing patents, and, therefore, treatment and control patents. The treatment proportions (column G) among the California clusters are much higher than those found in the Northeast Corridor. However, this is driven almost entirely by the cluster association with Silicon Valley. The control proportions (column J) are also larger than we found in the Northeast Corridor. The *t*-statistic for the difference in treatment and control proportions (column L) is highly significant for all the five-mile and 10-mile clusters. On average, a given patent citing an earlier patent in a five- or 10-mile cluster in California is four to four and a half times as likely to have a first inventor living in that cluster than would be expected by chance alone.

It is worth noting that there is significant cross-cluster variation. For 5-mile clusters in the Northeast, the location differentials for Philadelphia and Washington D.C. are more than twice the average. The largest location differential among our baseline results is 45.5 for the 5-mile Los Angeles cluster; this is ten times the average for 5-mile clusters in California.

To summarize, the clusters of R&D labs identified using our multicore approach appear to coincide with the geographic clustering of patent citations, an often-cited indicator of knowledge spillovers. The following section develops these results further and discusses a number of robustness checks.

6.5 Additional Results and Robustness Checks

6.5.1 The Relationship Between Citation Location Differentials and Spatial Scale

The statistics in the preceding tables suggest that there may be a systematic relationship between the size of clusters we study and the magnitude of location differentials we find. To explore this further, we extended our analysis to consider clusters at spatial scales of 20 miles. We summarize the results in Tables 5a and 5b.

A number of patterns are evident from the table. First, the increase in the number of originating patents associated with larger core clusters falls off because a number of clusters that are significant at smaller spatial scales are not significant at the larger spatial scales. The treatment and control proportions tend to increase as we consider larger core clusters. The difference between these proportions becomes more and more statistically significant as the sample size rises. At the same time, the location differential falls monotonically as the geographic size of the clusters increases. These results suggest that the core clusters are picking up knowledge spillovers over a variety of spatial scales. Nevertheless, the localization effects appear to be largest at spatial scales of five miles and perhaps less. This is also consistent with what we found in the results of our Global *K* analysis described earlier. And as already noted, the attenuation in

the localization differential as cluster size increases is a typical finding in studies examining localized knowledge spillovers.[44]

6.5.2 Are Patents Obtained in Our Clusters More Influential?

In this section, we investigate whether patents obtained by inventors living within our core clusters are somehow more important, or at least better known, than patents obtained outside of these clusters. We rely on a common metric of patent quality — the number of citations received.[45] We develop a "counterfactual" region for each of the 10-mile core clusters identified in Section 5. For example, the New York cluster is compared with the region outside of that cluster contained in states of New York, Connecticut, and northern New Jersey. The Boston cluster is compared with the region outside of the cluster in the states of Massachusetts, New Hampshire, and Rhode Island. In Table 6, we report a simple difference in means test for the number of citations per patents received by patents located inside or outside our clusters. For all our clusters, the average number of citations received by patents is greater inside the cluster compared with the average citations received outside the respective cluster; this difference in citations is statistically significant in all clusters except one (Philadelphia).

These results, combined with the results for the localization of citations, suggest there is prima facie evidence that the inventions developed within our clusters are more influential than inventions developed outside a cluster but within the same region of the country. An alternative explanation, which we cannot entirely rule out, is that patents within a cluster receive more citations because they are often cited by inventors living nearby. According to this reasoning, the

[44] See Carlino and Kerr (2015) for a review of studies documenting attenuation in knowledge spillovers as cluster size increases.

[45] Hall, Jaffe, and Trajtenberg (2005) show that a one-citation increase in the number of patents in a firm's portfolio increases its market value by 3 percent. For additional evidence, see Trajtenberg (1990).

inventions may not necessarily be better, but they are better known by researchers in the area. This interpretation only reinforces the evidence of localized knowledge spillovers in our clusters.

6.5.3 Alternative Approaches to Identifying Cluster Boundaries

In addition to clustering to take advantage of knowledge spillovers, it is also possible that R&D activity is geographically concentrated to take advantage of labor market pooling. As we have shown, one important concentration of R&D labs is found in around Cambridge, MA, and another important clustering is found in Silicon Valley. These labs are close to large pools STEM graduates and workers, the very workers R&D activity requires. Manufacturing activity tends to employ a more general workforce than does innovative activity and may therefore be more geographically dispersed compared with innovative activity.

To address the concern that we may be intermingling knowledge spillovers with labor market pooling, we first develop a measure of STEM workers by location.[46] For our backcloth, we replace the number of manufacturing employees in each zip code area with an estimate of the number of STEM workers. This is constructed using the proportion of STEM jobs in each four-digit NAICs industry multiplied by the number of jobs in each industry reported in the zip code business patterns data. We report the results of this alternative test for five- and 10-mile clusters in the Northeast Corridor (Tables 7a and 7b) and in California (Tables 8a and 8b). Note that the cluster definitions change when the backcloth changes, so the list of clusters in these tables differs from those in Tables 3 and 4. With the exception of the five-mile clusters in the Northeast Corridor, the average location differentials using the STEM worker backcloth are virtually the

[46] We use the taxonomy of STEM occupations found at http://www.bls.gov/oes/stem_list.xlsx. For details, see Watson (2014). This taxonomy is mapped to the 2010 vintage of the Standard Occupational Classifications (SOCs). We map back to the 2000 vintage of the SOCs so we can use the 2002 job counts from the Occupational Employment Statistics to calculate STEM employment "intensity" by industry.

same as for the baseline findings. The location differential falls from 6.0 for the five-mile clusters in the Northeast Corridor when considering the baseline results to 4.2 for the results when the clusters are based on STEM workers. For the most part, the findings reported for the location differentials in the baseline (and subsequent analysis) suggest little, if any, upwardly bias as a result of labor market pooling.

6.5.4 Alternative Approaches to Identifying Control Patents

As discussed in footnote 3, there has been some debate in the literature as to the best way of implementing a technological similarity requirement based on patent classifications. JTH identify potential control patents within the same three-digit primary patent class as the treatment patent. TFK suggest that the potential controls should be drawn more narrowly from patents that share the same patent class and subclass as the citing patent. They find that tests using this alternative approach reduce the size and significance of the localization ratios, especially at smaller geographies.

The results presented in Section 6.3 are based on the JTH approach of limiting potential control patents to ones that share the same three-digit primary class as the citing patent. As a robustness check, we implement one version of the matching requirements tested in TFK. We restrict potential control patents to ones that share the same primary class and subclass as the citing patent.[47] Our methodology is otherwise the same as we describe in Section 6.2. We report the results of this alternative test for five- and 10-mile clusters in the Northeast Corridor (Tables 9a and 9b) and in California (Tables 10a and 10b). Comparing these results with our baseline results (Tables 3a and 3b) and (4a and 4b), there are very small differences in the treatment and control

[47] This is analogous to the test reported in Table 3 column (6) in TFK.

proportions. The *t*-statistics using the TFK approach are only slightly smaller than they are when using the JTH approach, but they are nevertheless very large. We conclude that our results do not appear to be sensitive to the choice of technology controls.

More recently, methods for constructing a matched sample of treatment and control groups has evolved. Specifically, Coarsened Exact Matching, CEM, (Iacus, King, and Porro, 2011) can be used to improve the balance between the treated group (citing patents) and the control group.[48] In addition to matching on the application year of the patent and the patents three-digit technology classification, we also matched discrete bins on two additional variables: 1) the year the patent was granted; and 2) the number of citations a patent received (allcites). We relied upon the CEM algorithm in STATA to coarsen the matched bins based on an optimization of an objective function rather than arbitrarily assigning cut points to the bins.

We use the CEM matched controls in several ways. First, we follow the JTH location differential approach used in producing Tables 3 and 4, our baseline findings, but use the CEM controls. For this approach, we exclude patents with the same inventor or that were initially assigned to the same company as the originating patent.[49] The results are reported in Tables 11 (for the Northeast Corridor) and Table 12 (for California). The location differentials are uniformly smaller than we previously reported for the broad cluster in the Northeast Corridor and in California. On average, a given patent citing an earlier patent in a five-mile cluster in the Northeast Corridor is 4.5 times as likely to have a first inventor living in that cluster than would

[48] We thank an anonymous referee for suggesting CEM approach for selecting controls.

[49] For this approach, the set of potential control patents for a given treatment patent may overlap with the set of potential controls for other treatment patents. To rule out any possibility that this overlap may affect our tests we randomized the order in which treatment patents were matched to control patents, and we randomized the selection of a specific control patent when there was more than one potential control patent from which to choose. The results reported below allow for the selection of control patents with replacement.

be expected by chance alone, compared with a differential of 6.0 reported in our baseline results. The location differential in California's five-mile cluster falls to 2.5 when using the CEM matched controls from 4.5 reported for baseline. The location differential in the Northeast Corridor 10-mile cluster falls to 2.8 (when using the CEM-matched controls) from 3.6 reported for baseline. In the California 10-mile cluster, the location differential falls to 2.5 from 4.2 reported for baseline.

In our second approach, we estimate a logistic model of the likelihood that a patent in cluster h cites an originating patent in that cluster:

$$T_h = \alpha_0 + \beta_1 D_h + \varepsilon_h$$

where T_h is an indicator variable that equals one for observations corresponding to a treated patent (a patent that cites at least one originating patent in cluster *h*) and zero for the corresponding control patent; D_h is an indicator variable that equals one if the patent originates in cluster *h*, zero otherwise; and ε_h is a random error term. For this approach, we do not exclude patents with the same inventor or that were initially assigned to the same company as the originating patent. We report robust standard errors. The observations are weighted based on the number of CEM-matched controls found for each treated observation. The results are reported in Table 13. The estimated coefficients, $(\widehat{D}_h)$, are positive and significant supporting the findings reported in Tables 11 and 12.

Finally, to facilitate comparison, the main results found for location differentials are summarized in Table 14. The table shows the results when R&D clustering is analyzed with respect to manufacturing employment (baseline); or STEM workers; and when the controls are

alternatively selected to share the same patent class and subclass as the citing patents (disaggregated), or when the controls are selected using more stringently matched samples (CEM). Regardless of the specification chosen to construct the location differentials, we find that citations are at least about 2.5 times more likely to come from the same cluster as earlier patents than one would predict using a control sample of otherwise similar patents.

7. CONCLUDING REMARKS

In this article, we use a new data set on the location of R&D labs and several distance-based point pattern techniques to analyze the spatial concentration of the locations of more than 1,700 R&D labs in California and in a 10-state area in the Northeast Corridor of the United States. Rather than using a fixed spatial scale, we describe the spatial concentration of labs more precisely, by examining spatial structure at different scales using Monte Carlo tests based on Ripley's *K*-function. Geographic clusters at each scale are identified in terms of statistically significant departures from random locations reflecting the underlying distribution of economic activity. We present robust evidence that private R&D labs are indeed highly concentrated over a wide range of spatial scales.

We introduce a novel way to identify the spatial clustering of labs called *the multiscale core-cluster* approach. The analysis identifies four major clusters (one each in Boston, New York-northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.,) in the Northeast Corridor and three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

To verify that these local clusters are economically meaningful, we apply tests developed by JTH to measure the degree to which patent citations are localized in these clusters — tangible

evidence that knowledge spillovers are geographically mediated. For labs in the Northeast Corridor, we find, on average, that citations are about three to six times more likely to come from the same cluster as earlier patents than one would predict using a (control) sample of otherwise similar patents. In California, citations are roughly around three to five times more likely to come from the same cluster as earlier patents than one would predict using the control sample.

These localization ratios are at least as large as those reported by JTH, a conclusion that was in no way foregone since the spread of the Internet and patent databases drastically reduced the costs of searching patent applications by the early to mid-1990s. We also show that patents inside each cluster receive more citations on average than those outside the cluster in a suitably defined counterfactual area. In their study, JTH provide estimates of localization of knowledge spillovers that are averaged over the metro areas or states used in their study. But much information is lost regarding differences in the localization of knowledge spillovers in specific geographic areas. In this article, we show that such differences can be quite substantial. The results are robust to a number of alternative specifications for selecting control patents.

REFERENCES

Anselin, Luc. "Local Indicators of Spatial Association — LISA," *Geographical Analysis*, 27 (1995), pp. 93–115.

Arbia, Giuseppe, Giuseppe Espa, and Danny Quah. "A Class of Spatial Econometric Methods in the Empirical Analysis of Clusters of Firms in the Space," *Empirical Economics*, 34 (2008), pp. 81–103.

Audretsch, David B., and Maryann P. Feldman. "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review*, 86 (1996), pp. 630–40.

Berliant, Marcus, Robert R. Reed, III and Ping Wang. "Knowledge Exchange, Matching, and Agglomeration," *Journal of Urban Economics*, 60 (2006), pp. 69–95.

Besag, Julian, and James Newell. "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society*, 154 (1991), pp. 327–33.

Carlino, Gerald A. and William R. Kerr. "Agglomeration and Innovation," in: Henderson, J.Vernon, Duranton, Gilles, Strange, William (Eds.), *Handbook of Regional and Urban Economics*, Vol. 5A (2015), North Holland, Amsterdam.

Castro, Marcia C., and Burton H. Singer. "Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association," *Geographical Analysis*, 38 (2006), pp. 180–208.

Dalton, Donald Harold, and Manuel G. Serapio. "Globalizing Industrial Research and Development," Washington, D.C.: U.S. Department of Commerce, Office of Technology Policy (1995).

Directory of American Research and Technology, 23rd Ed. New York: R.R. Bowker (1999).

Duranton, Gilles, and Henry G. Overman. "Testing for Localization Using Micro-Geographic Data," *Review of Economic Studies*, 72 (2005), pp. 1077–106.

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

Ellison, Glenn, Edward L. Glaeser, and William Kerr. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review*, 100 (2010), pp. 1195–1213.

Getis, Arthur. "Interaction Modeling Using Second-Order Analysis," *Environment and Planning*, 16 (1984), pp. 173–83.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. "Market Value and Patent Citations," *RAND Journal of Economics*, 36 (2005), pp. 16–38.

Helsley, Robert, and William Strange. "Innovation and Input Sharing." *Journal of Urban Economics*, 51 (2002), pp. 25–45.

