Journal of Urban Economics 101 (2017) 14–26

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

The agglomeration of American R&D labs[☆]

Kristy Buzard[a], Gerald A. Carlino[b,*], Robert M. Hunt[b], Jake K. Carr[c], Tony E. Smith[d]

[a] Maxwell School, Syracuse University, Syracuse, NY 13244, United States
[b] Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106, United States
[c] Geography Department, The Ohio State University, Columbus OH 43210, United States
[d] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 29 June 2015
Revised 12 May 2017
Available online 26 May 2017

JEL classifications:
O31
R12

Keywords:
Spatial clustering
Geographic concentration
R&D labs
Innovation

ABSTRACT

We employ a unique data set to examine the spatial clustering of about 1700 private R&D labs in California and in the U.S. Northeast Corridor. 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, 1 mile, 5 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).

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 (Fig. 1). What is not immediately clear from the map is whether the spatial concentration of R&D is significantly greater than economic activity in general. Are the clustering of R&D labs in 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 whether the spatial pattern of R&D laboratories observed in Fig. 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, MSA, and state levels (see, for example, Ellison and Glaeser, 1997; Rosenthal and Strange, 2001; and Ellison et al., 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, metro areas, and states). While this aggregation of the data facilities computation, this approach leads to a number of aggregation issues. The first is known as the modifiable area unit

[☆] 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/.

[*] Corresponding author.
E-mail address: jerry.carlino@phil.frb.org (G.A. Carlino).

http://dx.doi.org/10.1016/j.jue.2017.05.007
0094-1190/© 2017 Elsevier Inc. All rights reserved.

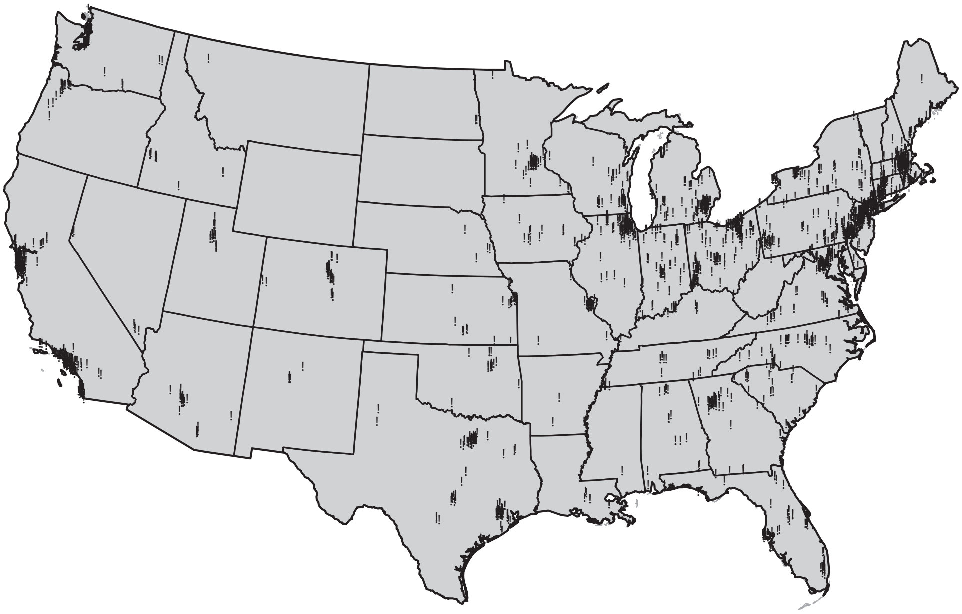

Fig. 1. Location of R&D labs. Source: *Directory of American Research and Technology* (1999) and authors' calculations. Each dash on the map represents the location of a single R&D lab. In areas with a dense cluster of labs, the dashes tend to sit on top of one another, representing a spatial cluster of labs.

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 County and Montgomery County border each other and have activity spilling across them, in a 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 metropolitan areas, we use continuous measures to identify the spatial structure of the concentrations of R&D labs. Specifically, we use point pattern methods to analyze locational patterns over a range of selected spatial scales (within a half mile, 1 mile, 5 miles, etc.). This approach allows us to consider the spatial extent of the agglomeration of R&D labs and to measure any attenuation of clustering with distance more accurately.[1]

Following 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 et al., 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 multiple-scale approach is similar in spirit to that of DO, our test statistics are based on Ripley's *K*-function rather than the "*K*-density" approach of DO. 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 1700 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 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 5 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

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

Table 1
Summary statistics.

Northeast (10-state)					
Variable	Mean	Std. dev.	Median	Minimum	Maximum
	All zip codes (6044)				
Land area, square miles	29.10	37.61	16.87	0.01	468.16
Radius*	2.55	1.66	2.32	0.06	12.21
Total Employment	4307.22	8994.78	1001.00	0.00	194114.00
Manufacturing employment	557.20	1213.46	76.30	0.00	22808.31
Total establishments	250.36	370.76	97.00	1.00	6962.00
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, square miles	20.95	29.46	12.04	0.06	361.79
Radius*	2.21	1.34	1.96	0.14	10.73
Total employment	15736.22	17620.83	11072.00	39.00	194114.00
Manufacturing employment	2057.08	2166.38	1356.30	0.00	22,808.31
Total establishments	697.51	574.58	568.50	6.00	6962.00
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 (1646)				
Land area, square miles	95.56	256.33	17.34	0.01	3806.05
Radius*	3.84	3.96	2.35	0.06	34.81
Total employment	5989.95	9758.35	1700.00	0.00	79766.00
Manufacturing employment	858.14	2394.39	64.50	0.00	27186.00
Total establishments	467.19	555.17	262.50	0.00	3527.00
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, square miles	18.78	37.75	8.19	0.07	385.98
Radius*	2.02	1.38	1.61	0.15	11.08
Total employment	19482.47	17300.91	15088.00	0.00	79766.00
Manufacturing employment	3607.79	5188.27	1569.00	0.00	27186.00
Total establishments	1173.13	677.45	1065.50	0.00	3527.00
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* (1999) and Zip Code Business Patterns.

