WORKING PAPER NO. 17-18
THE AGGLOMERATION OF AMERICAN RESEARCH AND DEVELOPMENT LABS

Kristy Buzard
Syracuse University

Gerald A. Carlino
Research Department
Federal Reserve Bank of Philadelphia

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

Jake K. Carr
Ohio State University

Tony E. Smith
University of Pennsylvania

July 2017

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

The Agglomeration of American Research and Development Labs[*]

Kristy Buzard
Maxwell School, Syracuse University, Syracuse, NY 13244

Gerald A. Carlino and Robert M. Hunt
Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106

Jake K. Carr
Geography Department, The Ohio State University, Columbus OH, 43210

Tony E. Smith
Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104

July 2017

We employ a unique data set to examine the spatial clustering of about 1,700 private research and development (R&D) labs in California and across the Northeast corridor of the United States. Using these data, which contain the R&D labs' complete addresses, we are able to more precisely locate innovative activity than with patent data, which only contain zip codes for inventors' residential addresses. We avoid the problems of scale and borders associated with using fixed spatial boundaries, such as zip codes, by developing a new point pattern procedure. Our multiscale core-cluster approach identifies the location and size of significant R&D clusters at various scales, such as a half mile, one mile, five miles, and more. Our analysis identifies four major clusters in the Northeast corridor (one each in Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.) and three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

Keywords: spatial clustering, geographic concentration, R&D labs, innovation

JEL Codes: O31, R12

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

1. INTRODUCTION

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

A key issue addressed in this paper is how to measure the spatial concentration of R&D labs. A number of previous 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 a spatial Gini approach 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 an industry's industrial structure. A number of recent studies have used the EG index to measure the geographic clustering of manufacturing employment at the zip code, county, metropolitan statistical area (MSA), and state levels (see,

for example, Ellison and Glaeser, 1997; Rosenthal and Strange, 2001; and Ellison, Glaeser, and Kerr, 2010). While the EG index accounts for the general tendency for economic activity to concentrate spatially, it nonetheless suffers from a number of important aggregation issues that result from using a fixed spatial scale. As has been pointed out by Duranton and Overman (2005) — hereafter, DO — EG indices transform points on a map (establishments) into units in boxes (such as zip codes, counties, MSAs, and states). While this aggregation of the data facilitates computation, this approach leads to a number of aggregation issues. The first is known as the modifiable areal unit problem (MAUP). These metrics depend upon the boundaries used to demarcate regions, and conclusions may differ if counties versus states, for example, are used as boundaries. The MAUP grows in severity as the level of aggregation increases. A related issue is referred to as "border effects": each region is considered an exclusive zone, and the closeness of activity in neighboring regions is not factored in. While Philadelphia and Montgomery counties border each other and have activity spilling across them, in county level analyses they are treated as being as distant from each other as they are from Los Angeles County. These partitions often lead to underestimations of concentration.

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

[1] Other studies that have used continuous measures of concentration include those by Marcon and Puech (2003) for French manufacturing firms; Arbia, Espa, and Quah (2008) for patents in Italy; and Murata, et al. (2015) for patent

Following 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 lab locations at each spatial scale from the distribution of an appropriately defined measure of economic activity (such as manufacturing employment) at that scale. This is important because studies have shown that manufacturing activity is agglomerated at various spatial scales (e.g., Ellison and Glaeser, 1997; Rosenthal and Strange, 2001; and Ellison, Glaeser, and Kerr, 2010) and the large majority of R&D activity is performed by manufacturing firms. Our main results take manufacturing employment as the benchmark, but our findings are robust to alternative benchmarks such as manufacturing establishments and the total employment of science, technology, engineering, and math (STEM) workers.

While this multiscale approach is similar in spirit to that of DO, our test statistics are based on Ripley's (1976) *K*-function rather than the "*K*-density" approach of DO. While the DO approach can reveal the spatial scale at which concentration occurs, it does not tell us where in space the concentration occurs. *K*-functions can easily be disaggregated to yield information about the *spatial locations* of clusters of R&D labs at various spatial scales. We take advantage of this feature of *K*-functions to perform the local cluster analysis in Section 4.

We begin the analysis by using global *K*-function statistics to test for the presence of significant clustering over a range of spatial scales. Our data set consists of almost 1,700 R&D labs in California and in a 10-state area in the Northeast corridor of the United States. We find strong evidence of spatial clustering at even very small spatial scales — distances as small as one-half

citations. Kerr and Kominers (2015) use continuous measures in a more general model, one application of which uses data on patent citations. See Carlino and Kerr (2015) for a recent review of this literature.

mile — and this clustering tends to exhibit rapid attenuation as scales increase. This pattern is consistent with empirical research on human capital spillovers and agglomeration economies.

Next, we focus on the question of *where* clustering occurs using a more refined procedure based on local *K*-functions. We introduce a novel procedure called the multiscale core-cluster approach to identify the location of clusters and the number of labs in these clusters. Core clusters at each scale are identified in terms of those points with the most significant local clustering at that scale. 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 five and 10 miles, these core clusters reveal the presence of the major agglomerations visible on any map. Our analysis identifies four major clusters in the Northeast corridor (one each in Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.,) and three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

Our work differs from past studies in a number of ways. Rather than looking at the geographic concentration of firms engaged in the production of goods (such as manufacturing), 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 also permits such clustering to be identified at a range of scales in continuous space rather than at a single predefined scale. Importantly, the use of the R&D lab data allows us to more accurately assign labs to locations since we have their complete addresses, an improvement on using patent data to measure the location of innovative activity. This allows us to implement tests for geographic concentration with very high precision at even the smallest of spatial scales. An important

limitation associated with patent data used in most past studies to analyze the spatial concentration of innovative activity is that only the zip codes of the inventors' residential addresses are listed on the patent. With patent data, one can only consider the geographic clustering of innovative activity at the average size of zip codes, and this is subject to measurement error if inventors live and work in different zip codes. As shown in Table 1, the typical size of a zip code in the Northeast corridor is about 30 square miles, while the average size in California is almost 100 square miles. Use of the patent information is further complicated in that many patents have multiple inventors who often reside in different locations. Patents do contain information on the assignee (usually the company that first owned the patent), but researchers typically do not use the assignee address 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). Finally, unlike the *K*-density approach, our local *K*-function method can be used to identify where in space clustering is occurring, something that is new to the agglomeration literature.

We also use the global *K*-function technique to examine the concentration of R&D labs in specific two-digit Standard Industrial Classification (SIC) industries relative to the concentration of labs across all industries. This sets a higher bar in our tests of spatial concentration as well as avoids a potential measurement issue at very small spatial scales that may occur when we use a benchmark that is not point pattern data. We find at small spatial scales (such as within a two- to three-block area) that 37 percent of the industries in the Northeast corridor are significantly more concentrated compared with overall R&D labs and that none are significantly more dispersed. In California, 50 percent are significantly more localized than R&D labs in general. The rapid attenuation of significant clustering of labs for many individual industries is consistent with the

view that at least one important component of agglomeration economies must be highly localized.

2. THEORY AND DATA

2.1 Data

We introduce a novel data set in this paper based on the 1998 edition of the *Directory of American Research and Technology*, which profiles the R&D activities of public and private enterprises in the United States. The directory includes virtually all nongovernment facilities engaged in any commercially applicable basic and applied research. For this paper, our data set contains the R&D establishments ("labs") associated with the top 1,000 publicly traded firms ranked in terms of R&D expenditure in Compustat.[2] These firms represent slightly less than 95 percent of all R&D expenditures reported in the 1999 edition of Compustat for 1998.[3] Thus, each lab in our data set is associated with its Compustat parent firm and information on its street address and a text description of its research specialization(s) to which we have assigned the corresponding four-digit SIC codes. Using the address information for each private R&D establishment, we geocoded the locations of more than 3,000 labs (shown in Figure 1).

[2] 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 U.S. Securities and Exchange Commission. The following company databases were also used to supplement or confirm our main sources: Hoover's Company Records database, Mergent Online, the Harris Selectory Online database, and the American Business and Service Directory.

[3] Although we cannot know for sure the impact on the analysis of including smaller labs, if these labs tend to cluster near larger labs as is widely believed, then we will underestimate the significance of clustering of R&D labs. Some clusters that fail our tests of significance may indeed be significantly clustered in that case as well, and some cluster boundaries may be slightly different than what we identify.

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

Even at the most aggregate level, it is easy to establish that R&D activity is relatively concentrated in these two regions. For example, in 1998, one-third of private R&D labs and 29 percent of private R&D expenditures were located within the Northeast corridor compared with 22 percent of total employment (21 percent of manufacturing employment) and 23 percent of the population. California accounted for almost 22 percent of all private R&D labs and 22 percent of private R&D expenditures in 1998 compared with 12 percent of total employment (11 percent of manufacturing employment) and 12 percent of the population. Together, these two regions accounted for the majority of all U.S. private R&D labs (and private R&D expenditures) in 1998.[4] This concentration is consistent with Audretsch and Feldman (1996), who report that the top four states in terms of innovation in their data are California, Massachusetts, New Jersey, and New York.

In our formal analysis that follows, 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. These data were obtained from the 1998 volume of Zip Code Business Patterns. Given that the vast majority of our R&D labs are owned by manufacturing firms, manufacturing

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

employment represents a good benchmark.[5] It is possible that owners of R&D labs locate these facilities using different factors than they use for locating manufacturing establishments. We address this concern by using total employment data at the census block level for 2002 from the Longitudinal Employer-Household Dynamics (LEHD) survey to identify feasible lab locations within each zip code.

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

2.2 Theory

How do we account for the geographic concentration of R&D activity observed in this paper? Much of the theoretical literature on urban agglomeration economies has focused on externalities in the production of goods and services rather than on invention itself. Nevertheless, the three

[5] In Section 5.1, we develop an alternative benchmark (or backcloth) for analyzing R&D clustering with respect to STEM workers. In Appendix A, we report results of our analyses using manufacturing establishments as an alternative benchmark. As we will see, our main findings are highly robust to the use of alternative backcloths.

formal mechanisms primarily explored in the literature —knowledge spillovers, sharing, and matching — are also relevant for innovative activity.[6]

2.2.1 Knowledge Spillovers

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

2.2.2 Sharing and Matching

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

2.2.3 Connection Between Theory and Evidence

[6] See Duranton and Puga (2003) for a more thorough discussion of the microfoundations of urban agglomeration economies.

In this paper, we do not attempt to identify the mechanism(s) underlying the geographic concentrations of labs we observe. We abstract from theoretical considerations and simply impose a statistical requirement on our tests for localization to determine whether R&D labs are clustered. This approach is based on a test of a simple location model (i.e., R&D locations are more clustered than would be expected from random draws from the distribution of overall manufacturing employment).

3. GLOBAL CLUSTER ANALYSIS

A key question is whether the overall patterns of R&D locations in the two regions we examine exhibit more clustering than would be expected from the spatial concentration of manufacturing in those regions. To address this question statistically, we start with the null hypothesis that R&D locations are mainly determined by the distribution of manufacturing employment within a zip code. Since the data are at the zip code level, it is necessary to assume that manufacturing employment is uniformly distributed within a zip code. This assumption is reasonable if zip codes are sufficiently small. Since we know the street addresses of the labs, then, at spatial scales smaller than the typical zip code size, these locations will tend to exhibit some degree of spurious clustering of labs relative to random locations.[7] In our sample, the radius of a typical zip code is about two miles for zip codes containing at least one lab (Table 1). 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 our null hypothesis. To do this, we obtained total employment data at the census block level for 2002 from the LEHD survey[8] and use these data to identify feasible lab

[7] We thank Gilles Duranton for this observation.

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

locations within each zip code area.[9] 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.[10]

Our basic null hypothesis is the following:

Hypothesis 1

> *Lab locations are no more concentrated than manufacturing employment at the zip code level and then no more concentrated than total employment within each zip code.*

In order to test whether the observed R&D lab locations are agglomerated relative to the benchmark identified in Hypothesis 1, we generate counterfactual locations consistent with Hypothesis 1 using a three-stage Monte Carlo procedure. In this procedure, (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 mentioned that actual locations are almost always along streets and cannot, of course, be random within blocks. But, as discussed in Section 3.2, blocks themselves are sufficiently small to allow such random effects to be safely ignored at the scales of most relevance for our purposes.

