WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER 17-32
LOCALIZED KNOWLEDGE SPILLOVERS: EVIDENCE FROM THE SPATIAL CLUSTERING OF R&D LABS AND PATENT CITATIONS

Kristy Buzard
Syracuse University

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

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

Jake K. Carr
The Ohio State University

Tony E. Smith
University of Pennsylvania

September 2017

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

Localized Knowledge Spillovers:
Evidence from the Spatial Clustering of R&D Labs and Patent Citations[*]

Kristy Buzard
Syracuse University

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

Jake K. Carr
The Ohio State University

Tony E. Smith
University of Pennsylvania

September 2017

Abstract

Patent citations are a commonly used indicator of knowledge spillovers among inventors, while clusters of research and development labs are locations in which knowledge spillovers are particularly likely to occur. In this paper, we assign patents and citations to newly defined clusters of American R&D labs to capture the geographic extent of knowledge spillovers. Our tests show that the localization of knowledge spillovers, as measured via patent citations, is strongest at small spatial scales and diminishes rapidly with distance. On average, patents within a cluster are about three to six times more likely to cite an inventor in the same cluster than one in a control group. At the same time, the strength of knowledge spillovers varies widely between clusters. The results are robust to the specification of patent technological categories, the method of citation matching and alternate cluster definitions.

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

JEL codes: O31, R12

[*] Correspondence: jerry.carlino@phil.frb.org. This paper has benefited from outstanding research assistance by Adam Scavette. 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

The primary activity at R&D establishments is knowledge-based, making concentrations of R&D labs indicative of places in which localized knowledge spillovers would occur. A recent study by Buzard et al. (2017) shows that R&D labs are, indeed, highly spatially concentrated even within a given metropolitan area. Buzard et al. (2017) introduce the multiscale core cluster procedure in which the boundaries of the core clusters are determined by interrelationships among the sample R&D labs in two major R&D regions: the Northeast corridor and California. These clusters should therefore reflect the appropriate boundaries in which knowledge spillovers are most likely to be at work more accurately than administrative boundaries. In that sense, the geography of their clusters are better suited for studying knowledge spillovers than are states, metropolitan areas, or other political or administrative boundaries. In this paper, we extend Buzard et al. (2017) by assigning patents and citations to the R&D clusters they identify and test for evidence of localized knowledge spillovers in patent citations.[1]

We provide evidence that the clustering of R&D labs is related to knowledge spillovers by studying the relative geographic concentration of citations to patents originating in the clusters we identify.[2] To do this, we construct treatment versus control tests for the localization of patent citations in the spirit of those found in Jaffe, Trajtenberg, and Henderson (1993), hereafter, JTH. For labs in the Northeast corridor, our baseline results indicate that citations are on average about four to six times more likely to come from the same cluster as earlier patents than one would predict using a (control) sample of otherwise similar patents. For California, the baseline results suggest that citations are on average roughly four to five times more likely to come from the same cluster as earlier patents than one would predict using the control sample.

We also find that patents inside each cluster receive more citations on average than those outside the cluster in a suitably defined counterfactual area. This suggests that the geography and scale

[1] Rather than using fixed geographic units, such as counties or metropolitan areas, Buzard et al. (2017) use continuous measures to identify the spatial structure of the concentrations of R&D labs. Specifically, they use point pattern methods to analyze locational patterns over a range of selected spatial scales (within 5 miles, 10 miles, 20 miles, etc.). This approach allows them to consider the spatial extent of the agglomeration of R&D labs and to measure any attenuation of clustering with distance more accurately.

[2] Earlier research — e.g., Jaffe, Trajtenberg, and Henderson (1993), Thompson and Fox-Kean (2005), Kerr and Kominers (2015), Murata et al. (2017, 2014) — document patterns of spatial concentration (often described as localization) in patent citations.

of the clusters identified by Buzard et al. (2017) is related to the extent of the localization of knowledge spillovers, at least as evidenced by patent citations. We can also speak to the question of whether the transmission of knowledge attenuates with distance. We add to the mounting evidence from studies using alternative data that knowledge spillovers begin to attenuate at distances ranging from just a few blocks to a few miles:[3] The localization of knowledge spillovers in our data appears strongest at small spatial scales (5 miles or less) and diminishes rapidly with distance. Given this attenuation, the magnitude of the localized spillovers documented by studies that use state and metropolitan area data may be understated, and the exact geography that is driving the spillovers may not be well identified.