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

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 is almost 100 square miles in California. 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 SIC industries relative to the concentration of labs across all industries. This both sets a higher bar in our tests of spatial concentration and 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 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 vintage of the *Directory of American Research and Technology* (1999), 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 1000 publicly traded firms ranked in terms of research and development expenditure in Compustat.[2] These firms represent slightly less than 95 percent of all R&D expenditures reported in the 1999 vintage 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 Standard Industrial Classification (SIC) codes. Using the address information for each private R&D establishment, we geocoded the locations of more than 3000 labs (shown in Fig. 1).

In this paper, we analyze two major regions of the U.S.: the Northeast Corridor and the state of California. There are 1035 R&D labs in 10 states comprising the Northeast Corridor of the United States (Connecticut, Delaware, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Virginia, including the District of Columbia – the Washington, D.C., cluster). There are 645 R&D labs in California.

Even at the most aggregate level, it is easy to establish that R&D activity is relatively concentrated in these two regions. For example, in 1998, one-third of private R&D labs and 29 percent of private R&D expenditures were located within the Northeast Corridor, compared with 22 percent of total employment (21 percent of manufacturing employment) and 23 percent of the population. California accounted for almost 22 percent of all private R&D labs and 22 percent of private R&D expenditures in 1998 compared with 12 percent of total employment (11 percent of manufacturing employment) and 12 percent of the population. Together, these two regions accounted for the majority of all U.S. private R&D labs (and private R&D expenditures) in 1998.[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 below, we assess the concentration of R&D establishments relative to a baseline of economic activity as reflected by the amount of manufacturing employment in the zip code. These data were obtained from the 1998 vintage of Zip Code Business Patterns. Given that the vast majority of our R&D labs are owned by manufacturing firms, manufacturing employment represents a good benchmark.[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 had about 29 square miles of land area with a radius of about 2.5 miles in 1998. Since there were approximately 6044 zip codes in the Northeast Corridor in 1998, there is, on average, one R&D facility for every six zip codes in this part of the country. The average zip code in the Northeast Corridor had about 4300 jobs in 1998, 13 percent of which were in manufacturing. In California, the average zip code consisted of about 96 square miles of land area with an average radius of slightly less than 4 miles. The average zip code in California had almost 6000 jobs in 1998, 14 percent of which were in manufacturing. Table 1 also provides descriptive statistics for those zip codes containing one or more R&D labs. These zip codes are physically smaller (with a radius of about 2 miles in each region) and contain three to four times more employment.

2.2. 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 formal mechanisms primarily explored in the literature – sharing, matching, and knowledge spillovers – 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 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 et al., 2006; Hunt, 2007). Also, specialized workers can readily find new positions without having to change locations (job hopping).

2.2.3. Connection between theory and evidence

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

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

[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.

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

[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.

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

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 our 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 2 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 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 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 below, blocks themselves are sufficiently small to allow such random effects to be safely ignored at the scales of most relevance for our purposes.

By repeating this procedure separately for the Northeast Corridor (with a set of $n = 1035$ location choices) and for California (with $n = 645$ location choices), one generates a pattern, $X = (x_i = (r_i, s_i) : i = 1, \dots, n)$, 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*, $X^0 = (x^0{}_i = (r^0{}_i, s^0{}_i) : i = 1, \dots, n)$, 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 $X = (x_i : i = 1, \dots, n)$, 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]

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

As described in Section 3, we draw a set of N point patterns, $X^s = (x^s{}_i : i = 1, \dots, n), s = 1, \dots, N$, and for a selection of radial distances, $D = (d_1, \dots, d_k)$, we calculate the resulting sample K-functions, $\{\hat{K}^s(d) : d \in D\}, s = 1, \dots, N$. 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

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

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

Our Monte Carlo test for clustering was carried out with $N = 999$ simulations at radial distances, $d \in D = \{0.25, 0.5, 0.75, 1, 2, \dots, 99, 100\}$, (i.e., at quarter-mile increments up to a mile and at one-mile increments from 1 to 100 miles). Before discussing these results, it should be noted that quarter-mile distances are approximately the smallest scale at which meaningful clustering can be detected within our present spatial framework. Recall that since locations consistent with the null hypothesis are distributed randomly within each census block, they cannot reflect any locational constraints inside such blocks. For example, if all observed lab locations are street addresses,

[7] We thank Duranton for this observation.

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

[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.

[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.

[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.

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 this data.[15]

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

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

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 Fig. 2(a), and those for California are shown in Fig. 2(b). Significance levels decrease nearly monotonically for California, while in the Northeast, we see a hump-shaped pattern. The high z-scores are consistent with the significance of the Monte Carlo results noted previously but add more detailed information about the patterns of significance.[16] Observe that in both figures, clustering is most significant at smaller scales but exhibits rapid attenuation as scales increase. This pattern is consistent with empirical research on human capital spillovers and agglomeration economies mentioned in the Theory Section 2.1.[17]

3.4. Relative clustering of R&D labs by industry

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

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

Such random samples are easily constructed by randomly permuting (reordering) the lab indices $1, \ldots, n$ and choosing the first n_j of these (as is also done in DO). With respect to clustering, one can then compare $\hat{K}(d)$ values for the observed pattern of labs in industry j with those for a set of N such randomly sampled patterns and derive both p-values, $P_j(d)$ and z-scores, $z_j(d)$ comparable with those in expressions (2) and (3), respectively. If $P_j(d)$ is sufficiently low [or $z_j(d)$ is sufficiently high], then it can be concluded that there is significantly more clustering at scale d for labs in industry j than would be expected under 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 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

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

[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.