Hunt, Robert. "Matching Externalities and Inventive Productivity." Federal Reserve Bank of Philadelphia Working Paper 07-07 (2007).

Iacus, Stefano, Gary King, and Giuseppe Porro. "Causal Inference without Balance Checking: Coarsened Exact Matching," *Political Analysis* (2011).

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108 (1993), pp. 577–98.

Kerr, William R., and Scott Duke Kominers. "Agglomerative Forces and Cluster Shapes," *Review of Economics and Statistics*, 97 (2015), pp. 877–99.

Kulldorff, Martin. "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, 26 (1997), pp. 1487–96.

Lai, Ronald, Alexander D'Amour, and Lee Fleming. "The Careers and Co-authorship Networks of U.S. Patent-Holders Since 1975," mimeo, Harvard Business School (2009).

Marcon, Eric, and Florence Puech. "Evaluating the Geographic Concentration of Industries Using Distance-Based Methods," *Journal of Economic Geography*, 3 (2003), pp. 409–28.

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura. "Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach," *Review of Economics and Statistics*, 96 (2015), pp. 967–985.

National Science Foundation. *Research and Development in Industry: 1998*, Arlington, VA: National Science Foundation, Division of Science Resources Studies (2000).

Ripley, Brian D. "The Second-Order Analysis of Stationary Point Patterns," *Journal of Applied Probability*, 13 (1976), pp. 255–66.

Rosenthal, Stuart, and William C. Strange. "The Determinants of Agglomeration," *Journal of Urban Economics*, 50 (2001), pp. 191–229.

Shapiro, S.S., and M.B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52 (1965), pp. 591–611.

Thompson, Peter, and Melanie Fox-Kean. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 95 (2005), pp. 461–64.

Trajtenberg, Manuel. "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *RAND Journal of Economics*, 21 (1990), pp. 172–87.

U.S. Patent and Trademark Office. Overview of the U.S. Patent Classification System (USPC). Washington, D.C. (2012), http://www.uspto.gov/patents/resources/classification/overview.pdf.

Watson, Audrey. "BLS Statistics By Occupation." Bureau of Labor Statistics (2014).

Table 1: Summary Statistics					
Northeast (10-State)					
Variable	Mean	Std. Dev.	Median	Minimum	Maximum
	All Zip Codes (6,044)				
Land Area, miles^2	29.10	37.61	16.87	0.01	468.16
Radius*	2.55	1.66	2.32	0.06	12.21
Total Employment	4,307.22	8,994.78	1,001.00	0.00	194,114.00
Manufacturing Employment	557.20	1,213.46	76.30	0.00	22,808.31
Total Establishments	250.36	370.76	97.00	1.00	6,962.00
Manufacturing Establishments	11.39	16.65	4.00	0.00	132.00
Labs	0.17	0.74	0.00	0.00	13.00
	Zip Codes with 1 or More Labs (549)				
Land Area, miles2	20.95	29.46	12.04	0.06	361.79
Radius*	2.21	1.34	1.96	0.14	10.73
Total Employment	15,736.22	17,620.83	11,072.00	39.00	194,114.00
Manufacturing Employment	2,057.08	2,166.38	1,356.30	0.00	22,808.31
Total Establishments	697.51	574.58	568.50	6.00	6,962.00
Manufacturing Establishments	32.40	23.49	26.00	0.00	132.00
Labs	1.89	1.68	1.00	1.00	13.00
California					
Variable	Mean	Std. Dev.	Median	Minimum	Maximum
	All Zip Codes (1,646)				
Land Area, miles2	95.56	256.33	17.34	0.01	3,806.05
Radius*	3.84	3.96	2.35	0.06	34.81
Total Employment	5,989.95	9,758.35	1,700.00	0.00	79,766.00
Manufacturing Employment	858.14	2,394.39	64.50	0.00	27,186.00
Total Establishments	467.19	555.17	262.50	0.00	3,527.00
Manufacturing Establishments	30.18	61.83	8.00	0.00	776.00
Labs	0.39	2.01	0.00	0.00	33.00
	Zip Codes with 1 or More Labs (204)				
Land Area, miles2	18.78	37.75	8.19	0.07	385.98
Radius*	2.02	1.38	1.61	0.15	11.08
Total Employment	19,482.47	17,300.91	15,088.00	0.00	79,766.00
Manufacturing Employment	3,607.79	5,188.27	1,569.00	0.00	27,186.00
Total Establishments	1,173.13	677.45	1,065.50	0.00	3,527.00
Manufacturing Establishments	94.52	96.32	62.00	0.00	636.00
Labs	3.16	4.90	1.50	1.00	33.00

Sources: Author's calculations using the 1998 editions of the *Directory of American Research and Technology* and Zip Code Business Patterns

* Calculated assuming a zip code of circular shape with an area as reported in the data

Table 2a: Concentration of Labs by Industry in Northeast Corridor (*P-values*)[†]									
			Miles						
INDUSTRY	**SIC**	**LABS**	**0.25**	**0.5**	**0.75**	**1**	**5**	**20**	**50**
Metal Mining	10	4	0.5021	0.5029	0.5044	0.5052	0.5227	0.1674	0.4149
Oil and Gas Extraction	13	3	0.5011	0.5019	0.5026	0.5034	0.5137	0.0906	0.2286
Food	20	25	0.5825	0.6278	0.6750	0.7081	0.0984	0.2097	**0.0480**
Textile Mill	22	14	**0.0267**	**0.0465**	0.0690	0.0859	0.3468	0.7839	0.6446
Apparel	23	5	0.5036	0.5063	0.5082	0.5101	0.5399	0.7230	0.9088
Paper	26	28	0.6029	0.6596	0.7103	0.7460	0.4685	0.2833	0.3058
Printing & Publishing	27	3	0.5009	0.5012	0.5019	0.5024	0.5111	0.5837	0.7040
Chemicals	28	420	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0020**	**0.0001**
Petroleum Refining	29	24	0.0844	0.1380	0.1980	0.2425	0.3012	**0.0079**	**0.0358**
Rubber Products	30	38	0.6728	0.7493	0.8135	0.8544	0.5710	0.7974	0.9965
Stone, Clay, Glass, and Concrete Products	32	36	**0.0002**	**0.0008**	**0.0032**	**0.0011**	0.1041	0.7385	0.6886
Primary Metal Industries	33	36	0.6555	0.7284	0.7921	0.8327	0.7848	0.2592	0.4881
Fabricated Metal Products	34	44	**0.0004**	**0.0026**	**0.0101**	**0.0200**	0.0911	0.6985	0.8571
Industrial and Commercial Machinery	35	140	0.6024	0.7659	0.4192	0.4052	0.9910	0.9898	0.9867
Electronics	36	242	0.1958	0.5789	0.5825	0.7329	0.7058	0.8030	0.7423
Transportation Equipment	37	40	0.2277	0.3575	0.4867	0.5711	0.9594	0.9989	0.9744
Measuring, Analyzing, and Controlling Instruments	38	243	**0.0334**	0.1509	0.3838	0.3983	0.8171	0.8937	0.8778
Miscellaneous Manufacturing Industries	39	18	**0.0468**	0.0789	0.1126	0.1380	**0.0378**	0.1672	0.1093
Business Services	73	137	**0.0004**	**0.0052**	**0.0166**	**0.0055**	**0.0004**	**0.0001**	**0.0022**

[†]Concentration is conditional on the location of overall R&D labs. Bold indicates significantly more concentrated than overall labs at the 5 percent level of significance. Light gray indicates significantly more dispersed than overall labs at the 5 percent level of significance.

Source: Author's calculations using the 1998 editions of the Directory of American Research and Technology.

Table 2b: Concentration of Labs by Industry in California (*P-values*)[†]									
			Miles						
INDUSTRY	**SIC**	**LABS**	**0.25**	**0.5**	**0.75**	**1**	**5**	**20**	**50**
Oil and Gas Extraction	13	2	0.5015	0.5025	0.5040	0.5060	0.5455	0.6275	0.7010
Heavy Construction	16	2	0.5010	0.5015	0.5035	0.5055	0.5330	0.6210	0.1910
Food	20	3	0.5055	0.5100	0.5150	0.5185	0.5990	0.7700	0.4925
Paper	26	2	0.5020	0.5035	0.5045	0.5080	0.5340	0.6175	0.1970
Chemicals	28	129	**0.0025**	**0.0100**	**0.0170**	0.0705	0.9670	0.9920	0.9480
Petroleum Refining	29	2	0.5005	0.5025	0.5040	0.5065	0.5385	0.6105	0.6875
Rubber Products	30	8	**0.0235**	0.0535	0.0980	0.1320	0.4020	0.3660	0.1630
Stone, Clay, Glass, and Concrete Products	32	6	0.5125	0.5290	0.5515	0.5695	0.7950	0.7075	0.4215
Primary Metal Industries	33	11	**0.0435**	0.1130	0.1780	0.2455	0.8770	0.7235	0.2865
Fabricated Metal Products	34	16	0.5925	0.6840	0.7670	0.8235	0.9890	0.4555	0.1765
Industrial and Commercial Machinery	35	99	**0.0140**	**0.0100**	**0.0105**	**0.0120**	**0.0020**	**0.0010**	**0.0205**
Electronics	36	211	**0.0450**	**0.0030**	**0.0075**	**0.0030**	**0.0010**	**0.0030**	0.1040
Transportation Equipment	37	36	**0.0010**	**0.0030**	**0.0030**	**0.0030**	0.4635	0.2635	0.1570
Measuring, Analyzing, and Controlling Equipment	38	134	**0.0010**	**0.0480**	0.2165	0.4610	0.8845	0.9960	1.0000
Miscellaneous Manufacturing Industries	39	8	0.5285	0.5620	0.5980	0.6280	0.9000	0.7310	0.7205
Business Services	73	147	**0.0300**	**0.0150**	**0.0105**	**0.0045**	**0.0020**	**0.0010**	**0.0010**

[†]Concentration is conditional on the location of overall R&D labs. Bold indicates significantly more concentrated than overall labs at the 5 percent level of significance. Light gray indicates significantly more dispersed than overall labs at the 5 percent level of significance.

Source: Author's calculations using the 1998 editions of the Directory of American Research and Technology.

Table 3a: Five-Mile Clusters in the Northeast Corridor, Baseline Results

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L		
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic		
Framingham–Marlborough–Westborough, MA	323	3,498	104	2.97%	2,941	87	2.96%	2,941	0	0.00%	**N/A**	**9.5**		
Boston–Cambridge–Waltham–Woburn, MA	2,634	27,664	1,717	6.21%	23,614	1,468	6.22%	23,614	256	1.08%	**5.7**	**30.0**		
Silver Spring–Bethesda, MD–McLean, VA	367	3,424	89	2.60%	2,843	70	2.46%	2,843	3	0.11%	**23.3**	**7.9**		
Trenton–Princeton, NJ	889	9,022	260	2.88%	7,547	224	2.97%	7,547	23	0.30%	**9.7**	**13.0**		
Parsippany–Morristown–Union, NJ	1,710	14,567	358	2.46%	12,337	314	2.55%	12,337	69	0.56%	**4.6**	**12.7**		
Greenwich–Stamford, CT–Scarsdale, NY	1,205	11,218	141	1.26%	9,477	115	1.21%	9,477	36	0.38%	**3.2**	**6.5**		
Stratford–Milford, CT	235	1,484	12	0.81%	1,280	10	0.78%	1,280	0	0.00%	**N/A**	**3.2**		
Conshohocken–King of Prussia–West Chester, PA	539	2,352	68	2.89%	2,111	59	2.79%	2,111	4	0.19%	**14.8**	**7.0**		
Wilmington–New Castle, DE	624	3,501	72	2.06%	3,055	61	2.00%	3,055	11	0.36%	**5.5**	**5.9**		
All Five-Mile Clusters	8,526	76,730	2,821	3.68%	65,205	2,408	3.69%	65,205	402	0.62%	**6.0**	**38.5**		

Table 3b: 10-Mile Clusters in the Northeast Corridor, Baseline Results

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L		
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic		
Boston, MA	4,719	48,315	4,263	8.82%	41,082	3,679	8.96%	41,082	747	1.82%	**4.9**	**45.9**		
Washington, DC	926	9,741	327	3.36%	8,089	270	3.34%	8,089	31	0.38%	**8.7**	**14.0**		
New York, NY	7,768	67,982	4,738	6.97%	57,626	3,997	6.94%	57,626	1,493	2.59%	**2.7**	**34.8**		
Philadelphia, PA	1,594	9,028	409	4.53%	7,851	343	4.37%	7,851	35	0.45%	**9.8**	**16.2**		
All 10-Mile Clusters	15,007	135,066	9,737	7.21%	114,648	8,289	7.23%	114,648	2,306	2.01%	**3.6**	**60.0**		

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Source: NBER Patent Data Project and authors' calculations.

Table 4a: Five-Mile Clusters in California, Baseline Results

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego	444	3,434	77	2.24%	2,914	67	2.30%	2,914	9	0.31%	**7.4**	**6.7**
Los Angeles	454	3,646	104	2.85%	3,143	91	2.90%	3,143	2	0.06%	**45.5**	**9.4**
Palo Alto–San Jose	11,318	145,471	26,684	18.34%	121,455	22,407	18.45%	121,455	4,986	4.11%	**4.5**	**114.7**
Dublin–Pleasanton	283	3,899	127	3.26%	3,257	110	3.38%	3,257	5	0.15%	**22.0**	**10.0**
All Five-Mile Clusters	12,499	156,450	26,992	17.25%	130,769	22,675	17.34%	130,769	5,002	3.83%	**4.5**	**115.2**

Table 4b: 10-Mile Clusters in California, Baseline Results

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego	2,099	20,079	970	4.83%	16,951	844	4.98%	16,951	176	1.04%	**4.8**	**21.4**
Los Angeles	1,266	10,685	609	5.70%	9,264	537	5.80%	9,264	62	0.67%	**8.7**	**19.9**
San Francisco	14,963	188,943	44,215	23.40%	157,997	37,184	23.53%	157,997	8,907	5.64%	**4.2**	**147.3**
All 10-Mile Clusters	18,328	219,707	45,794	20.84%	184,212	38,565	20.94%	184,212	9,145	4.96%	**4.2**	**148.6**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Source: NBER Patent Data Project and authors' calculations.