It's possible that technologically related activities may cluster to benefit from agglomeration forces other than knowledge spillovers, such as sharing and better matching of workers and firms. These other sources of agglomeration potentially explain some of the geographic concentration of technologically related research activity.[4] To address this issue, our basic approach is to follow JTH in constructing a control sample of patents that have the same technological and temporal distribution as the citations to account for these other agglomeration forces. Our test for knowledge spillovers is whether the citation matching frequency is significantly greater than the control matching frequency. Put differently, our test is whether citations are more localized relative to what would be expected given the existing distribution of technologically related activity. To further control for sharing and matching externalities, we perform a robustness check using the alternative cluster definitions developed by Buzard et al. (2017) in which the backcloth is based on STEM workers and find that the baseline results are qualitatively unchanged.

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

[3] See, for example, Kerr and Kominers (2015); Elvery and Sveikauskas (2010), Arzaghi and Henderson (2008), Agrawal, Kapur, and McHale (2008), Keller (2002), Rosenthal and Strange (2001), Adams and Jaffe (1996), and Audretsch and Feldman (1996).

[4] See Carlino and Kerr (2015) for more discussion on the theory of the agglomeration of innovative activity.

results are also robust to drawing the controls more narrowly from patents that share the same patent class and subclass as the citing patents.

Finally, we show that our results persist when we use coarsened exact matching as an alternative method to select the controls. In this case, the tests for the localization of patent citations are at the lower end of our findings, particularly in California.

2. SPATIAL CLUSTERS

We use R&D cluster definitions from Buzard et al. (2017), which cover California and a 10-state area in the Northeast U.S.[5] They use continuous methods (based on Ripley's (1976) *K*-function) to assess the concentration of R&D labs relative to a baseline of manufacturing employment.[6] Roughly speaking, a lab is *locally agglomerated* at a scale of *d* miles, if it has more neighboring labs within distance *d* than would be expected (statistically) based only on the distribution of manufacturing employment. Of special relevance are *core points* at scale, *d*, defined as those labs exhibiting maximally significant[7] local agglomeration at this scale, and also having at least four neighboring labs within distance *d* (to avoid small clusters of little practical significance).

To identify distinct clusters at scale *d*, Buzard et al. (2017) create buffers of radius *d* around each core point in ArcMap and designate the set of labs in each connected component of these buffer zones as a *core cluster* of points. Each distinct cluster thus contains a given set of "connected" core points along with all other points that contributed to their maximal statistical significance. Buzard et al. (2017) refer to this procedure as the *multiscale core-cluster approach*. Although this approach is meaningful for any set of scale choices, we focus primarily on 5- and 10-mile clusters that correspond closely to intuition about the size and location of research clusters.

[5] The ten states in the Northeast are Connecticut, Delaware, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Virginia, plus the District of Columbia; they contain 1,035 R&D labs. There are 645 R&D labs in California.

[6] Buzard et al. (2017) develop an alternative benchmark or backcloth for analyzing R&D clustering with respect to STEM workers to address the concern that R&D labs may follow knowledge workers. We will provide patent citation results using this alternative backcloth as well as for the manufacturing employment backcloth.

[7] Using *K*-function permutation tests based on 1,000 permutations, Buzard et al. (2017) define maximal significance to be the smallest p-value obtainable under that test, namely $p = 0.001$.

An overall depiction of core clusters for both the Northeast corridor and California is shown in Figures 1a and 1b, respectively. Figure 1a 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 1b shows the three major clusters in California (one each in the Bay Area, Los Angeles, and San Diego).