[9] 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.

[10] 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.

By repeating this procedure separately for the Northeast corridor (with $n = 1,035$ 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 Hypothesis 1, 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 randomly drawn according to the distribution of manufacturing employment.

3.1 *K*-Functions

The most popular measure of clustering for point processes is Ripley's (1976) *K*-function, $K(d)$, which (for any given mean density of points) is essentially the expected number of additional points within distance d of any given point.[11] In particular, if $K(d)$ is higher than would be expected under Hypothesis 1, 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 (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:[12]

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

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

[12] 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.

As described in Section 3, we draw a set of *N* point patterns, $X^s = (x_i^s : i = 1,..,n)$, $s = 1,..,N$, and for a selection of radial distances, $D = (d_1,..,d_k)$, we 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 Hypothesis 1.

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,[13] and if the number of these $N+1$ sample values at least as large as $\hat{K}^0(d)$ is denoted by $N^0(d)$, then the fraction

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

is a (maximum likelihood) estimate of the *p-value* for a one-sided test of Hypothesis 1.

For example, if $N = 999$ and $N^0(d)$ = 10 so that $P(d) = 0.01$, then under Hypothesis 1, 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.

3.2 Test Results for Global Clustering

[13] At this point it should be noted that, since all sample *K*-functions are subject to the same "edge effects" as the observed sample, the presence of edge effects should not influence our test results.

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

Given this range of possible spatial scales, our results show that clustering in the Northeast corridor is so strong (relative to manufacturing employment) that the estimated *p*-values are 0.001 for all scales considered. The results are the same for California up to about 60 miles, and 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 these data.[15]

---

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

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

3.3 Variations in Global Clustering by Spatial Scale

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

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

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 2a and those for California are shown in Figure 2b. Significance levels decrease nearly monotonically for California, while in the Northeast corridor, we see a hump-shaped pattern. The high *z*-scores are consistent with the significance of the Monte Carlo results noted previously but add more detailed information about the patterns of significance.[16] Observe that, in both figures, clustering is most significant at 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 Section 2.2.[17]

3.4 Relative Clustering of R&D Labs by Industry

[16] The benchmark value of $z = 1.65$, shown as a dashed line in both Figures 2a and 2b, corresponds to a *p*-value of 0.05 for the one-sided tests of Hypothesis 1in expression (2).

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

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

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

Hypothesis 2

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.*

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

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 the $\hat{K}(d)$ values for the observed pattern of labs in industry j with those for a set of N 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 the null hypothesis that the probability of finding a randomly selected R&D lab associated with a particular industry is proportional to the total number of R&D labs in that area.

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

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

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

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

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

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

[21] The eight industries are Chemicals and Allied Products; Rubber and Miscellaneous Plastics Products; Primary Metal Industries; Industrial and Commercial Machinery; Electronic and Other Electrical Equipment; Transportation Equipment; Measuring, Analyzing, and Controlling Instruments; and Business Services.

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

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 two miles. In California, the clustering of R&D labs is most significant at very small spatial scales for four of the seven industries shown in Table 3b. Two of the other industries, Electronic and Other Electrical Equipment and Business Services, have local peaks at one-half mile and at one mile, respectively.

In addition, Figure 3a shows rapid attenuation of *z*-scores at small scales for all seven industries in the Northeast 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 Allied Products 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 the computer programming or data processing subcategories.) Turning to California, Figure 3b shows rapid attenuation of *z*-scores at small scales for four of these seven industries. The other three industries, Industrial and Commercial Machinery, Electronic and Other Electrical Equipment, and Business Services (mostly in the computer programming and data processing subcategory), exhibit an opposite trend in which relative clusters become more significant at larger scales.

Finally, it is of interest to note that three industries are among the most significantly clustered industries in both the Northeast corridor and California, namely Chemicals and Allied Products, Business Services, and Measuring, Analyzing, and Controlling Instruments. The Chemical and Allied Products 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 and Allied Products industry accounts for about 16 percent of the labs we study. Thus, at least within the geographic area under study, this industry is seen to be a major contributor to the overall clustering pattern of R&D shown in Figures 2a and 2b. But it should be equally clear from Figures 3a and 3b that significant clustering occurs in many other industries as well. So, clustering of R&D labs is by no means specific to drugs and chemicals.

4. LOCAL CLUSTER ANALYSIS

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

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

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

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

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

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 in Equation 4.

4.1 Local Testing Procedure

For the local testing procedure, we use Hypothesis 1 from Section 3: R&D labs are distributed in a manner proportional to manufacturing employment at the zip code level and proportional to total employment at the block level.[24] 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 Hypothesis 1 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* as in Figures 4a and 4b. 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

[24] We replace manufacturing employment with STEM workers in Section 5.1 and with manufacturing establishments in Appendix A as robustness checks.

p-values necessarily suffer from "multiple testing" problems, which we address in later sections and more systematically in Appendix B.

4.2 Test Results for Local Clustering

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

A visual inspection of the *p*-value maps generated by our test results showed that the clearest patterns of distinct clustering can be captured by the three representative distances, $D = \{1, 5, 10\}$. Of these three, the single best distance for revealing the overall clustering pattern in the entire data set appears to be five miles, as illustrated for the Northeast corridor and California in Figures 4a and 4b, respectively. As seen in the legend, those R&D locations, i, exhibiting

maximally significant clustering — $P_i(5) = 0.001$— are shown as 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.[25] In California, there are again three distinct groups, roughly described (from north to south) as the "San Francisco Bay Area," "Los Angeles area (mainly Irvine)," and "San Diego." While these patterns are visually compelling, it is important to establish such results more formally.

5. IDENTIFYING SPATIAL CLUSTERS: THE MULTISCALE CORE-CLUSTER APPROACH

The global cluster analysis in Section 3 identified the *scales* at which clustering is most significant (relative to manufacturing employment). The local cluster analysis in Section 4 provided information about *where* clustering is most significant at each spatial scale. But neither of these methods formally identifies or defines specific "clusters" of labs. In this section, we apply some additional techniques to identify clusters, which we call the *multiscale core-cluster* approach.

As discussed in Appendix B, a number of cluster-identification techniques have been developed to identify sequences of clusters that are individually "most significant" in an appropriate sense.[26] The present approach is based more directly on the *K*-function methods and, in particular, focuses on the *multiscale* nature of local *K*-functions. More specifically, this

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

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

clustering procedure starts with the local point-wise clustering results in Section 4.1 and seeks to identify subsets of points that can serve as "core" cluster points at a given selection of relevant scales, *d*. Here, we again focus on the three scales, $D = \{1,5,10\}$, used in Section 4.1. At each scale, $d \in D$, we define a *core point* to be a maximally significant R&D lab (i.e., with a local *K*-function *p*-value of 0.001 using the 999 simulations of *K* at distance *d* in Section 4). In order to exclude "isolated" points that simply happen to be in areas with little or no manufacturing, we also require that there be at least *four* other R&D labs within this *d*-mile radius. Finally, to identify distinct clusters of such points, we create a *d*-mile-radius buffer around each core point (in ArcMap). We designate the set of points (labs) in each connected component of these buffer zones as a *core cluster* of points at scale *d*. Hence, each such cluster contains a given set of "connected" core points along with all other points that contributed to their maximal statistical significance at scale *d*. These concepts are best illustrated by examples.

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

A second example, from the Northeast corridor, is provided by the hierarchical complex of R&D clusters in the Boston area, shown in Figure 6a. Here again, the entire Boston area is itself a single 10-mile cluster. Moreover, within this area, there is again a dominant five-mile core cluster containing the five major one-mile clusters in the Boston area. The largest of these is concentrated around the university complex in Cambridge, while the others are centered at points along Route 128 surrounding Boston. This is seen more clearly in Figure 6b,[27] which also shows that most R&D labs in the Boston area are located in close proximity to major transportation routes, including Interstate Routes 90, 93, 95, and 495.

Note, finally, that while the clusters in both Figures 5 and 6a tend to be nested by scale, this is not always the case.[28] For example, the five-mile "Livermore Lab" cluster in Figure 5 is seen to be mostly outside the major 10-mile cluster. Here, there is a concentration of six R&D labs within two miles of each other, although Livermore is relatively far from the Bay Area. So, while this concentration is picked up at the five-mile scale, it is too small by itself to be picked up at the 10-mile scale.

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

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

[28] The area of five-mile clusters in the Northeast corridor is on average 277 square miles, while the area of 10-mile clusters in the Northeast corridor is on average 2,498 square miles. In California, the corresponding areas for five- and 10-mile clusters are 319 and 1,326 square miles, respectively.

maximizing methods.[29] In addition, since all core clusters are determined simultaneously, the path-dependency problem of sequential methods does not arise.

In summary, an overall depiction of core clusters for both the Northeast corridor and California (at scales, $d = 5,\ 10$) is shown in Figures 7a and 7b, respectively. Figure 7a shows the four major clusters identified for the Northeast corridor (one each in Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.), while Figure 7b shows the three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

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).[30] Thus, the primary objective of this more heuristic procedure is to produce explicit representations of clusters that capture both their relative shapes and concentrations in a natural way.

[29] This point is demonstrated in Appendix B.

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

Finally, in Buzard et al. (2016), we document that patent citations are more highly geographically localized within these clusters of R&D labs than outside them. We argue that this demonstrates that these clusters are associated with economically meaningful outcomes.

5.1. Alternate Cluster Boundaries: Employment in STEM Industries as Benchmark

Firms' desires to take advantage of knowledge spillovers is one mechanism that could explain spatial clustering of innovative activity, and the specific clusters identified in this paper are consistent with a knowledge spillover explanation. It is also possible that R&D activity is geographically concentrated to take advantage of labor market pooling. As we have shown, one important concentration of R&D labs is found in Cambridge, MA, and another important clustering is found in the Silicon Valley. These labs are close to large pools of STEM graduates and workers, the very workers R&D activity requires. Manufacturing activity tends to employ a more general workforce than does innovative activity and may therefore be more geographically dispersed compared with innovative activity.

To address this concern, we first develop a measure of STEM workers by location. For our backcloth, we replace the number of manufacturing employees in each zip code area with an estimate of the number of STEM workers. This is constructed using the proportion of STEM jobs in each four-digit North American Industry Classification System (NAICS) industry multiplied by the number of jobs in each industry reported in the Zip Code Business Patterns data. Hypothesis 1 becomes:

Hypothesis 3

> *Lab locations are no more concentrated than STEM worker employment at the zip code level and then no more concentrated than total employment within each zip code.*

We report the results of this alternative test for five- and 10-mile clusters in the Northeast corridor (Figure 8a) and in California (Figure 8b). The clusters identified using STEM workers as a reference are in remarkable agreement with the clusters obtained when using manufacturing employment as the backcloth. The four major clusters in the Northeast corridor (Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.),) previously identified in Figure 7a resurface when using the STEM worker backcloth. Similarly, the three major clusters identified in Figure 7b for California (one each in the Bay Area, Los Angeles, and San Diego) reemerge using the STEM worker backcloth.

However, there are certain differences between the results using the different backcloths. Notice first that the STEM worker clusters appear to be larger than those found when using the manufacturing employment backcloth. This is true for the clusters in the Northeast corridor and in California. In addition, a number of additional smaller clusters emerge under the STEM worker backcloth. Five additional 10- mile clusters are found in the Northeast corridor (one each in Lancaster, PA; Hagerstown, MD; Binghamton, NY; Syracuse, NY; and Rochester, NY) and also in Richmond, VA. Three additional 10-mile clusters are found in California (one each in Santa Rosa, Santa Barbara, and Malibu).

6. CONCLUDING REMARKS

In this paper, we use a new data set on the location of R&D labs and several distance-based point-pattern techniques to analyze the spatial concentration of the locations of more than 1,700 R&D labs in California and in a 10-state area in the Northeast corridor of the United States. Rather than using a fixed spatial scale, we describe the spatial concentration of labs more precisely, by examining spatial structure at different scales using Monte Carlo tests based on

Ripley's *K*-function. Geographic clusters at each scale are identified in terms of statistically significant departures from random locations reflecting the underlying distribution of economic activity. We present robust evidence that private R&D labs are indeed highly concentrated over a wide range of spatial scales.