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

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

[18] We assign labs to an industry based on information contained in the *Directory of American Research and Technology* (1999). In the Northeast Corridor, there are 19 industrial groupings corresponding to SICs 10, 13, 20–23, 26–30, 32–39, and 73. In California, there are 16 industrial groupings corresponding to SICs 13, 16, 20, 26, 28–30, 32–39, and 73. The industry names of these SICs are included in Tables 2a and 2b.

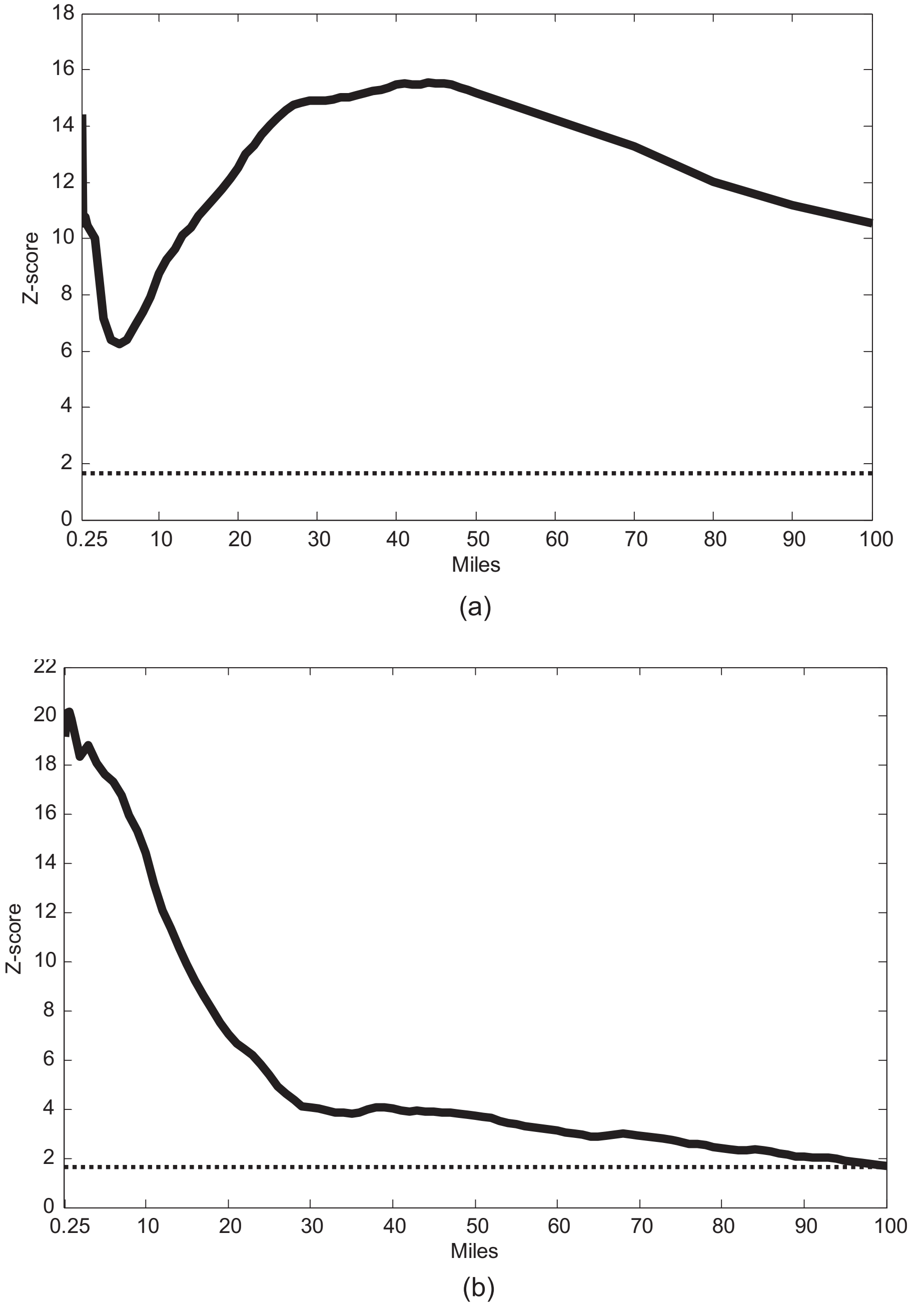

Fig. 2. (a) z-Scores for Northeast Corridor dotted line $Z = 1.65$. (b) z-Scores for California dotted line $Z = 1.65$.

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 Fig. 3, where the z-scores for each of the seven industries in the Northeast with most significant clustering is shown in Fig. 3(a), and those for seven of the eight most significant California industries are shown in Fig. 3(b).[22] Because we are especially interested in the attenuation of z-scores at small scales, these z-scores are calculated in increments of 0.25 miles up to five miles. For all but one of these industries in the Northeast, the clustering of R&D labs is by far most significant at very small spatial scales — a quarter mile or less. The lone exception is Miscellaneous Manufacturing

[19] The seven industries are Textile Mill Products; Stone, Clay, Glass and Concrete; Fabricated Metals; 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 Products; Primary Metal Products; Industrial and Commercial Machinery; Electronics; Transportation Equipment; Measuring, Analyzing, and Controlling Equipment; and Business Services.