Table 5a: Citation Location Differentials and Spatial Scale (Northeast Corridor)

Cluster Size	# of Clusters	Originating Patents	Citing Patents	Treatment Proportion (%)	Control Proportion (%)	Localization Differential	*t*-statistic
5-Mile	9	8,526	76,737	3.69	0.60	6.2	41.8
10-Mile	4	15,007	135,075	7.23	2.44	3.0	58.0
20-Mile	3	21,941	191,685	9.82	4.82	2.0	59.4

Source: NBER Patent Data Project

Table 5b: Citation Location Differentials and Spatial Scale (California)

Cluster Size	# of Clusters	Originating Patents	Citing Patents	Treatment Proportion (%)	Control Proportion (%)	Localization Differential	*t*-statistic
5-Mile	4	12,499	156,450	17.30	1.48	11.7	156.7
10-Mile	3	18,328	219,705	20.89	2.12	9.8	202.7
20-Mile	2	18,523	223,285	22.55	2.52	9.0	210.9

Source: NBER Patent Data Project and authors' calculations

Control Patents are chosen to have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Table 6: Citation Differentia Between Labs Inside Clusters vs. Labs Outside Clusters (Difference in Means† Test)

	Inside Cluster[1]			Outside Cluster[2]			
Area	Mean	Std. Dev.	n	Mean	Std. Dev.	n	*t*-statistic
Boston	12.888	18.148	4,704	9.949	14.895	2,644	7.491
New York	11.065	16.338	8,279	9.491	14.410	10,600	6.912
Philadelphia	8.030	9.657	1,598	7.654	10.515	3,655	1.262
Washington, D.C.	11.707	17.457	1,273	7.825	10.371	1,741	7.073
Southern California	11.464	15.734	3,668	9.087	12.074	6,716	7.956
Northern California	15.532	19.845	15,106	10.811	15.110	2,680	14.155

Source: NBER Patent Data Project and authors' calculations

†: Citations per Patent Granted, 1996-1997

1: Inside Cluster refers to all patents in one or more 10-mile clusters in the region.

2: Outside Cluster refers to all patents outside of the 10-mile clusters in the regions defined as follows: Boston (Massachusetts/New Hampshire/Rhode Island), New York (New York/Connecticut/Northern NJ), Philadelphia (Delaware/Eastern Pennsylvania/Southern NJ), Washington, D.C. (Maryland/D.C./Virginia), Southern California (10 southern counties), and Northern California (remaining counties).

Table 7a: Five-Mile Clusters in the Northeast Corridor, STEM Worker Clusters

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Bethesda–Rockville, MD–Vienna, VA	414	4,291	100	2.33%	3,499	75	2.14%	3,499	9	0.26%	**8.3**	**7.3**
Columbia–Laurel, MD	53	497	3	0.60%	453	3	0.66%	453	0	0.00%	**N/A**	**1.7**
Phoenix–Cockeysville, MD	72	419	0	0.00%	363	0	0.00%	363	0	0.00%	**N/A**	**N/A**
Wilmington, DE	539	2,352	68	2.89%	2,093	57	2.72%	2,093	5	0.24%	**11.4**	**6.7**
King of Prussia, PA	974	5,535	242	4.37%	4,848	207	4.27%	4,848	15	0.31%	**13.8**	**13.2**
Philadelphia, PA	81	617	6	0.97%	544	5	0.92%	544	0	0.00%	**N/A**	**2.2**
Princeton, NJ–New York, NY	5,124	46,014	2,323	5.05%	38,804	1,960	5.05%	38,804	684	1.76%	**2.9**	**25.4**
Long Island, NY	270	1,913	18	0.94%	1,692	17	1.00%	1,692	1	0.06%	**17.0**	**3.8**
Danbury, CT	347	4,410	162	3.67%	3,772	126	3.34%	3,772	2	0.05%	**63.0**	**11.1**
Stratford, CT	240	1,501	12	0.80%	1,309	12	0.92%	1,309	1	0.08%	**12.0**	**3.1**
North Haven, CT	105	457	13	2.84%	411	13	3.16%	411	0	0.00%	**N/A**	**3.7**
Hartford, CT	87	503	8	1.59%	452	7	1.55%	452	0	0.00%	**N/A**	**2.7**
Hudson–Westborough, MA	255	2,841	84	2.96%	2,368	77	3.25%	2,368	3	0.13%	**25.7**	**8.4**
Boston–Cambridge, MA	2,958	30,920	2,059	6.66%	26,437	1,780	6.73%	26,437	326	1.23%	**5.5**	**32.7**
Nashua, NH	295	2,966	54	1.82%	2,521	44	1.75%	2,521	1	0.04%	**44.0**	**6.5**
Binghamton, NY	23	332	0	0.00%	300	0	0.00%	300	0	0.00%	**N/A**	**N/A**
Syracuse, NY	40	238	15	6.30%	212	12	5.66%	212	0	0.00%	**N/A**	**3.6**
Buffalo, NY	91	410	1	0.24%	377	1	0.27%	377	0	0.00%	**N/A**	**1.0**
Pittsburgh, PA	42	165	2	1.21%	148	2	1.35%	148	0	0.00%	**N/A**	**1.4**
Pittsburgh–Verona, PA	70	426	4	0.94%	381	4	1.05%	381	0	0.00%	**N/A**	**2.0**
All Five-Mile Clusters	12,080	106,807	5,174	4.84%	90,984	4,402	4.84%	90,984	1,047	1.15%	**4.2**	**46.4**

Table 7b: 10-Mile Clusters in the Northeast Corridor, STEM Worker Clusters

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Richmond, VA	154	668	71	10.63%	604	68	11.26%	604	0	0.00%	**N/A**	**8.8**
Washington, DC–Baltimore, MD	1,376	12,724	538	4.23%	10,655	462	4.34%	10,655	71	0.67%	**6.5**	**17.3**
Hagerstown, MD	17	40	1	2.50%	39	1	2.56%	39	0	0.00%	**N/A**	**1.0**
Lancaster, PA	104	566	8	1.41%	514	7	1.36%	514	0	0.00%	**N/A**	**2.7**
Philadelphia, PA–Wilmington, DE–Cherry Hill, NJ	2,601	14,166	992	7.00%	12,424	870	7.00%	12,424	109	0.88%	**8.0**	**25.1**
Pittsburgh, PA	921	5,804	400	6.89%	5,101	351	6.88%	5,101	17	0.33%	**20.6**	**18.0**
Binghamton, NY	329	3,128	31	0.99%	2,640	29	1.10%	2,640	2	0.08%	**14.5**	**4.9**
Syracuse, NY	130	678	44	6.49%	615	41	6.67%	615	0	0.00%	**N/A**	**6.6**
Rochester, NY	1,571	7,983	391	4.90%	6,853	345	5.03%	6,853	23	0.34%	**15.0**	**17.2**
Buffalo, NY	122	632	3	0.47%	578	3	0.52%	578	0	0.00%	**N/A**	**1.7**
Boston, MA	4,682	47,968	3,901	8.13%	40,735	3,356	8.24%	40,735	737	1.81%	**4.6**	**42.5**
New York, NY–Northern NJ–CT	9,514	80,971	6,239	7.71%	68,831	5,313	7.72%	68,831	2,286	3.32%	**2.3**	**35.9**
All 10-Mile Clusters	21,521	175,328	12,619	7.20%	149,589	10,846	7.25%	149,589	3,245	2.17%	**3.3**	**66.1**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

The clusters identified in the above table are based on STEM workers as the backcloth. Note that the cluster definitions change because the backcloth changed to STEM workers instead of manufacturing workers used in Tables 3 and 4.

Source: NBER Patent Data Project and authors' calculations.

Table 8a: Five-Mile Clusters in California, STEM Worker Clusters

						Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego–La Jolla	563	4,134	119	2.88%	3,518	111	3.16%	3,518	9	0.26%	**12.3**	**9.5**
Carlsbad	261	1,628	43	2.64%	1,443	36	2.49%	1,443	0	0.00%	**N/A**	**6.1**
Irvine	946	7,466	375	5.02%	6,589	325	4.93%	6,589	33	0.50%	**9.8**	**15.8**
Camarillo	199	1,943	39	2.01%	1,704	30	1.76%	1,704	1	0.06%	**30.0**	**5.3**
Santa Barbara	82	1,401	55	3.93%	1,222	52	4.26%	1,222	1	0.08%	**52.0**	**7.2**
San Jose–Santa Clara	14,220	182,445	42,563	23.33%	152,229	35,803	23.52%	152,229	7,956	5.23%	**4.5**	**149.0**
Pleasanton	283	3,899	127	3.26%	3,284	111	3.38%	3,284	8	0.24%	**13.9**	**9.6**
Santa Rosa	127	1,013	29	2.86%	903	27	2.99%	903	0	0.00%	**N/A**	**5.3**
All Five-Mile Clusters	16,681	203,929	43,350	21.26%	170,892	36,495	21.36%	170,892	8,008	4.69%	**4.6**	**149.4**

Table 8b: Ten-Mile Clusters in California, STEM Worker Clusters

						Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego	2,146	20,504	1,056	5.15%	17,232	926	5.37%	17,232	171	0.99%	**5.4**	**23.3**
Anaheim–Irvine	1,911	15,353	1,063	6.92%	13,410	929	6.93%	13,410	115	0.86%	**8.1**	**26.0**
Oxnard–Camarillo	76	475	15	3.16%	432	13	3.01%	432	0	0.00%	**N/A**	**3.7**
Santa Barbara	288	3,299	129	3.91%	2,871	118	4.11%	2,871	4	0.14%	**29.5**	**10.5**
San Francisco–Palo Alto–San Jose	14,564	185,644	44,114	23.76%	154,996	37,127	23.95%	154,996	8,314	5.36%	**4.5**	**151.6**
Santa Rosa	144	1,197	54	4.51%	1,061	48	4.52%	1,061	0	0.00%	**N/A**	**7.1**
All 10-Mile Clusters	19,129	226,472	46,431	20.50%	190,002	39,161	20.61%	190,002	8,604	4.53%	**4.6**	**154.1**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.
The clusters identified in the above table are based on STEM workers as the backcloth. Note that the cluster definitions change because the backcloth changed to STEM workers instead of manufacturing workers used in Tables 3 and 4.
Source: NBER Patent Data Project and authors' calculations.

Table 9a: Five-Mile Clusters in the Northeast Corridor, Disaggregated Subclasses

						Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Framingham–Marlborough–Westborough, MA	323	3,498	104	2.97%	2,915	90	3.09%	2,915	2	0.07%	**45.0**	**9.3**
Boston–Cambridge–Waltham–Woburn, MA	2,634	27,664	1,717	6.21%	23,126	1,470	6.36%	23,126	235	1.02%	**6.3**	**30.8**
Silver Spring–Bethesda, MD–McLean, VA	367	3,424	89	2.60%	2,765	74	2.68%	2,765	10	0.36%	**7.4**	**7.1**
Trenton–Princeton, NJ	889	9,022	260	2.88%	7,420	226	3.05%	7,420	15	0.20%	**15.1**	**13.8**
Parsippany–Morristown–Union, NJ	1,710	14,567	358	2.46%	11,889	303	2.55%	11,889	78	0.66%	**3.9**	**11.7**
Greenwich-Stamford, CT–Scarsdale, NY	1,205	11,218	141	1.26%	9,222	104	1.13%	9,222	31	0.34%	**3.4**	**6.3**
Stratford–Milford-CT	235	1,484	12	0.81%	1,262	8	0.63%	1,262	1	0.08%	**8.0**	**2.3**
Conshohocken–King of Prussia-West Chester, PA	539	2,352	68	2.89%	1,929	54	2.80%	1,929	7	0.36%	**7.7**	**6.1**
Wilmington–New Castle, DE	624	3,501	72	2.06%	2,940	61	2.07%	2,940	6	0.20%	**10.2**	**6.8**
All Five-Mile Clusters	8,526	76,730	2,821	3.68%	63,468	2,390	3.77%	63,468	385	0.61%	**6.2**	**38.7**

Table 9b: 10-Mile Clusters in the Northeast Corridor, Disaggregated Subclasses

						Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Boston, MA	4,719	48,315	4,263	8.82%	40,317	3,612	8.96%	40,317	722	1.79%	**5.0**	**45.7**
Washington, DC	926	9,741	327	3.36%	7,849	266	3.39%	7,849	42	0.54%	**6.3**	**13.0**
New York, NY	7,768	67,982	4,738	6.97%	55,955	3,751	6.70%	55,955	1,426	2.55%	**2.6**	**33.3**
Philadelphia, PA	1,594	9,028	409	4.53%	7,497	344	4.59%	7,497	41	0.55%	**8.4**	**15.8**
All 10-Mile Clusters	15,007	135,066	9,737	7.21%	111,618	7,973	7.14%	111,618	2,231	2.00%	**3.6**	**58.6**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same six-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Source: NBER Patent Data Project and authors' calculations.

Table 10a: Five-Mile Clusters in California, Disaggregated Subclasses

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego	444	3,434	77	2.24%	2,887	54	1.87%	2,887	5	0.17%	**10.8**	**6.4**
Los Angeles	454	3,646	104	2.85%	3,005	86	2.86%	3,005	2	0.07%	**43.0**	**9.1**
Palo Alto–San Jose	11,318	145,471	26,684	18.34%	119,907	22,116	18.44%	119,907	4,974	4.15%	**4.4**	**113.5**
Dublin–Pleasanton	283	3,899	127	3.26%	3,269	108	3.30%	3,269	4	0.12%	**27.0**	**10.0**
All 5-Mile Clusters	12,499	156,450	26,992	17.25%	129,068	22,364	17.33%	129,068	4,985	3.86%	**4.5**	**113.9**

Table 10b: 10-Mile Clusters in California, Disaggregated Subclasses

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
San Diego	2,099	20,079	970	4.83%	16,629	819	4.93%	16,629	159	0.96%	**5.2**	**21.6**
Los Angeles	1,266	10,685	609	5.70%	8,897	484	5.44%	8,897	43	0.48%	**11.3**	**19.7**
San Francisco	14,963	188,943	44,215	23.40%	155,861	36,534	23.44%	155,861	8,803	5.65%	**4.2**	**145.6**
All 10-Mile Clusters	18,328	219,707	45,794	20.84%	181,387	37,837	20.86%	181,387	9,005	4.96%	**4.2**	**146.9**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control Patents are chosen to have the same six-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Source: NBER Patent Data Project and authors' calculations.