We introduce a novel way to identify the spatial clustering of labs called the *multiscale core-cluster* approach. The analysis identifies four major clusters in the Northeast corridor (one each in Boston, New York–Northern New Jersey, Philadelphia–Wilmington, and Washington, D.C.,) and three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego). Work by Buzard et al. (2016) demonstrates that these clusters are associated with economically meaningful outcomes such as patenting.

REFERENCES

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

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

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

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

Besag, Julian, and James Newell. "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society*, 154 (1991), pp. 143–55.

Buzard, Kristy, Gerald A. Carlino, Robert M. Hunt, Jake K. Carr, and Tony E. Smith. "Localized Knowledge Spillovers: Evidence from the Agglomeration of American R&D Labs and Patent Data," Federal Reserve Bank of Philadelphia Working Paper No. 16-25 (2016).

Carlino, Gerald A. and William R. Kerr. "Agglomeration and Innovation," in: Duranton, Gilles, J. Vernon Henderson, and William Strange (Eds.), *Handbook of Regional and Urban Economics*, Vol. 5A, pp. 349–404 (2015), North Holland, Amsterdam.

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

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

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. 1077–1106.

Duranton, Gilles, and Diego Puga. "Mirco-Foundations of Urban Agglomeration Economies," in: Henderson, J. Vernon, and Jacques-Francoise Thisse (Eds.), *Handbook of Regional and Urban Economics*, Vol. 4, pp. 2063–2118 (2003), North Holland, Amsterdam.

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

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

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

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

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

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

Kulldorff, Martin. "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, 26 (1997), pp. 1481–96.

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

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

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

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

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

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

Table 1: Summary Statistics					
Northeast Corridor (10-State)					
Variable	Mean	Std. Dev.	Median	Minimum	Maximum
	All Zip Codes (6,044)				
Land Area, miles^2	29.10	37.61	16.87	0.01	468.16
Radius*	2.55	1.66	2.32	0.06	12.21
Total Employment	4,307.22	8,994.78	1,001.00	0.00	194,114.00
Manufacturing Employment	557.20	1,213.46	76.30	0.00	22,808.31
Total Establishments	250.36	370.76	97.00	1.00	6,962.00
Manufacturing Establishments	11.39	16.65	4.00	0.00	132.00
Labs	0.17	0.74	0.00	0.00	13.00
	Zip Codes with 1 or More Labs (549)				
Land Area, miles^2	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, miles^2	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, miles^2	18.78	37.75	8.19	0.07	385.98
Radius*	2.02	1.38	1.61	0.15	11.08
Total Employment	19,482.47	17,300.91	15,088.00	0.00	79,766.00
Manufacturing Employment	3,607.79	5,188.27	1,569.00	0.00	27,186.00
Total Establishments	1,173.13	677.45	1,065.50	0.00	3,527.00
Manufacturing Establishments	94.52	96.32	62.00	0.00	636.00
Labs	3.16	4.90	1.50	1.00	33.00

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

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

Table 2a: Concentration of Labs by Industry in the Northeast Corridor (*P-values*)†									
			Miles						
INDUSTRY	**SIC**	**LABS**	**0.25**	**0.5**	**0.75**	**1**	**5**	**20**	**50**
Metal Mining	10	4	0.5021	0.5029	0.5044	0.5052	0.5227	0.1674	0.4149
Oil and Gas Extraction	13	3	0.5011	0.5019	0.5026	0.5034	0.5137	0.0906	0.2286
Food 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 and Other Finished Products	23	5	0.5036	0.5063	0.5082	0.5101	0.5399	0.7230	0.9088
Paper and Allied Products	26	28	0.6029	0.6596	0.7103	0.7460	0.4685	0.2833	0.3058
Printing, Publishing, and Allied Industries	27	3	0.5009	0.5012	0.5019	0.5024	0.5111	0.5837	0.7040
Chemicals and Allied Products	28	420	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0020**	**0.0001**
Petroleum Refining and Related Industries	29	24	0.0844	0.1380	0.1980	0.2425	0.3012	**0.0079**	**0.0358**
Rubber and Miscellaneous Plastics 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 and Computer Equipment	35	140	0.6024	0.7659	0.4192	0.4052	0.9910	0.9898	0.9867
Electronic and Other Electrical Equipment	36	242	0.1958	0.5789	0.5825	0.7329	0.7058	0.8030	0.7423
Transportation Equipment	37	40	0.2277	0.3575	0.4867	0.5711	0.9594	0.9989	0.9744
Measuring, Analyzing, and Controlling Instruments	38	243	**0.0334**	0.1509	0.3838	0.3983	0.8171	0.8937	0.8778
Miscellaneous Manufacturing Industries	39	18	**0.0468**	0.0789	0.1126	0.1380	**0.0378**	0.1672	0.1093
Business Services	73	137	**0.0004**	**0.0052**	**0.0166**	**0.0055**	**0.0004**	**0.0001**	**0.0022**

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

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

Table 2b: Concentration of Labs by Industry in California (*P-values*)†									
			Miles						
INDUSTRY	**SIC**	**LABS**	**0.25**	**0.5**	**0.75**	**1**	**5**	**20**	**50**
Oil and Gas Extraction	13	2	0.5015	0.5025	0.5040	0.5060	0.5455	0.6275	0.7010
Heavy Construction	16	2	0.5010	0.5015	0.5035	0.5055	0.5330	0.6210	0.1910
Food 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 and Related Industries	29	2	0.5005	0.5025	0.5040	0.5065	0.5385	0.6105	0.6875
Rubber and Miscellaneous Plastics 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 and Computer Equipment	35	99	**0.0140**	**0.0100**	**0.0105**	**0.0120**	**0.0020**	**0.0010**	**0.0205**
Electronic and Other Electrical Equipment	36	211	**0.0450**	**0.0030**	**0.0075**	**0.0030**	**0.0010**	**0.0030**	0.1040
Transportation Equipment	37	36	**0.0010**	**0.0030**	**0.0030**	**0.0030**	0.4635	0.2635	0.1570
Measuring, Analyzing, and Controlling Instruments	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 type indicates significantly more concentrated than overall labs at the 5 percent level of significance. Light gray type indicates significantly more dispersed than overall labs at the 5 percent level of significance.

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

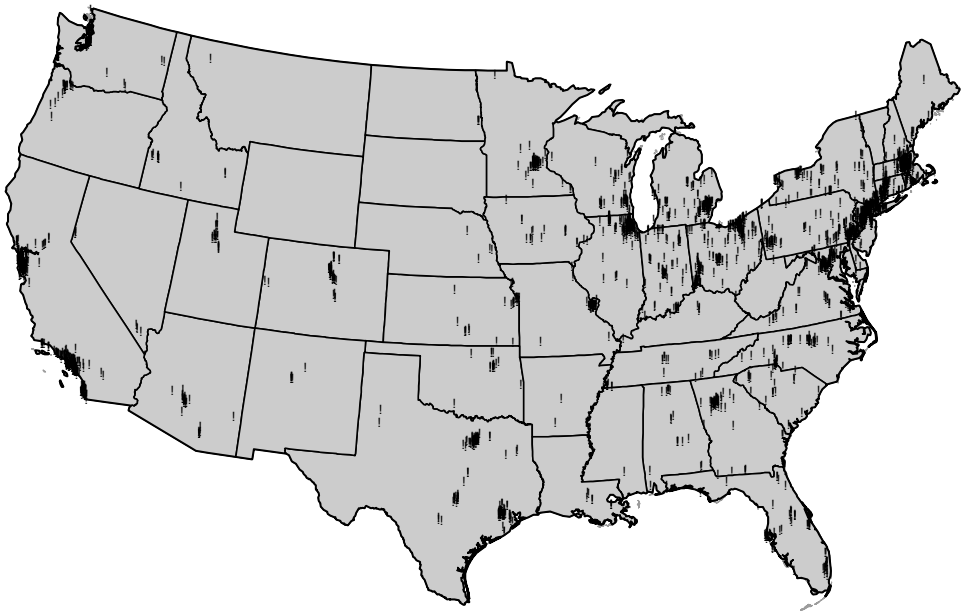


Figure 1: Location of R&D Labs

Source: The 1998 edition of the *Directory of American Research and Technology* and authors' calculations

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

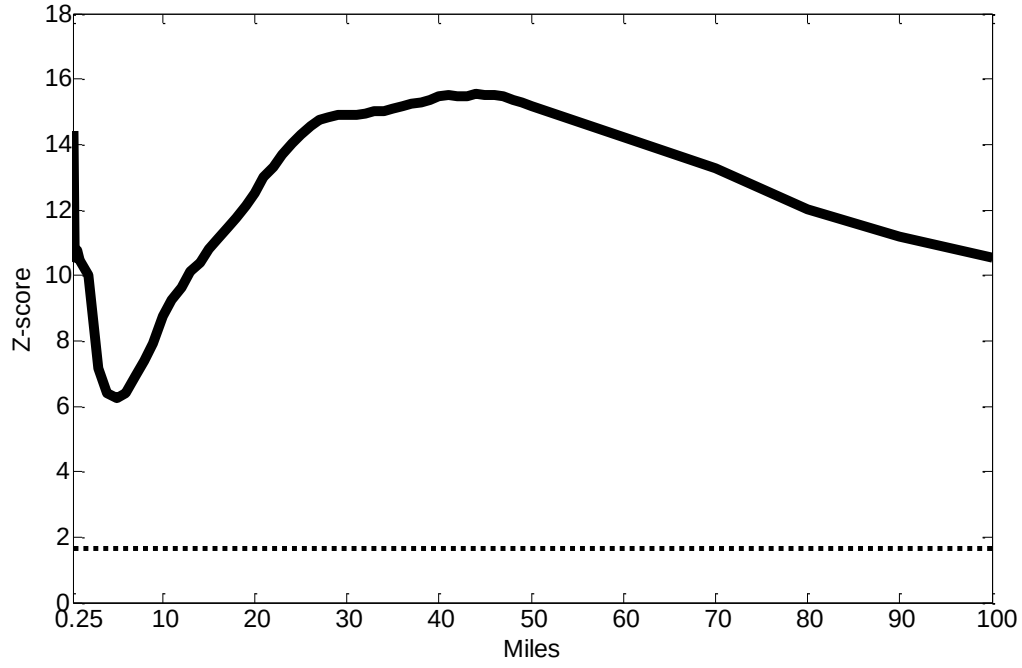


Figure 2a: *Z*-scores for Northeast Corridor

Dotted line *Z* = 1.65

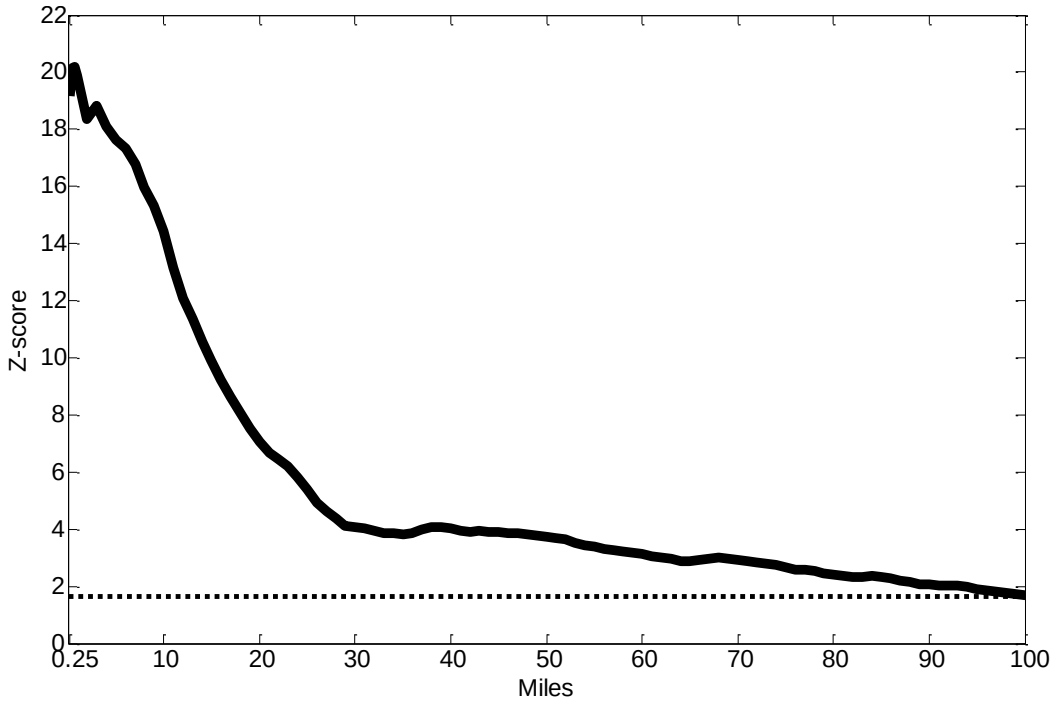
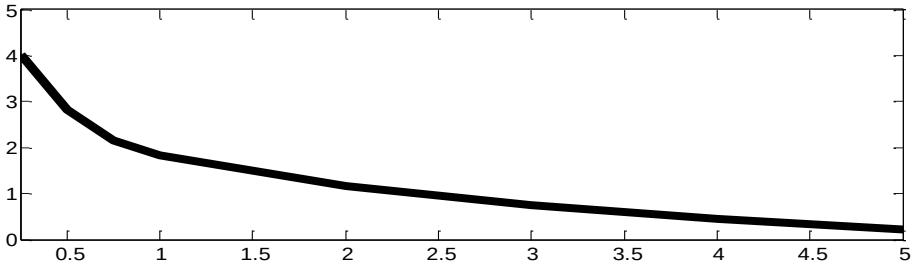