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

Table 2a
Concentration of labs by industry in Northeast Corridor (*p-values*)[†].

Industry	Miles							
	SIC	LABS	0.25	0.5	0.75	1	5	20	50
Metal mining	10	4	0.5021	0.5029	0.5044	0.5052	0.5227	0.1674	0.4149
Oil and gas extraction	13	3	0.5011	0.5019	0.5026	0.5034	0.5137	0.0906	0.2286
Food	20	25	0.5825	0.6278	0.6750	0.7081	0.0984	0.2097	**0.0480**
Textile mill	22	14	**0.0267**	**0.0465**	0.0690	0.0859	0.3468	0.7839	0.6446
Apparel	23	5	0.5036	0.5063	0.5082	0.5101	0.5399	0.7230	0.9088
Paper	26	28	0.6029	0.6596	0.7103	0.7460	0.4685	0.2833	0.3058
Printing and publishing	27	3	0.5009	0.5012	0.5019	0.5024	0.5111	0.5837	0.7040
Chemicals	28	420	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0020**	**0.0001**
Petroleum refining	29	24	0.0844	0.1380	0.1980	0.2425	0.3012	**0.0079**	**0.0358**
Rubber products	30	38	0.6728	0.7493	0.8135	0.8544	0.5710	0.7974	0.9965
Stone, clay, glass, and concrete products	32	36	**0.0002**	**0.0008**	**0.0032**	**0.0011**	0.1041	0.7385	0.6886
Primary metal industries	33	36	0.6555	0.7284	0.7921	0.8327	0.7848	0.2592	0.4881
Fabricated metal products	34	44	**0.0004**	**0.0026**	**0.0101**	**0.0200**	0.0911	0.6985	0.8571
Industrial and commercial machinery	35	140	0.6024	0.7659	0.4192	0.4052	0.9910	0.9898	0.9867
Electronics	36	242	0.1958	0.5789	0.5825	0.7329	0.7058	0.8030	0.7423
Transportation equipment	37	40	0.2277	0.3575	0.4867	0.5711	0.9594	0.9989	0.9744
Measuring, analyzing, and controlling instruments	38	243	**0.0334**	0.1509	0.3838	0.3983	0.8171	0.8937	0.8778
Miscellaneous manufacturing Industries	39	18	**0.0468**	0.0789	0.1126	0.1380	**0.0378**	0.1672	0.1093
Business services	73	137	**0.0004**	**0.0052**	**0.0166**	**0.0055**	**0.0004**	**0.0001**	**0.0022**

Source: Author's calculations using the 1998 editions of the Directory of American Research and Technology (1999).
[†] Concentration is conditional on the location of overall R&D labs. Bold indicates significantly more concentrated than overall labs at the 5 percent level of significance. Light gray indicates significantly more dispersed than overall labs at the 5 percent level of significance.

Table 2b
Concentration of labs by Industry in California (*p-values*)[†].

Industry	Miles							
	SIC	LABS	0.25	0.5	0.75	1	5	20	50
Oil and gas extraction	13	2	0.5015	0.5025	0.5040	0.5060	0.5455	0.6275	0.7010
Heavy construction	16	2	0.5010	0.5015	0.5035	0.5055	0.5330	0.6210	0.1910
Food	20	3	0.5055	0.5100	0.5150	0.5185	0.5990	0.7700	0.4925
Paper	26	2	0.5020	0.5035	0.5045	0.5080	0.5340	0.6175	0.1970
Chemicals	28	129	**0.0025**	**0.0100**	**0.0170**	0.0705	0.9670	0.9920	0.9480
Petroleum refining	29	2	0.5005	0.5025	0.5040	0.5065	0.5385	0.6105	0.6875
Rubber products	30	8	**0.0235**	0.0535	0.0980	0.1320	0.4020	0.3660	0.1630
Stone, clay, glass, and concrete products	32	6	0.5125	0.5290	0.5515	0.5695	0.7950	0.7075	0.4215
Primary metal industries	33	11	**0.0435**	0.1130	0.1780	0.2455	0.8770	0.7235	0.2865
Fabricated metal products	34	16	0.5925	0.6840	0.7670	0.8235	0.9890	0.4555	0.1765
Industrial and commercial machinery	35	99	**0.0140**	**0.0100**	**0.0105**	**0.0120**	**0.0020**	**0.0010**	**0.0205**
Electronics	36	211	**0.0450**	**0.0030**	**0.0075**	**0.0030**	**0.0010**	**0.0030**	0.1040
Transportation equipment	37	36	**0.0010**	**0.0030**	**0.0030**	**0.0030**	0.4635	0.2635	0.1570
Measuring, analyzing, and controlling equipment	38	134	**0.0010**	**0.0480**	0.2165	0.4610	0.8845	0.9960	1.0000
Miscellaneous manufacturing industries	39	8	0.5285	0.5620	0.5980	0.6280	0.9000	0.7310	0.7205
Business services	73	147	**0.0300**	**0.0150**	**0.0105**	**0.0045**	**0.0020**	**0.0010**	**0.0010**

Source: Author's calculations using the 1998 editions of the Directory of American Research and Technology (1999).
[†] Concentration is conditional on the location of overall R&D labs. Bold indicates significantly more concentrated than overall labs at the 5 percent level of significance. Light gray indicates significantly more dispersed than overall labs at the 5 percent level of significance.

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, Electronics and Business Services have local peaks at one-half mile and at one mile, respectively.

