WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 15-03

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

Kristy Buzard
Syracuse University

Gerald A. Carlino
Federal Reserve Bank of Philadelphia

Robert M. Hunt
Federal Reserve Bank of Philadelphia

Jake K. Carr
The Ohio State University

Tony E. Smith
University of Pennsylvania

January 2015

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

January 2015

ABSTRACT

We employ a unique data set to examine the spatial clustering of private R&D labs, and, using patent citations data, we provide evidence of localized knowledge spillovers within these clusters. Jaffe, Trajtenberg, and Henderson (1993, hereafter JTH) provide an aggregate measure of the importance of knowledge spillovers at either the state or metropolitan area level. However, 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. Instead of using fixed spatial boundaries, we develop a new procedure — the multiscale core-cluster approach — for identifying the location and size of specific R&D clusters. This approach allows us to better capture the geographic extent of knowledge spillovers. We examine the evidence for knowledge spillovers within R&D clusters in two regions: the Northeast Corridor and California. In the former, we find that citations are from three to six times more likely to come from the same cluster as earlier patents than in comparable control samples. Our results are even stronger for labs located in California: Citations are roughly 10 to 12 times more likely to come from the same cluster. Our tests reveal evidence of the attenuation of localization effects as distance increases: The localization of knowledge spillovers is strongest at small spatial scales (5 miles or less) and diminishes rapidly with distance. At the smallest spatial scales, our localization statistics are generally much larger than JTH report for the metropolitan areas included in their tests.

JEL Codes: O31, R12

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

*Research Department, Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106. Phone: (215) 574-6434. Fax: (215) 574-4303. E-mail: jerry.carlino@phil.frb.org.

1. INTRODUCTION

For many decades, researchers have argued that investments in research and development (R&D), and the resulting innovations, are a key component of long-run productivity growth. A theoretical basis is established in the models of endogenous growth developed by Romer (1990), Grossman and Helpman (1991), and Aghion and Howitt (1992). Similarly, there is a large empirical literature documenting the relationships between investment in R&D, productivity, and economic growth (see Akcay, 2011, for a recent survey of this literature). R&D is primarily an input. Its fertility depends on many factors, but surely an important one is the exchange of ideas. Even though we live in an era of global commerce and ubiquitous electronic communications networks, as shown in this paper, physical proximity is an important ingredient in the innovation process.

Despite the central role knowledge spillovers play in innovation and growth, they are particularly hard to verify empirically. Krugman (1991) noted that "knowledge flows are invisible; they leave no paper trail by which they may be measured and tracked." But Jaffe, Trajtenberg, and Henderson (1993, hereafter JTH) point out "knowledge flows do sometimes leave a paper trail" in the form of patented inventions. Patents contain information about the location of inventors as well as citations to prior patents that are related to the claimed invention. As with citations to academic articles, patent citations often represent tangible evidence of knowledge spillovers. JTH document a pattern of spatial concentration (often described as localization) in patent citations. All else being equal, patents are several times more likely to cite earlier patents from inventors located within the same metropolitan area (or state) than ones obtained by inventors living farther away. Such a pattern is consistent with the hypothesis that at least some important knowledge spillovers decrease with distance.

One concern with the JTH approach is that states, or even metropolitan areas, may not represent the spatial scale over which localized knowledge spillovers operate or, at least, where they are most powerful. There is mounting evidence from studies using alternative data that the transmission of knowledge begins to attenuate at distances ranging from a few blocks to just a few miles (e.g., Kerr and Kominers, 2014; Elvery and Sveikauskas, 2010; Arzaghi and Henderson, 2008; Agrawal, Kapur, and McHale, 2008; Keller, 2002; Rosenthal and Strange, 2001; Adams and Jaffe, 1996; and Audretsch and Feldman, 1996). If that is the case, the

magnitude of the localized spillovers documented by JTH may be understated, and the exact geography that is driving the spillovers may not be well identified. The techniques developed in this paper shed light on both of these concerns.

If knowledge spillovers are important for innovative activity, we would expect creative activity to be spatially concentrated to take advantage of these externalities. In this paper, we employ a unique data set to examine the spatial clustering of private R&D labs in two regions: the Northeast Corridor of the U.S. and the state of California. These two regions accounted for more than 55 percent of total private R&D in 1998 (the most recent year for which our data are available). 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.[1] Specifically, we use point pattern methods to analyze locational patterns over a range of selected spatial scales (within one-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.

Following Duranton and Overman (2005) — hereafter DO — we look for geographic clusters of labs that represent statistically significant departures from spatial randomness using simulation 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 manufacturing employment at that scale.[2]

In the first phase of the analysis, we employ Ripley's (1976) 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. These results are consistent with evidence reported in both strands of the literature described above. Interestingly, the extent of clustering measured by our test statistics varies with distance in different ways in our tests of labs located on the East and West Coasts: There is clear evidence of attenuation with distance among

[1] As noted in Section 2, this approach avoids a measurement problem associated with relying on fixed boundaries.

[2] As a robustness check, we verify that we obtain qualitatively similar results using an alternative counterfactual based on the spatial distribution of manufacturing establishments.

labs in California. Among labs located along the Northeast Corridor, the clustering statistics have an inverse-U shape.

We also use the global *K*-function technique to examine the *relative* concentration of R&D labs in specific two-digit SIC industries. By relative, we mean relative to the spatial concentration of R&D labs as a whole. 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, 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). At this scale, R&D labs in 37 percent of the industries studied in the Northeast Corridor are significantly more concentrated than are labs overall, and none are significantly more dispersed. Tests for the labs in California show that one-half of the industries exhibit significantly more concentration, and none are significantly more dispersed.

The second part of our analysis returns to the question of clustering at different scales but now focuses on the question of *where* such clustering occurs. To identify where clustering is occurring, we use a more refined procedure based on local *K*-functions. We call this procedure the multiscale core-cluster approach. Core clusters at each scale are identified in terms of those points with the most significant local clustering at that scale. 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. As a consistency check, these results are replicated using the significance-maximizing procedures developed by Besag and Newell (1991) and Kulldorff (1997).

In the final part our analysis, we document that knowledge spillovers are indeed more highly 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 JTH. In their study, JTH use cross-sectional analysis at either the state or metropolitan area levels to provide estimates of knowledge spillovers that are averaged over the metro areas or states. But, much information is lost regarding differences in the localization of knowledge spillovers in specific geographic areas. We show that such differences can be quite substantial.

For labs in the Northeast Corridor, we find that citations are 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. Our results are even stronger for labs located in California: Citations are roughly 10 to 12 times more likely to come from the same cluster as earlier patents than one would predict using the control sample. 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 main result. This is strong evidence 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. In addition, 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. Finally, we show that at the smallest spatial scales, our localization statistics are on average much larger than JTH report for the metropolitan areas included in their tests.

We also verify that our results for localized spillovers using patent citations are little affected when we use a variety of alternative approaches to identifying control patents suggested by the literature (e.g., Thompson and Fox-Kean, 2005 — hereafter TFK). JTH develop a matching rate analysis of the localization of patent citations in which they control for the preexisting spatial concentration of three-digit technological categories. Our main results are consistent with JTH and are based on three-digit controls. TFK argue that three-digit controls are too broad and may induce spurious evidence of localization. As a robustness check, we follow TFK and draw our potential controls more narrowly from patents that share the same patent class and subclass as the citing patent — six-digit controls. These results are found to be highly robust with respect to such controls, suggesting that they are not solely a consequence of technical aggregation.[3]

The remainder of the paper is organized as follows. Section 2 provides a brief review of the literature on geographic concentration. Section 3 describes the data used in this paper. In Section 4, we present the statistical methodology and test results for our global analyses of spatial clustering. In Section 5, we present our local analyses of clustering, and in Section 6, we

[3] Murata et al. (2014a) use a distance-based approach and find substantial evidence supporting the localization of patent citations even when six-digit technological classifications are used to select the control patents.

introduce a new approach (the multiscale core-cluster approach) for identifying explicit R&D clusters. Section 7 contains our analysis of the geographic concentration of patent citations, and we conclude in Section 8. The appendices document a number of robustness tests and provide additional information about the character and composition of research clusters identified in the paper.

2. LITERATURE REVIEW

A number of papers have used a spatial Gini coefficient to measure the geographical concentration of economic activity. Audretsch and Feldman (1996) were among the first to use this measure to show that innovative activity at the state level tends to be considerably more concentrated than is manufacturing employment. Ellison and Glaeser (1997) — hereafter EG — extended the spatial Gini coefficient to condition not only on the location of manufacturing employment but also on industrial structure. A number of researchers used the EG index to measure the clustering of manufacturing employment at the zip code, county, MSA, and state levels (see, for example, Rosenthal and Strange, 2001; and Ellison, Glaeser, and Kerr, 2010).

The EG index suffers from a number of important aggregation issues that result from using a fixed spatial scale. One aggregation issue is known as the modifiable area unit problem (MAUP). The problem is that conclusions reached when the underlying data are aggregated to a particular set of boundaries (say, counties) may differ markedly from conclusions reached when the same underlying data are aggregated to a different set of boundaries (say, MSAs). And the MAUP is more severe as the level of aggregation increases. Another problem is that researchers sometimes construct indexes of localization but do not report any indication of the statistical significance of their results. Without further statistical analyses, it is not clear whether the concentrations reported are significantly different from ones that might result even if the locations of economic activity resulted from random draws.[4]

To address these issues, DO used microdata to identify the postal codes for each manufacturing plant in the U.K., thus allowing these data to be geocoded. Geocoding is important since DO are not bound by a fixed geographical classification but base their approach on the actual distance

[4] Recently, Cassey and Smith (2014) developed a procedure to simulate confidence intervals for statistical tests of EG indices.

between firms. Rather than using a specific index to measure geographic concentration, DO take a nonparametric approach (based on kernel densities). Essentially, DO construct frequency distributions of the pair-wise distances between plants in a given industry. When the mass of the distribution is concentrated at short distances, this represents a spatial concentration of plants in the industry. If the mass of the distribution is concentrated at longer distances, this represents a more dispersed spatial pattern. DO consider whether the number of plants at a given distance is *significantly* different from the number that would result if their locations were randomly chosen.

A few other studies have used continuous measures of concentration. Marcon and Puech (2003) use distance-based methods to evaluate the spatial concentration of French manufacturing firms and find that some industries are concentrated, while other industries are dispersed. Arbia, Espa, and Quah (2008) use a *K*-function approach to study the spatial distribution of patents in Italy during the 1990s. Kerr and Kominers (2014) develop a model where the costs of interaction among agents define the distance over which forces for agglomeration of activity operate. In one application, Kerr and Kominers (2014) use data on patent citations and show that technologies with short distances over which agents interact are characterized by smaller and denser concentrations relative to technologies allowing for interactions over longer distances.

Our work differs from past studies in a number of ways. Unlike the papers mentioned previously, we use a new location-based data set that allows us to consider the spatial concentration of private R&D establishments. Rather than focusing on the overall concentration of R&D *employment*, we analyze the clustering of individual R&D *labs*. Our analytical approach permits such clustering to be identified at a range of scales in continuous space, rather than at a single predefined scale. 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.[5] Our tests for the localization of R&D labs also control for industrial concentration and, in particular, the concentration of manufacturing employment.[6] One advantage of *K*-functions is that they can

[5] Murata, et al. (2014b) show that the *K*-density approach can be subject to a downward bias due to edge effects. 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.

[6] Duranton and Overman (2005) suggest five properties for a good index of concentration. The index should (1) be comparable across industries, (2) control for overall concentration of industry, (3) control for industrial concentration, (4) be unbiased with respect to scale and aggregation, and (5) test for the significance of the results. It can be shown that the index of concentration used in this study satisfies these conditions.

easily be disaggregated to yield information about the *spatial locations* of clusters of R&D labs at various scales.

Finally, in Section 7, we employ a similar approach to JTH except that we replace states/metropolitan areas as the unit of analysis with the R&D clusters identified using the techniques developed in Section 6 of the paper. We show that patents from these clusters generate citations that are more localized than patent citations in general.

As with our paper, Murata et al. (2014a) use continuous approaches to derive global measures of clustering at various spatial scales. But the implementation is quite different. In their paper, they document the clustering of patent citations at various spatial scales relative to the distribution of a suitably defined set of control patents. But their approach cannot reveal *where* in space this clustering of citations occurs. In this paper, we use a related statistical approach to construct global and local measures of the clustering of R&D labs relative to the more general pattern of economic activity. We then use the set of local clusters identified by our technique to test for the localization of citations to patents obtained by inventors living in those clusters.

3. 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 research and development 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.[7] These firms represent just under 95 percent of all R&D expenditures reported in the 1999 vintage of Compustat for 1998. 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

[7] 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, MergentOnline, the Harris Selectory Online Database, and the American Business Directory (ABD).

specialization(s) to which we have assigned the corresponding four-digit SIC code(s). 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.[8] 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. The locations of these labs are shown in Figures 2a and 2b, respectively.[9]

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, as compared with 22 percent of total employment (21 percent of manufacturing employment) and 23 percent of the population. California itself accounted for almost 22 percent of all private R&D labs (and 22 percent of private R&D expenditures) in 1998 as 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.[10] This concentration is consistent with Audretsch and Feldman (1996), who report that the top four states in terms of innovation in their data include 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 code. The spatial patterns of employment levels for the Northeast Corridor and California are shown in Figures 3a and 3b, respectively (where zip code boundaries have been omitted to allow

[8] In a companion paper, we analyze similar patterns among 650 private R&D labs located in the Midwest. We find patterns of clustering comparable with those described in this paper.

[9] In some cases, a company reported multiple labs at the same address. For the analysis presented in this paper, we treated these cases as separate labs. As a robustness check, we also generated a map in which multiple labs owned by the same company and with a common street address were treated as a single lab. This reduces our lab count to 951 in the Northeast corridor and to 651 in California. We repeated our analyses using this alternative map and found essentially the same results.