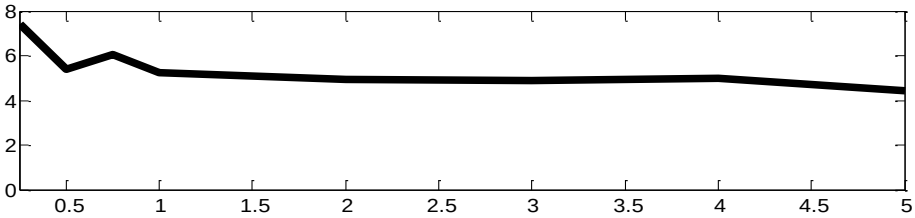
Figure 2b: *Z*-scores for California

Dotted line *Z* = 1.65

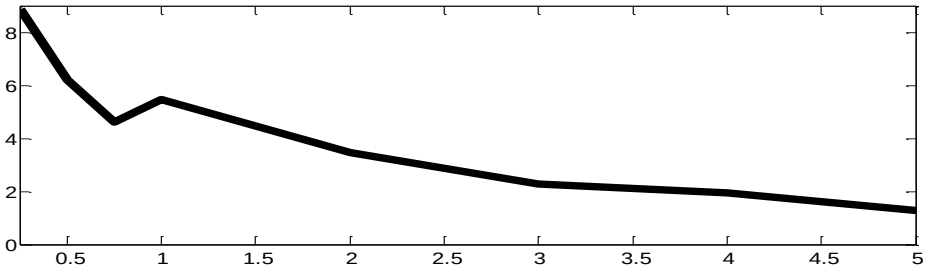
Textile Mill Products



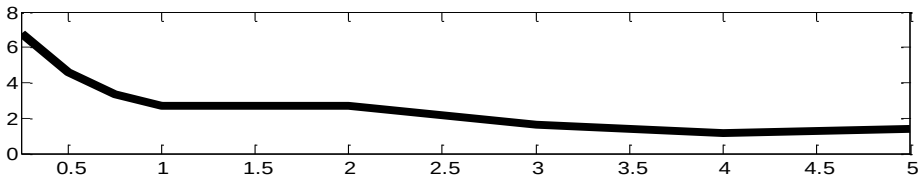
Chemicals and Allied Products



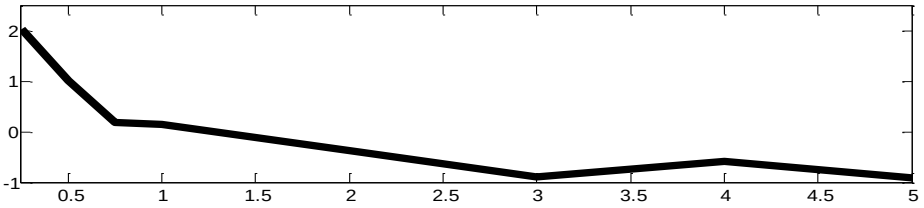
Stone, Clay, Glass, and Concrete Products



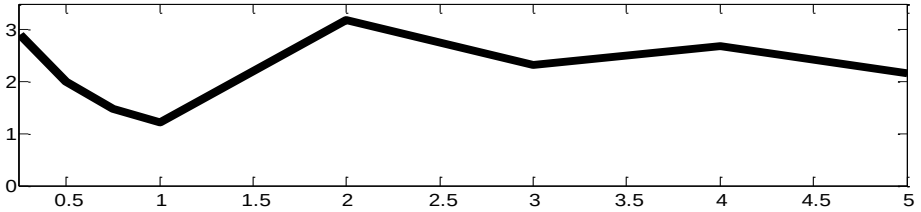
Fabricated Metal Products



Measuring, Analyzing, and Controlling Instruments



Miscellaneous Manufacturing Industries



Business Services

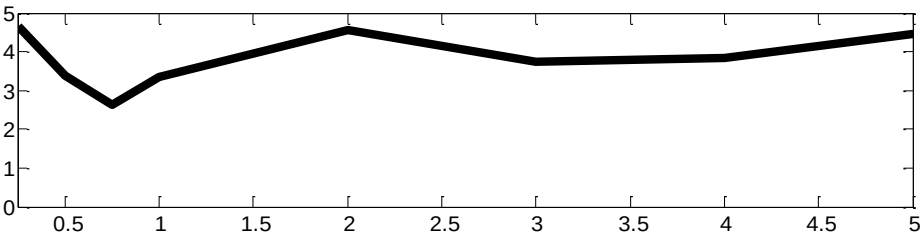
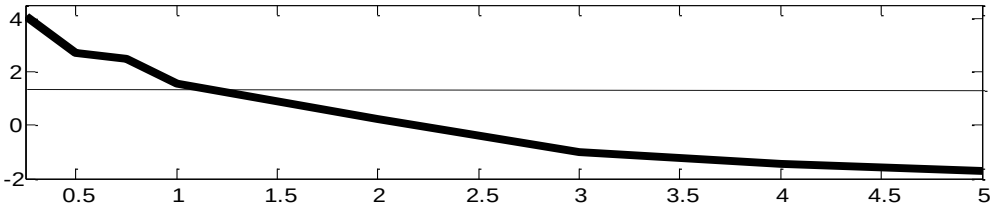
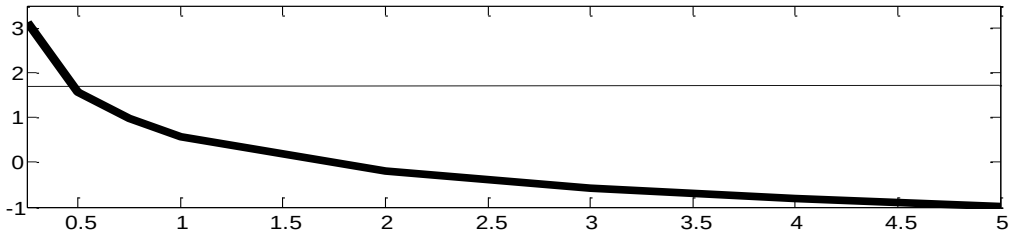


Figure 3a: Northeast Corridor Industry *Z*-Scores

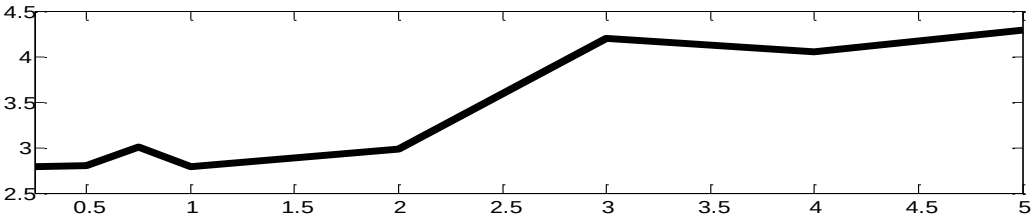
Chemicals and Allied Products



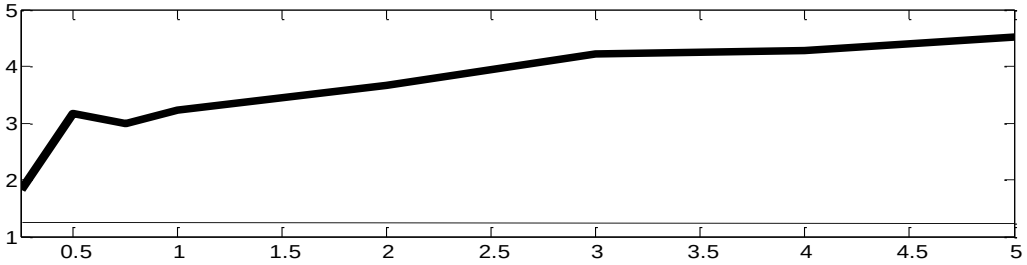
Primary Metal Industries



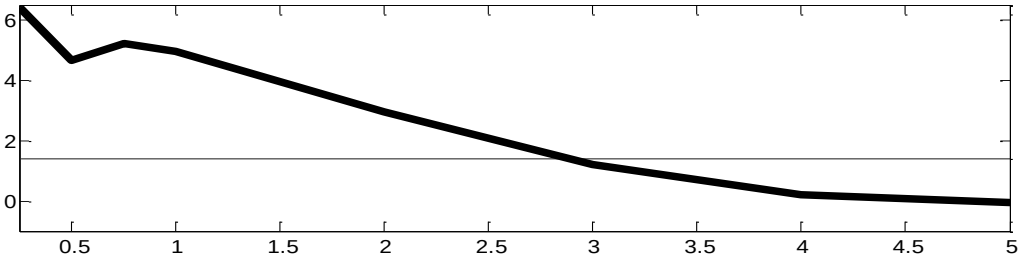
Industrial and Commercial Machinery



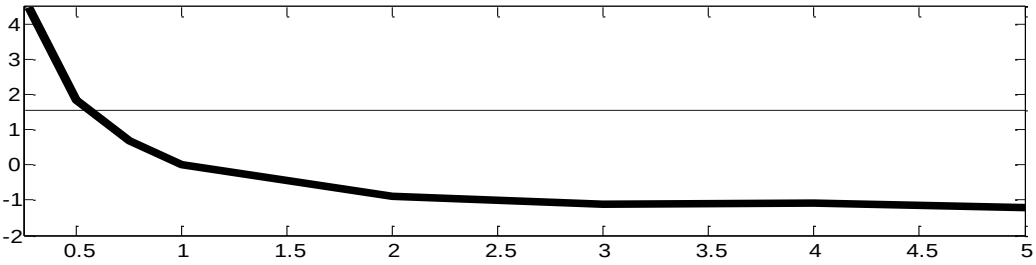
Electronic and Other Electrical Equipment