Table 11a: Five-Mile Clusters in the Northeast Corridor, Coarsened Exact Matching

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Framingham–Marlborough–Westborough, MA	323	3,498	104	2.97%	2,845	80	2.81%	2,845	9	0.32%	**8.9**	**7.6**
Boston–Cambridge–Waltham–Woburn, MA	2,634	27,664	1,717	6.21%	22,937	1,400	6.10%	22,937	284	1.24%	**4.9**	**27.9**
Silver Spring–Bethesda, MD–McLean, VA	367	3,424	89	2.60%	2,779	69	2.48%	2,779	15	0.54%	**4.6**	**6.0**
Trenton–Princeton, NJ	889	9,022	260	2.88%	7,453	207	2.78%	7,453	25	0.34%	**8.3**	**12.1**
Parsippany–Morristown–Union, NJ	1,710	14,567	358	2.46%	11,912	282	2.37%	11,912	91	0.76%	**3.1**	**10.0**
Greenwich–Stamford, CT–Scarsdale, NY	1,205	11,218	141	1.26%	9,277	109	1.17%	9,277	49	0.53%	**2.2**	**4.8**
Stratford–Milford, CT	235	1,484	12	0.81%	1,228	11	0.90%	1,228	2	0.16%	**5.5**	**2.5**
Conshohocken–King of Prussia–West Chester, PA	539	2,352	68	2.89%	1,964	53	2.70%	1,964	13	0.66%	**4.1**	**5.0**
Wilmington–New Castle, DE	624	3,501	72	2.06%	2,940	53	1.80%	2,940	11	0.37%	**4.8**	**5.3**
All 5-Mile Clusters	8,526	76,730	2,821	3.68%	63,335	2,264	3.57%	63,335	499	0.79%	**4.5**	**34.1**

Table 11b: 10-Mile Clusters in the Northeast Corridor, Coarsened Exact Matching

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Boston, MA	4,719	48,315	4,263	8.82%	39,760	3,493	8.79%	39,760	896	2.25%	**3.9**	**40.7**
Washington, DC	926	9,741	327	3.36%	7,851	250	3.18%	7,851	58	0.74%	**4.3**	**11.1**
New York, NY	7,768	67,982	4,738	6.97%	55,989	3,706	6.62%	55,989	1,710	3.05%	**2.2**	**27.9**
Philadelphia, PA	1,594	9,028	409	4.53%	7,603	327	4.30%	7,603	68	0.89%	**4.8**	**13.3**
All 10-Mile Clusters	15,007	135,066	9,737	7.21%	111,203	7,776	6.99%	111,203	2,732	2.46%	**2.8**	**50.7**

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control patents are selected using the coarsened exact matching procedure Control patents must have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Control patents must have the same application year and three-digit technology classification as the treatment patents, in addition to having the same grant year and the number of citations that the treatment patent receives.

Source: NBER Patent Data Project and authors' calculations

Table 12a: Five-Mile Clusters in California, Coarsened Exact Matching

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L		
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic		
San Diego	444	3,434	77	2.24%	2,811	58	2.06%	2,811	14	0.50%	**4.1**	**5.2**		
Los Angeles	454	3,646	104	2.85%	3,019	79	2.62%	3,019	5	0.17%	**15.8**	**8.2**		
Palo Alto–San Jose	11,318	145,471	26,684	18.34%	118,537	21,223	17.90%	118,537	8,962	7.56%	**2.4**	**76.5**		
Dublin–Pleasanton	283	3,899	127	3.26%	3,199	87	2.72%	3,199	9	0.28%	**9.7**	**8.1**		
All 5-Mile Clusters	12,499	156,450	26,992	17.25%	127,566	21,447	16.81%	127,566	8,990	7.05%	**2.4**	**77.0**		

Table 12b: 10-Mile Clusters in California, Coarsened Exact Matching

						Treatment Group				Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L		
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents†	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic		
San Diego	2,099	20,079	970	4.83%	16,392	801	4.89%	16,392	335	2.04%	**2.4**	**14.1**		
Los Angeles	1,266	10,685	609	5.70%	8,915	457	5.13%	8,915	90	1.01%	**5.1**	**16.1**		
San Francisco	14,963	188,943	44,215	23.40%	154,195	35,457	22.99%	154,195	14,455	9.37%	**2.5**	**104.5**		
All 10-Mile Clusters	18,328	219,707	45,794	20.84%	179,502	36,715	20.45%	179,502	14,880	8.29%	**2.5**	**105.5**		

*The subset of citing patents for which we obtained a similar control patent. See text for details.

†Control patents are selected using the coarsened exact matching procedure Control patents must have the same three-digit technology classification as the citing patent and its application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to whom the originating patent is assigned.

Control patents must have the same application year and three-digit technology classification as the treatment patents, in addition to having the same grant year and the number of citations that the treatment patent receives.

Source: NBER Patent Data Project and authors' calculations.

Table 13†		
Northeast		
Cluster Name	Coefficient on Originating Patent ($\widehat{D}_i$)	Standard Errors
Boston5A	2.82	0.1062*
Boston5B	1.5	0.0300*
NY5A	2.17	0.0737*
NY5B	1.26	0.0603*
NY5C	0.8	0.0967*
NY5D	2.26	0.3235*
Philly5A	3.13	0.1321*
Philly5B	2.28	0.1335*
Boston10	1.37	0.0199*
DC10	1.65	0.0652*
NY10	0.79	0.0192*
Philly10	2.13	0.0574*
	Broad Regions	
NE5	0.77	0.0167*
NE10	0.68	0.0113*
California		
Cluster Name	Coefficient on Originating Patent ($\widehat{D}_i$)	Standard Errors
SD5	2.34	0.1251*
LA5	2.52	0.1137*
SF5A	1.06	0.0107*
SF5B	2.81	0.1098*
SD10	1.56	0.0381*
LA10	2.06	0.0493*
SF10	1.09	0.0093*
Broad Regions		
CA5	1.01	0.0103*
CA10	0.99	0.0086*

†The California regressions included 1,390,727 observations. The Northeast Corridor regressions included 1,444,272 observations. Robust standard errors are reported.
*Indicates significance at the 1 percent level.

Table 14: Summary of Location Differentials					
	Northeast Corridor			California	
	Five-Mile Cluster	10-Mile Cluster		Five-Mile Cluster	10-Mile Cluster
Baseline	6.0	3.6		4.5	4.2
STEM	4.2	3.3		4.6	4.6
Disaggregated	6.2	3.6		4.5	4.2
CEM	4.5	2.8		2.4	2.5

[†]Baseline results from column K in Tables 3 and 4; STEM results from column K in Tables 7 and 8; Disaggregated results from column K in Tables 9 and 10; and CEM results from column K in Tables 11 and 12.

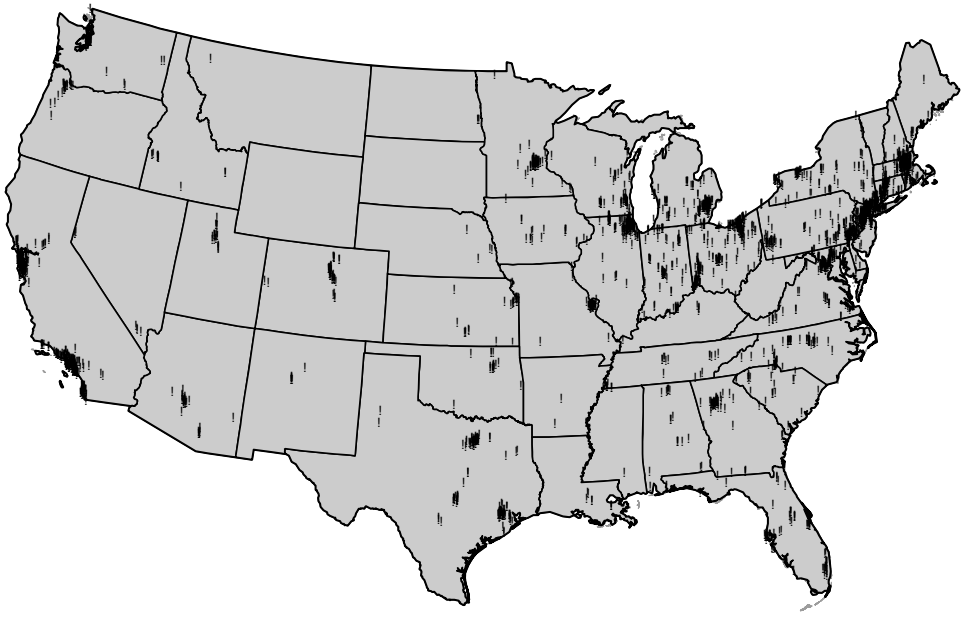


Figure 1: Location of R&D Labs

Source: *Directory of American Research and Technology* and authors' calculations

Each dot on the map represents the location of a single R&D lab. In areas with a dense cluster of labs, the dots tend to sit on top of one another, representing a spatial cluster of labs.