To see how the multiscale core clustering approach works, examine the San Francisco Bay Area in California shown in Figure 2. The approach identified one 10-mile cluster that covers almost all of the Bay Area. There is a dominant 5-mile core cluster that is completely nested in the 10-mile cluster; this is the area most commonly referred to as Silicon Valley. Finally, as the figure shows, there are numerous one-mile clusters running from the Stanford Research Park area to San Jose at the center of Silicon Valley.[8] Buzard et al. (2017) similarly identify other clusters of R&D labs such as the one centered around Cambridge, MA, and the Route 128 corridor that correspond closely to the most well-known R&D concentrations in the study area (Figure 3).

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. See Buzard et al. (2017) for a discussion of the benefits of the multiscale core cluster approach relative to significance-maximizing methods.

The ultimate value of such clusters for our purposes can be determined only by testing their economic significance — to which we now turn.

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

In this section, we test for evidence of localized knowledge spillovers by assigning patents and citations to the core clusters identified by Buzard et al. (2017). More specifically, we study the

[8] Note that while the clusters in Figure 3 tend to be nested by scale, this is not always the case. For example, the 5-mile "Livermore Lab" cluster in Figure 2 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 5-mile scale, it is too small by itself to be picked up at the 10-mile scale.

relative geographic concentration of citations to patents originating in the clusters. These citations are a concrete indication of the transmission of information from one inventor to another.

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

We construct a comparable test statistic, with several refinements, and we substitute the R&D clusters identified in Buzard et al. (2017) for the state and metropolitan area geography used by JTH. This provides us with an alternative way to test for possible localized knowledge spillovers at much smaller spatial scales than are found in much of the preceding literature. Recall that the boundaries of the core clusters are determined by interrelationships among the R&D labs in our sample and, therefore, should more accurately reflect the appropriate boundaries in which knowledge spillovers are most likely to be at work. In that sense, the geography of our clusters should be better suited for studying localized knowledge spillovers than states, metropolitan areas, or other political or administrative boundaries.

3.1 Construction of the Citations Data Set

For this analysis, we use data obtained from the NBER Patent Data Project.[9] The data span the years 1996–2006. We identify the inventors on a patent using data on inventor codes found in the Patent Network Dataverse (Lai, D'Amour, and Fleming, 2009). Patents are assigned to locations based on the zip code associated with the *residential* address of the first inventor on the patent.[10] We do not use the address of the assignee (typically the company that first owned the patent) because this may not reflect the location where the research was conducted (e.g., it may be the address of the corporate headquarters and not the R&D facility). While it's possible that an inventor's home lies outside a cluster while his professional work takes place inside a cluster, this type of measurement error would bias our results against finding significant location differentials. As a robustness check, we repeated our main analysis using the zip code of the second inventor on the patent. While the sample size is smaller because not all patents list two or more inventors, the results were virtually the same as we report below.[11]

For our tests, we rely primarily on the boundaries identified by the 5-mile and 10-mile core clusters located in the Northeast corridor and in California.[12] For each core cluster at a given scale, we assemble four sets of patents. The first set, which we call *originating patents*, represent those patents granted in the years 1996–1997 by an inventor living in the cluster.[13] We call the second set of patents *citing patents*. These consist of all subsequent patents — including patents for which the residential address of the first inventor is located outside the U.S. — that cite one or more of the originating patents, after excluding patents with the same inventor or that were initially assigned to the same company as the originating patent. We exclude these self-citations because these are unlikely to represent the knowledge spillovers we seek to identify.[14]

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

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

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

[12] In Section 4.1 that follows, we report comparable tests for larger and smaller clusters.

[13] 1996-1997 is chosen because the labs data on which the clusters are based is from 1998.

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

For every citing patent, we attempt to match it to an appropriate control patent. When we are successful, we include the citing patent in a set we call *treatment patents* and the matched patent in a set we call *control patents*. We select control patents using the following approach. For a given citing patent, the set of potential control patents must have an application date after the grant date of the originating patent that is cited. Potential control patents also cannot cite the originating patent. The application date of potential control patents must be within one year (six months on either side) of the application date of the treatment patent. Finally, as was done by JTH, potential control patents must have the same three-digit primary patent class as the treatment patent.[15] In this way, potential controls are drawn from patents in the same technological field.