Transportation Equipment



Measuring, Analyzing, and Controlling Instruments



Business Services

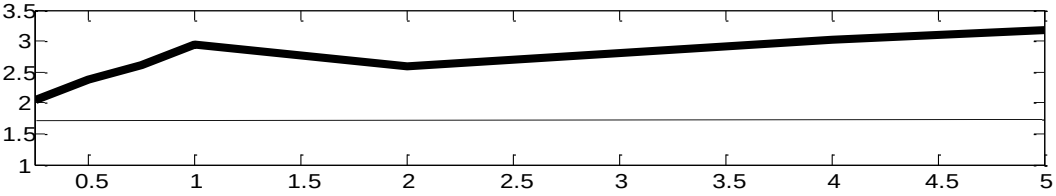


Figure 3b: California Industry Z-Scores

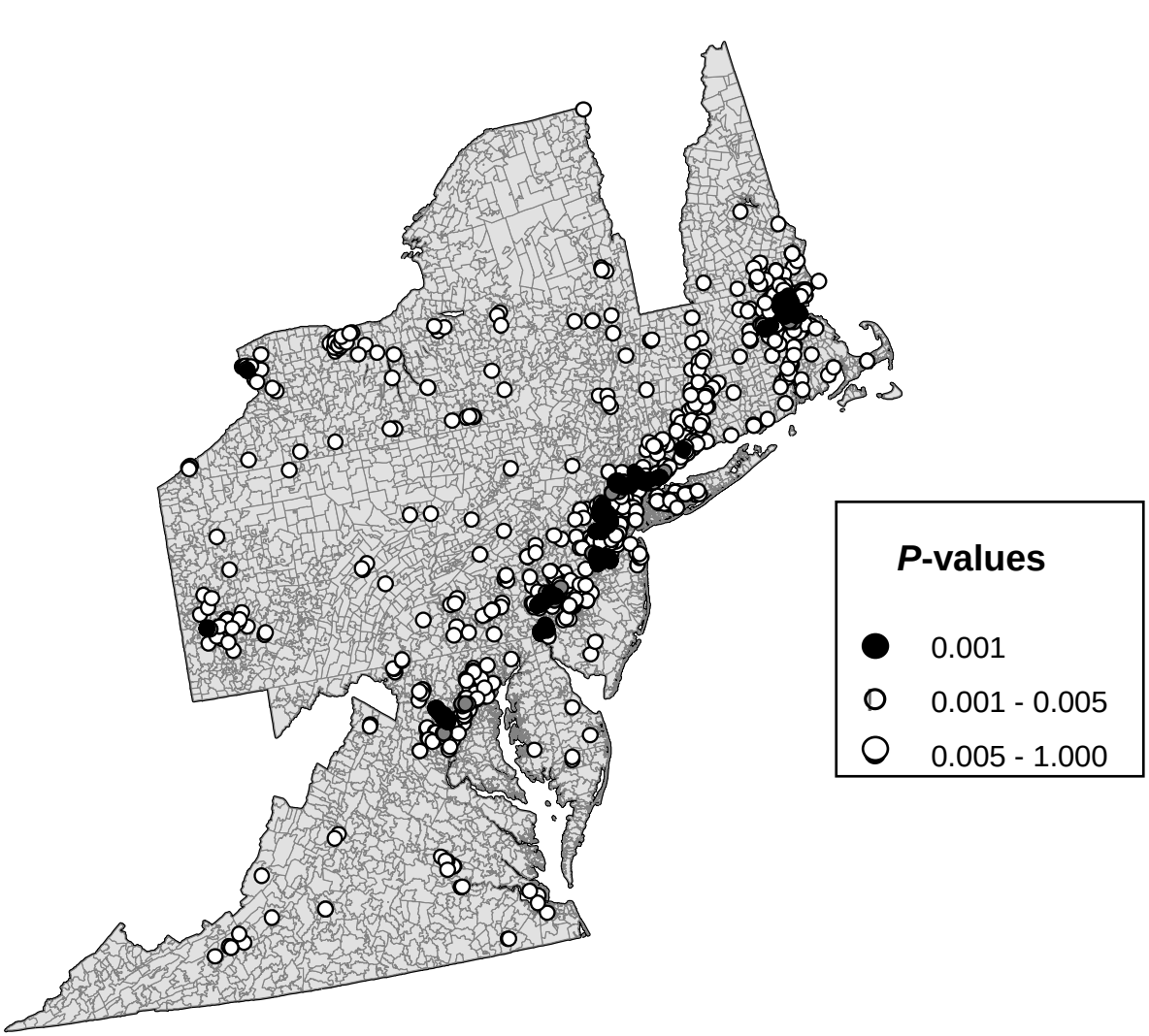


Figure 4a: Northeast Corridor *P*-values at *d* = 5 miles

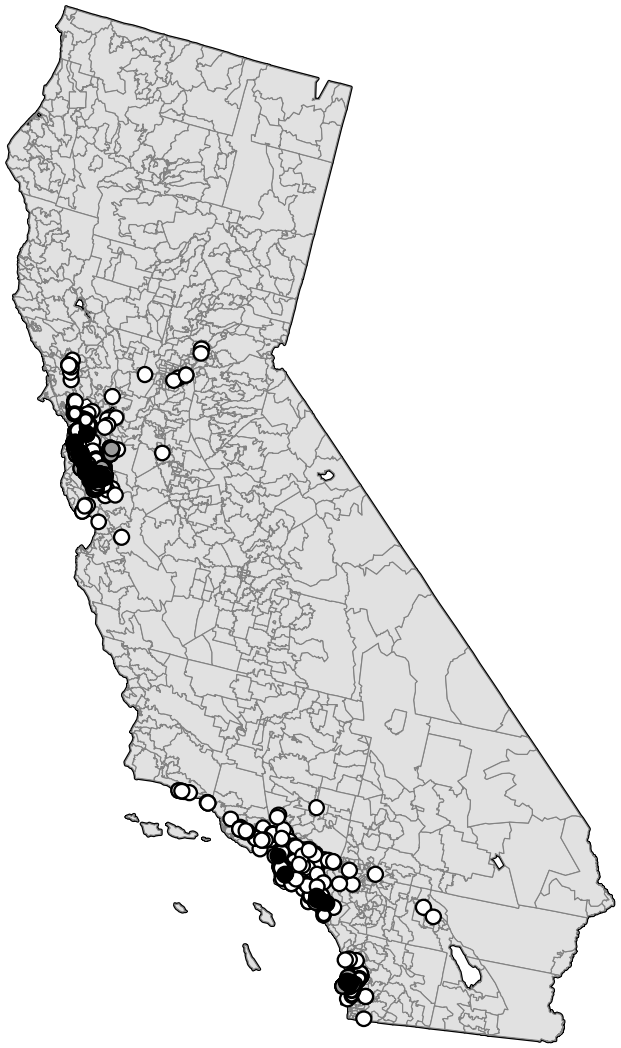


Figure 4b: California *P*-values at *d* = 5 miles

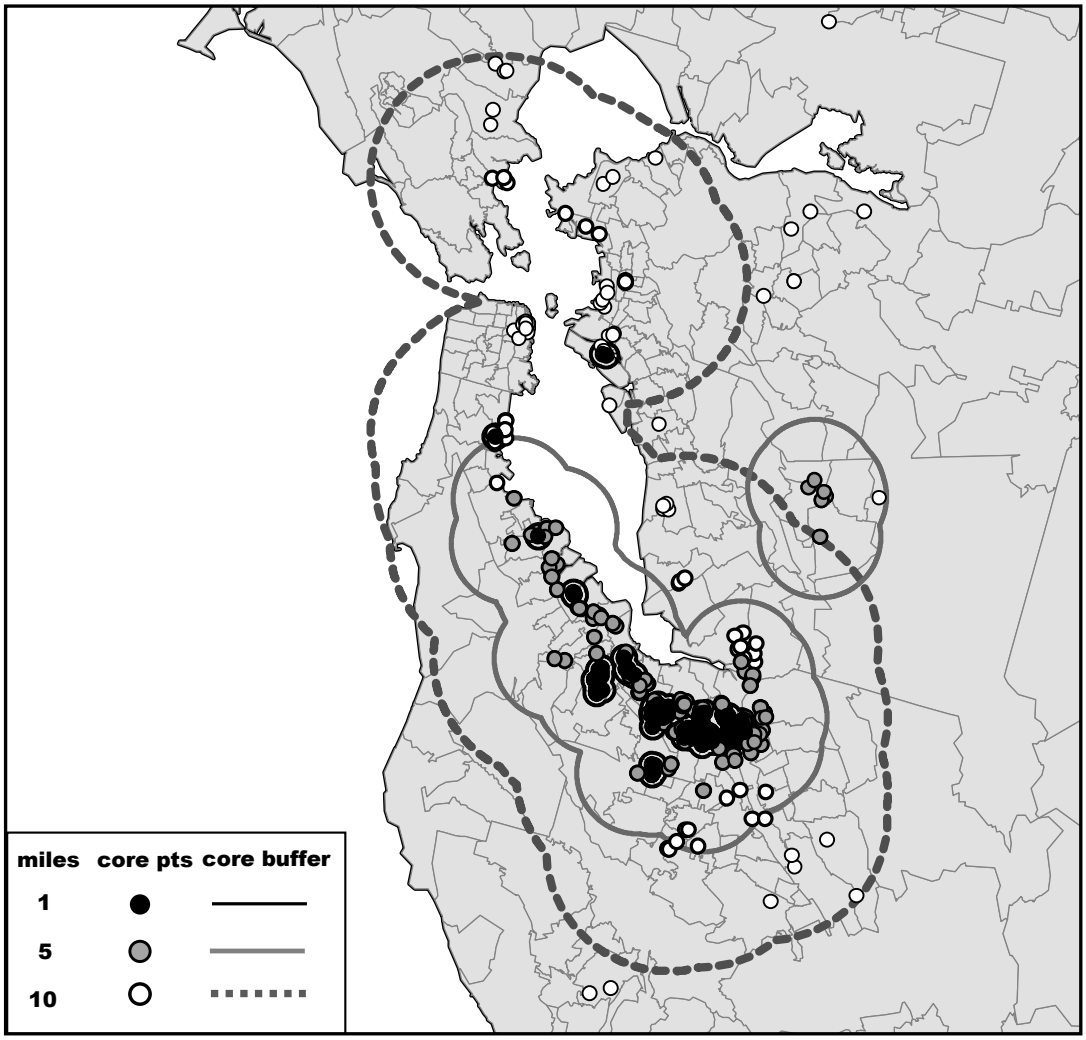


Figure 5: Multiscale Core Clusters in the San Francisco Bay Area

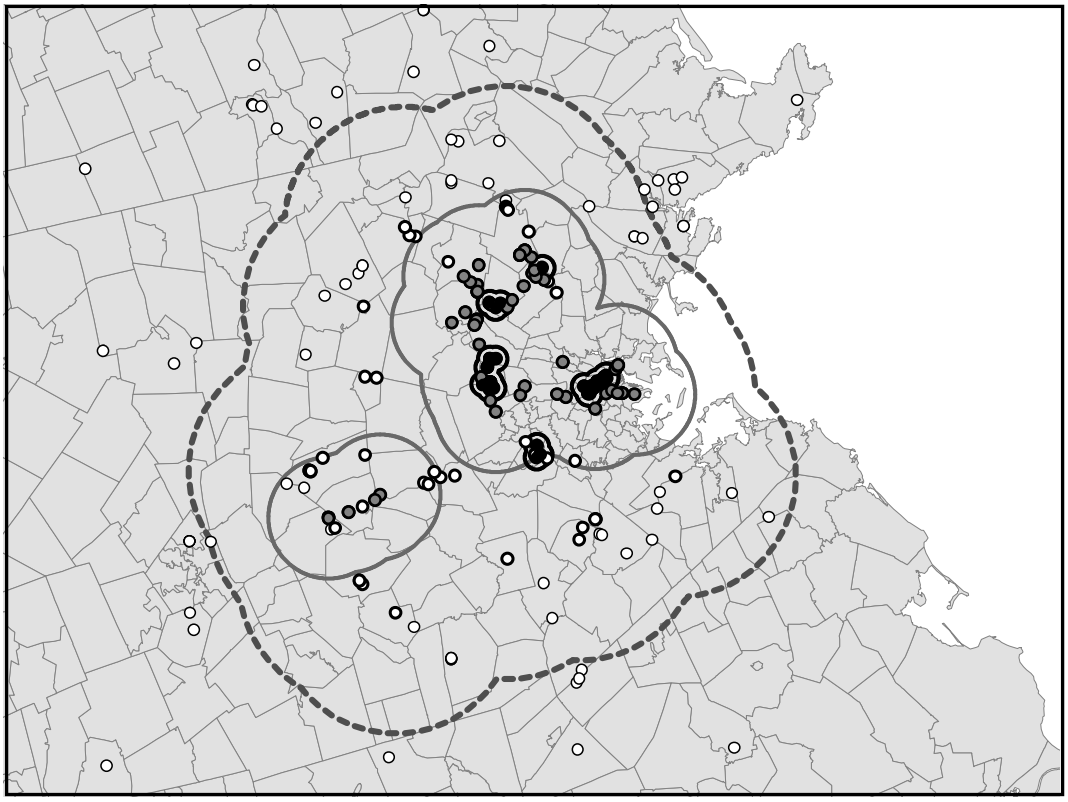


Figure 6a: Multiscale Core Clusters in Boston

Figure 6b: Proximity to Major Routes in Boston

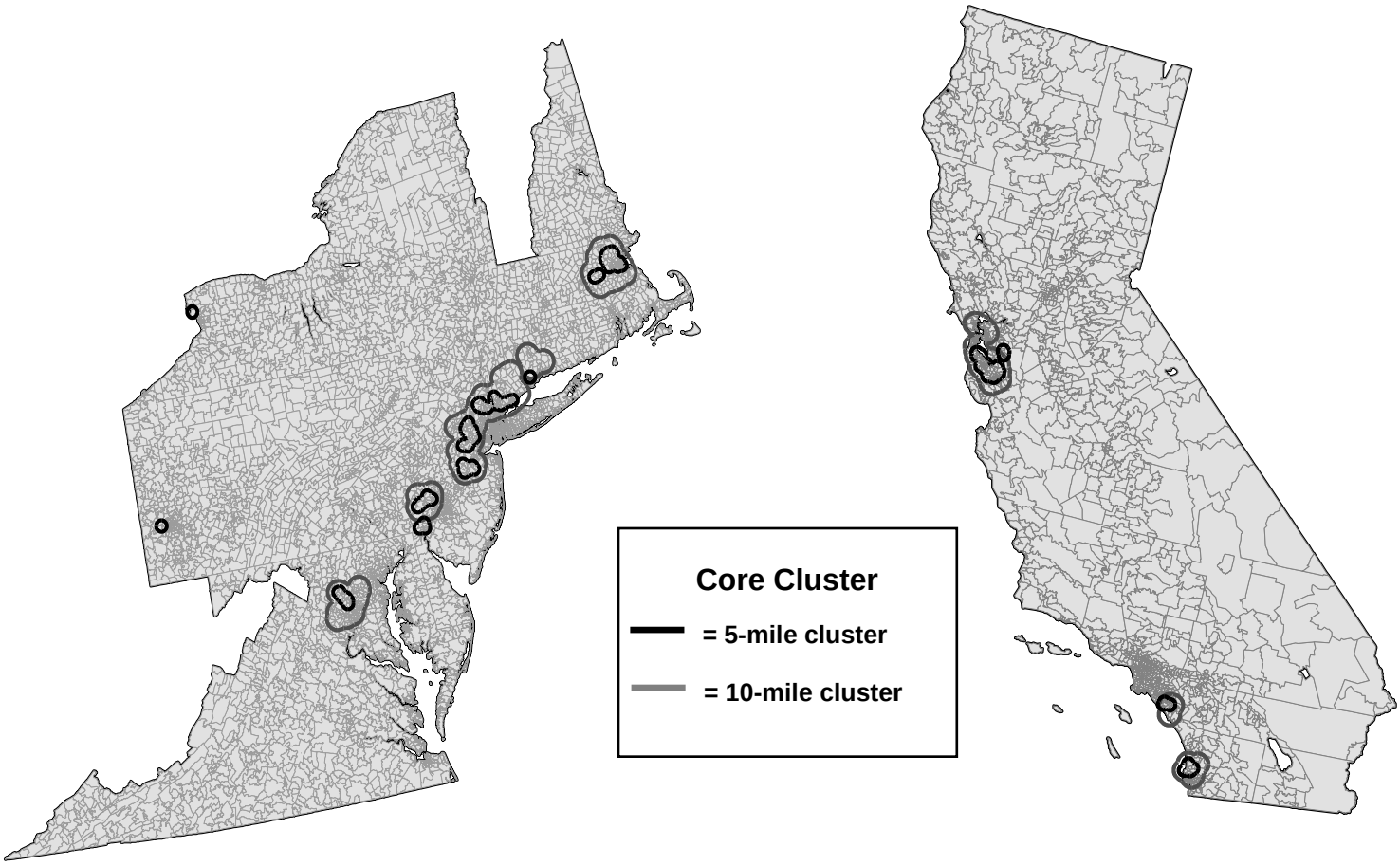


Figure 7a: Northeast Corridor Core Clusters
***d* = 5, 10**

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

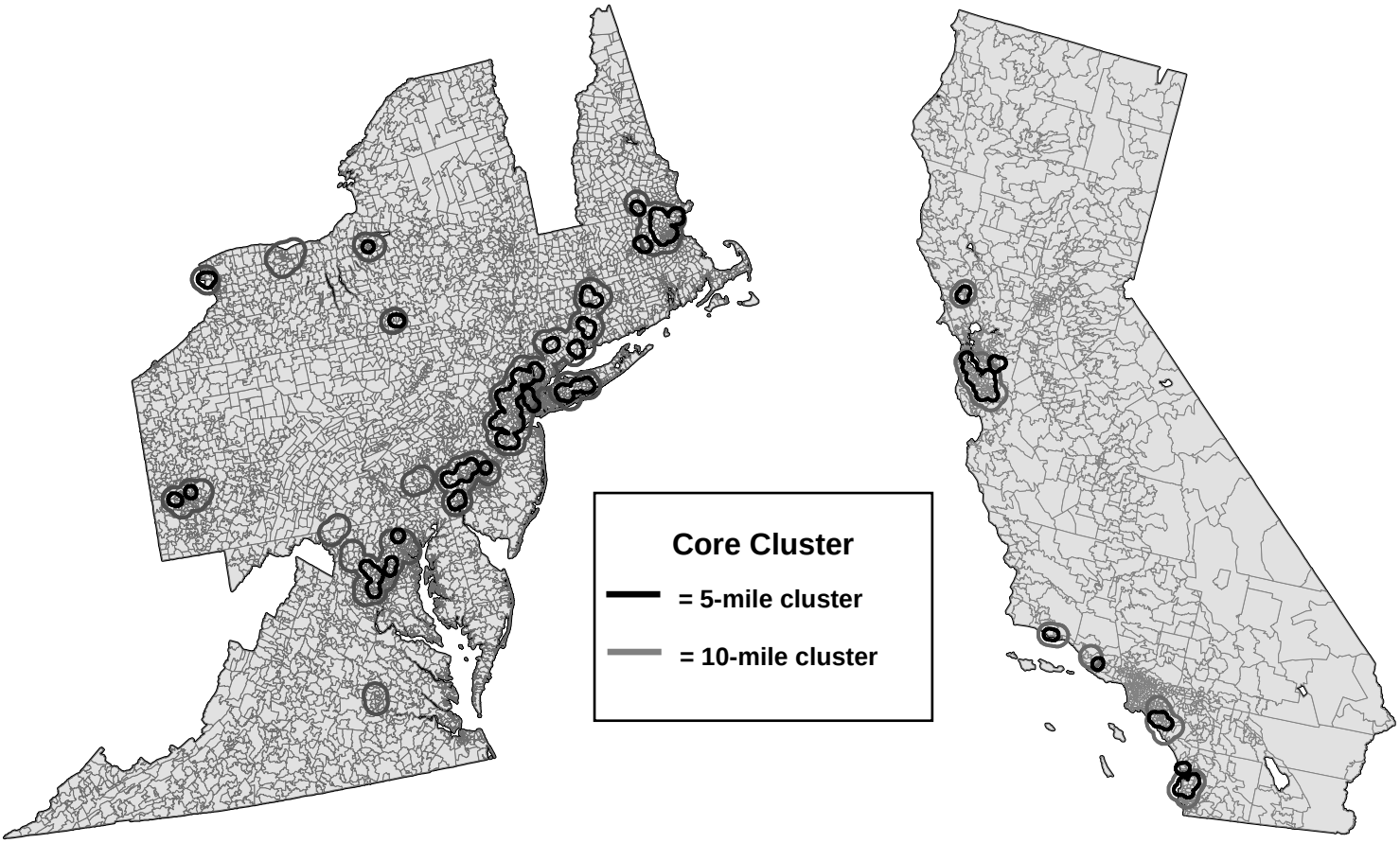


Figure 8a: Northeast Corridor Core Clusters
***d* = 5, 10 (STEM Workers)**

Figure 8b: California Core Clusters
***d* = 5, 10 (STEM Workers)**

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 in 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 corridor manufacturing industries combined. Figure A1 shows the comparison of typical counterfactual lab patterns in the lower half of Manhattan generated by the manufacturing establishment distribution on the left with the manufacturing employment distribution on the right (where zip codes 10018 and 10001 are the darkest pair in the left panel). So, while manufacturing employment appears to be quite concentrated in this area, it is clear that manufacturing establishments are relatively far more concentrated. Because this area constitutes such an extreme outlier in our data, we have run the simulation analyses both with and without the lower half of Manhattan (where the latter excludes the 20 R&D labs in the lower half of Manhattan as well). T he 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 2a 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 the lower half of Manhattan in Figure A1, the patterns of clustering significance for both manufacturing establishments and employment are now qualitatively similar and that, 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 labs 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. Here it is evident that, except for small scales up to about three miles, the shape and levels of significance for both manufacturing establishments and manufacturing employment in Figure 2b 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 stablishments, 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 Core-Cluster Results

As discussed in Section 4 of the paper, our method of identifying core clusters is, by construction, based on the results of local *K*-function analyses. Because such analyses involve separate tests at multiple locations (some nearby) and at multiple scales (some quite large), we must address certain aspects of the well-known "multiple testing" problem.[31] In this Appendix, we first discuss the multiple-testing problem itself and then compare our multiscale core-cluster approach with significance-maximizing approaches to resolving this problem.