[10] Data for private R&D expenditures are from Table A.39 of NSF (2000).

levels in smaller zip code areas to be seen). 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 in two ways: First, we use 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. Second, we also examine the concentration of labs conducting R&D in *specific* industries, as compared with the locations of *all* R&D labs.[11]

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 just under 4 miles. The average zip code in California had almost 6,000 jobs in 1998, of which 14 percent 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.

For the analysis in Section 7 of this paper, we use patent and citation data obtained from the NBER Patent Data Project.[12] We use data for patents granted in the years 1996-2006. In particular, we are interested in the geographic distribution of citations to patents obtained by inventors living within one of the R&D clusters we identify in Section 6 of the paper. We use the zip code associated with the residential address of the first inventor on the patent to geolocate each patent.

[11] In our data, there are two significant nonmanufacturing industries with R&D labs. These are electronics wholesaling (which includes firms such as Apple computers) and software. In the Appendix A, we report results of our analyses using manufacturing establishments as an alternative benchmark.

[12] See https://sites.google.com/site/patentdataproject/.

4. 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 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[13] and used this to identify feasible lab locations within each zip code area.[14] 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.[15]

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 employment levels, and (iii) point locations within blocks are selected randomly. It should be

[13] 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.

[14] 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.

[15] 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.

mentioned that actual locations are almost always along streets and can, of course, not be random within blocks. But, as discussed further in Section 4.2 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).

In the next section, we introduce the appropriate test statistics in terms of *K*-functions. In Section 4.2 and Section 4.3, we summarize our test results for global clustering. In Section 4.4, we consider the *relative* concentration of labs conducting R&D in specific (two-digit SIC) industries as compared with the locations of all R&D labs. In other words, we investigate whether labs in some industries exhibit more clustering than R&D labs in general.

4.1 *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.[16] 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

[16] 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[17]

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

As described in the preceding section, 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

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

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

4.2 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-

[17] 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.

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 taken to be 0.5 miles.[18]

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 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.[19]

4.3 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

[18] 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, and this percentage is reduced to about 4 percent for California.

[19] 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 appeared to be sufficiently large to obtain reliable estimates of sampling distributions under H_0.

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

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 4a, and those for California are shown in Figure 4b. These high *z*-scores are consistent with the significance of the Monte Carlo results above but add more detailed information about the patterns of significance.[20] Observe that in both figures, clustering is most significant at 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.

Turning finally to clustering at larger scales, the most striking difference between these figures is that while the *z*-scores for California decrease over the full range of distances, those for the Northeast exhibit a strong local maximum at around 50 miles. Further insight can be gained by comparing the spatial relation of manufacturing employment with R&D labs in each case. First, a visual comparison of Figures 2b and 3b for California shows that the main concentrations of both manufacturing employment and R&D labs are essentially the same, namely the San Francisco Bay Area together with the Los Angeles and San Diego areas. In fact, more than 81 percent of all manufacturing employment is within 5 miles of R&D labs.[21] This implies that as *K*-function scales increase, the differences between observed R&D locations and random locations (proportional to employment manufacturing) are systematically reduced. So the strongest clustering of R&D labs relative to manufacturing is seen only at smaller scales.

By way of contrast, it can be seen from Figures 2a and 3a that there is much less agreement between concentrations of manufacturing employment and R&D labs in the Northeast Corridor. As is apparent from Figure 2a (and will be verified statistically below), R&D labs are mostly concentrated in the linear Corridor between Washington, D.C. and Boston. But Figure 3a shows that while this Corridor also contains most of the manufacturing employment, there are

[20] The benchmark value of $z = 1.65$, shown as a dashed line in both Figures 4a and 4b, corresponds to a *p*-value of 0.05 for the one-sided tests of H_0 in expression (2) above.

[21] More precisely, the centroids of zip code areas containing 81.4 percent of manufacturing employment are within 5 miles of at least one R&D lab.

substantial concentrations elsewhere.[22] More generally, the portion of manufacturing employment within 5 miles of R&D labs now drops to less than 65 percent. So unlike California, there is no simple monotone relation between clustering at different scales. As will be more evident below (see Figure 6a), the peak in significance at around 50 miles is largely attributable to the scale of clustering along the Washington-Boston corridor. Here, clusters are so close together that larger radii about their individual members tend to gain increasing numbers of labs from adjacent clusters up to a scale of about 50 miles.

4.4 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 colocate 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.[23] 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., that 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 $n = 1,035$ for the Northeast Corridor and $n = 645$ for California), let n_j denote the number of these labs associated with industry j. Our

[22] It should be noted that some of the largest employment zip code areas (such as in Manhattan and Boston) are too small to be seen properly on this map of the entire Northeast.

[23] 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, 22, 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.

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 a 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.

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, they allow us to say nothing about clustering of R&D labs in general. But the benefits of this approach are twofold. 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 above, here 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.[24] Moreover, none are

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

significantly more dispersed.[25] 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. Again, none are significantly more dispersed.[26]

A graphical representation of these results is presented in Figures 5a and 5b, where the *z*-scores for each of the seven industries in the Northeast Corridor with most significant clustering are shown in Figure 5a, and those for the seven (of eight) most significant California industries are shown in Figure 5b.[27] 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 5 miles. For all but one of these industries in the Northeast Corridor, the clustering of R&D labs is by far most significant at very small spatial scales — 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 2 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 Figure 5b. Two of the other industries, Electronics and Business Services, have local peaks at one-half mile and one mile, respectively.

In addition, Figure 5a shows rapid attenuation of *z*-scores at small scales for all seven industries in the Northeast Corridor. 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 of these labs are associated with firms engaged in computer programming or data processing.) In California, Figure 5b 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

[25] With respect to dispersion, two of the 19 industries are found to be significantly more dispersed starting at a distance of 5 miles, and a third industry exhibits some degree of relative dispersion at 50 miles.

[26] 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.

[27] To conserve on space, the graph of the *z*-scores for Rubber Products is not shown in Figure 5b since the labs doing R&D in this industry accounted for less than 1 percent of all labs in California.

Processing) exhibit an opposite trend in which relative clusters become more significant at larger scales.

Finally, note that three industries are among the most significantly clustered in both the Northeast Corridor and California: Chemicals, Business Services, and the Measuring, Analyzing, and Controlling Instruments. 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 Corridor, 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 4a and 4b. But it should be equally clear from Figures 5a and 5b 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.

5. 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).[28] 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[29]

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

[28] 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).

[29] 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.

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

5.1 Local Testing Procedure

For the remainder of the paper, we use the same null hypothesis employed in Section 4.1 (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). 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.

5.2 Test Results for Local Clustering

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 above 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 (say, 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 (like 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 5 miles, as illustrated for the Northeast Corridor and California in Figures 6a and 6b, respectively.[30] 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 roughly described (from north to south) as the "Boston," "New York City," "Philadelphia," and "Washington, D.C.," agglomerations.[31] 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.

6. IDENTIFYING SPATIAL CLUSTERS

The global cluster analysis in Section 4 identified the *scales* at which clustering is most significant (relative to manufacturing employment). The local cluster analysis in Section 5.1 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

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

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

apply some additional techniques to identify clusters. In the process of doing so, we will address statistical issues that could potentially influence our present (and previous) results.

6.1 The Multiple-Testing Problem

Our method of identifying clusters is, by construction, a local cluster analysis. Because we are testing over multiple locations (some nearby) and multiple scales (some quite large), we must address certain aspects of the well-known "multiple testing" problem.[32]

Suppose first that there was, in fact, no local clustering of R&D labs (so that the observed pattern X^0 of R&D locations could not be distinguished statistically from the patterns generated under our null hypothesis). Suppose also that all local *K*-function tests were 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 standard 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." That is one reason why we focus only on *p*-values no greater than 0.005 in Figures 6a and 6b.

But this adjustment alone is not sufficient in instances where 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.[33]

As discussed in Appendix B, we have investigated two econometric approaches for resolving the multiple testing problem in a spatial setting, namely the Kulldorff (1997) SATSCAN procedure and the earlier Besag and Newell (1991) approach. Both methods attempt to resolve the

[32] 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. See Appendix B for a discussion of the robustness of the local cluster analysis.

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

multiple-testing problem by employing sequential testing procedures, where only single "maximally significant" clusters are identified in each step. For appropriate choices of threshold parameters, both of these methods yield clusters that are qualitatively in close agreement with Figures 6a and 6b and help to confirm that our local clustering results are not heavily influenced by multiple-testing problems.

However, as discussed in Appendix B, the SATSCAN and the Besag and Newell testing procedures suffer from other limitations, namely the "shape" restrictions in defining feasible clusters for significance-maximization procedures and the "path-dependencies" created among successively generated clusters.

6.2 A Multiscale Core-Cluster Approach

Given these limitations of significance-maximizing approaches, it is useful to consider an alternative approach to cluster identification that explicitly uses the *multiscale* nature of local *K*-functions. This clustering procedure starts with the results of the local point-wise clustering procedure in Section 5.1 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 5.1. 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*-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 define 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 7. Starting at the 10-mile level, we see that there are two core clusters (represented by dotted curves), which together essentially cover the entire Bay Area. The major core cluster essentially includes all of the Bay Area south of San

Francisco, and in particular, all of Silicon Valley. At the 5-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. There is again a smaller secondary cluster of labs, approximately centered on the Lawrence Livermore National Laboratory complex. Finally, at the 1-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 8a. Here again, the entire Boston area is itself a single 10-mile cluster. Moreover, within this area, there is again a dominant 5-mile core cluster containing the five major 1-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 8b,[34] 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 7 and 8a tend to be nested by scale, this is not always the case. For example, the 5-mile "Livermore Lab" cluster in Figure 7 is seen to be mostly outside the major 10-mile cluster. Here, there is a concentration of six R&D labs within 2 miles of each other, though Livermore itself is relatively far from the Bay Area. So while this concentration is picked up at the 5-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-

[34] For visual clarity, only core cluster points (and not their associated buffers) are shown in Figure 8b.

maximizing methods.[35] In addition, since all core clusters are determined simultaneously, the problem of path dependencies among clusters does not arise. By way of summary, an overall depiction of core clusters for both the Northeast Corridor and California (at scales, $d = 5,\ 10$) is shown in Figures 9a and 9b, respectively. Figure 9a shows the four major clusters identified for the Northeast Corridor (one each in Boston, New York-Northern New Jersey, Philadelphia-Wilmington, and Washington, D.C.), while Figure 9b 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).[36]

One objective of this procedure is to produce explicit representations of clusters that capture both their relative shapes and concentrations in a natural way. We can then test the economic relevance of these clusters, which is the objective of Section 7. Since there is no universally accepted definition of clusters, it seems prudent to analyze this problem from many viewpoints and look for areas of substantial agreement among them.

7. 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

[35] This point is demonstrated in Appendix B.

[36] It should be noted that certain, more systematic procedures may be possible. For example, the selection of "best representative" *d*-scales could in principle be 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 in Section 2, 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 6.2 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.

7.1 Construction of the Citations Data Set

For this analysis, we use data obtained from the NBER Patent Data Project.[37] 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.[38] 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). 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 (not all patents list two or more inventors), the results were virtually the same as we report below. [39]

For our tests, we rely primarily on the boundaries identified by our 5-mile and 10-mile core clusters located in the Northeast Corridor and in California.[40] For each core cluster at a given scale, we assemble four sets of patents. The first set, which we call *originating patents*, represent 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 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.[41]

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. The set of potential control patents excludes patents where the residential address of the first inventor is located outside the U.S. 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

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

[38] 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 our inventors are located at the centroid of the zip code where they live.

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

[40] In Section 7.4.1 below, we report comparable tests for larger and smaller clusters.

[41] 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).

patents also cannot cite the originating patent. The application date of potential control patents must be within three years 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.[42] 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 one potential control patent from which to choose.[43] 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, in Section 7.4.4, we repeat the analysis by sampling potential controls *without* replacement.[44] In this case, a potential control patent can be matched with one citing patent at most. As is shown later, while this significantly reduces the rate at which we can match control patents to citing patents, it does not materially affect the test statistics.

7.2 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.[45] 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},

[42] 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.

[43] 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.

[44] 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.

[45] The following formulation of the proportions used for testing purposes is based largely on Murata et al. (2014a).

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 three years of the treatment patent (see above). 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[46]

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

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

$$\tilde{p} = \frac{\sum_{i=1}^{\eta_o} \tilde{m}_i}{\sum_{i=1}^{\eta_o} \tilde{\eta}_i} = \frac{1}{\tilde{\eta}} \sum_{i=1}^{n_o} \tilde{m}_i \,. \tag{7}$$

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.[47]

7.3 Main Results

Table 3a presents the results of our localization or matching rate tests among 5-mile clusters in the Northeast Corridor, while Table 3b shows the results for the 10-mile clusters. As the last row

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

[47] 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.

of Table 3a shows, inventors living in the 5-mile clusters obtained 8,526 patents in 1996-1997 (column A). Those patents subsequently received 76,737 citations from other patents during the sample period (column B). Our matching algorithm, with replacement, was able to match nearly all of the citing patents with an appropriate control patent (column H). Among the treatment patents, 3.7 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 0.60 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 last row of the table, on average, a given patent citing an earlier patent in a 5-mile cluster is a little over 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. And given that the values of the localization differentials we find for California (see below) are considerably higher than those reported by JTH suggests that we find larger effects than those reported by JTH. 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 5-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 three 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 level of SMSAs/CMSAs. 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 almost twice the average in the Boston cluster.

Tables 4a and 4b present the results of our localization tests among 5- 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 5-mile and 10-mile clusters. The overall location differentials are generally higher than we found among clusters in the Northeast Corridor. On average, a given patent citing an earlier patent in a 5- or 10-mile cluster in California is almost 10 or 12 times as likely to have a first inventor living in that cluster than would be expected by chance alone.