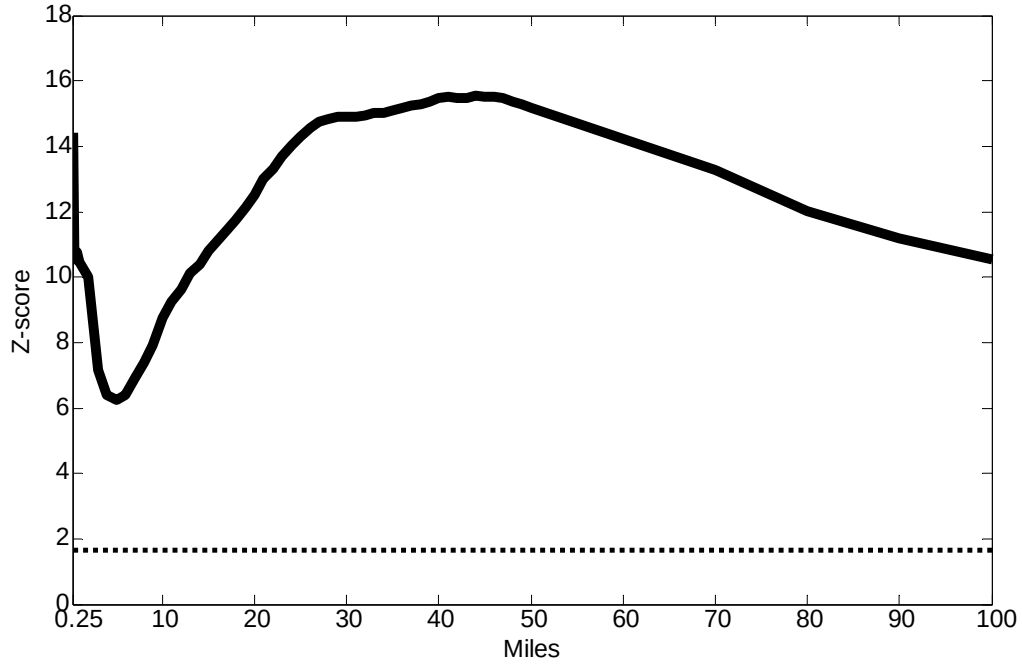


Figure 2a: *Z*-scores for Northeast Corridor

Dotted line *Z* = 1.65

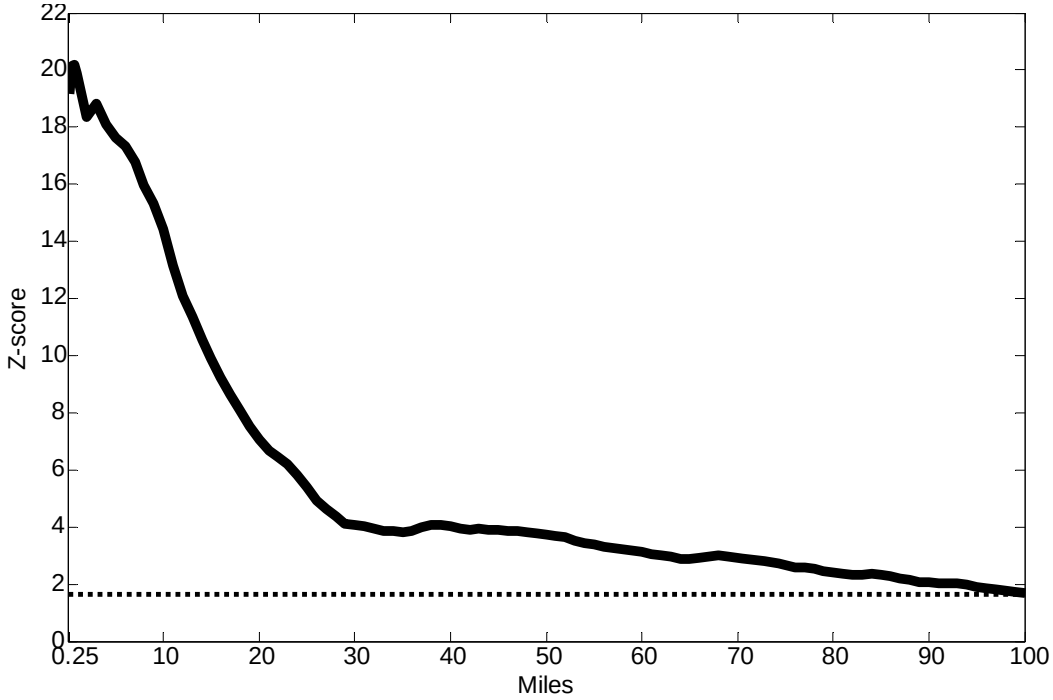


Figure 2b: *Z*-scores for California

Dotted line *Z* = 1.65

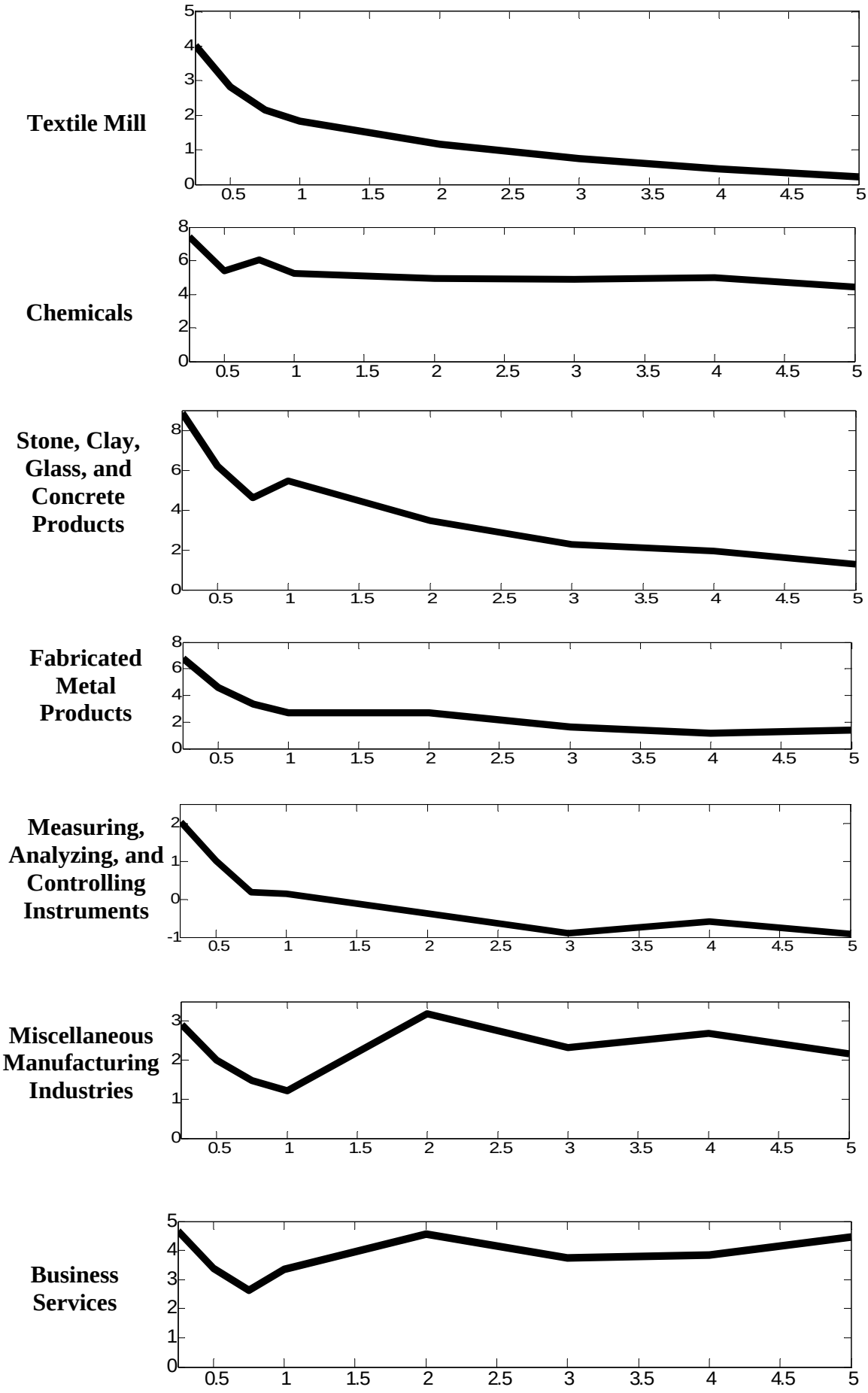


Figure 3a. Northeast Corridor Industry *Z*-Scores

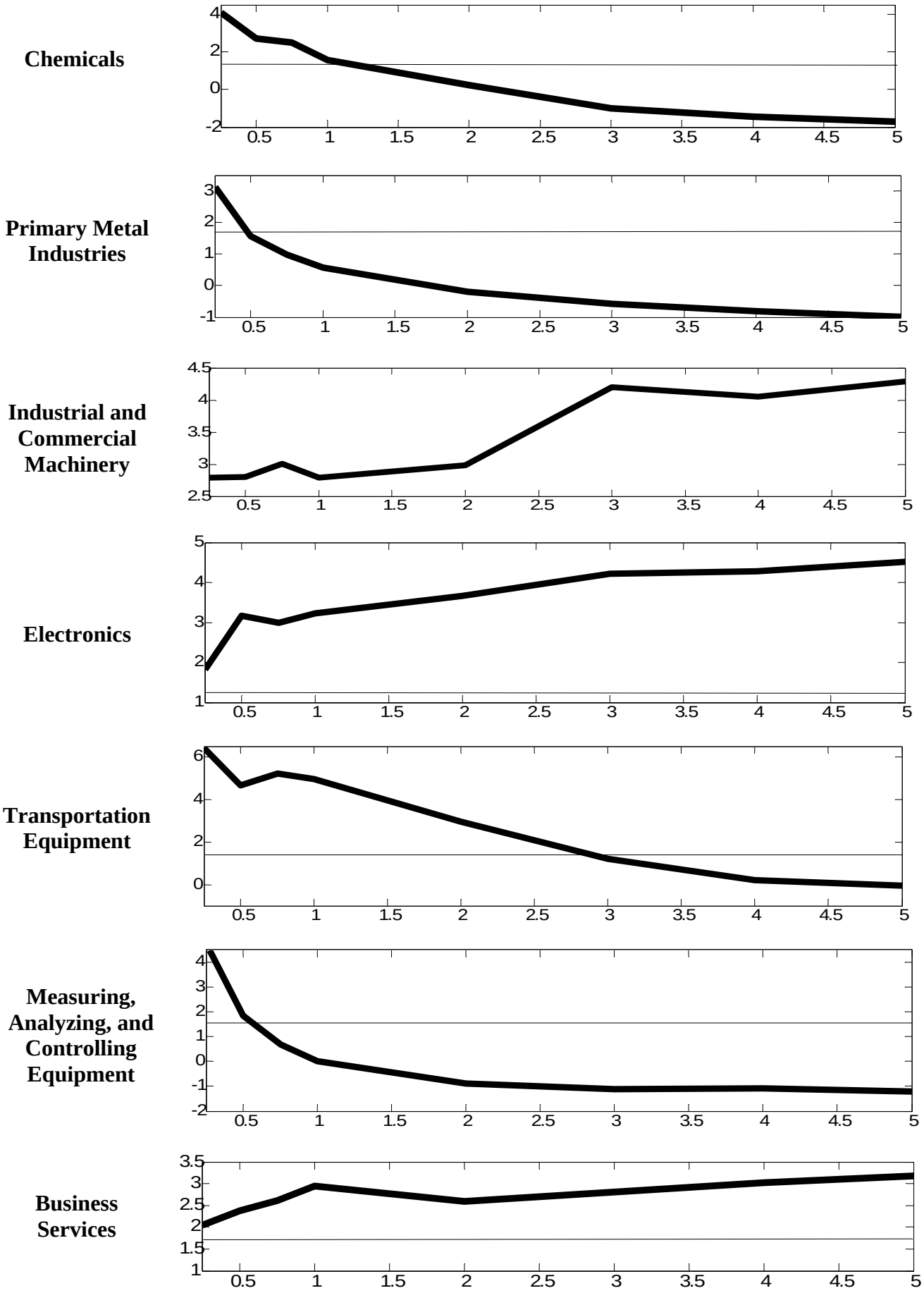
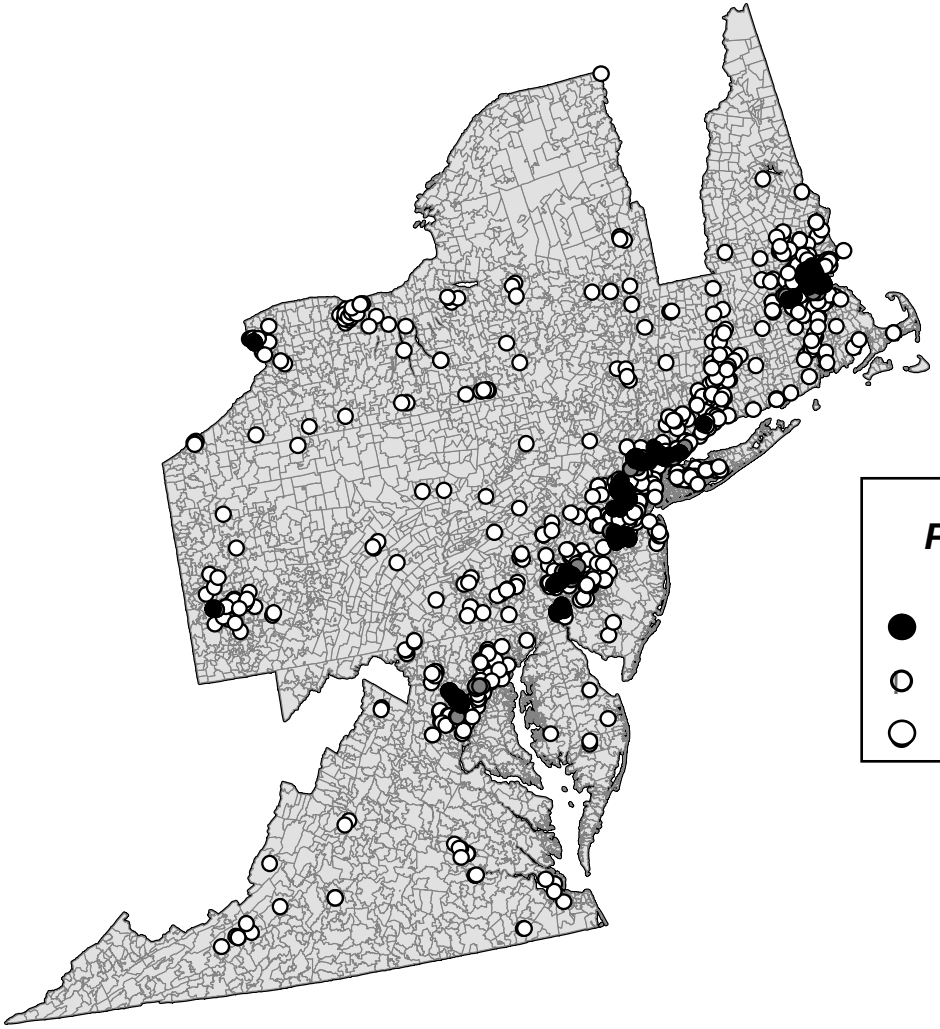