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

Finally, it is of interest to note that three industries are among the most significantly clustered industries in both the Northeast and California, namely Chemicals, Business Services, and the Manufacturing, Analyzing, and Controlling Equipment industry. The Chemical industry (SIC 28) merits some special attention, if for no other reason than this category includes labs engaged in pharmaceutical R&D, a very important segment of the U.S. economy. In our data, this category of labs accounts for about 40 percent of all labs in the Northeast, a share more than twice as large as any other two-digit SIC industry. In California, the Chemicals industry accounts for about 16 percent of the labs we study. Thus, at least within the geographic area under study, this industry is seen to be a major contributor to the overall clustering pattern of R&D shown in Fig. 2(a) and (b). But it should be equally clear from Fig. 3(a) and (b) 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.

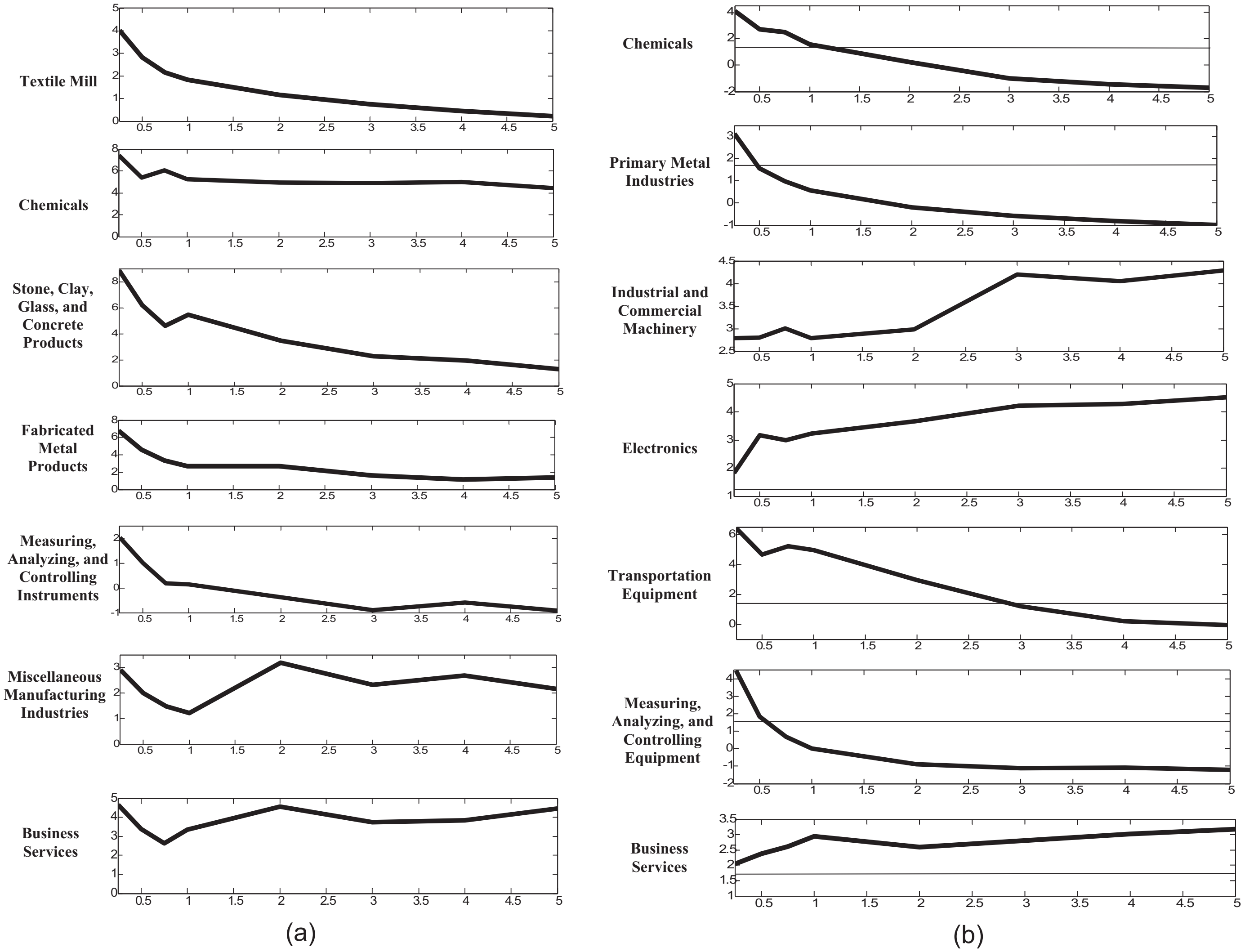

Fig. 3. (a) Northeast Corridor industry *z*-scores. (b) California Industry *z*-scores.

4. Local cluster analysis

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

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

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 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 Eq. (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, $\hat{K}_i^s(d), s = 1, \dots, N$, under H_0 are obtained by generating point patterns, $X^s = (x^s{}_j : j = 1, \dots, n-1), s = 1, \dots, N$, 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,