The set of potential control patents for a given treatment patent may overlap with the set of potential controls for other treatment patents. To rule out any possibility that this overlap may affect our tests, we randomized the order in which treatment patents were matched to control patents, and we randomized the selection of a specific control patent when there was more than one potential control patent from which to choose.[16] The main results reported below allow for the selection of control patents with replacement. In other words, a given control patent may be matched to more than one citing patent. As a robustness check (not shown), we repeat the analysis by sampling potential controls *without* replacement.[17] In this case, a potential control

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

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

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

patent can be matched with at most one citing patent. While this reduces the rate at which we can match control patents to citing patents, it does not materially affect the test statistics.[18]

3.2 The Test Statistics

For any given cluster scale, $d\ \ (=5,\ 10)$, let η_o denote the number of *originating patents* indexed $\{o_i : i=1,\cdots,\eta_0\}$ that were granted to inventors living in one of the core clusters at scale *d* in the years 1996–1997.[19] Let η_i denote the number of subsequent citations $\{c_{ij} : j=1,\cdots,\eta_i\}$ to o_i (after removing self-citations) over the years 1996–2006. For each of these citing patents, c_{ij}, designated as *treatment patents*, we attempted to identify a unique *control patent*, $\tilde{c}_{ij}$, with the same three-digit patent class and with an application date within one year of the treatment patent (see previous description). We are not always successful in doing so. Let $\tilde{\eta}_i (\leq \eta_i)$ denote the number of treatment patents, c_{ij}, for which a control, $\tilde{c}_{ij}$, was found.

Among these $\tilde{\eta}_i$ treatment patents, we count the number of patents, m_i, for which the residential address of the first inventor on the citing patent is located in the *same* core cluster as the originating patent it cites. The fraction of all such patents at scale *d*, i.e., the *treatment proportion*, is given by[20]

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

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

$$\tilde{p} = \frac{\sum_{i=1}^{\eta_o} \tilde{m}_i}{\sum_{i=1}^{\eta_o} \tilde{\eta}_i} = \tfrac{1}{\tilde{\eta}} \sum_{i=1}^{n_o} \tilde{m}_i\ . \tag{2}$$

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

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

[20] The dependency of fraction, *p* (and all other quantities in (1)) is taken to be implicit.

The resulting test statistic is simply the difference between these proportions, i.e., $p-\tilde{p}$. Under the null hypothesis of "no localization of knowledge spillovers," this difference of independent proportions is well known to be asymptotically normal with mean zero and thus provides a well-defined test statistic.[21]

3.3 Main Results

Table 1a presents the results of our localization or matching rate tests for the nine 5-mile clusters identified in by Buzard, et al. (2017) for the Northeast corridor, while Table 1b shows the results for the four 10-mile clusters they identify. As the last row of Table 1a shows, inventors living in the 5-mile clusters obtained 8,526 patents in 1996–1997 (column A). Those patents subsequently received 76,730 citations from other patents during the sample period (column B). Our matching algorithm, with replacement, was able to match 85 percent of the citing patents with an appropriate control patent (column H). Among the treatment patents, 3.69 percent (column G) had a first inventor living in the same cluster as the patent it cited; this is the treatment proportion. Among the control patents, only 0.62 percent (column J) had a first inventor living in the same cluster as the patent cited by the treatment patent; this is the control proportion. As shown in the next to the last column of the table, on average, a given patent citing an earlier patent in a 5-mile cluster is six times as likely to have a first inventor living in that cluster than would be expected by chance alone. This value is on the higher side of the range reported by JTH for their test of localization at the metropolitan-area level. As the last row of the Table 1a shows, the difference between the treatment and control proportions is highly statistically significant (column L). In addition, the location differential — defined as the ratio of treatment and control proportions — is at least around 3.0 for every 5-mile cluster.