Figure 3b: California Industry Z-Scores



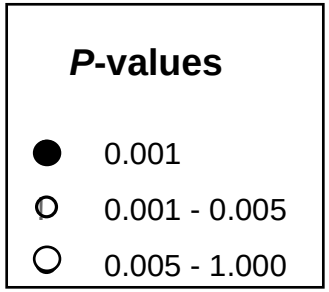


Figure 4a: Northeast Corridor *P*-values at *d* = 5 miles

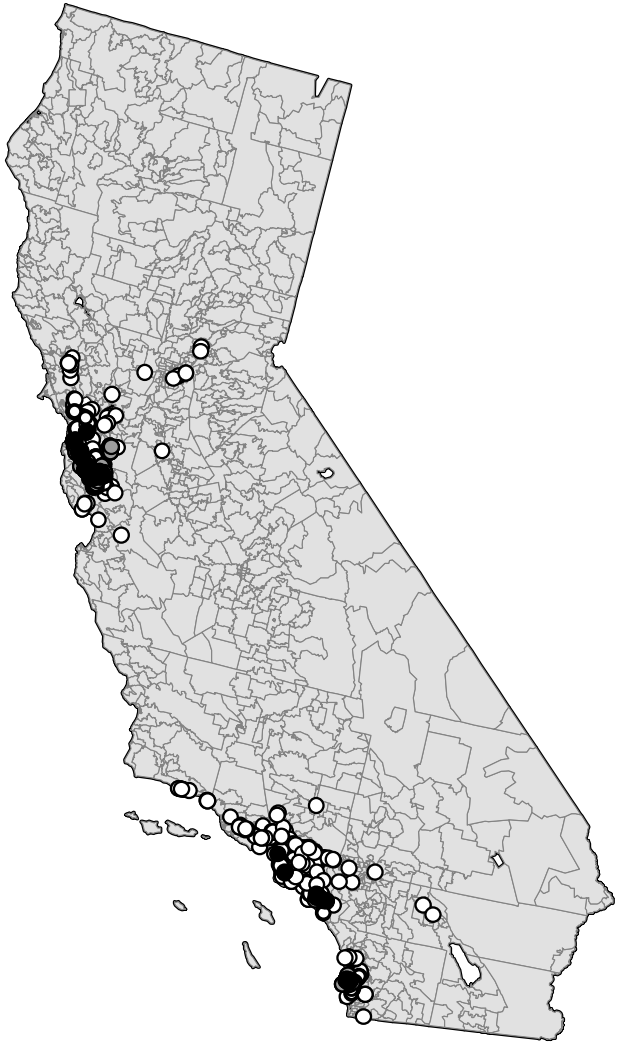


Figure 4b: California *P*-values at *d* = 5 miles

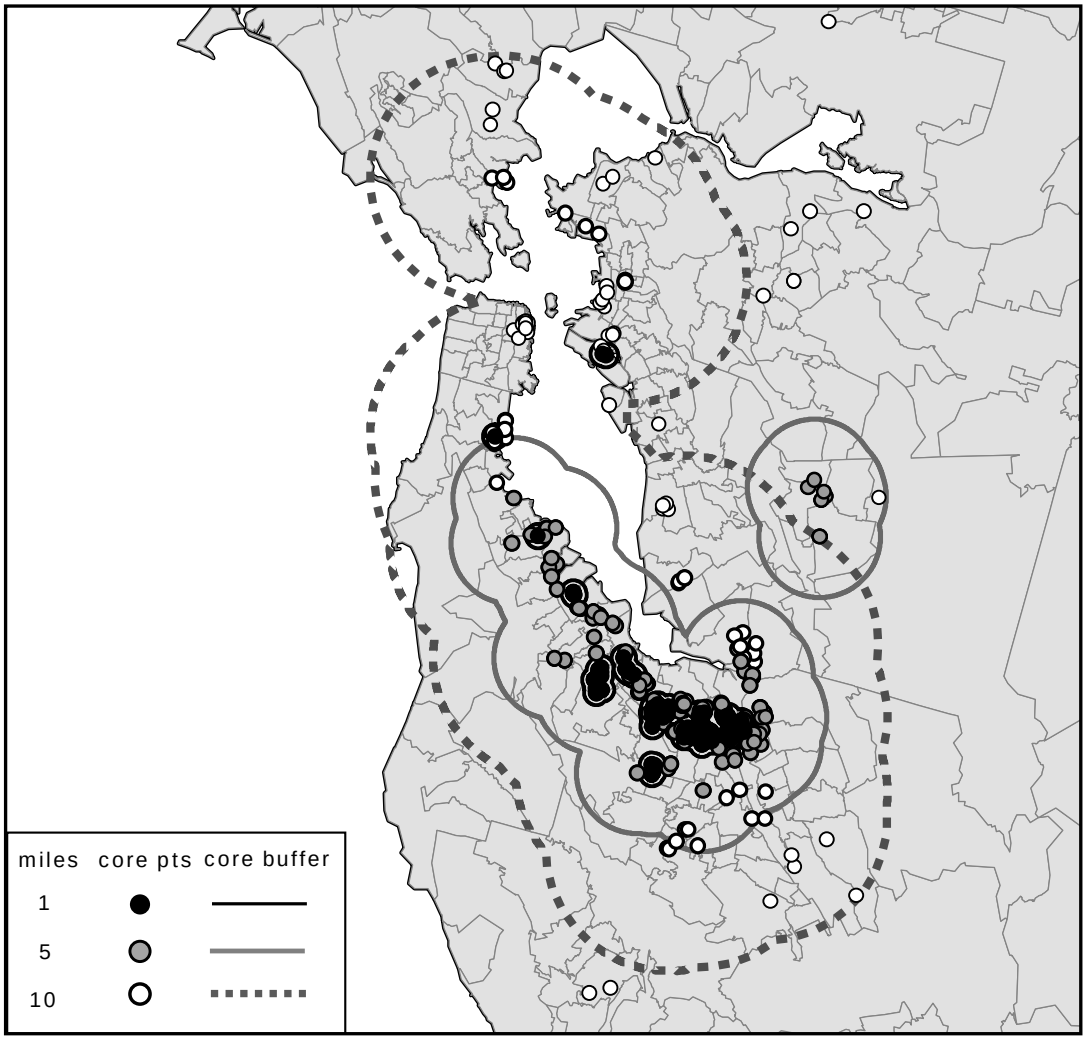


Figure 5: Multiscale Core Clusters in the San Francisco Bay Area

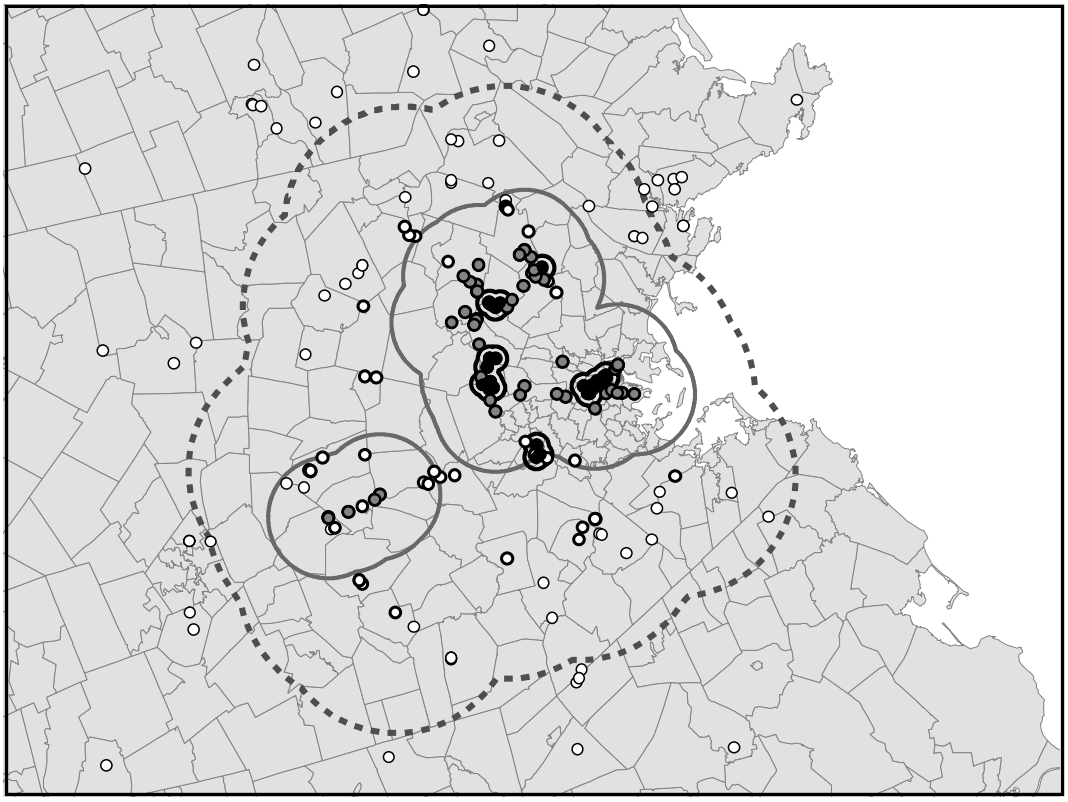


Figure 6a: Multiscale Core Clusters in Boston

Figure 6b: Proximity to Major Routes in Boston

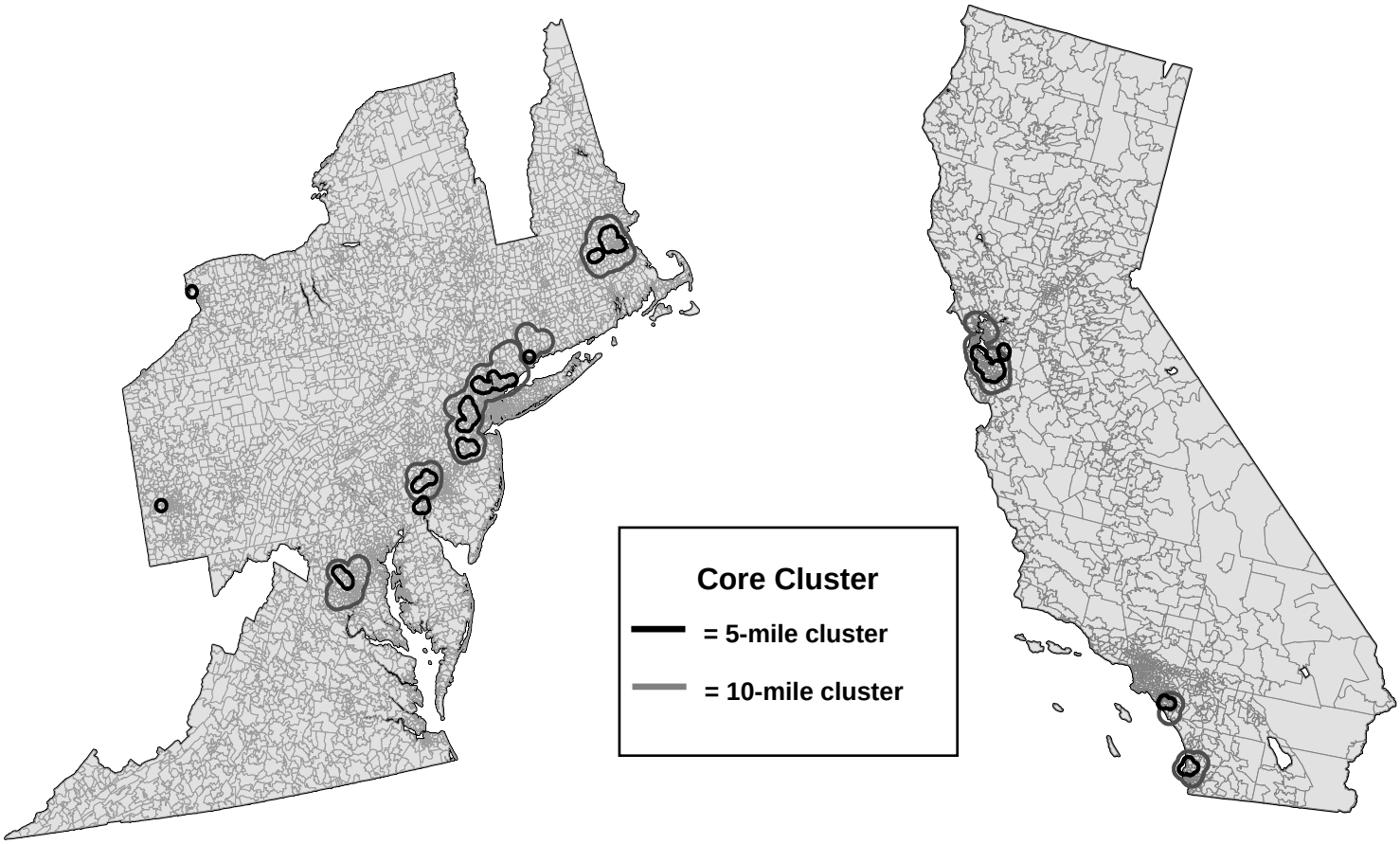


Figure 7a: Northeast Corridor Core Clusters
***d* = 5, 10**

Figure 7b: California Core Clusters
***d* = 5, 10**

For Online Publication

Appendix A: Robustness of Global *K*-Cluster Results

For completeness, we have analyzed R&D clustering with respect to Manufacturing Establishments as well as Manufacturing Employment. To do so, the number of manufacturing employees in each zip code area was simply replaced with the number of manufacturing establishments. In both the Northeast Corridor and California, the only substantive differences in global clustering with respect to these two reference distributions was due to certain anomalies arising from clusters of small establishments in industries not closely related to R&D activity.

The single most dramatic example is for the Northeast Corridor, where the Garment District in South Manhattan is so strongly concentrated (more than 2,000 establishments in two adjacent zip codes: 10018 and 10001) that it far outweighs the clustering of establishments in all other Northeast manufacturing industries combined. Figure A1 shows the comparison between a typical counterfactual lab patterns in South Manhattan generated by the manufacturing establishment distribution on the left, with the manufacturing employment distribution on the right (where zip codes 10018 and 10001 are the darkest pair in the left panel). So, while manufacturing employment appears to be quite concentrated in this area, it is clear that manufacturing establishments are relatively far more concentrated. Because this area constitutes such an extreme outlier in our data, we have run the simulation analyses both with and without South Manhattan (where the latter excludes the 20 R&D labs in South Manhattan as well), and the resulting global *Z*-scores are shown in Figures A2 and A3, respectively.

Notice first that the overall shape of the curve in Figure A2 is qualitatively very similar to that for manufacturing employment in Figure 4a of the text. But the values of the curve in Figure A2 are drastically lower and fail to yield significant clustering for essentially all scales less than 20 miles. But in Figure A3, it is seen that by removing only the small area of South Manhattan in Figure A1, the patterns of clustering significance for both manufacturing establishments and employment are now qualitatively similar, and indeed clustering at small scales is more significant with respect to the distribution of establishments. So, the influence of the garment industry is seen to be quite dramatic. Moreover, since it is reasonable to assume that the location of manufacturing R&D is relatively insensitive to this particular industry, the removal of this outlier seems reasonable.

Turning next to California, a similar anomaly was found with respect to the Jewelry District in Central Los Angeles, which again represents a strong clustering of small manufacturers not closely related to R&D. But because the effect of this cluster is much smaller in scope, we present only the full set of results for all manufacturing establishments in Figure A4 below. Here it is evident that except for small scales up to about three miles, the shape and levels of significance for both manufacturing establishments and manufacturing employment in Figure 4b of the text are remarkably similar.

Finally, it should be mentioned that a similar analysis was done using Total Employment as the reference distribution. Clustering anomalies for this distribution were even more severe than for Manufacturing Establishments, and the anomalies appear to have little relation to manufacturing

R&D. So, results for this distribution are deemed to have little relevance for the present analysis and are not reported.

For Online Publication

Appendix B: Robustness of Core-Cluster Results

As discussed in Section 6 of the paper, our method of identifying core clusters is, by construction, based on the results of local *K*-function analyses. Because such analyses involve separate tests at multiple locations (some nearby) and at multiple scales (some quite large), we must address certain aspects of the well-known "multiple testing" problem.[50] In this Appendix, we first discuss the multiple-testing problem itself, and then compare our core-cluster approach with "significance-maximizing" approaches to resolving this problem.

To motivate the multiple-testing problem in the setting of Section 5 in the text, we start by supposing that there is no discernible local clustering of R&D labs (i.e., that the observed pattern X^0 of R&D locations cannot be distinguished statistically from the patterns generated under our null hypothesis). In addition, suppose that all local *K*-function tests were in fact statistically independent of one another. Then, by construction, we should expect 5 percent of our resulting test statistics to be statistically significant at the 0.05 percent level. So, when many such tests are involved (there are 1,035 tests at each scale, $d \in D$, in the Northeast Corridor and 645 tests at each scale in California), one is bound to find some degree of "significant clustering" using such testing procedures. As is well known, this type of "false positive rate" can be mitigated by reducing the *p*-value threshold level deemed to be "significant." In fact, that is one reason why we focused only on *p*-values no greater than 0.005 in Figure 6 of the text.