The results reported in JTH represent an average across the metropolitan areas in their study, including metro areas in California. The fact that we get roughly comparable values for the locational differentials for the clusters in the Northeast Corridor as those JTH find for the average SMSA, coupled with our finding of much larger locational differentials in our California clusters, suggest that our locational differentials are much higher on average than those reported by JTH.

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.

7.4 Additional Results and Robustness Checks

7.4.1 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 is 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 5 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.

7.4.2 Evidence of Patent Quality

In this section, we investigate whether patents obtained by inventors living within our core clusters are somehow more important than patents obtained outside of these clusters. We rely on a common metric of patent quality — the number of citations received.[48] We develop a "counterfactual" region for each of the 10-mile core clusters identified in Section 6. 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 of 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 either more valuable or at least 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 inventions may not necessarily be better, but they are better

[48] Hall et al. (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).

known by researchers in the area. This interpretation only reinforces the evidence of localized knowledge spillovers in our clusters.

7.4.3 Alternative Approaches to Identifying Control Patents

As was mentioned in Section 2, 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 7.3 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.[49] Our methodology is otherwise the same as we describe in Section 7.2. We report the results of this alternative test for 5- and 10- mile clusters in the Northeast Corridor (Tables 7a and 7b) and in California (Tables 8a and 8b). Comparing these results with Tables 3a and 3b and 4a and 4b, there are very small differences in the treatment and control 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.

7.4.4 Sampling Without Replacement

To rule out the possibility that sampling control patents with replacement is somehow affecting our results, we repeated the exercise above sampling potential controls without replacement (see Tables 9 and 10). The ability to match a citing patent to a potential control is reduced when using this approach (to about 53 percent). Otherwise, the results are very comparable with what is reported in Tables 3 and 4 above. This suggests that regardless of the approach used, there is not

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

a substantial difference between the treatment proportion and the share of all citing patents with a first inventor living in the same cluster as the first inventor of the originating patent.

8. CONCLUDING REMARKS

In this article, we use a new data set on the location of R&D labs and U.S. Patent and Trademark Office patent data and find strong evidence of knowledge spillovers among labs in close geographic proximity. This is accomplished by using several distance-based econometric techniques to analyze the spatial concentration of the locations of over 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 attempt to 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 manufacturing activity (employment).

We present robust evidence that private R&D labs are indeed highly concentrated over a wide range of spatial scales. We also find that R&D labs in some industries (e.g., chemicals, including drugs) are substantially more spatially concentrated than are R&D labs as a whole. This is especially the case when we focus on very small spatial scales (1 mile or less).

We introduce a novel way to identify clusters, called *the multiscale core-cluster* approach. The analysis identified 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).

In the final section of the paper, we verify that these local clusters are economically meaningful. We do so by applying 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 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 10 to 12 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, and often much larger, than those

reported by JTH. In their study, JTH use cross-sectional analysis at either the state or metropolitan area levels. Thus, they 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

Adams, J., and A. Jaffe. "Bounding the Effects of R&D: An Investigation Using Matched Establishment-Firm Data," *RAND Journal of Economics*, 27 (1996), pp. 700-21.

Agrawal, Ajay, Devesh Kapur, and John McHale. "How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data," *Journal of Urban Economics*, 64 (2008), pp. 258-69.

Aghion, Philippe, and Peter Howitt. "A Model of Growth Through Creative Destruction," *Econometrica*, 60 (1992), pp. 323-51.

Akcay, Selcuk. "Causality Relationship Between Total R&D Investment and Economic Growth: Evidence from United States," *Journal of Faculty of Economics and Administrative Sciences*, 16 (2011), pp. 79-92.

Anselin, L. "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.

Arzaghi, Mohammad, and J. Vernon Henderson. "Networking Off Madison Avenue," *Review of Economic Studies*, 75 (2008), pp. 1011-38.

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

Cassey, A., and B. Smith. "Simulating Confidence for the Ellison–Glaeser Index," *Journal of Urban Economics*, 81 (2014), pp. 85-103.

Castro, M.C., and B.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, D.H., and M.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.

Elvery, Joel A., and Leo Sveikauskas. "How Far Do Agglomeration Effects Reach?" Unpublished Paper, Cleveland State University (2010).

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

Grossman, G., and E. Helpman. *Innovation and Growth in the Global Economy*, Boston: MIT Press (1991).

Guimarães, Paulo, Octávio Figueiredo, and Douglas Woodward. "Measuring the Localization of Economic Activity: A Parametric Approach," *Journal of Regional Science*, 47 (2007), pp. 753-44.

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

Henderson, Rebecca, Adam B. Jaffe, and Manuel Trajtenberg. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment – Comment," *American Economic Review*, 95 (2005), pp. 461-64.

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

Keller, W. "Geographic Localization of International Technology Diffusion," *American Economic Review*, 92 (2002), pp. 120-42.

Kerr, William R., and Scott Duke Kominers. "Agglomerative Forces and Cluster Shapes," *Review of Economics and Statistics*, (2014) Forthcoming.

Krugman, P. *Geography and Trade*, Cambridge, MA: MIT Press (1991).

Kulldorff, M. "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, 26 (1997), pp. 1487-1496.

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

Marshall, A. *Principles of Economics*, London: Macmillan (1890).

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura."Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach," *Review of Economics and Statistics*, (2014a) Forthcoming.

Murata, Yasusada, Ryo Nakajima, and Ryuichi Tamura. "Testing for Localization using Micro-geographic Data: A New Approach," mimeo, Nihon University (2014b) .

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

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

Romer, P. "Endogenous Technological Change," *Journal of Political Economy*, 98 (1990), pp. S71-S102.

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

Rosenthal, Stuart, and William C. Strange. "The Attenuation of Human Capital Spillovers," *Journal of Urban Economics*, 64 (2008), pp. 373-89.

Shapiro, S.S., and Wilk, M.B. "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.

United States 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.

Table 1: Summary Statistics					
Northeast (10-State)					
Variable	Mean	Std. Dev.	Median	Minimum	Maximum
	All Zip Codes (6,044)				
Land Area	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	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	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	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: Authors' calculations based on the 1998 editions of the *Directory of American Research and Technology* and Zip Code Business Patterns

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

Table 2a: Concentration of Labs by Industry 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	0.5277	0.1674	0.4149
Oil and Gas Extraction	13	3	0.5011	0.5019	0.5026	0.5034	0.5337	0.0906	0.2286
Food and Kindred Products	20	25	0.5825	0.6278	0.6750	0.7081	0.0984	0.2097	**0.0480**
Textile Mill Products	22	14	**0.0267**	**0.0465**	0.0690	0.0859	0.3468	0.7839	0.6446
Apparel	23	5	0.6029	0.6596	0.7103	0.7460	0.4685	0.2833	0.3058
Paper and Allied Products	26	28	0.6029	0.6596	0.7103	0.7460	0.4685	0.2833	0.3058
Printing and Publishing	27	3	0.5009	0.5012	0.5019	0.5024	0.5111	0.5837	0.0740
Chemicals and Allied Products	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
Electrical and Electronic Equipment Except Computers	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
Instruments and Related Products	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 shading indicates significantly more dispersed than overall labs at the 5 percent level of significance.

Source: Authors' calculations using the *Directory of Research and Technology* (1998).

Table 2b: Concentration of Labs by Industry 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 and Kindred Products	20	3	0.5055	0.5100	0.5150	0.5185	0.5990	0.7700	0.4925
Paper and Allied Products	26	2	0.5020	0.5035	0.5045	0.5080	0.5340	0.6175	0.1970
Chemicals and Allied Products	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**
Electrical and Electronic Equipment Except Computers	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
Instruments and Related Products	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 shading indicates significantly more dispersed than overall labs at the 5 percent level of significance.

Source: Authors' calculations using the *Directory of Research and Technology* (1998).

Table 3a: Five-Mile Buffer Clusters in the Northeast Corridor

					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%	3,458	104	3.01%	3,458	5	0.14%	**20.8**	**9.6**
Boston-Cambridge-Waltham-Woburn, MA	2,634	27,665	1,718	6.21%	27,407	1,702	6.21%	27,407	228	0.83%	**7.5**	**34.5**
Silver Spring-Bethesda, MD-McLean, VA	367	3,426	91	2.66%	3,347	89	2.66%	3,347	4	0.12%	**22.3**	**8.9**
Trenton-Princeton, NJ	889	9,023	260	2.88%	8,917	260	2.92%	8,917	33	0.37%	**7.9**	**13.4**
Parsippany-Morristown-Union, NJ	1,710	14,568	358	2.46%	14,427	355	2.46%	14,427	114	0.79%	**3.1**	**11.2**
Greenwich-Stamford, CT-Scarsdale, NY	1,205	11,220	141	1.26%	11,032	141	1.28%	11,032	50	0.45%	**2.8**	**6.6**
Stratford-Milford-CT	235	1,484	12	0.81%	1,480	12	0.81%	1,480	4	0.27%	**3.0**	**2.0**
Conshohocken-King of Prussia-West Chester, PA	539	2,352	68	2.89%	2,344	68	2.90%	2,344	8	0.34%	**8.5**	**7.0**
Wilmington-New Castle, DE	624	3,501	72	2.06%	3,471	72	2.07%	3,471	9	0.26%	**8.0**	**7.1**
All 5-Mile Clusters	8,526	76,737	2,824	3.68%	75,883	2,803	3.69%	75,883	455	0.60%	**6.2**	**41.8**

Source: NBER Patent Data Project

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 3b: Ten-Mile Buffer Clusters in the Northeast Corridor

					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	48,316	4,264	8.83%	47,853	4,222	8.82%	47,853	752	1.57%	**5.6**	**51.2**
Washington, D.C.	926	9,744	329	3.38%	9,575	326	3.40%	9,575	43	0.45%	**7.6**	**15.0**
New York, NY	7,768	67,986	4,740	6.97%	67,215	4,700	6.99%	67,215	2,407	3.58%	**2.0**	**28.0**
Philadelphia, PA	1,594	9,029	409	4.53%	8,952	405	4.52%	8,952	61	0.68%	**6.6**	**16.3**
All 10-Mile Clusters	15,007	135,075	9,742	7.21%	133,595	9,653	7.23%	133,595	3,263	2.44%	**3.0**	**58.0**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 4a: Five-Mile Buffer Clusters in California

					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%	3,403	76	2.23%	3,403	8	0.24%	**9.5**	**7.5**
Los Angeles	454	3,646	104	2.85%	3,632	104	2.86%	3,632	9	0.25%	**11.6**	**9.1**
Palo Alto-San Jose	11,318	145,471	26,684	18.34%	143,647	26,431	18.40%	143,647	2,266	1.58%	**11.7**	**156.6**
Dublin-Pleasonton	283	3,899	127	3.26%	3,874	127	3.28%	3,874	2	0.05%	**63.5**	**11.2**
All 5-Mile Clusters	12,499	156,450	26,992	17.25%	154,556	26,738	17.30%	154,556	2,285	1.48%	**11.7**	**156.7**

Source: NBER Patent Data Project

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 4b: Ten-Mile Buffer Clusters in California

					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	20,077	970	4.83%	19,857	960	4.83%	19,857	129	0.65%	**7.4**	**25.7**
Los Angeles	1,266	10,685	609	5.70%	10,640	607	5.70%	10,640	54	0.51%	**11.2**	**22.1**
San Francisco	14,963	188,943	44,215	23.40%	186,672	43,789	23.46%	186,672	4,429	2.37%	**9.9**	**202.3**
All 10-Mile Clusters	18,328	219,705	45,794	20.84%	217,169	45,356	20.89%	217,169	4,612	2.12%	**9.8**	**202.7**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

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

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 Buffer Clusters in the Northeast Corridor (Disaggregated Patent Classes)

					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%	3,388	101	2.98%	3,388	4	0.12%	**25.25**	**9.6**
Boston-Cambridge-Waltham-Woburn, MA	2,634	27,665	1,718	6.21%	26,738	1,659	6.20%	26,738	233	0.87%	**7.1**	**33.7**
Silver Spring-Bethesda, MD-McLean, VA	367	3,426	91	2.66%	3,278	86	2.62%	3,278	7	0.21%	**12.3**	**8.3**
Trenton-Princeton, NJ	889	9,023	260	2.88%	8,674	249	2.87%	8,674	27	0.31%	**9.2**	**13.5**
Parsippany-Morristown-Union, NJ	1,710	14,568	358	2.46%	13,977	332	2.38%	13,977	108	0.77%	**3.1**	**10.8**
Greenwich-Stamford, CT-Scarsdale, NY	1,205	11,220	141	1.26%	10,696	137	1.28%	10,696	48	0.45%	**2.9**	**6.6**
Stratford-Milford-CT	235	1,484	12	0.81%	1,424	9	0.63%	1,424	0	0.00%	**N/A**	**3.0**
Conshohocken-King of Prussia-West Chester, PA	539	2,352	68	2.89%	2,175	63	2.90%	2,175	12	0.55%	**5.3**	**6.0**
Wilmington-New Castle, DE	624	3,501	72	2.06%	3,343	70	2.09%	3,343	6	0.18%	**11.7**	**7.4**
All 5-Mile Clusters	8,526	76,737	2,824	3.68%	73,693	2,706	3.67%	73,693	445	0.60%	**6.1**	**40.9**

Source: NBER Patent Data Project

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 7b: Ten-Mile Buffer Clusters in the Northeast Corridor (Disaggregated Patent Classes)

					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	48,316	4,264	8.83%	46,687	4,114	8.81%	46,687	732	1.57%	**5.6**	**50.6**
Washington, D.C.	926	9,744	329	3.38%	9,365	318	3.40%	9,365	49	0.52%	**6.5**	**14.3**
New York, NY	7,768	67,986	4,740	6.97%	65,027	4,500	6.92%	65,027	2,344	3.60%	**1.9**	**26.8**
Philadelphia, PA	1,594	9,029	409	4.53%	8,610	379	4.40%	8,610	46	0.53%	**8.2**	**16.5**
All 10-Mile Clusters	15,007	135,075	9,742	7.21%	129,689	9,311	7.18%	129,689	3,171	2.45%	**2.9**	**56.7**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 8a: Five-Mile Buffer Clusters in California (Disaggregated Patent Classes)