Table 1b presents the results of our localization tests among 10-mile clusters in the Northeast corridor. At a somewhat larger spatial scale, we find there are more originating patents, more citing patents, and, thus, more treatment and control patents. Both the treatment and control proportions (columns G and J) are higher than was found among the 5-mile clusters. The *t*

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

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

Tables 2a and 2b present the results of our localization tests among 5- and 10-mile clusters, respectively, in California identified by Buzard, et al. (2017). Compared with the Northeast corridor, we find many more originating patents, citing patents, and, therefore, treatment and control patents. The treatment proportions (column G) among the California clusters are much higher than those found in the Northeast corridor. However, this is driven almost entirely by the cluster association with Silicon Valley. The control proportions (column J) are also larger than we found in the Northeast corridor. The *t*-statistic for the difference in treatment and control proportions (column L) is highly significant for all the 5-mile and 10-mile clusters. On average, a given patent citing an earlier patent in a 5- or 10-mile cluster in California is four to four-and-a-half times as likely to have a first inventor living in that cluster than would be expected by chance alone.

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

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

4. ADDITIONAL RESULTS AND ROBUSTNESS CHECKS

4.1 The Relationship Between Citation Location Differentials and Spatial Scale

The statistics in the preceding tables suggest that there may be a systematic relationship between the size of the clusters we study and the magnitude of the location differentials we find. To explore this further, we extended our analysis to consider clusters at spatial scales of 20 miles. We summarize the results in Table 3a and Table 3b.

A number of patterns are evident from the table. First, the rate of increase in the number of originating patents associated with larger core clusters falls off because a number of clusters that are significant at smaller spatial scales are not significant at the larger spatial scales. The treatment and control proportions tend to increase as we consider larger core clusters. The difference between these proportions becomes more and more statistically significant as the sample size rises. At the same time, the location differential falls monotonically as the geographic size of the clusters increases. These results suggest that the core clusters are picking up knowledge spillovers over a variety of spatial scales. Nevertheless, the localization effects appear to be largest at spatial scales of 5 miles and perhaps less. The attenuation in the localization differential as cluster size increases is a typical finding in studies examining localized knowledge spillovers.[22]

4.2 Are Patents Obtained in the Core Clusters More Influential?

In this section, we investigate whether patents obtained by inventors living within a core cluster are somehow more important, or at least better known, than patents obtained outside of these clusters. We rely on a common metric of patent quality: the number of citations received.[23] We develop a “counterfactual” region for each of the 10-mile core clusters identified in Section 2. For example, the New York cluster is compared with the region outside of that cluster contained in the states of New York, Connecticut, and northern New Jersey. The Boston cluster is compared with the region outside the cluster in the states of Massachusetts, New Hampshire, and

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

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

Rhode Island. In Table 4, we report a simple difference in means test for the number of citations per patents received by patents located inside or outside our clusters. For all our clusters, the average number of citations received by patents is greater inside the cluster compared with the average citations received outside the respective cluster; this difference in citations is statistically significant in all clusters except one (Philadelphia).

These results, combined with the results for the localization of citations, suggest there is prima facie evidence that the inventions developed within a core cluster are more influential than inventions developed outside a cluster but within the same region of the country. An alternative explanation, which we cannot entirely rule out, is that patents within a cluster receive more citations because they are often cited by inventors living nearby. According to this reasoning, the inventions may not necessarily be better, but they are better known by researchers in the area. This interpretation only reinforces the evidence of localized knowledge spillovers in our clusters.

4.3 Alternative Approaches to Identifying Cluster Boundaries

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

To address the concern that we may be intermingling knowledge spillovers with labor market pooling, we use an alternative set of clusters developed by Buzard et al. (2017) based on a measure of STEM workers by location.[24] For the backcloth of these clusters, they 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

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

Patterns. We report the results of this alternative test for 5- and 10-mile clusters in the Northeast corridor (Tables 5a and 5b) and in California (Tables 6a and 6b). Note that the cluster definitions change when the backcloth changes, so the list of clusters in these tables differs from those in Tables 1 and 2. With the exception of the 5-mile clusters in the Northeast corridor, the average location differentials using the STEM worker backcloth are virtually the same as for the baseline findings. The location differential falls from 6.0 for the 5-mile clusters in the Northeast corridor when considering the baseline results to 4.2 for the results when the clusters are based on STEM workers. For the most part, the findings reported for the location differentials in the baseline (and subsequent analysis) suggest little, if any, upward bias as a result of labor market pooling.