But such adjustments are by themselves not sufficient in instances in which the assumption of statistical independence is violated. This is quite likely when radial neighborhoods around different test points are large enough to intersect and thus contain common points (either observed or counterfactual). In such cases, the resulting *p*-values at these test points must necessarily exhibit positive spatial autocorrelation, much in the same way that kernel smoothing of spatial data induces autocorrelation.[51]

Several statistical approaches have been developed for resolving such problems. Most prominent among these are the Kulldorff (1997) SATSCAN approach and the earlier Besag and Newell (1991) approach. Both methods employ sequential testing procedures, in which only single "maximally significant" clusters are identified in each step. To describe this sequential procedure in the present setting, we now focus on zip code areas (cells) and replace individual locations with counts of R&D labs in each area (cell counts). Using centroid distance between cells, candidate clusters are then defined as unions of *m*-nearest neighbors to given "seed" cells, and a test statistic is constructed to determine the single most significant cluster. In both of these significance-maximizing procedures, the notion of "significance" is defined with respect to tests that are based essentially on the original hypothesis, H_0, namely that R&D labs are distributed

[50] While global cluster analyses may also suffer from multiple testing over a range of spatial scales, this problem is particularly severe when conducting tests of local clustering that spatially overlap.

[51] For a full discussion of these issues in a spatial context, see, for example, Castro and Singer (2006).

(at the zip code level) in a manner proportional to manufacturing employment. One key difference is that counterfactual locations are implicitly assumed to be randomly distributed inside each zip code (i.e., are distributed proportional to area rather than total employment at the block level). To determine a second most significant cluster, the zip code areas in the most significant cluster are removed, and the same procedure is then applied to the remaining zip code areas. This procedure is typically repeated until some significance threshold (such as a *p*-value exceeding 0.05) is reached.

While this repeated series of tests might appear to reintroduce multiple testing, such tests are by construction defined over successively smaller spatial domains and hence are not directly comparable. Notice also that at each step of this procedure, the cluster identified has an explicit form, namely, a seed zip code area together with its current nearest neighbors. So, both the multiple-testing and cluster-identification problems raised for *K*-function analyses noted previously are at least partially resolved by this significance-maximizing approach.

We applied both the Besag-Newell procedure and Kulldorff's SATSCAN procedure to our data and found them to be in remarkable agreement with each other. Thus, we present only the results of the (more popular) SATSCAN procedure. In this setting, we ran the maximum of 10 iterations allowed by the SATSCAN software, and the results from the union of these 10 clusters are plotted in Figure B1 for labs in California, and in Figure B2 for labs in the Northeast Corridor. By comparing these results with Figures 6a and 6b in the text, it is evident that both procedures are identifying essentially the same areas. These comparisons thus serve as one type of robustness check on our core-cluster results.

However, there are certain differences between these results. Notice first that the SATSCAN clusters appear to be more circular in form than the corresponding core clusters. This is particularly evident in the Northeast Corridor, where isolated clusters such as Boston, Philadelphia and Washington, D.C., appear to be very circular. As mentioned previously, this particular SATSCAN procedure only considers circular (nearest-neighbor) clusters when identifying a "most significant" one. While it is possible to extend this restriction to certain classes of elliptical clusters, the key point is that prior restrictions must be placed on the set of "potential clusters" to keep search times within reasonable bounds. By way of contrast, our present core-cluster approach involves no prior restrictions on cluster shapes, and in this sense is more flexible in nature.

A second limitation of these significance-maximizing approaches that is less evident by visual inspection is the path-dependent nature of cluster formation. As mentioned previously, the zip code areas defining clusters created at each step of the procedure are removed before considering each new cluster. When clusters are very distinct (such as Boston, Philadelphia, and Washington in Figure B2), this removal process creates no difficulties. But when subsequent clusters are in the same area as previous clusters (such as the Bay Area in Figure B1 and the New York area in Figure B2), the formation of early clusters modifies the neighborhood relations among the remaining zip codes at later stages. So, at a minimum, these modifications require careful "conditional" interpretations of all clusters beyond the first cluster. Thus, a second advantage of the present core-cluster approach is the simultaneous formation of all clusters, which naturally avoids any type of sequential constraints.

For Online Publication

Appendix C: Description of the Major Areas of Agglomeration[52]

C.1 Northeast Corridor

Of the 1,035 R&D labs in the Northeast Corridor, 34 percent conduct research in chemicals; 17 percent conduct research in electronic equipment except computer equipment; 16 percent do research in measuring, analyzing, and control equipment; 9 percent conduct research in computer programming and data processing; and another 9 percent do research in industrial, commercial machinery, and computer equipment.

The Boston Agglomeration

There are 182 R&D labs within Boston's single 10-mile cluster, as shown in Figure 8a.[53] Most of these labs conduct R&D in five three-digit SIC code industries — computer programming and data processing, drugs, lab apparatus and analytical equipment, communications equipment, and electronic equipment. The largest five-mile cluster shown in Figure 8a contains 109 labs, which account for 60 percent of all labs in the larger 10-mile cluster. At the one-mile scale, Boston has five clusters, all of which are centered in the largest five-mile cluster. The largest of these one-mile clusters contains 27 labs, half of which conduct research on drugs.

The New York City Agglomeration

The single largest cluster identified within our 10-state study area is the 10-mile cluster above New York City (shown in Figure C1) that stretches from Connecticut to New Jersey. This cluster contains a total of 287 R&D labs. There are 134 (47 percent) labs in this cluster that conduct research on chemicals and allied products, 62 of which focus on drugs. Labs in this cluster also conduct research based on electrical equipment and industrial machinery. Within this highly elongated 10-mile cluster, four distinct 5-mile clusters were identified. Most of the concentration is seen to occur in the two clusters west of New York City, which, in particular, contain five of the nine one-mile clusters identified. Among these one-mile clusters, the largest is the "Central Park" cluster shown in Figure A1. About two-thirds of the 17 labs in this cluster are conducting research on drugs, perfumes, and cosmetics, or computer programming and data processing.

The Philadelphia Agglomeration

As seen in Figure C2, there is a large 10-mile cluster mostly to the west of Philadelphia (the city of Philadelphia is shown in darker gray), where there are a total of 44 labs. Of these 44 labs, 16 conduct research on drugs, and another 15 labs conduct research in the areas of computers, electronics, and instruments and related products. This cluster, in turn, contains a five-mile cluster centered in the King of Prussia area directly west of Philadelphia and contains 29 labs, with 40 percent doing research on drugs. There is a second five-mile cluster, containing 17 labs, centered in the city of Wilmington to the southwest. Here, 88 percent of the labs are doing research on chemicals and allied products.

---

[52] In addition to the four major areas of agglomeration discussed in what follows, there are two smaller agglomerations: one in Pittsburgh and another in Buffalo.

[53] The map legend in Figure 7 in the text applies to all map figures in this section.

The Washington, D.C., Agglomeration

The final area of concentration in the Northeast Corridor is the 10-mile cluster around Washington, D.C., which contains 74 R&D labs as shown in Figure C3 (with the city of Washington, D.C., in darker gray), where one five-mile cluster can also be seen. About one-quarter of the labs in the 10-mile cluster do research in the areas of computer programming and data processing. Furthermore, another 20 percent of the labs conduct research on communications equipment. In turn, this cluster contains two one-mile clusters, the largest of which (to the north) contains 16 labs with one-half conducting research on drugs.

C.2 California

Turning to California, 27 percent of 645 private R&D labs in the state conduct research in electronic equipment except computers; 18 percent do research in computer and data processing services; another 18 percent carry out research in chemicals, and 16 percent perform R&D in measuring, analyzing, and controlling equipment.

California's Bay Area

Of the 645 labs in California, 340 (slightly more than 50 percent) are located in the single 10-mile cluster in the Bay Area. This cluster stretches from Novato in the north to San Jose in the south and from Dublin–Pleasanton in the east to the Pacific Ocean in the west (Figure 7). Research in these labs is concentrated in three SIC industries: electronic equipment except computers; computer and data processing services; and chemicals and allied products. The Bay Area has two five-mile clusters, the most prominent of which is in the Palo Alto–San Jose area, consisting of 282 labs. The 10-mile cluster also contains seven one-mile clusters. The most prominent one-mile cluster is in Silicon Valley and consists of 138 labs (accounting for 41 percent of all labs in the Bay Area), with 30 percent conducting research in computer and data processing services.

San Diego

The largest five-mile cluster in Southern California consists of 56 labs found in San Diego. Of these 56 labs, 20 conduct research on chemicals; 11 perform research in the computer and data processing service; and 10 do research in measuring equipment. This cluster, in turn, contains a five-mile cluster consisting of 44 labs, and within it is a one-mile cluster consisting of 33 labs.

The Los Angeles Area

The most prominent cluster of labs in the Los Angeles area consists of 51 labs located in the Irvine–Santa Ana–Newport Beach area. Within this five-mile cluster, there are two separate one-mile clusters, one comprising 20 labs, and the other consisting of 10 labs. Electronic equipment except computers is the main area of research for these labs followed by measuring, analyzing, and controlling equipment; and transportation equipment. In addition, there are two separate one-mile clusters to the north of the 10-mile cluster. One of the clusters is in Torrance with nine labs, and the other in Santa Monica has seven labs.

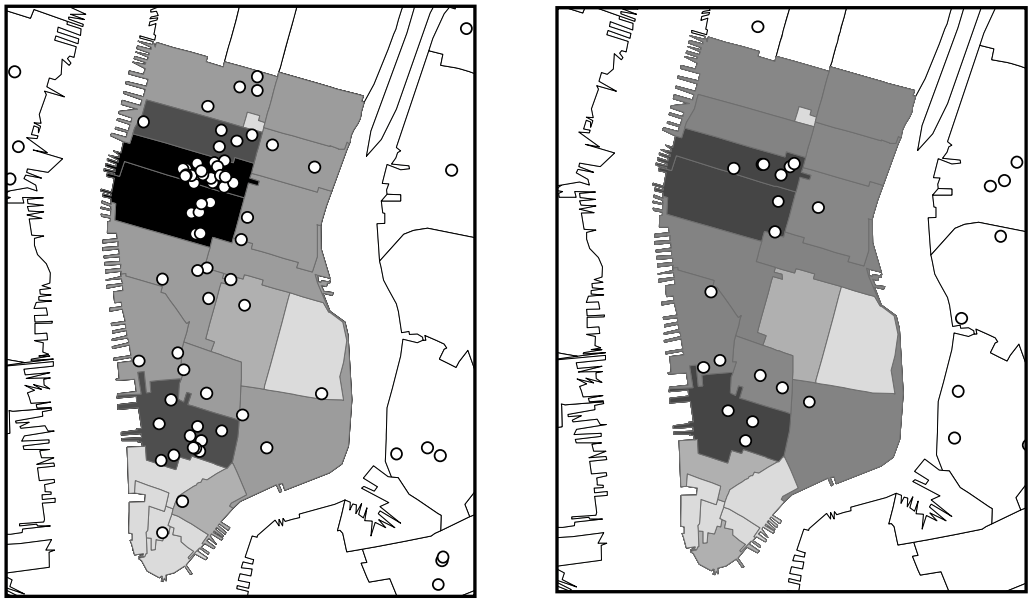


Figure A1. Manufacturing Establishment Counterfactuals (left panel) and Manufacturing Employment Counterfactuals (right panel)

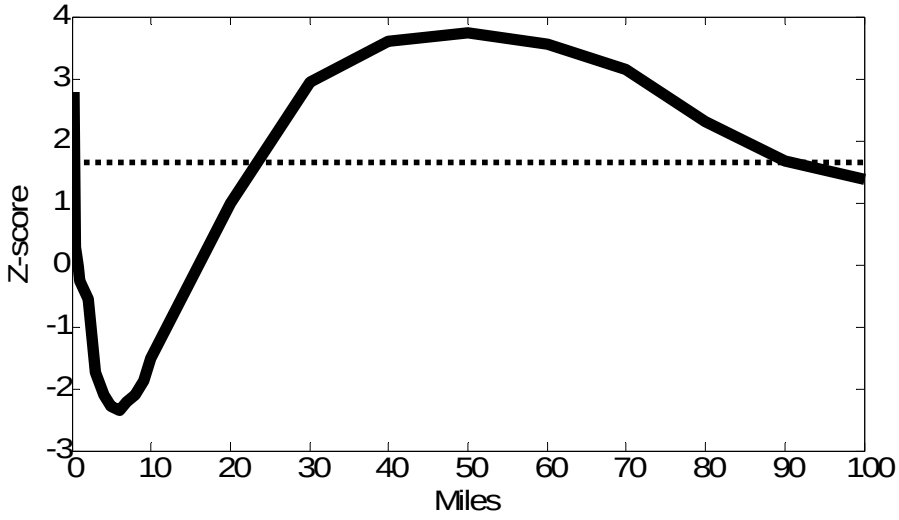


Figure A2: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Including South Manhattan