					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%	3,314	63	1.90%	3,314	10	0.30%	**6.3**	**6.3**
Los Angeles	454	3,646	104	2.85%	3,477	98	2.82%	3,477	7	0.20%	**14.0**	**9.0**
Palo Alto-San Jose	11,318	145,471	26,684	18.34%	141,219	26,076	18.46%	141,219	2,363	1.67%	**11.0**	**154.4**
Dublin-Pleasonton	283	3,899	127	3.26%	3,794	127	3.35%	3,794	4	0.11%	**31.8**	**10.9**
All 5-Mile Clusters	12,499	156,450	26,992	17.25%	151,804	26,364	17.37%	151,804	2,384	1.57%	**11.1**	**154.4**

Source: NBER Patent Data Project

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 8b: Ten-Mile Buffer Clusters, 2 Date Restrictions in California (Disaggregated Patent Classes)

					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	20,077	970	4.83%	19,283	925	4.80%	19,283	135	0.70%	**6.9**	**24.8**
Los Angeles	1,266	10,685	609	5.70%	10,268	568	5.53%	10,268	43	0.42%	**13.2**	**21.8**
San Francisco	14,963	188,943	44,215	23.40%	183,310	43,137	23.53%	183,310	4,559	2.49%	**9.5**	**199.4**
All 10-Mile Clusters	18,328	219,705	45,794	20.84%	212,861	44,630	20.97%	212,861	4,737	2.23%	**9.4**	**199.7**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 9a: Five-Mile Buffer Clusters in the Northeast Corridor (Control Patents Selected Without Replacement)

					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%	1,823	43	2.36%	1,823	2	0.11%	**21.5**	**6.2**
Boston-Cambridge-Waltham-Woburn, MA	2,634	27,665	1,718	6.21%	14,879	918	6.17%	14,879	149	1.00%	**6.2**	**24.2**
Silver Spring-Bethesda, MD-McLean, VA	367	3,426	91	2.66%	1,795	59	3.29%	1,795	2	0.11%	**29.5**	**7.4**
Trenton-Princeton, NJ	889	9,023	260	2.88%	4,663	132	2.83%	4,663	19	0.41%	**6.9**	**9.3**
Parsippany-Morristown-Union, NJ	1,710	14,568	358	2.46%	7,838	195	2.49%	7,838	49	0.63%	**4.0**	**9.4**
Greenwich-Stamford, CT-Scarsdale, NY	1,205	11,220	141	1.26%	5,657	85	1.50%	5,657	20	0.35%	**4.3**	**6.4**
Stratford-Milford-CT	235	1,484	12	0.81%	808	6	0.74%	808	2	0.25%	**3.0**	**1.4**
Conshohocken-King of Prussia-West Chester, PA	539	2,352	68	2.89%	1,306	40	3.06%	1,306	4	0.31%	**10.0**	**5.5**
Wilmington-New Castle, DE	624	3,501	72	2.06%	1,964	38	1.93%	1,964	2	0.10%	**19.0**	**5.7**
All 5-Mile Clusters	8,526	76,737	2,824	3.68%	40,733	1,516	3.72%	40,733	249	0.61%	**6.1**	**30.7**

Source: NBER Patent Data Project

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 9b: Ten-Mile Buffer Clusters in the Northeast Corridor (Control Patents Selected Without Replacement)

					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	48,316	4,264	8.83%	26,006	2,313	8.89%	26,006	432	1.66%	**5.4**	**37.4**
Washington, D.C.	926	9,744	329	3.38%	5,152	192	3.73%	5,152	25	0.49%	**7.7**	**11.5**
New York, NY	7,768	67,986	4,740	6.97%	35,788	2,559	7.15%	35,788	1,221	3.41%	**2.1**	**22.4**
Philadelphia, PA	1,594	9,029	409	4.53%	4,982	221	4.44%	4,982	29	0.58%	**7.6**	**12.4**
All 10-Mile Clusters	15,007	135,075	9,742	7.21%	71,928	5,285	7.35%	71,928	1,707	2.37%	**3.1**	**44.2**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 10a: Five-Mile Buffer Clusters in California (Control Patents Selected Without Replacement)

					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%	1,790	47	2.63%	1,790	2	0.11%	**23.5**	**6.5**
Los Angeles	454	3,646	104	2.85%	2,049	68	3.32%	2,049	7	0.34%	**9.7**	**7.2**
Palo Alto-San Jose	11,318	145,471	26,684	18.34%	74,197	13,856	18.67%	74,197	1,193	1.61%	**11.6**	**113.5**
Dublin-Pleasonton	283	3,899	127	3.26%	2,098	75	3.57%	2,098	0	0.00%	**N/A**	**8.8**
All 5-Mile Clusters	12,499	156,450	26,992	17.25%	80,134	14,046	17.53%	80,134	1,202	1.50%	**11.7**	**113.7**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

Table 10b: Ten-Mile Buffer Clusters, Two Date Restrictions in California (Control Patents Selected Without Replacement)

					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	20,077	970	4.83%	10,364	511	4.93%	10,364	63	0.61%	**8.1**	**19.1**
Los Angeles	1,266	10,685	609	5.70%	5,977	367	6.14%	5,977	31	0.52%	**11.8**	**17.3**
San Francisco	14,963	188,943	44,215	23.40%	96,403	22,785	23.64%	96,403	2,294	2.38%	**9.9**	**146.2**
All 10-Mile Clusters	18,328	219,705	45,794	20.84%	112,744	23,663	20.99%	112,744	2,388	2.12%	**9.9**	**146.7**

Source: NBER Patent Data Project and authors' calculations

*The subset of citing patents for which we obtained a similar control patent. Control patents are selected with replacement and based on three-digit patent technology classifications.

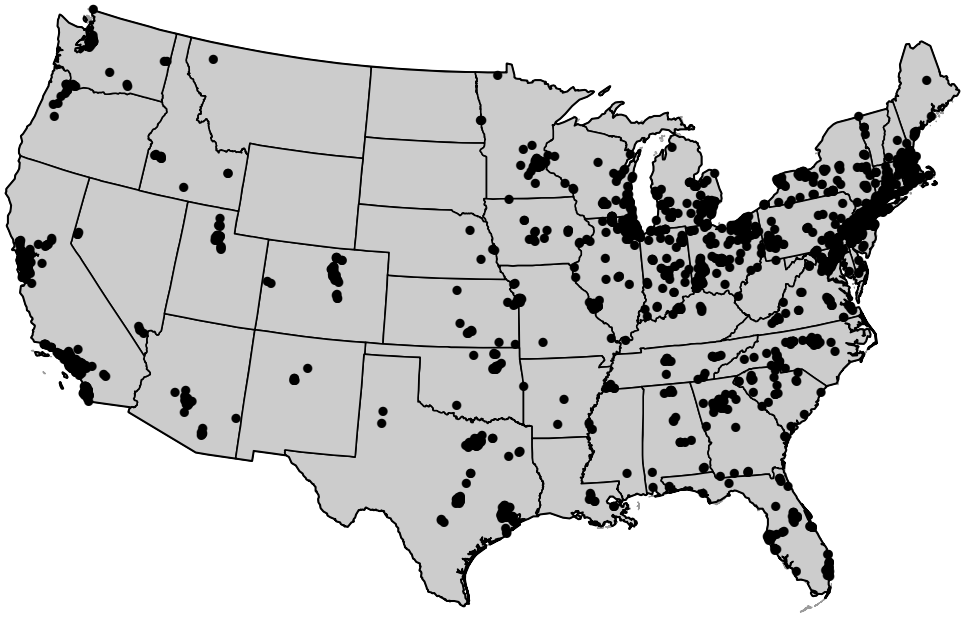

Figure 1: Location of R&D Labs

Source: *Directory of American Research and Technolog*y 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.