4.4 Alternative Approaches to Identifying Control Patents

4.4.1 Disaggregated Subclasses

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

The results presented thus far are based on the JTH approach of limiting potential control patents to ones that share the same three-digit primary class as the citing patent. As a robustness check, we implement one version of the matching requirements tested in TFK. We restrict potential control patents to ones that share the same primary class and subclass as the citing patent.[25] Our methodology is otherwise the same as we describe in Section 3.2. We report the results of this alternative test for 5- and 10-mile clusters in the Northeast corridor (Tables 7a and 7b) and in California (Tables 8a and 8b). Comparing these results with our baseline results (Tables 1a and 1b) and (2a and 2b), there are very small differences in the treatment and control proportions. The *t*-statistics using the TFK approach are only slightly smaller than they are when using the

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

JTH approach, but they are nevertheless very large. We conclude that our results do not appear to be sensitive to the choice of technology controls.

4.4.2 Coarsened Exact Matching

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

We use the CEM matched controls in several ways. First, we follow the JTH location differential approach used in producing Tables 1 and 2, our baseline findings, but use the CEM controls. For this approach, we exclude patents with the same inventor or that were initially assigned to the same company as the originating patent.[26] The results are reported in Table 9 (for the Northeast corridor) and Table 10 (for California). The location differentials are uniformly smaller than we previously reported for the broad cluster in the Northeast corridor and in California. On average, a given patent citing an earlier patent in a 5-mile cluster in the Northeast corridor is 4.5 times as likely to have a first inventor living in that cluster than would be expected by chance alone, compared with a differential of 6.0 reported in our baseline results. The location differential in California's 5-mile cluster falls to 2.5 when using the CEM matched controls from 4.5 reported for the baseline. The location differential in the Northeast corridor 10-mile cluster falls to 2.8 when using the CEM-matched controls from 3.6 reported for the baseline. In the California 10-mile cluster, the location differential falls to 2.5 from 4.2 reported for the baseline.

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

In our second approach, we estimate a logistic model of the likelihood that a patent in cluster h cites an originating patent in that cluster. More formally, if for any given patent, we let T_h denote the indicator variable that this patent cites at least one originating patent in cluster *h*, and similarly, let D_h indicate whether this patent itself originates in cluster *h*, then the conditional likelihood, $\Pr(T_h = 1 | D_h)$, of citing patents in cluster *h* given D_h is postulated to be of the logit form:

$$\Pr(T_h = 1 | D_h) = \frac{\exp(\alpha_h + \beta_h D_h)}{1 + \exp(\alpha_h + \beta_h D_h)}$$

In this setting, it should be clear that citations of patents in cluster *h* are more likely for (treatment) patents in cluster *h* than for (control) patents not in cluster *h*, i.e., $\Pr(T_h = 1 | D_h = 1) > \Pr(T_h = 1 | D_h = 0)$, if and only if $\beta_h > 0$. The estimated coefficients, $(\hat{\beta}_h)$, are reported in Table 11 (along with robust standard errors for these estimates).[27] As seen from the table, the estimated coefficients for *all* clusters are significantly positive (at the 1% level), and thus provide strong support for the findings in Tables 9 and 10.

Finally, to facilitate comparison, the main results found for location differentials are summarized in Table 12. The table shows the results when R&D clustering is analyzed with respect to (i) manufacturing employment (baseline), (ii) STEM workers, (iii) when the controls are alternatively selected to share the same patent class and subclass as the citing patents (disaggregated), and (iv) when the controls are selected using more stringently matched samples (CEM). Regardless of the specification chosen to construct the location differentials, we find that citations are at least about 2.5 times more likely to come from the same cluster as earlier patents than one would predict using a control sample of otherwise similar patents.

[27] For this approach, we do not exclude patents with the same inventor or that were initially assigned to the same company as the originating patent. The observations are weighted based on the number of CEM-matched controls found for each treated observation.

5. CONCLUDING REMARKS

In this paper, we verify that the local clusters identified in Buzard et al. (2017) are economically meaningful by applying tests developed by JTH to measure the degree to which patent citations are localized in these clusters — tangible evidence that knowledge spillovers are geographically mediated. For labs in the Northeast corridor, we find, on average, that citations are about three to six times more likely to come from the same cluster as earlier patents than one would predict using a (control) sample of otherwise similar patents. In California, citations are around three to five times more likely to come from the same cluster as earlier patents than one would predict using the control sample.