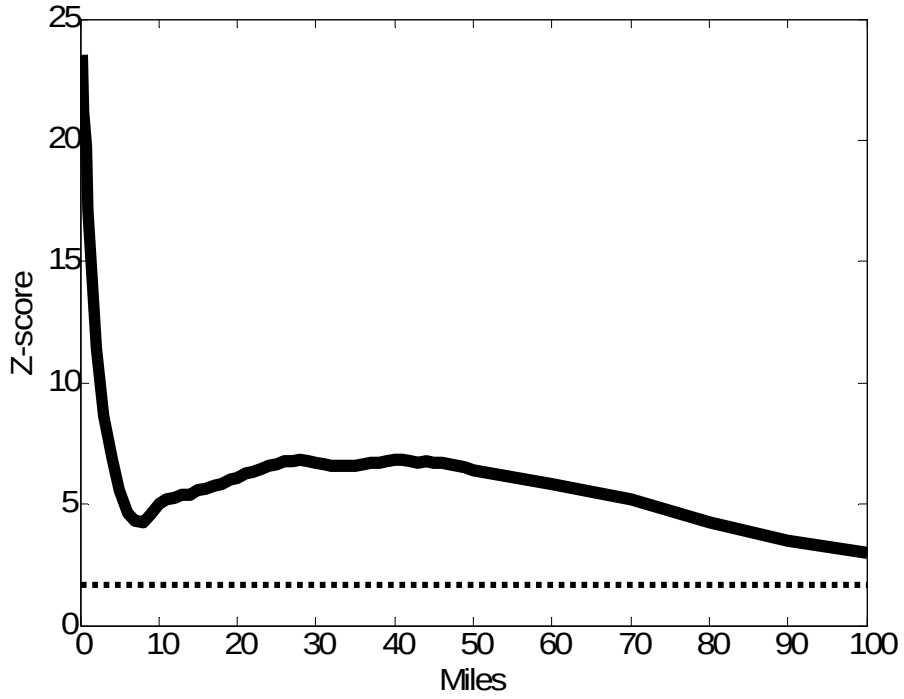
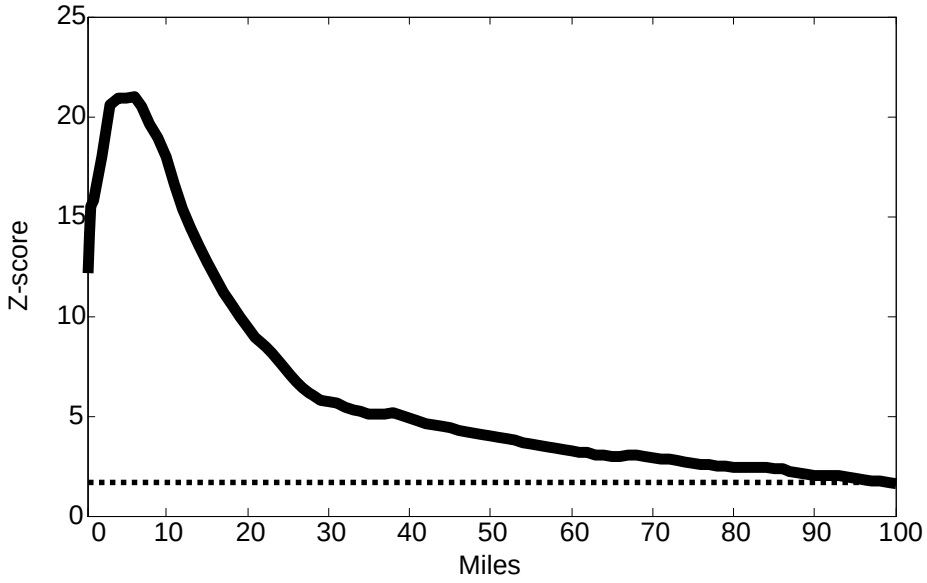


Figure A3: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Excluding South Manhattan



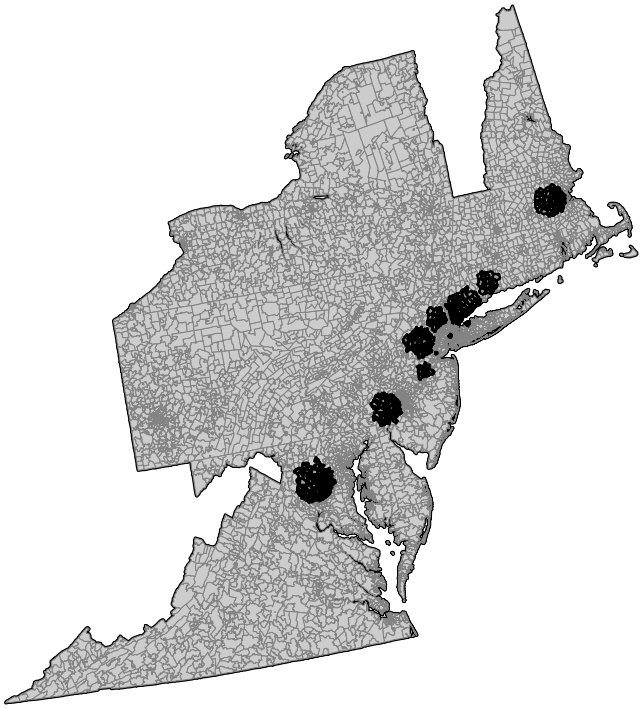


Figure B1: SATSCAN Clusters for the Northeast Corridor

Figure B2: SATSCAN Clusters for California

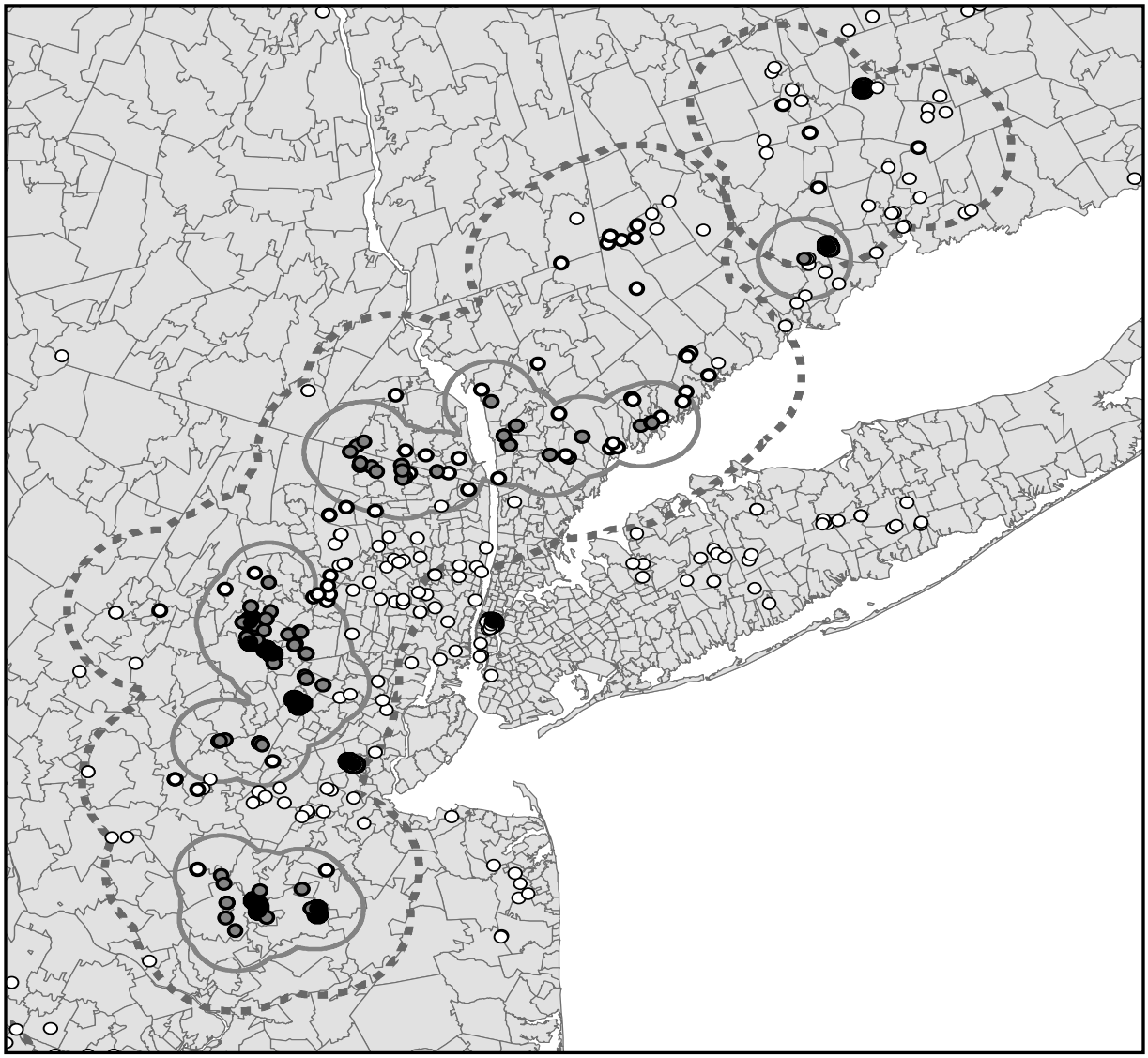


Figure C1: New York Core Clusters

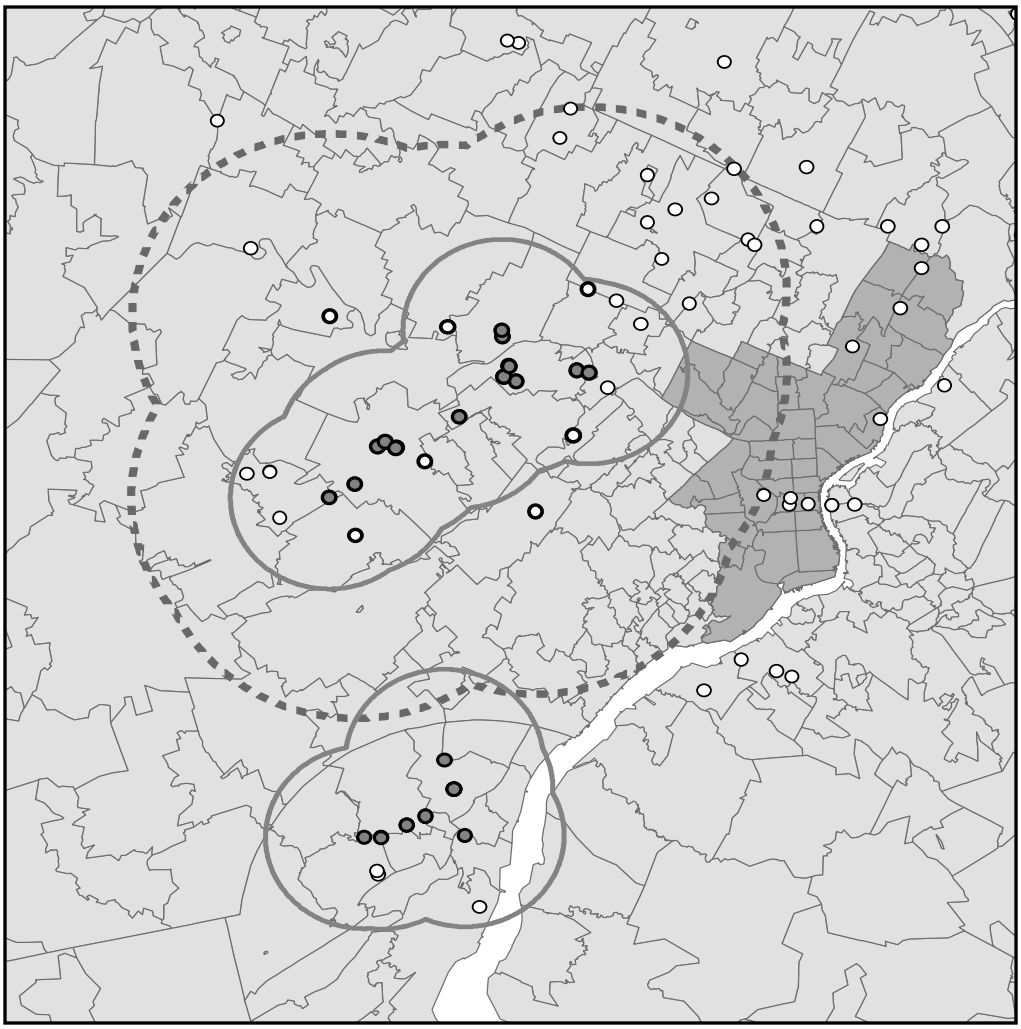


Figure C2: Philadelphia Core Clusters

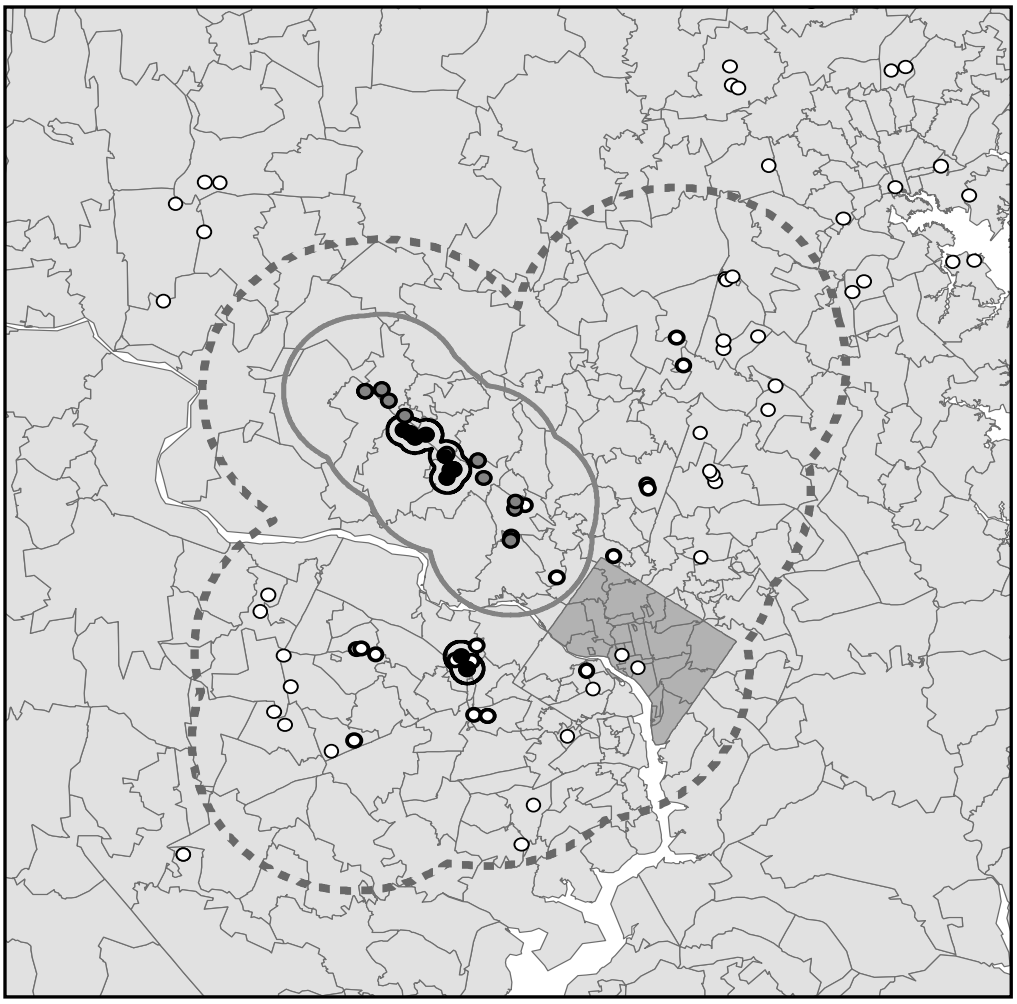


Figure C3: Washington, D.C., Core Clusters