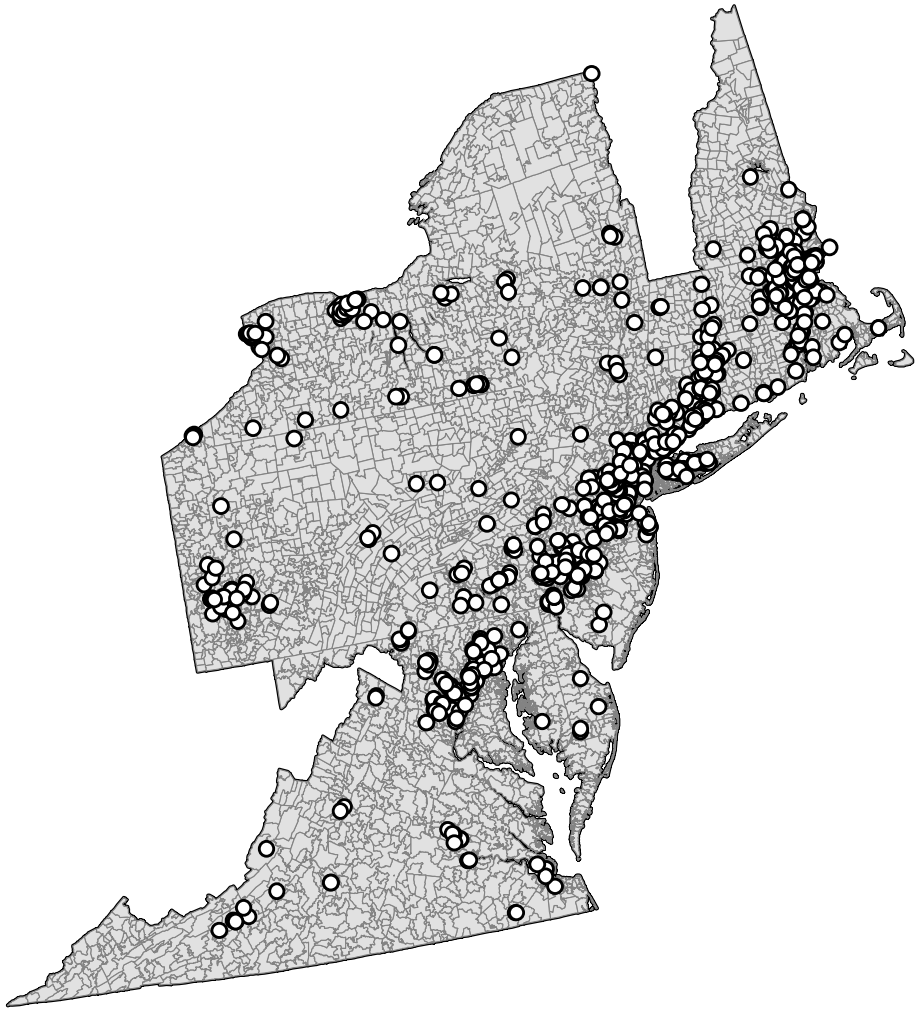

Figure 2a: R&D Labs in the Northeast Corridor

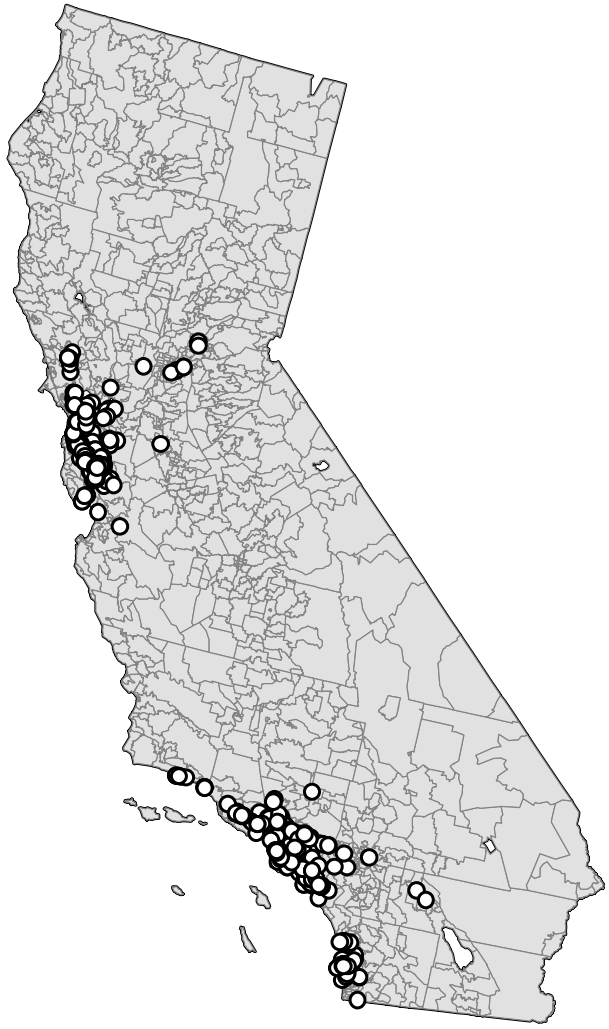

Figure 2b: R&D Labs in California

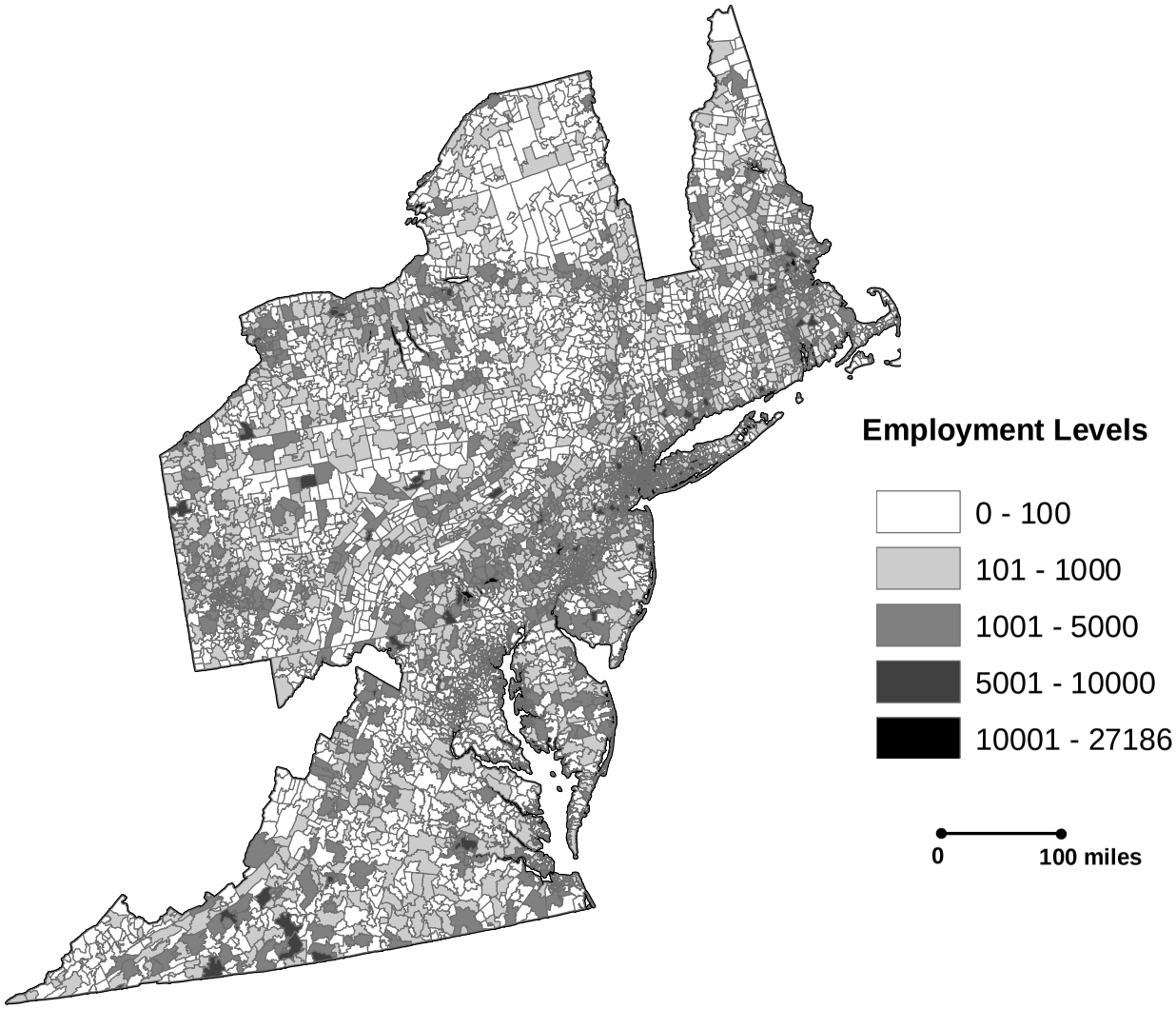

Figure 3a: Manufacturing Employment in Northeast Corridor

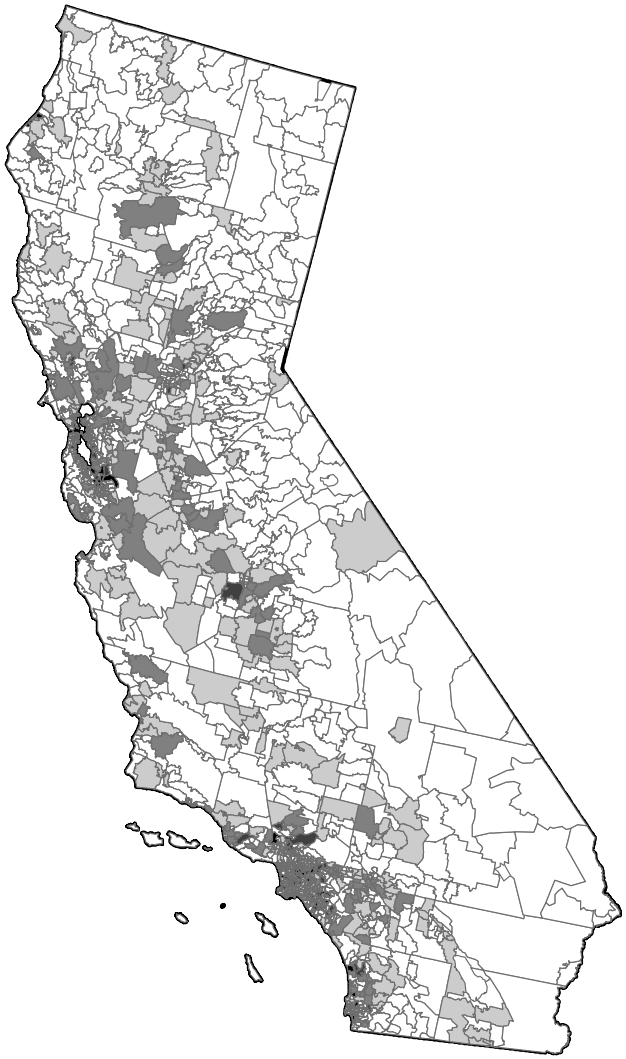

Figure 3b: Manufacturing Employment in California

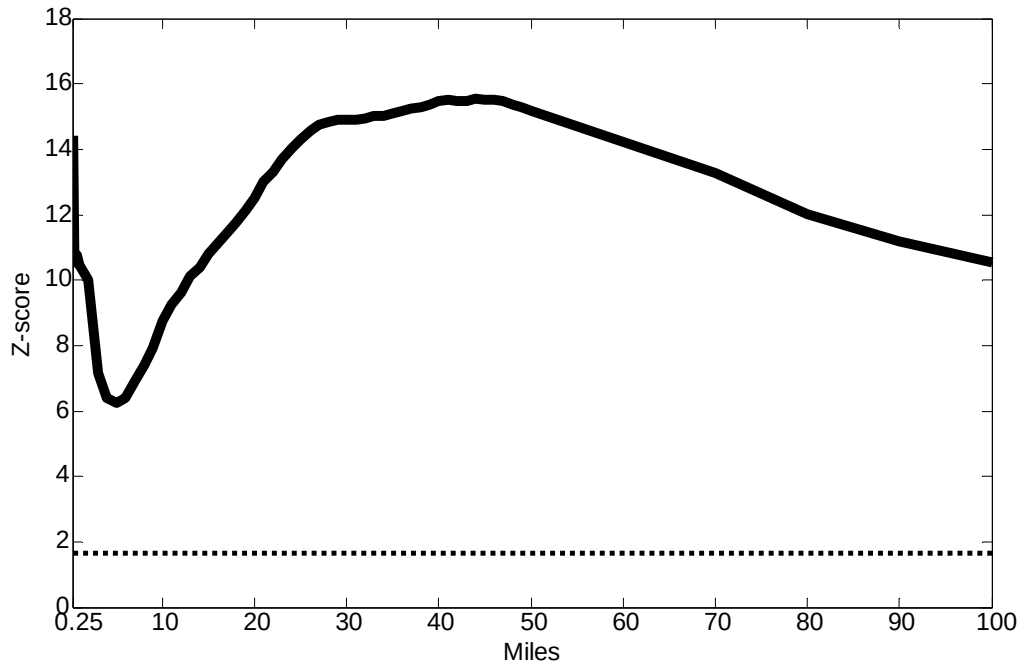

Figure 4a: *Z*-scores for Northeast Corridor

Dotted line *Z* = 1.65

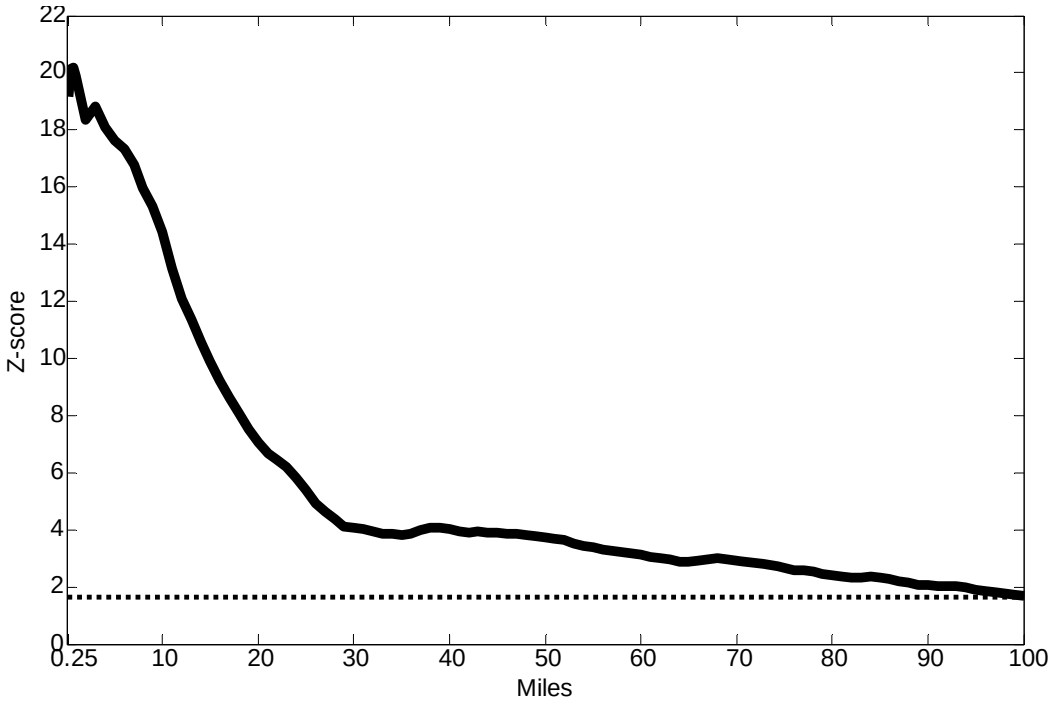

Figure 4b: *Z*-scores for California

Dotted line *Z* = 1.65

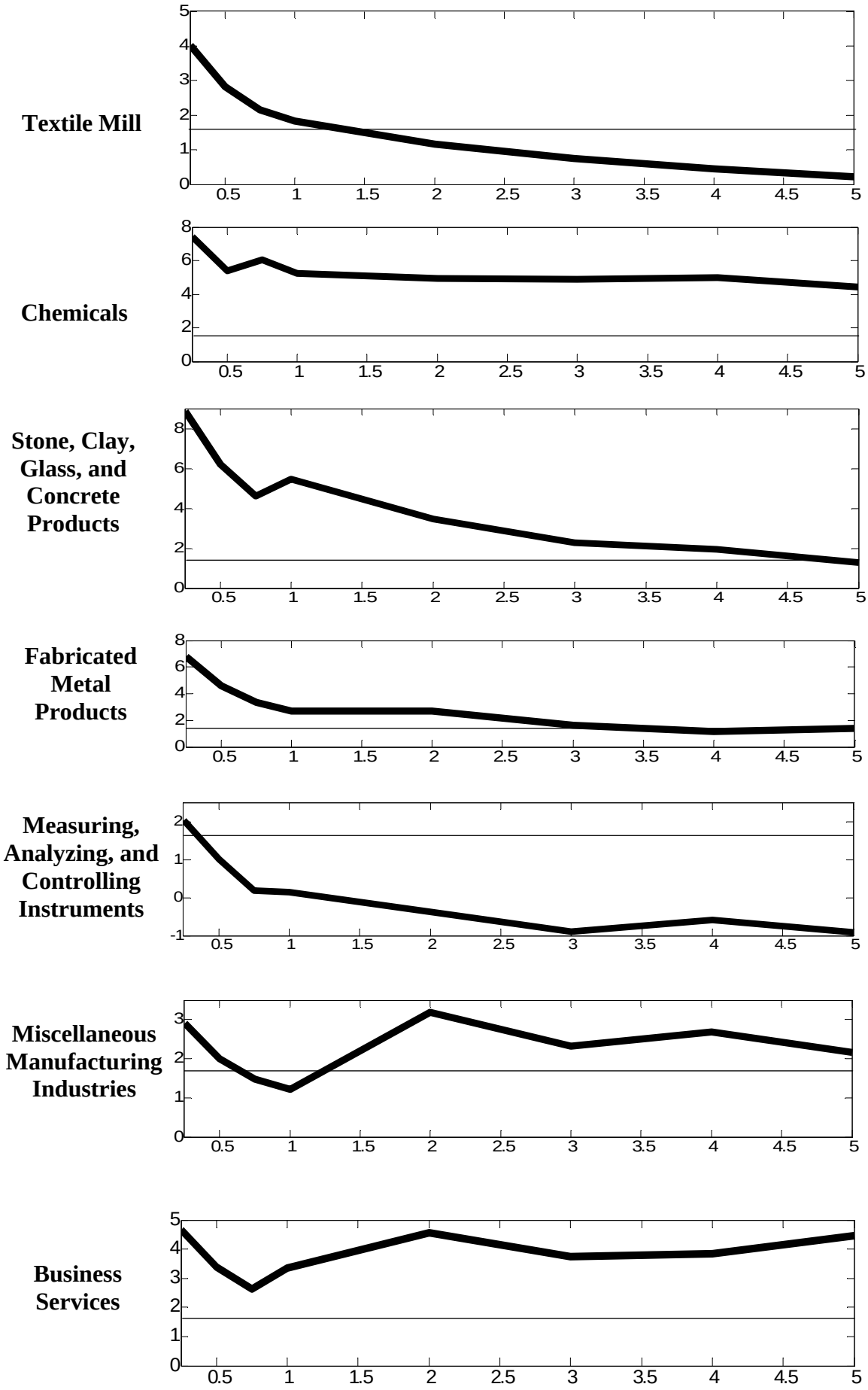

Figure 5a: Northeast Corridor Industry *Z*-scores

Solid line *Z* = 1.65

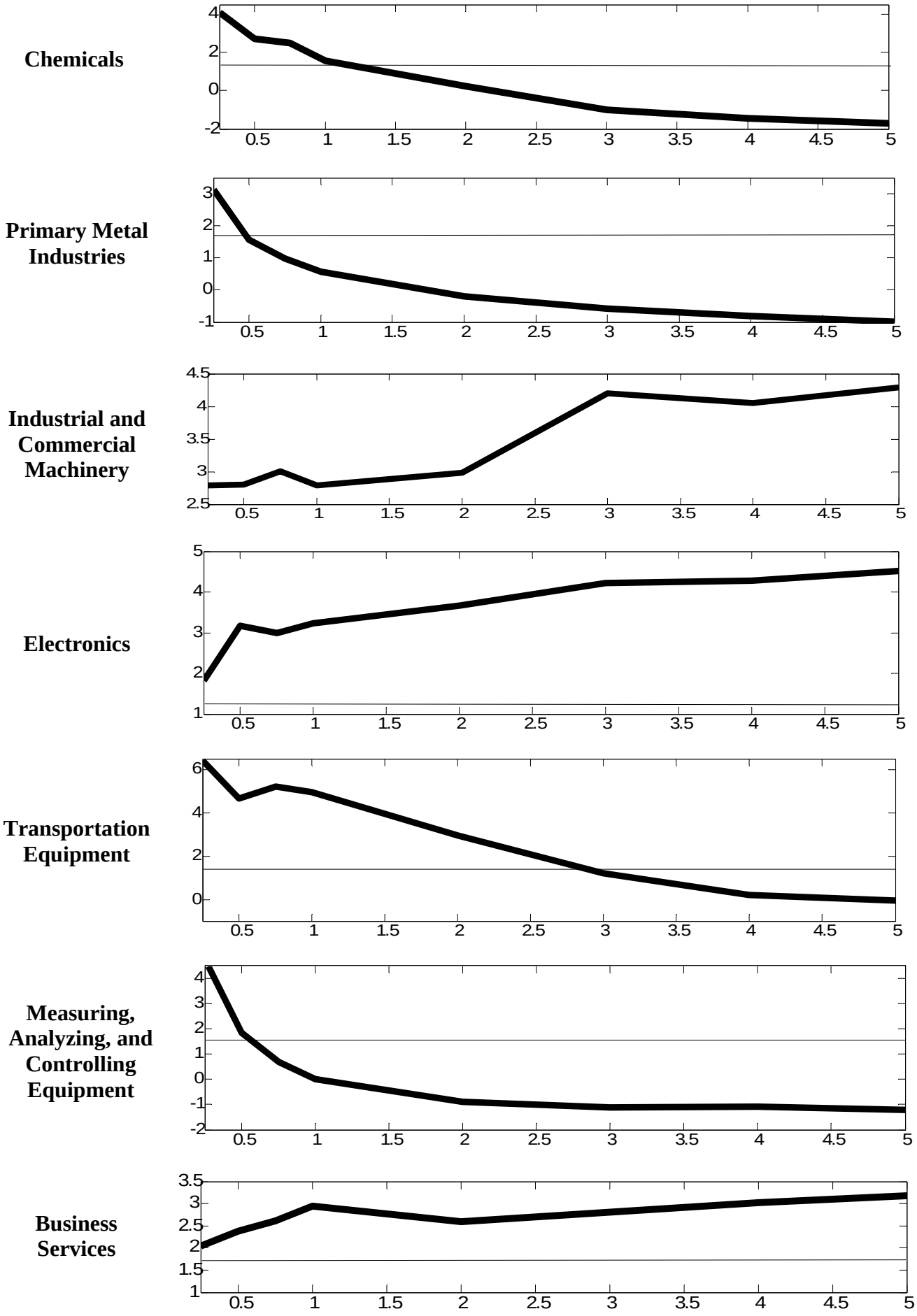

Figure 5b: CA Industry *Z*-scores

Solid Line *Z* = 1.65

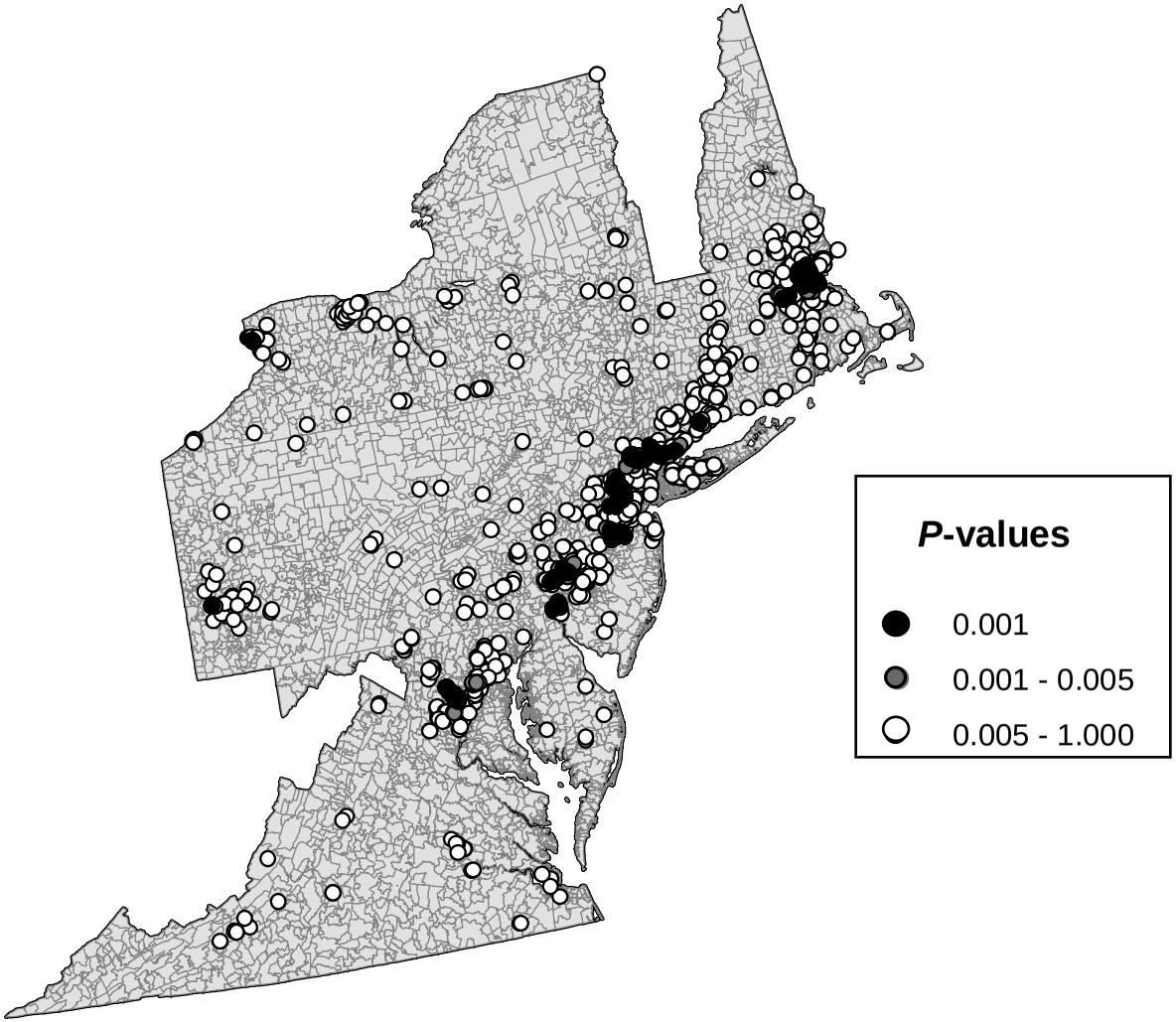

Figure 6a: Northeast Corridor *P*-values at *d* = 5 miles

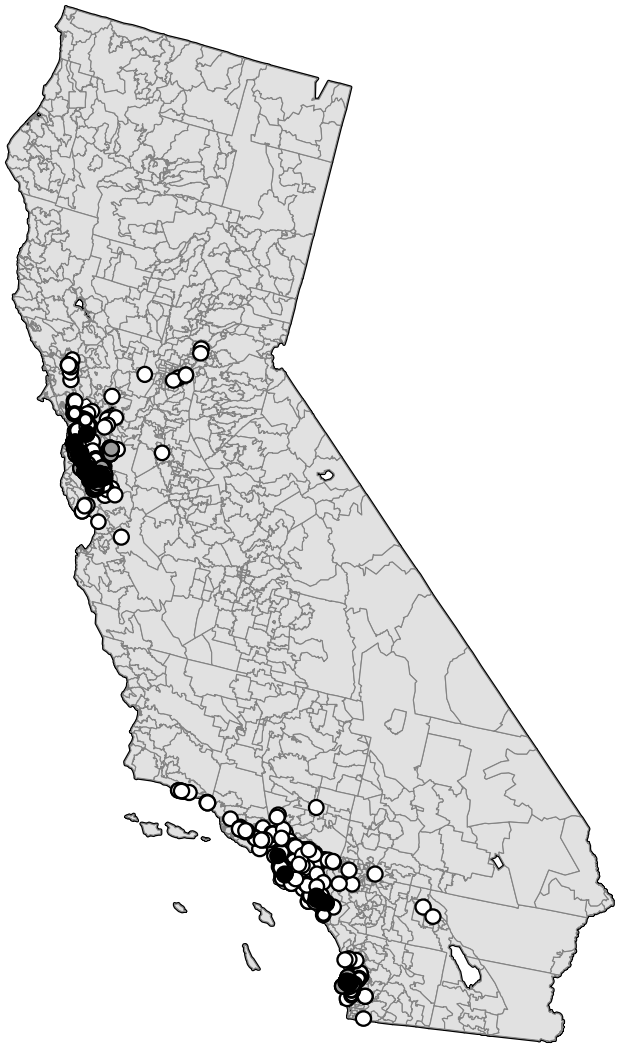

Figure 6b: California *P*-values at *d* = 5 miles

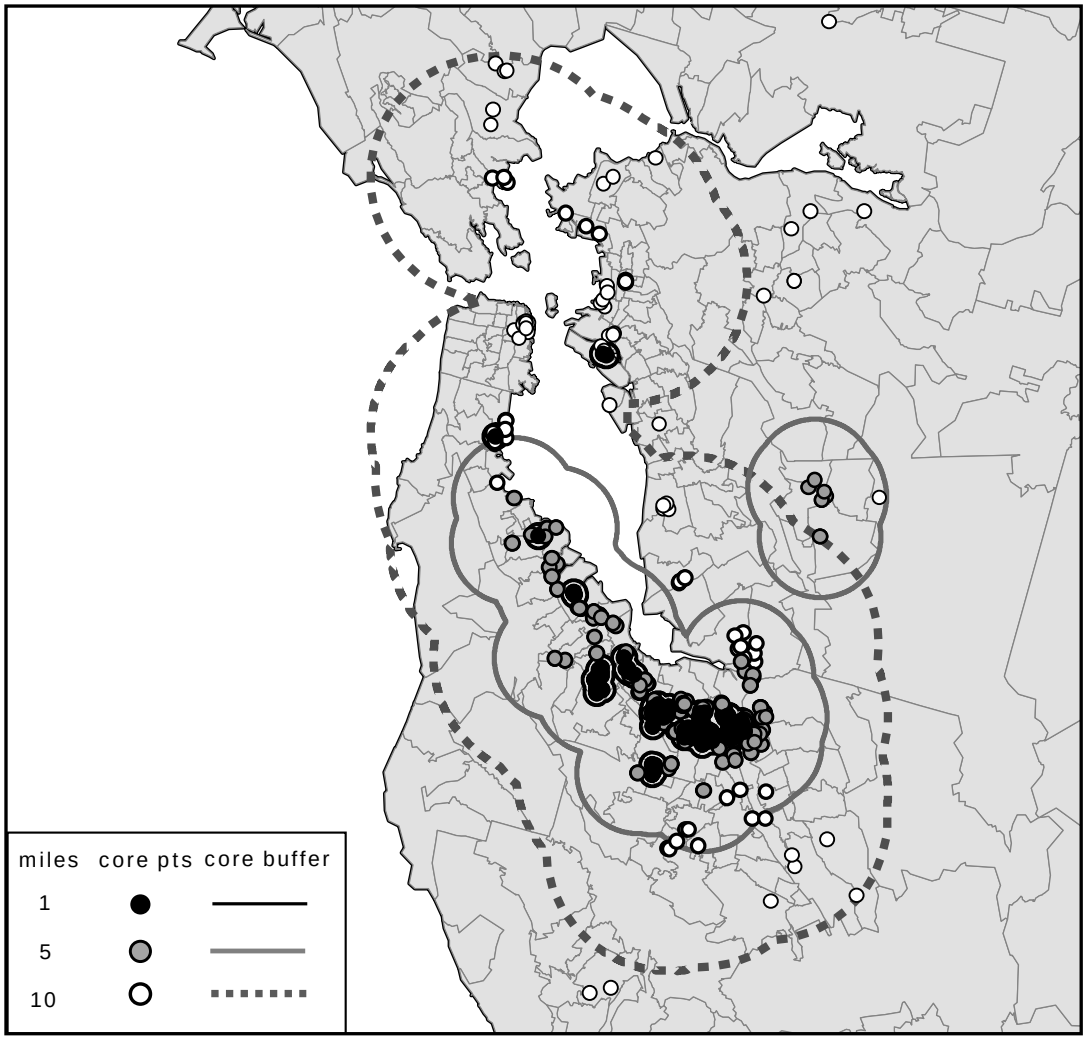

Figure 7: Multiscale Core Clusters in the San Francisco Bay Area

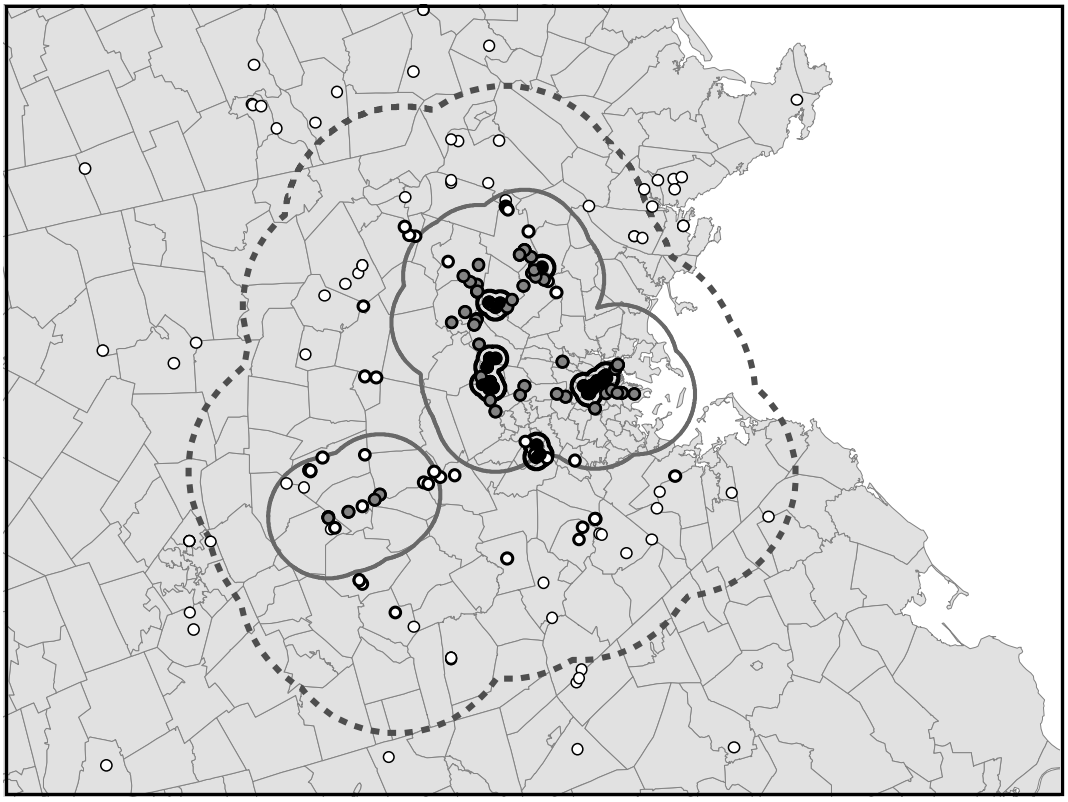

Figure 8a: Multiscale Core Clusters in Boston

Figure 8b: Proximity to Major Routes in Boston

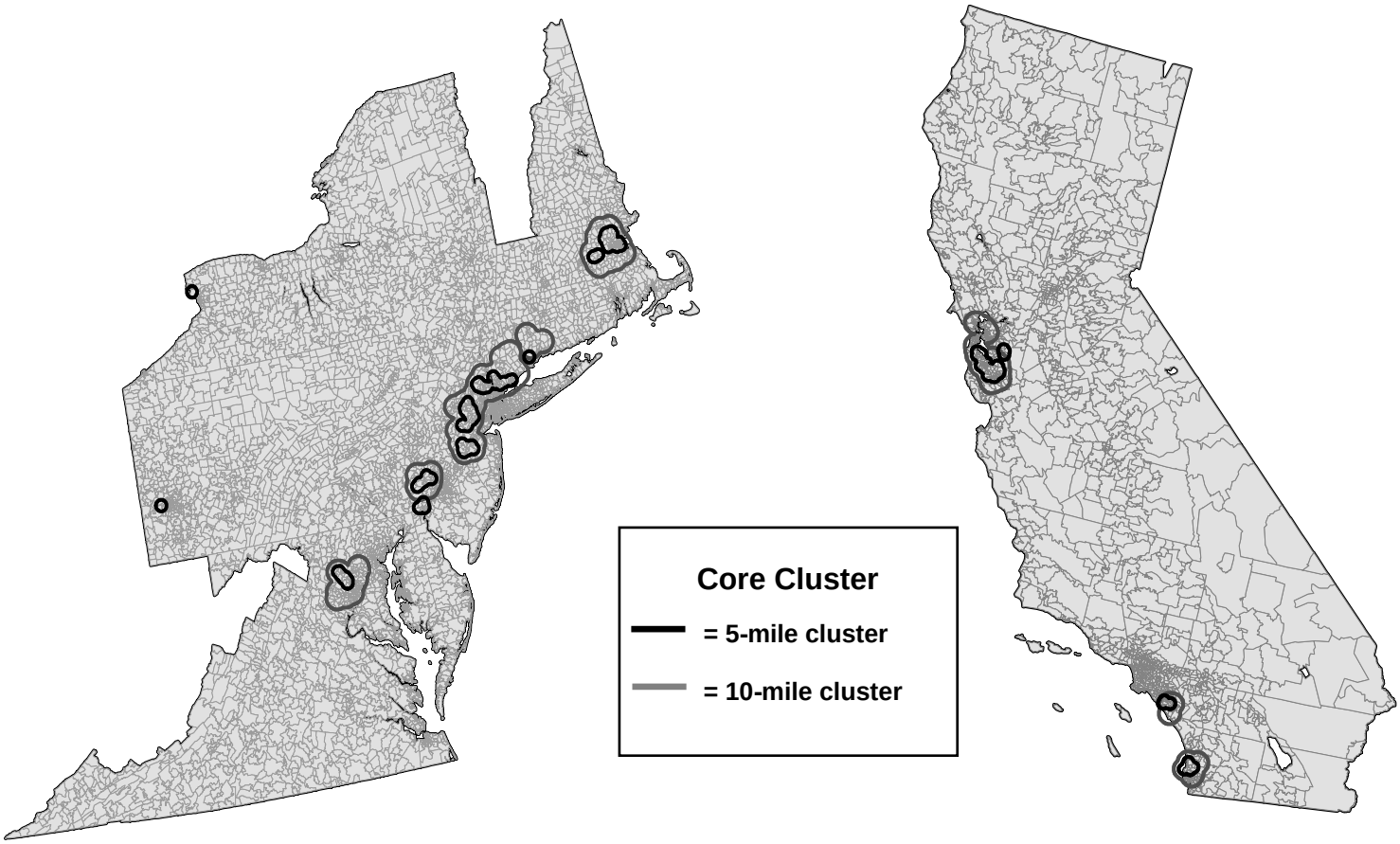

Figure 9a: Northeast Corridor Core Clusters

***d* = 5, 10**

Figure 9b: California Core Clusters
***d* = 5, 10**

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 is seen 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 3 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.

Appendix B: Robustness of Local *K*-Cluster Results