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

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

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

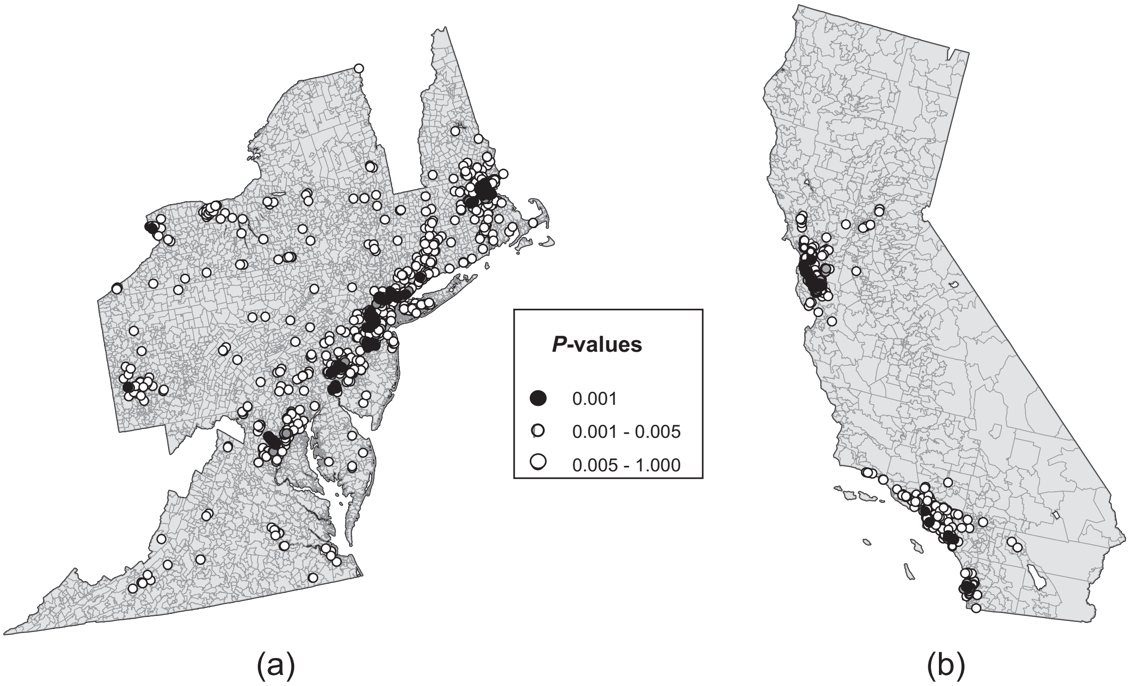

Fig. 4. (a) Northeast Corridor p-values at $d = 5$ miles. (b) California p-values at $d = 5$ miles.

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 Fig. 4(a) and (b). This allows one to check visually for *regions* of significant clustering. In particular, groupings of very low p-values serve to indicate not only the location but also the *approximate size* of possible clusters. Such groupings based on p-values necessarily suffer from "multiple testing" problems, which we address in later sections and more systematically in Appendix B.

4.2. Test results for local clustering

For our local cluster analyses, simulations were again performed using $N = 999$ test patterns of size $n - 1$ for each of the n (=1035 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 $D = \{0.25, 0.5, 0.75, 1, 2, \ldots, 99, 100\}$. 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 Fig. 4(a) and (b), respectively. As seen in the legend, those R&D locations, i, exhibiting maximally significant clustering $[P_i(5) = 0.001]$ are shown in black, and those with p-values not exceeding 0.005 are shown as dark gray. Here, it is evident that essentially all of the most significant locations occur in four distinct groups in the Northeast Corridor, which can be roughly described (from north to south) as the "Boston," "New York City," "Philadelphia," and "Washington, D.C.," agglomerations.[25] In California, there are again three distinct groups, roughly described (from north to south) as "San Francisco Bay Area," "Los Angeles area (mainly Irvine)," and "San Diego." While these patterns are visually compelling, it is important to establish such results more formally.

5. Identifying spatial clusters: the 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 previously, and in particular, focuses on the *multiscale* nature of local K-functions. More specifically, this clustering procedure starts with the local point-wise clustering results in Section 4.1 and seeks to identify subsets of points that can serve as "core"

[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.

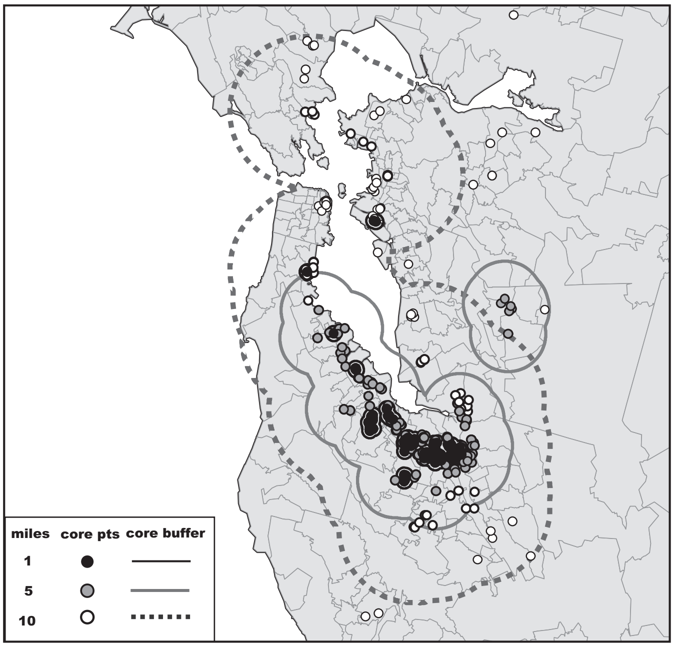

Fig. 5. Multiscale core clusters in the San Francisco Bay Area.

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 clustering in our data set, namely the San Francisco Bay Area in California shown in Fig. 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 Fig. 6(a). 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 Fig. 6(b),[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.