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

REFERENCES

Adams, J., and A. Jaffe. “Bounding the Effects of R&D: An Investigation Using Matched Establishment-Firm Data.” *RAND Journal of Economics*, 27 (1996), pp. 700–721.

Agrawal, Ajay, Devesh Kapur, and John McHale. “How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data,” *Journal of Urban Economics*, 64 (2008), pp. 258–269.

Arzaghi, Mohammad, and J. Vernon Henderson. “Networking Off Madison Avenue,” *Review of Economic Studies,* 75 (2008), pp. 1,011–1,038.

Audretsch, David B., and Maryann P. Feldman. “R&D Spillovers and the Geography of Innovation and Production,” *American Economic Review*, 86 (1996), pp. 630–640.

Buzard, Kristy, Gerald A. Carlino, Robert M. Hunt, Jake K. Carr and Tony E. Smith, “The Agglomeration of American R&D Labs,” *Journal of Urban Economics*, 101 (2017), pp. 14-26.

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

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

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

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

Jaffe, Adam, Manuel Trajtenberg, and Rebecca Henderson. “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations,” *Quarterly Journal of Economics,* 108 (1993), pp. 577–598.

Keller, W. “Geographic Localization of International Technology Diffusion,” *American Economic Review,* 92 (2002), pp. 120-142.

Kerr, William R., and Scott Duke Kominers. “Agglomerative Forces and Cluster Shapes,” *Review of Economics and Statistics*, 97 (2015), pp. 877–899.

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

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 (2014), pp. 967–985.

Murata, Yasusada, Ryo Nakajima, and Ryuichi Tamura. “Testing for Localization: A New Approach,” Keio-IES Discussion Paper Series No. DP 2017-017 (2017).

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. 255–266.

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

Thompson, Peter, and Melanie Fox-Kean. “Patent Citations and the Geography of Knowledge Spillovers: A Reassessment,” *American Economic Review*, 95 (2005), pp. 450–460.

Trajtenberg, Manuel. “A Penny for Your Quotes: Patent Citations and the Value of Innovations,” *RAND Journal of Economics*, 21 (1990), pp. 172–187.

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

Table 1a: Five-Mile Clusters in the Northeast corridor, Baseline Results

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

Table 1b: 10-Mile Clusters in the Northeast corridor, Baseline Results

					Treatment Group			Control Group				
Column	A	B	C	D	E	F	G	H	I	J	K	L
Cluster	Originating Patents	Citing Patents	From Same Cluster	Percent (C/B)	Matched Citing Patents*	From Same Cluster*	Percent (F/E)	Control Patents[†]	From Same Cluster	Percent (I/H)	Location Differential (G/J)	*t* Statistic
Boston, MA	4,719	48,315	4,263	8.82%	41,082	3,679	8.96%	41,082	747	1.82%	**4.9**	**45.9**
Washington, DC	926	9,741	327	3.36%	8,089	270	3.34%	8,089	31	0.38%	**8.7**	**14.0**
New York, NY	7,768	67,982	4,738	6.97%	57,626	3,997	6.94%	57,626	1,493	2.59%	**2.7**	**34.8**
Philadelphia, PA	1,594	9,028	409	4.53%	7,851	343	4.37%	7,851	35	0.45%	**9.8**	**16.2**
All 10-Mile Clusters	15,007	135,066	9,737	7.21%	114,648	8,289	7.23%	114,648	2,306	2.0 1%	**3.6**	**60.0**

Sources: NBER Patent Data Project and authors' calculations.

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

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

Table 2a: Five-Mile Clusters in California, Baseline Results

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

Table 2b: 10-Mile Clusters in California, Baseline Results

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

Sources: NBER Patent Data Project and authors' calculations.

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

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

Table 3a: Citation Location Differentials and Spatial Scale (Northeast corridor)

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

Source: NBER Patent Data Project

Table 3b: Citation Location Differentials and Spatial Scale (California)

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

Sources: NBER Patent Data Project and authors' calculations.