For completeness, we compare our local *K*-cluster results with two alternative econometric approaches that have been developed for resolving multiple testing problems in spatial applications, namely, the seminal approach of Besag and Newell (1991) and the more recent refinement of this approach by Kulldorff (1997). Both approaches resolve the multiple-testing problem by conducting only a single test.

In the present setting, one focuses on zip code areas (cells) and replaces 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 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 the *K*-function analyses above 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 located in the Northeast Corridor and Figure B2 for labs in California. By comparing these results with Figures 6a and 6b in the text, it is evident that both procedures are identifying essentially the same areas. This comparison thus serves as one type of robustness check on our local *K*-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 previously mentioned, 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" in order 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 it is much more flexible in nature.

A second limitation of these significance-maximizing approaches, which is less evident by visual inspection, is the path-dependent nature of cluster formation. As previously mentioned, 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, D.C. in Figure B1), this removal process creates no difficulties. But when subsequent clusters are in the same area as previous clusters (such as the New York area in Figure B1 and the Bay 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.

Appendix C: Description of the Major Areas of Agglomeration[50]

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.[51] 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 5-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 1-mile scale, Boston has five clusters, all of which are centered in the largest 5-mile cluster. The largest of these 1-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

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

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

the nine 1-mile clusters identified. Among these 1-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 5-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 5-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.

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, DC, in darker gray), where one 5-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 1-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 5-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 1-mile clusters. The most prominent 1-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 5-mile cluster consisting of 44 labs, and within it is a 1-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 5-mile cluster, there are two separate 1-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 1-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.

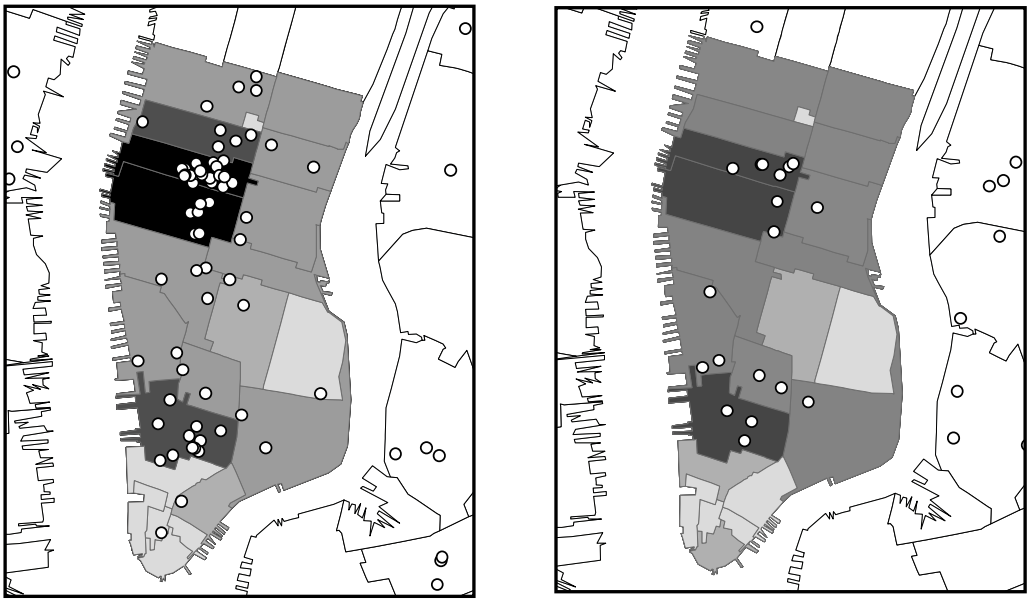

Figure A1. Manufacturing Establishment Counterfactuals (left panel) and Manufacturing Employment Counterfactuals (right panel)

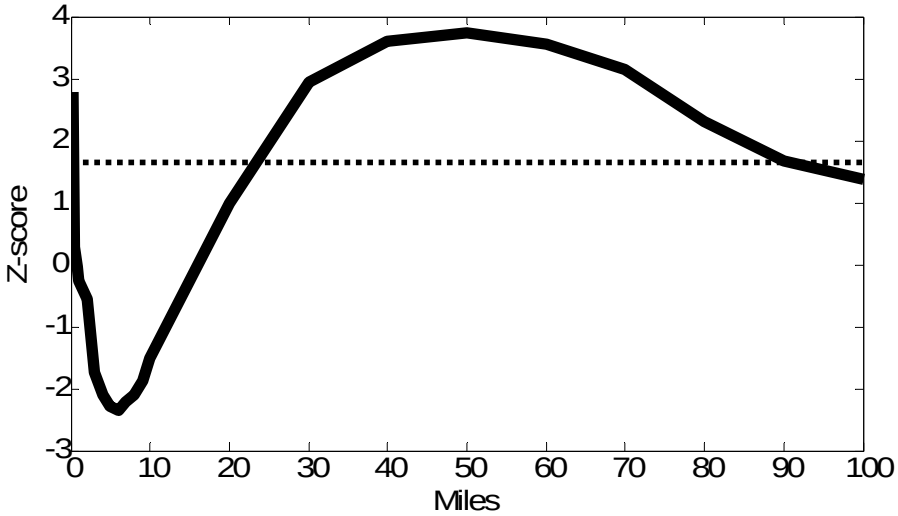

Figure A2: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Including South Manhattan

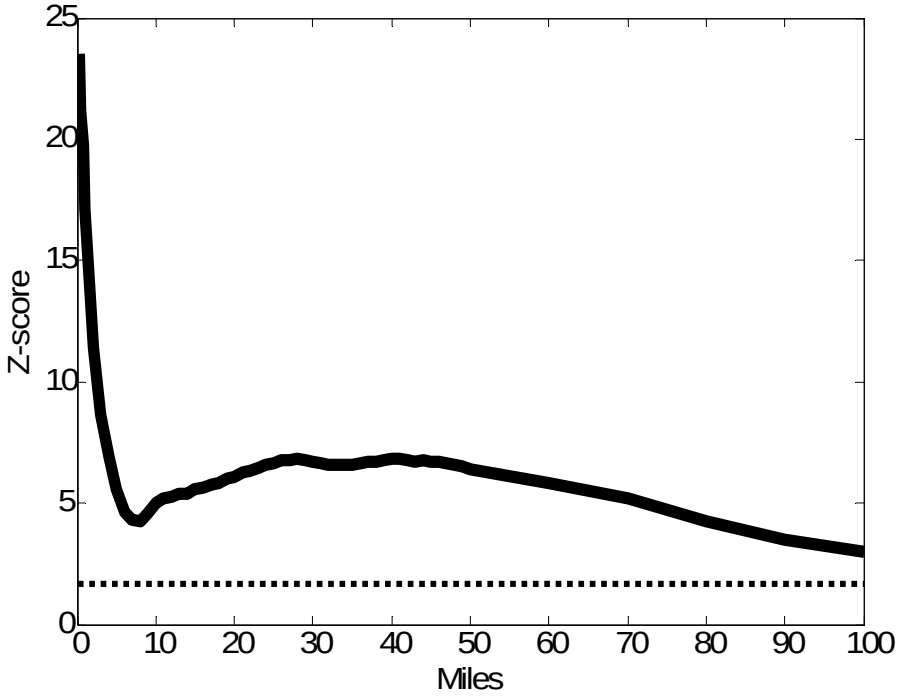

Figure A3: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Excluding South Manhattan