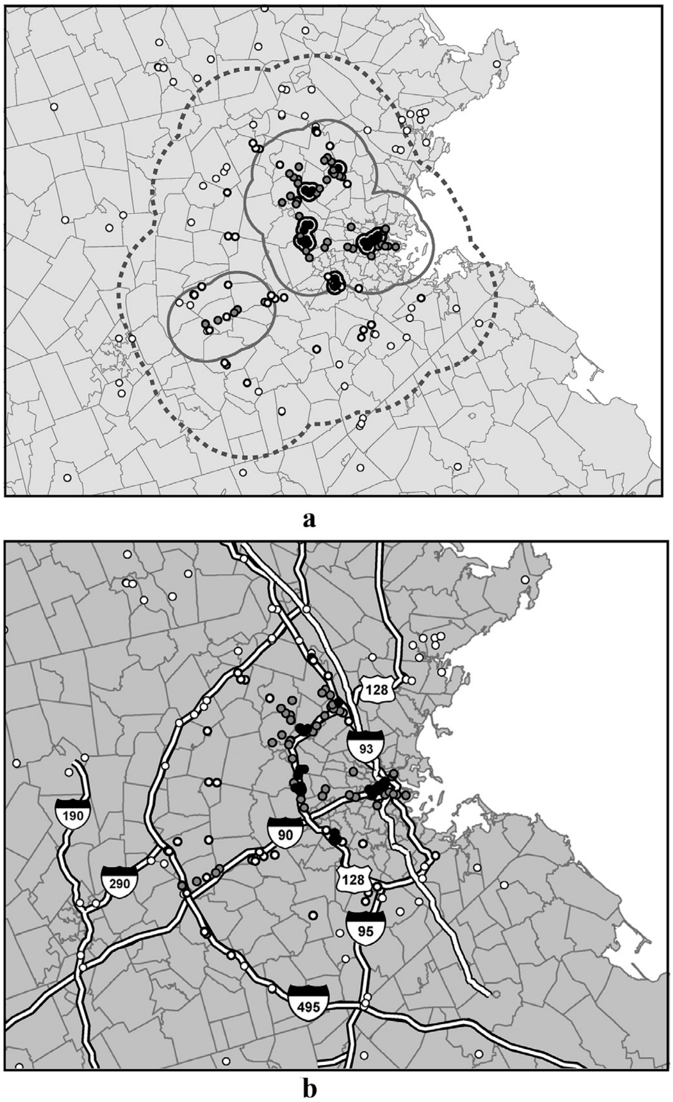

Fig. 6. (a) Multiscale core clusters in Boston. (b) Proximity to major routes in Boston.

Note, finally, that while the clusters in both Figs. 5 and 6(a) tend to be nested by scale, this is not always the case.[28] For example, the five-mile "Livermore Lab" cluster in Fig. 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

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

[28] The area of 5-mile clusters in the Northeast is on average 277 square miles, while the area of 10-mile clusters in the Northeast is on average 2498. In California, the corresponding areas for 5- and 10-miles clusters are 319 and 1326 square miles, respectively.

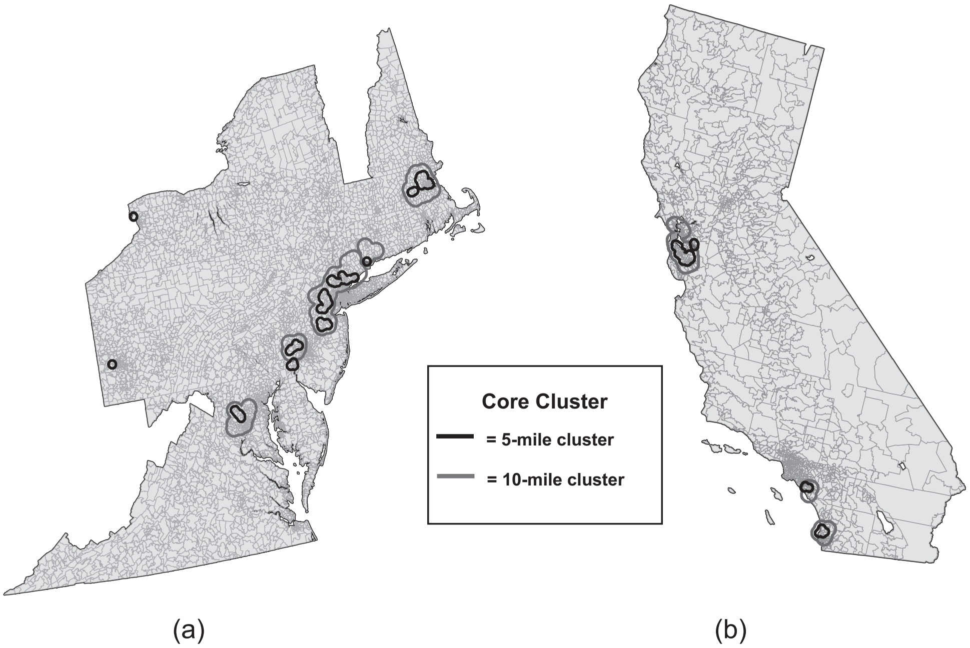

Fig. 7. (a) Northeast corridor core clusters $d = 5, 10$. (b) California Core Clusters $d = 5, 10$.

helps to overcome the particular limitations of significance-maximizing approaches mentioned previously. First, the shapes of individual core clusters are seen to be more sensitive to the actual configuration of points than those found in significance-maximizing methods.[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 Fig. 7(a) and (b), respectively. Fig. 7(a) 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 Fig. 7(b) 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.

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' desire 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 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 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 (Fig. 8(a)) and in California (Fig. 8(b)). 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, Philadelphia and Washington, D.C.) previously identified in Fig. 7(a) resurface when using the STEM worker backcloth. Similarly, the three major clusters identified in Fig. 7(b) for California (one

[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.