To motivate the multiple-testing problem in the setting of Section 5 in the text, we start by supposing that there is no discernible local clustering of R&D labs (i.e., that the observed pattern X^0 of R&D locations cannot be distinguished statistically from the patterns generated under our null hypothesis). In addition, suppose that all local *K*-function tests were in fact statistically independent of one another. Then, by construction, we should expect 5 percent of our resulting test statistics to be statistically significant at the 0.05 percent level. So, when many such tests are involved (there are 1,035 tests at each scale, $d \in D$, in the Northeast corridor and 645 tests at each scale in California), one is bound to find some degree of significant clustering using such testing procedures. As is well known, this type of false positive rate can be mitigated by reducing the *p*-value threshold level deemed to be significant. In fact, that is one reason why we focused only on *p*-values no greater than 0.005 in Figure 4 of the text.

But such adjustments are by themselves not sufficient in instances in which the assumption of statistical independence is violated. This is quite likely when radial neighborhoods around different test points are large enough to intersect and thus contain common points (either

[31] 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.

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.[32]

Several statistical approaches have been developed for resolving such problems. Most prominent among these are the Kulldorff (1997) SaTScan approach and that of Besag and Newell (1991) approach. Both methods employ sequential testing procedures, in which only single maximally significant clusters are identified in each step. To describe this sequential procedure in the present setting, we now focus on zip code areas (cells) and replace individual locations with counts of R&D labs in each area (cell counts). Using centroid distance between cells, candidate clusters are then defined as unions of *m*-nearest neighbors to given "seed" cells, and a test statistic is constructed to determine the single most significant cluster. In both of these significance-maximizing procedures, the notion of "significance" is defined with respect to tests that are based essentially on Hypothesis 1, namely that R&D labs are distributed (at the zip code level) in a manner proportional to manufacturing employment. One key difference is that counterfactual locations are implicitly assumed to be randomly distributed inside each zip code (i.e., are distributed proportional to area rather than total employment at the block level). To determine a second most significant cluster, the zip code areas in the most significant cluster are removed, and the same procedure is then applied to the remaining zip code areas. This procedure is typically repeated until some significance threshold (such as a *p*-value exceeding 0.05) is reached.

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

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

We applied both the Besag and Newell procedure and Kulldorff's SaTScan procedure to our data and found them to be in remarkable agreement with each other. Thus, we present only the results of the (more popular) SaTScan procedure. In this setting, we ran the maximum of 10 iterations allowed by the SaTScan software, and the results from the union of these 10 clusters are plotted in Figure B1 for labs in California and in Figure B2 for labs in the Northeast corridor. By comparing these results with Figures 4a and 4b in the text, it is evident that both procedures are identifying essentially the same areas. These comparisons thus serve as one type of robustness check on our core-cluster results.