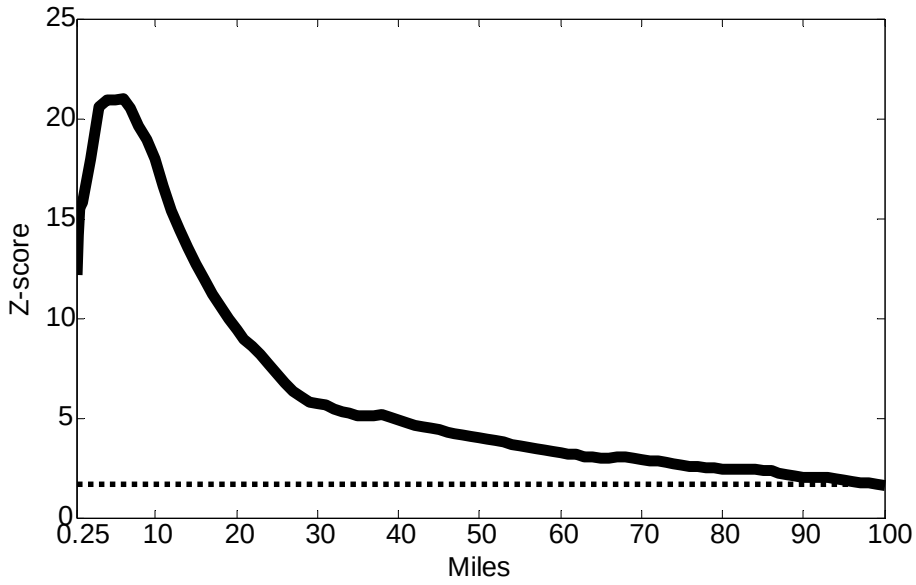

Figure A4: *Z*-scores Relative to Manufacturing Establishments for California

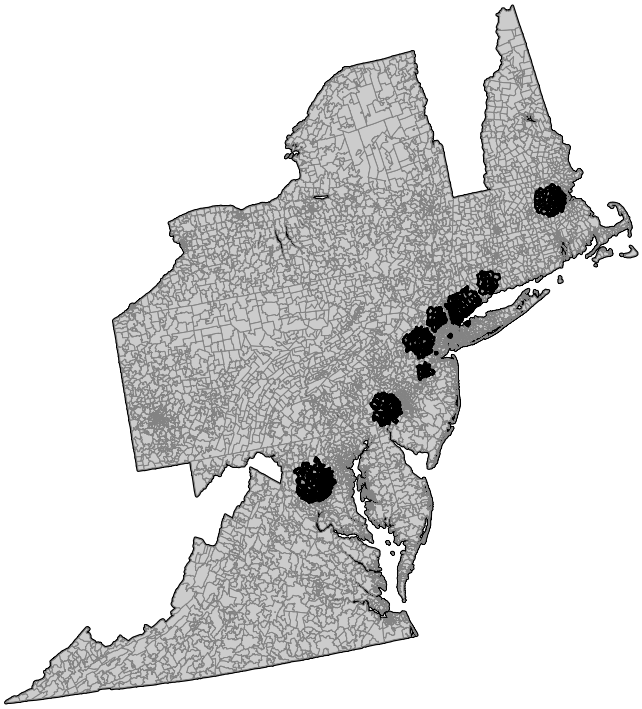

Figure B1: SATSCAN Clusters for the Northeast corridor

Figure B2: SATSCAN Clusters for California

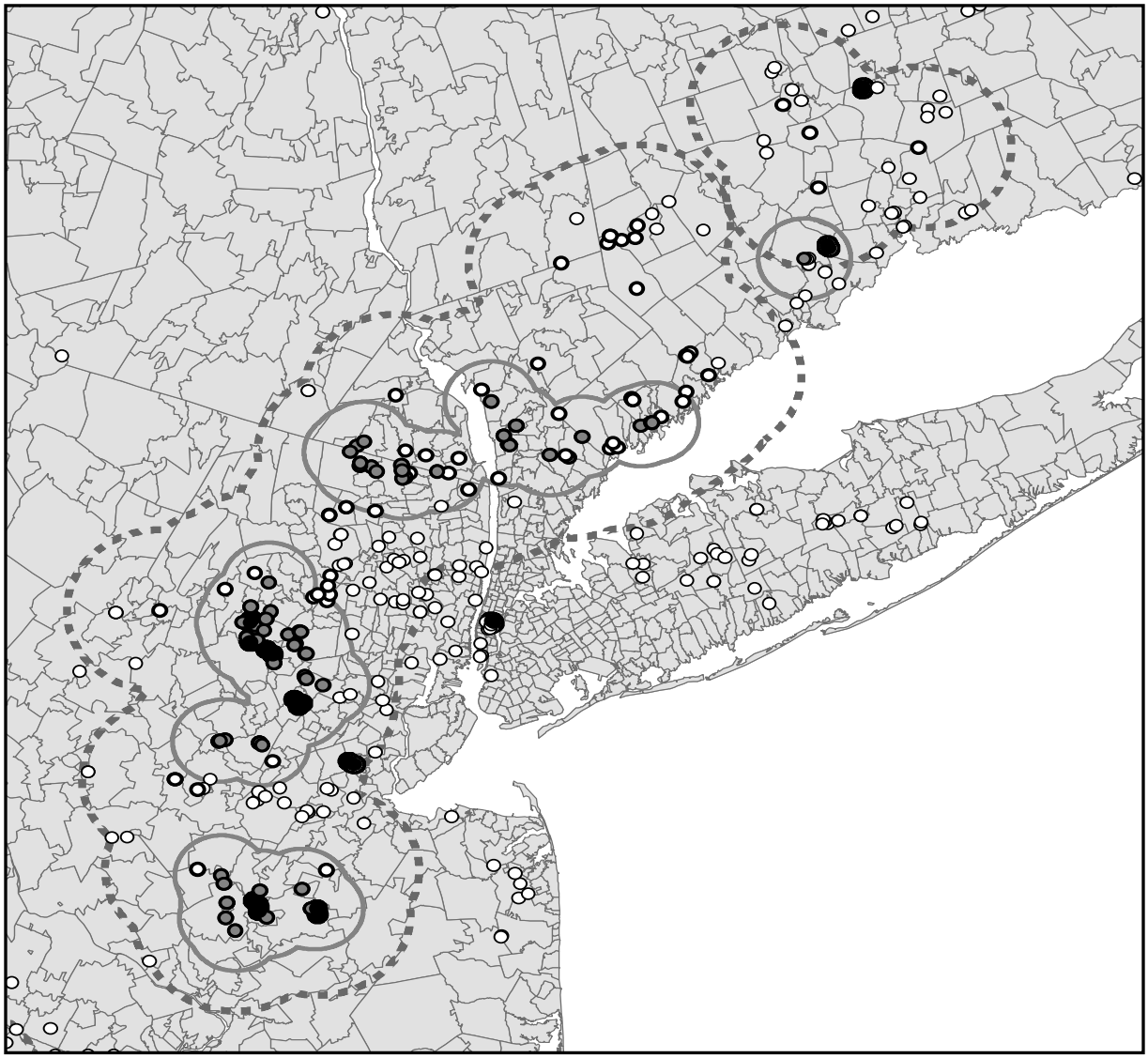

Figure C1: New York Core Clusters

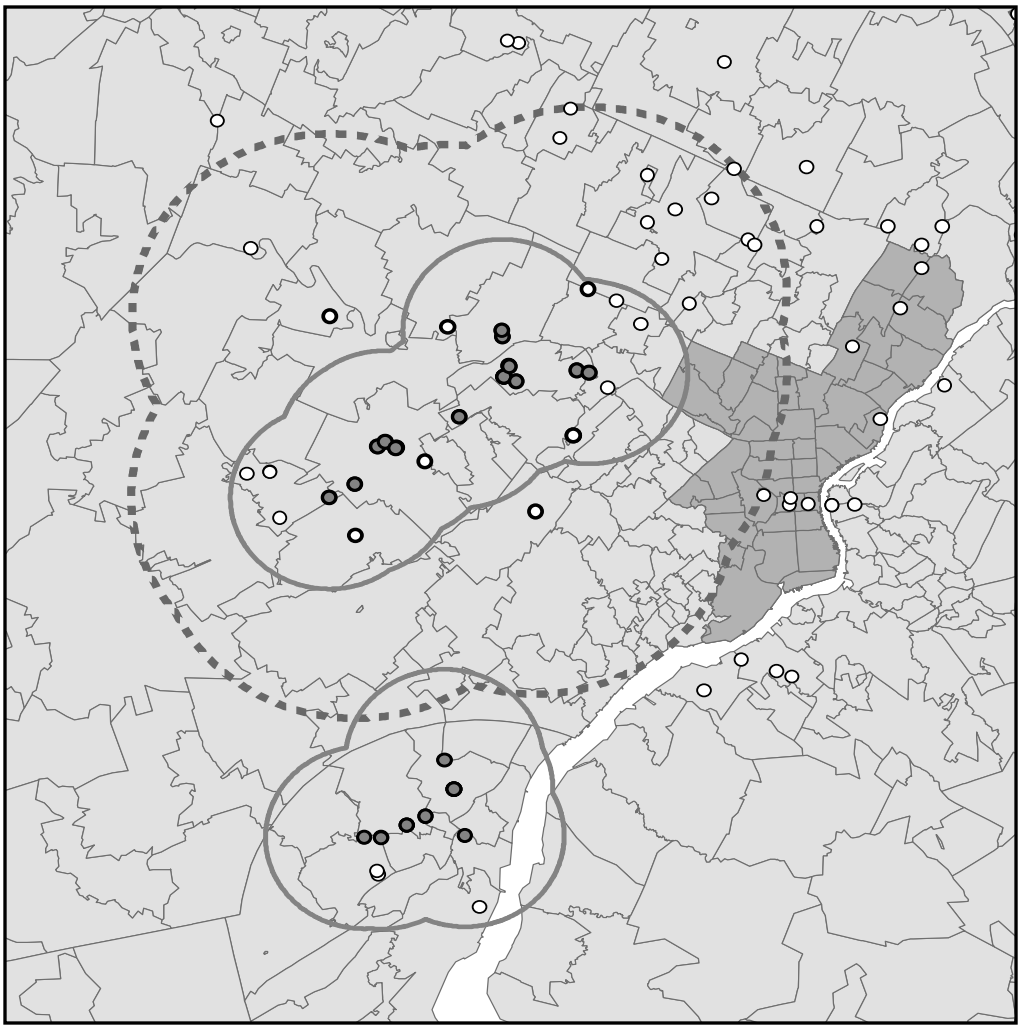

Figure C2: Philadelphia Core Clusters

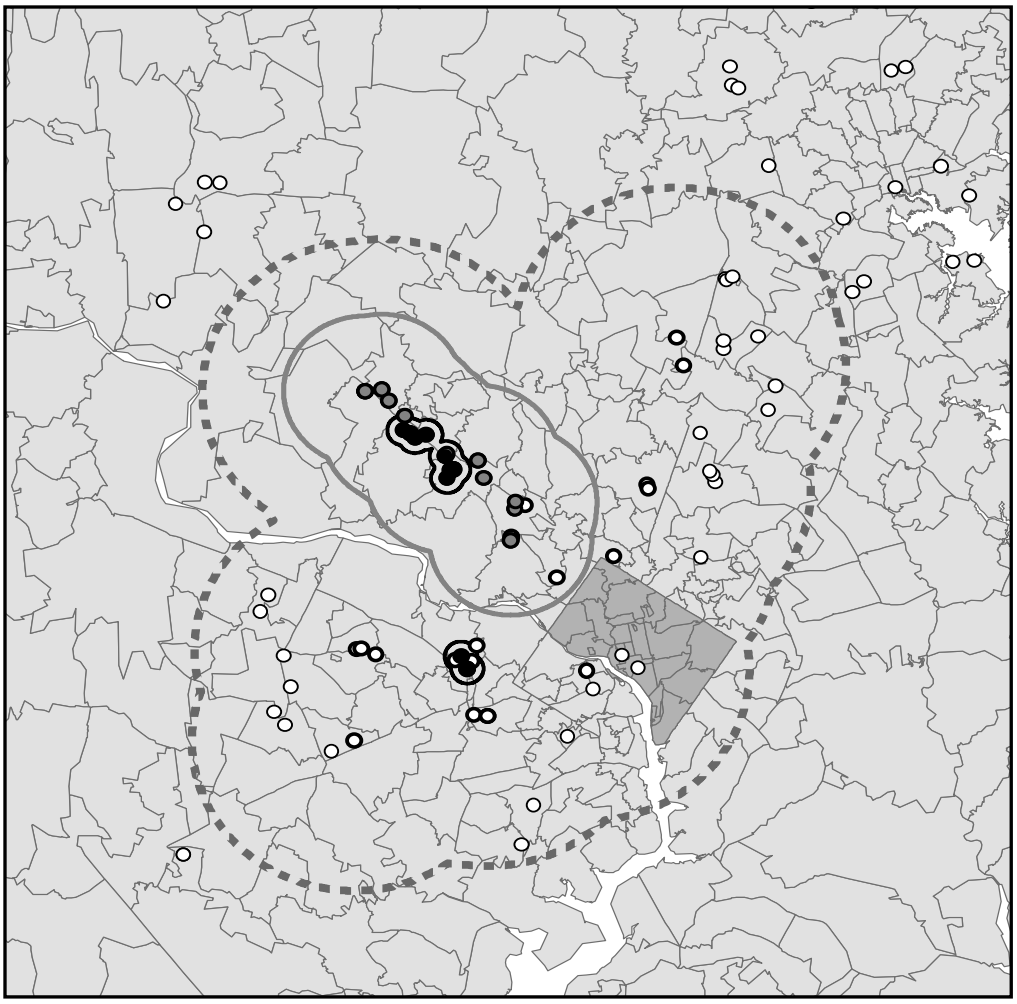

Figure C3: Washington, D.C. Core Clusters