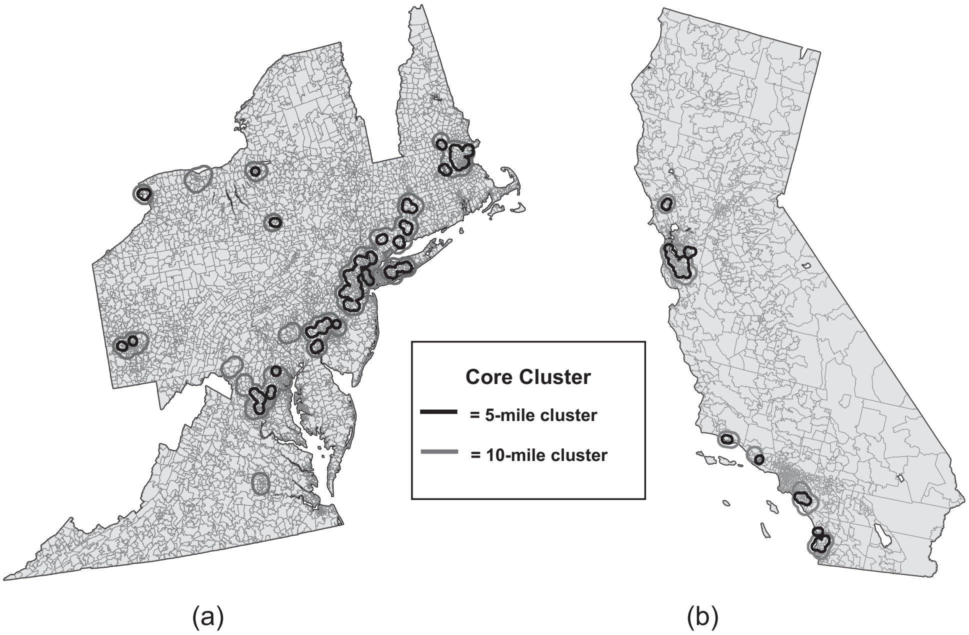

Fig. 8. (a) Northeast corridor core clusters $d = 5$, 10 (STEM workers). (b) California core clusters $d =$ 5, 10 (STEM workers).

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 cluster emerge under the STEM worker backcloth. Five additional ten mile cluster are found in the Northeast Corridor (one each in Lancaster, PA, Hagerstown, MD, Binghamton, NY, Syracuse, NY, Rochester, NY, and in Richmond, VA). Three additional ten 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 1700 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.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jue.2017.05.007.

References

Anselin, L., 1995. Local indicators of spatial association – LISA. Geogr. Anal. 27, 93–115.
Arbia, G., Espa, G., Quah, D., 2008. A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. Empir. Econ. 34, 81–103.
Audretsch, D.B., Feldman, M.P., 1996. R&D spillovers and the geography of innovation and production. Am. Econ. Rev. 86, 630–640.
Berliant, M., Reed III, R.R., Wang, P., 2006. Knowledge exchange, matching, and agglomeration. J. Urban Econ. 60, 69–95.
Buzard, K., Carlino, G.A., Hunt, R.M., Carr, J.K., Smith, T.E., 2016. 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.
Carlino, G.A., Kerr, W.R., 2015. Agglomeration and innovation. In: Henderson, J.Vernon, Duranton, Gilles, Strange, William (Eds.). Handbook of Regional and Urban Economics, 5A. North Holland, Amsterdam.
Dalton, D.H., Serapio, M.G., 1995. Globalizing Industrial Research and Development. U.S. Department of Commerce, Office of Technology Policy, Washington, D.C.
Directory of American Research and Technology, 23rd ed. New York: R.R. Bowker (1999).
Duranton, G., Puga, D., 2003. Mirco-foundations of urban agglomeration economies. In: Henderson, J.Vernon, Thisse, J.F. (Eds.). Handbook of Regional and Urban Economics, 4. North Holland, Amsterdam.
Duranton, G., Overman, H.G., 2005. Testing for localization using micro-geographic data. Rev. Econ. Stud. 72, 1077–1106.
Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. J. Polit. Econ. 105, 889–927.
Ellison, G., Glaeser, E.L., Kerr, W., 2010. What causes industry agglomeration? evidence from coagglomeration patterns. Am. Econ. Rev. 100, 1195–1213.
Getis, A., 1984. Interaction modeling using second-order analysis. Environ. Plan. 16, 173–183.
Helsley, R., Strange, W., 2002. Innovation and input sharing. J. Urban Econ. 51, 25–45.
Hunt, R., Matching Externalities and Inventive Productivity. Federal Reserve Bank of Philadelphia Working Paper 07-07 (2007).
Kerr, W.R., Kominers, S.D., 2015. Agglomerative forces and cluster shapes. Rev. Econ. Stat. 97, 877–899.
Marcon, E., Puech, F., 2003. Evaluating the geographic concentration of industries using distance-based methods. J. Econ. Geogr. 3, 409–428.
Murata, Y., Nakajima, R., Okamoto, R., Tamura, Ryuichi, 2015. Localized knowledge spillovers and patent citations: a distance-based approach. Rev. Econ. Stat. 96, 967–985.
National Science Foundation, 2000. Research and development in industry: 1998. National Science Foundation, Division of Science Resources Studies, Arlington, VA.
Ripley, B.D., 1976. The second-order analysis of stationary point patterns. J. Appl. Probab. 13, 255–266.
Rosenthal, S., Strange, W.C., 2001. The determinants of agglomeration. J. Urban Econ. 50, 191–229.
Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611.