However, there are certain differences between these results. Notice first that the SaTScan clusters appear to be more circular in form than the corresponding core clusters. This is particularly evident in the Northeast corridor, where isolated clusters such as Boston, Philadelphia, and Washington, D.C. appear to be very circular. As mentioned previously, this particular SaTScan procedure only considers circular (nearest-neighbor) clusters when identifying a most significant one. While it is possible to extend this restriction to certain classes of elliptical clusters, the key point is that prior restrictions must be placed on the set of potential clusters to keep search times within reasonable bounds. By way of contrast, our present core-

cluster approach involves no prior restrictions on cluster shapes and, in this sense, is more flexible in nature.

A second limitation of these significance-maximizing approaches that is less evident by visual inspection is the path-dependent nature of cluster formation. As mentioned previously, the zip code areas defining clusters created at each step of the procedure are removed before considering each new cluster. When clusters are very distinct (such as Boston, Philadelphia, and Washington, 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 Bay Area in Figure B2 and the New York area in Figure B1), the formation of early clusters modifies the neighborhood relations among the remaining zip codes at later stages. So, at a minimum, these modifications require careful conditional interpretations of all clusters beyond the first cluster. Thus, a second advantage of the present core-cluster approach is the simultaneous formation of all clusters, which naturally avoids any type of sequential constraints.

For Online Publication

Appendix C: Description of the Major Areas of Agglomeration[33]

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 instruments; 9 percent conduct research in

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

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 6a.[34] Most of these labs conduct R&D in five three-digit SIC code industries — computer programming and data processing, drugs, lab apparatus and analytical equipment, communications equipment, and electronic equipment. The largest five-mile cluster shown in Figure 8a contains 109 labs, which accounts for 60 percent of all labs in the larger 10-mile cluster. At the one-mile scale, Boston has five clusters, all of which are centered in the largest five-mile cluster. The largest of these one-mile clusters contains 27 labs, half of which conduct research on drugs.

The New York City Agglomeration

The single largest cluster identified within our 10-state study area is the 10-mile cluster above New York City (shown in Figure C1) that stretches from Connecticut to New Jersey. This cluster contains a total of 287 R&D labs. There are 134 labs (47 percent) 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 five-mile clusters were identified. Most of the concentration is seen to occur in the two clusters west of New York City, which, in particular, contain five of the nine one-mile clusters identified. Among these one-mile clusters, the largest is the Central Park cluster shown in Figure A1. About two-thirds of the 17 labs in this cluster are

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

conducting research on drugs, perfumes, and cosmetics, or computer programming and data processing.

The Philadelphia Agglomeration

As seen in Figure C2, there is a large 10-mile cluster mostly to the west of Philadelphia (the city of Philadelphia is shown in darker gray) where there are a total of 44 labs. Of these 44 labs, 16 conduct research on drugs, and another 15 labs conduct research in the areas of computers, electronics, and instruments and related products. This cluster, in turn, contains a five-mile cluster centered in the King of Prussia area directly west of Philadelphia, in which there are 29 labs, with 40 percent doing research on drugs. There is a second five-mile cluster, containing 17 labs, centered in the city of Wilmington, DE, 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, D.C., in darker gray), where one five-mile cluster can also be seen. About one-quarter of the labs in the 10-mile cluster do research in the areas of computer programming and data processing. Furthermore, another 20 percent of the labs conduct research on communications equipment. In turn, this cluster contains two one-mile clusters, the largest of which (to the north) contains 16 labs with one-half conducting research on drugs.

C.2 California

Turning to California, 27 percent of 645 private R&D labs in the state conduct research in electronic equipment except computers, 18 percent do research in computer and data processing

services, another 18 percent carry out research in chemicals, and 16 percent perform R&D in measuring, analyzing, and controlling instruments.

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 5). Research in these labs is concentrated in three SIC industries: electronic equipment except computers; computer and data processing services; and chemicals and allied products. The Bay Area has two five-mile clusters, the most prominent of which is in the Palo Alto–San Jose area, consisting of 282 labs. The 10-mile cluster also contains seven one-mile clusters. The most prominent one-mile cluster is in the 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 computer and data processing services, and 10 do research in measuring instruments. This cluster, in turn, contains a five-mile cluster consisting of 44 labs, and within it is a one-mile cluster consisting of 33 labs.

The Los Angeles Area

The most prominent cluster of labs in the Los Angeles area consists of 51 labs located in the Irvine–Santa Ana–Newport Beach area. Within this five-mile cluster, there are two separate one-mile clusters, one comprising 20 labs and the other consisting of 10 labs. Electronic equipment

except computers is the main area of research for these labs followed by measuring, analyzing, and controlling instruments; and transportation equipment. In addition, there are two separate one-mile clusters to the north of the 10-mile cluster. One of the clusters is in Torrance with nine labs, and the other in Santa Monica has seven labs.

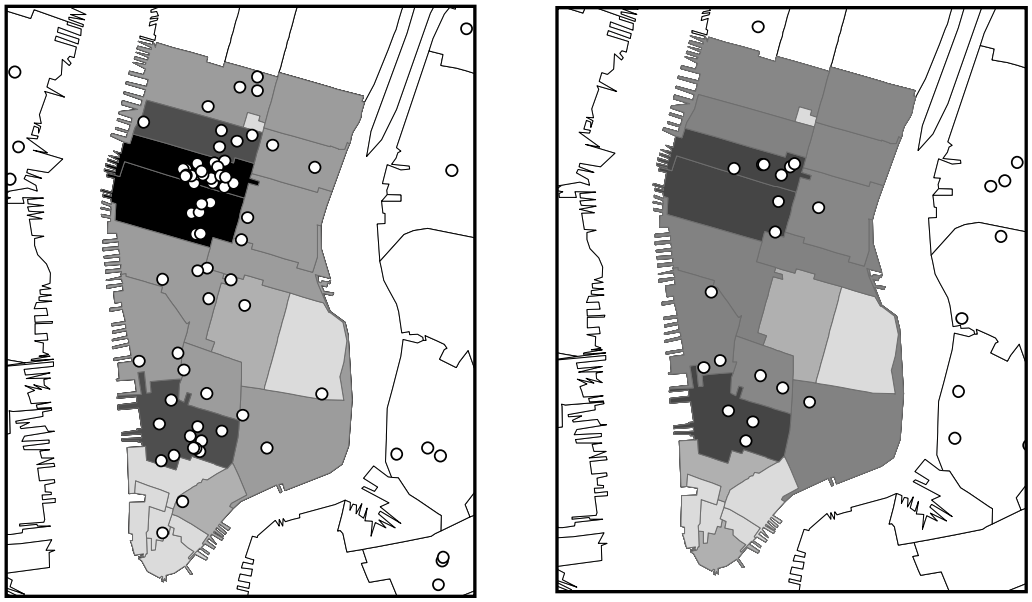


Figure A1: Manufacturing Establishment Counterfactuals (left panel) and Manufacturing Employment Counterfactuals (right panel) in Lower Half of Manhattan

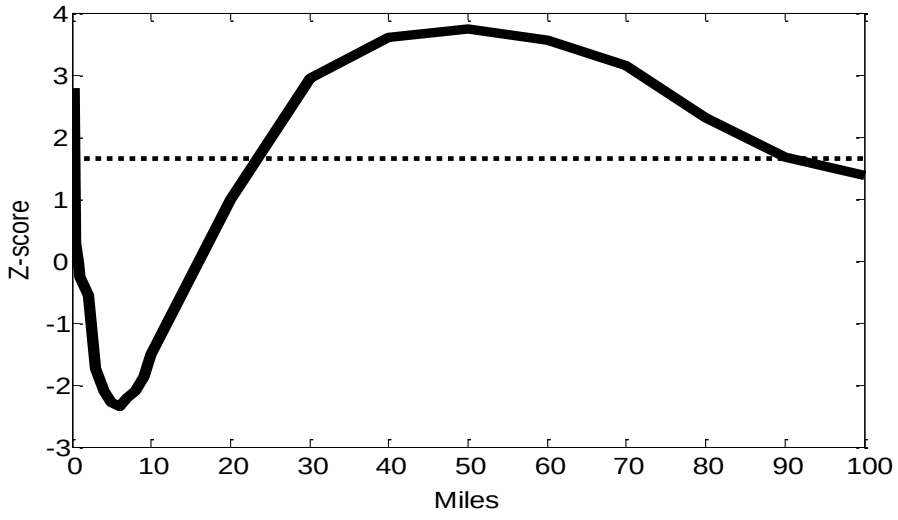


Figure A2: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Including the Lower Half of Manhattan

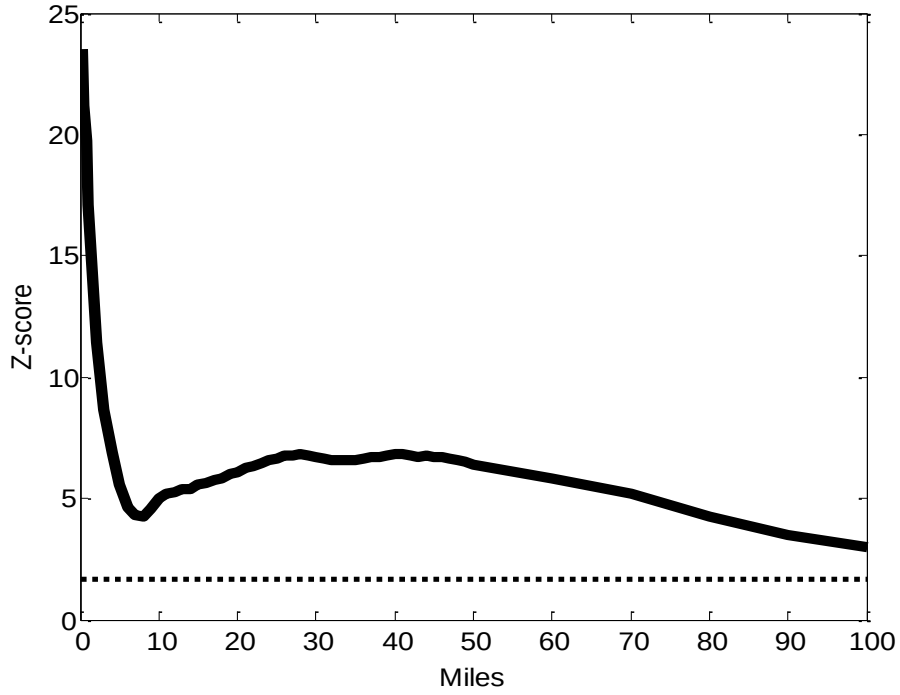


Figure A3: *Z*-scores Relative to Manufacturing Establishments for the Northeast Corridor Excluding the Lower Half of Manhattan

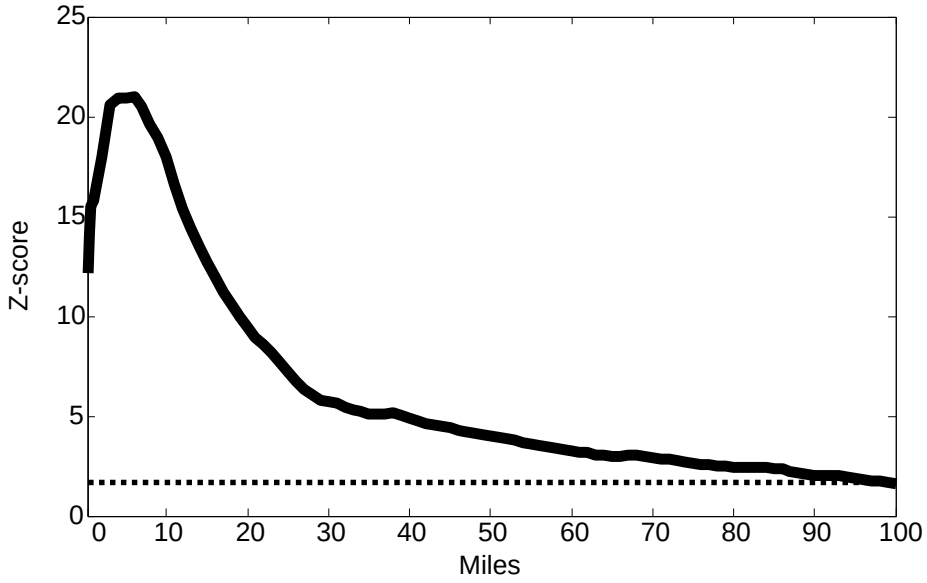


Figure A4: *Z*-scores Relative to Manufacturing Establishments for California

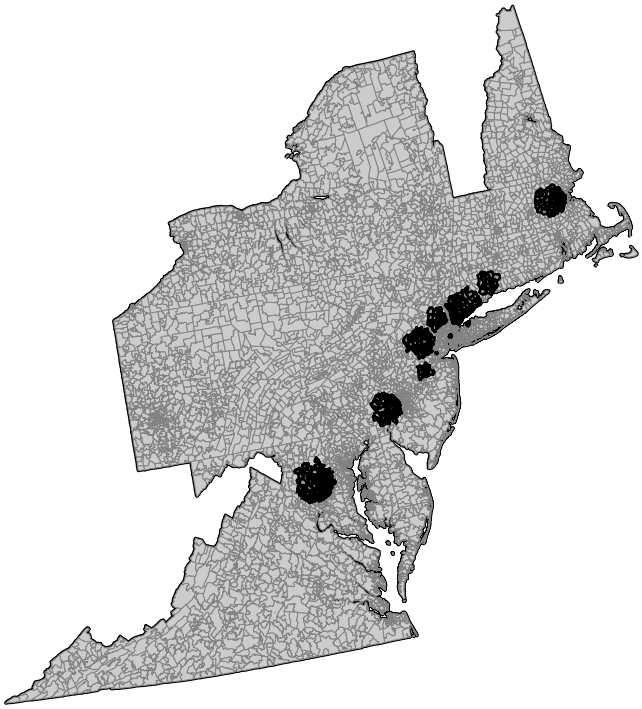


Figure B1: SaTScan Clusters for the Northeast Corridor

Figure B2: SaTScan Clusters for California

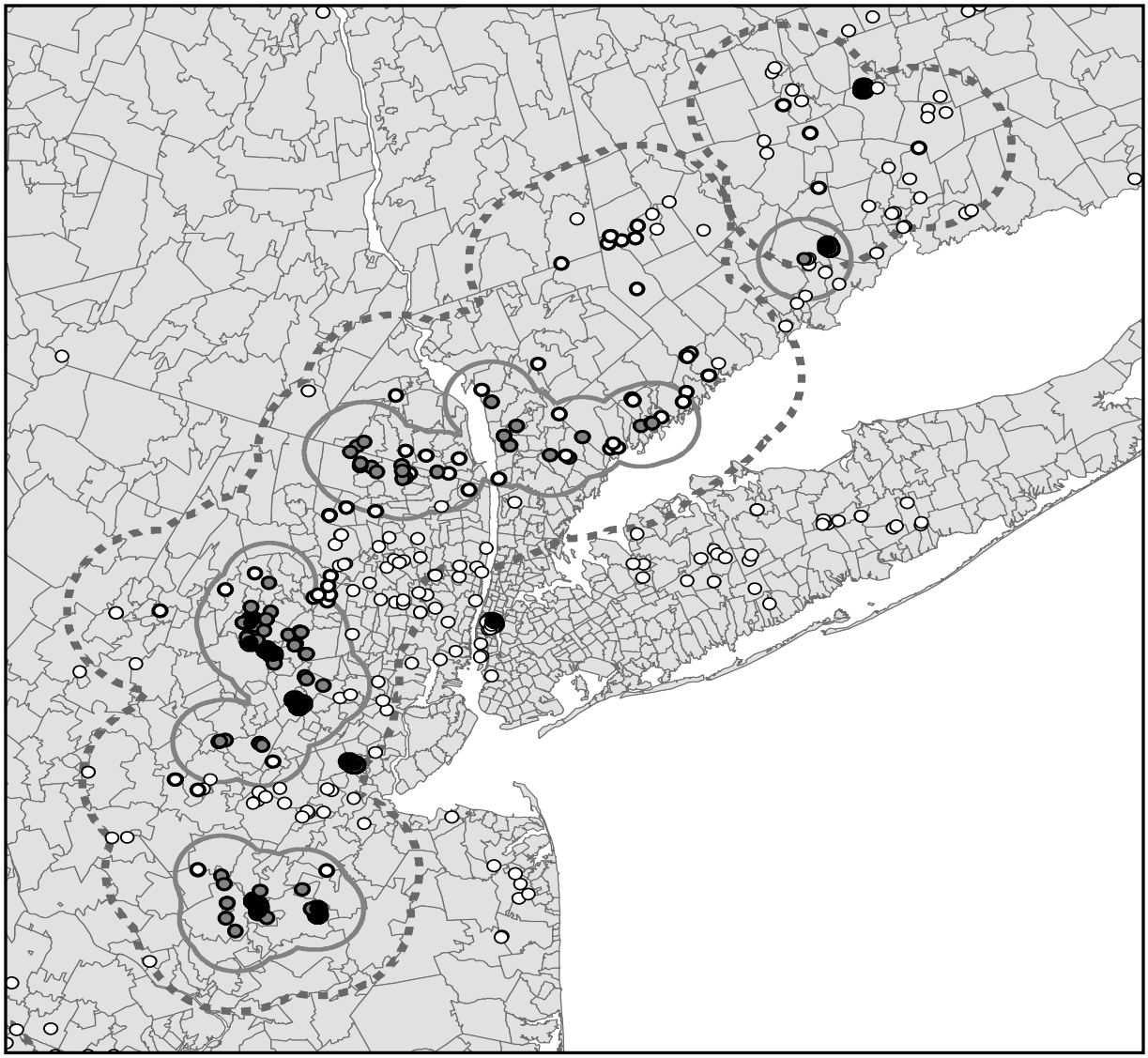


Figure C1: New York City Core Clusters

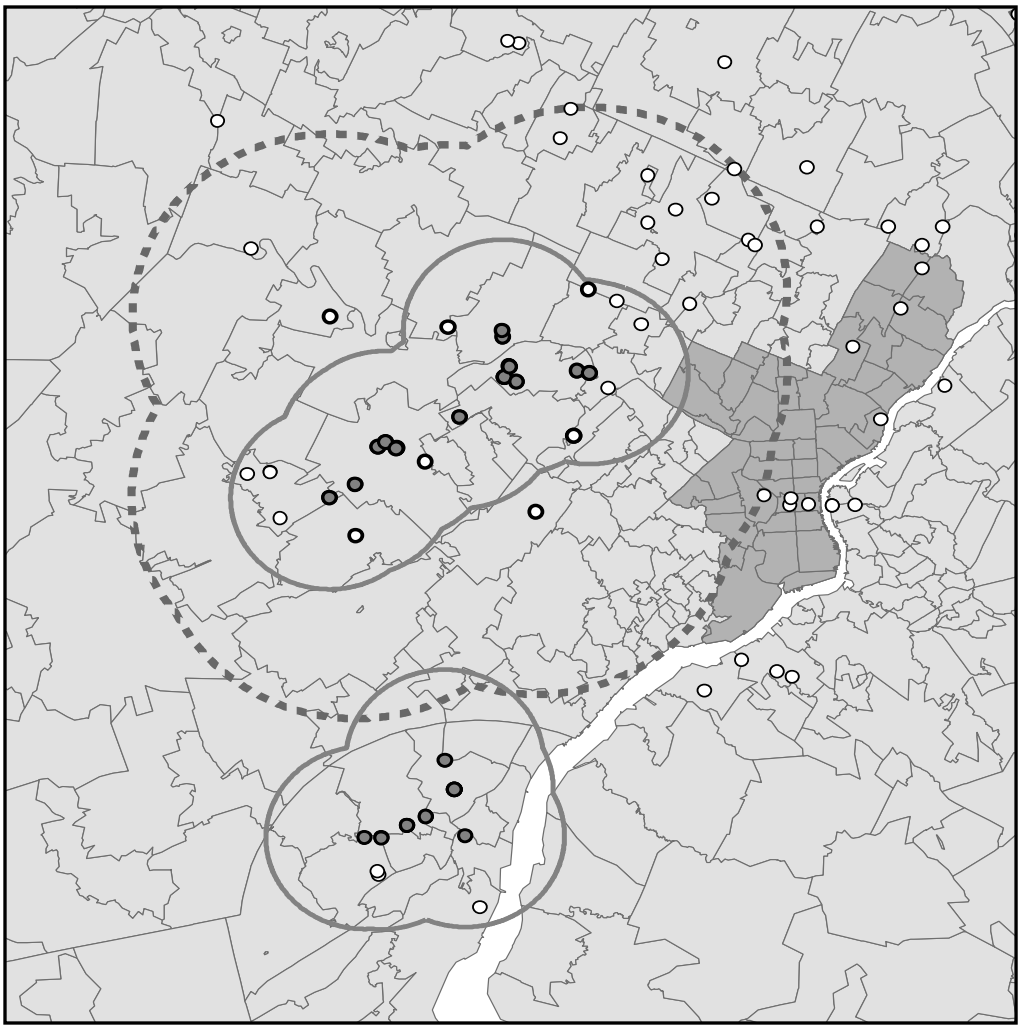


Figure C2: Philadelphia Core Clusters

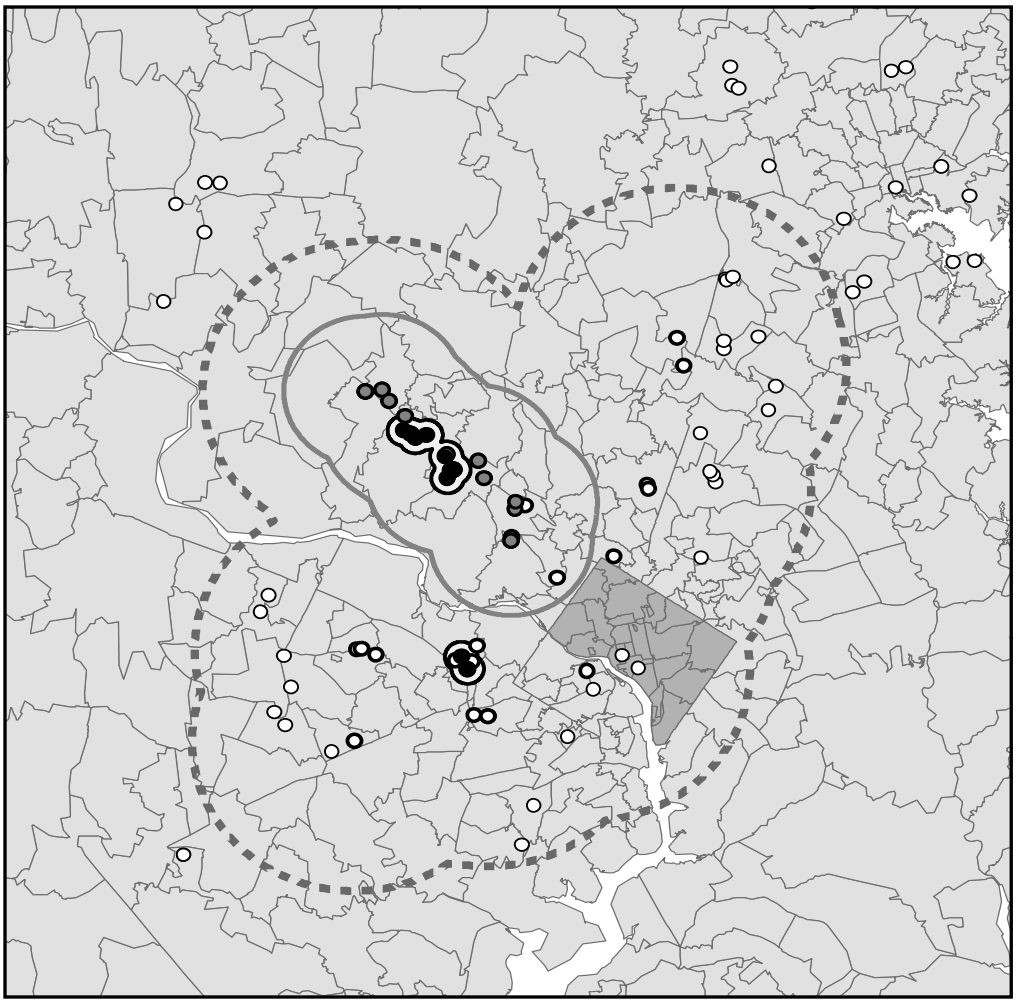


Figure C3: Washington, D.C. Core Clusters