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

Table 4: Citation Differentia Between Labs Inside Clusters vs. Labs Outside Clusters (Difference in Means†Test)

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

Sources: NBER Patent Data Project and authors' calculations

†: Citations per Patent Granted, 1996–1997

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

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

Table 5a: Five-Mile Clusters in the Northeast corridor, STEM Worker Clusters

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

Table 5b: 10-Mile Clusters in the Northeast corridor, STEM Worker Clusters

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

Sources: NBER Patent Data Project and authors' calculations.

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

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

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

Table 6a: Five-Mile Clusters in California, STEM Worker Clusters

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

Table 6b: Ten-Mile Clusters in California, STEM Worker Clusters

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

Sources: NBER Patent Data Project and authors' calculations.

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

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

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

Table 7a: Five-Mile Clusters in the Northeast corridor, Disaggregated Subclasses

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

Table 7b: 10-Mile Clusters in the Northeast corridor, Disaggregated Subclasses

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

Sources: NBER Patent Data Project and authors' calculations.

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

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

Table 8a: Five-Mile Clusters in California, Disaggregated Subclasses

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

Table 8b: 8-Mile Clusters in California, Disaggregated Subclasses

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

Sources: NBER Patent Data Project and authors' calculations.

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

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

Table 9a: Five-Mile Clusters in the Northeast corridor, Coarsened Exact Matching

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

Table 9b: 10-Mile Clusters in the Northeast corridor, Coarsened Exact Matching

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

Sources: NBER Patent Data Project and authors’ calculations.

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

†Control patents are selected using the coarsened exact matching procedure Control patents must have the same three-digit technology classification as the citing patent, and their application date must be within a one-year window of the citing patent’s application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to which the originating patent is assigned. Control patents must have the same application year and three-digit technology classification as the treatment patents, in addition to having the same grant year and the number of citations that the treatment patent receives.

Table 10a: Five-Mile Clusters in California, Coarsened Exact Matching

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

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

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

Source: NBER Patent Data Project and authors' calculations.

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

[†]Control patents are selected using the coarsened exact matching procedure Control patents must have the same three-digit technology classification as the citing patent, and their application date must be within a one-year window of the citing patent's application date. These control patents are chosen with replacement sampling. We eliminate self-citations and do not allow controls to be drawn from patents assigned to the same firm to which the originating patent is assigned. Control patents must have the same application year and three-digit technology classification as the treatment patents, in addition to having the same grant year and the number of citations that the treatment patent receives.

Table 11[†]

Northeast

Cluster Name	Coefficient on Originating Patent ($\hat{\beta}_h$)	Standard Errors
Boston5A	2.82	0.1062*
Boston5B	1.5	0.0300*
NY5A	2.17	0.0737*
NY5B	1.26	0.0603*
NY5C	0.8	0.0967*
NY5D	2.26	0.3235*
Philly5A	3.13	0.1321*
Philly5B	2.28	0.1335*
Boston10	1.37	0.0199*
DC10	1.65	0.0652*
NY10	0.79	0.0192*
Philly10	2.13	0.0574*
	Broad Regions	
NE5	0.77	0.0167*
NE10	0.68	0.0113*

California

Cluster Name	Coefficient on Originating Patent ($\hat{\beta}_h$)	Standard Errors
SD5	2.34	0.1251*
LA5	2.52	0.1137*
SF5A	1.06	0.0107*
SF5B	2.81	0.1098*
SD10	1.56	0.0381*
LA10	2.06	0.0493*
SF10	1.09	0.0093*
	Broad Regions	
CA5	1.01	0.0103*
CA10	0.99	0.0086*

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

Table 12: Summary of Location Differentials

	Northeast Corridor			California	
	Five-Mile Cluster	10-Mile Cluster		Five-Mile Cluster	10-Mile Cluster
Baseline	6.0	3.6		4.5	4.2
STEM	4.2	3.3		4.6	4.6
Disaggregated	6.2	3.6		4.5	4.2
CEM	4.5	2.8		2.4	2.5

[†] Baseline results from column K in Tables 1 and 2; STEM results from column K in Tables 5 and 6; Disaggregated results from column K in Tables 7 and 8; CEM results from column K in Tables 9 and 10.

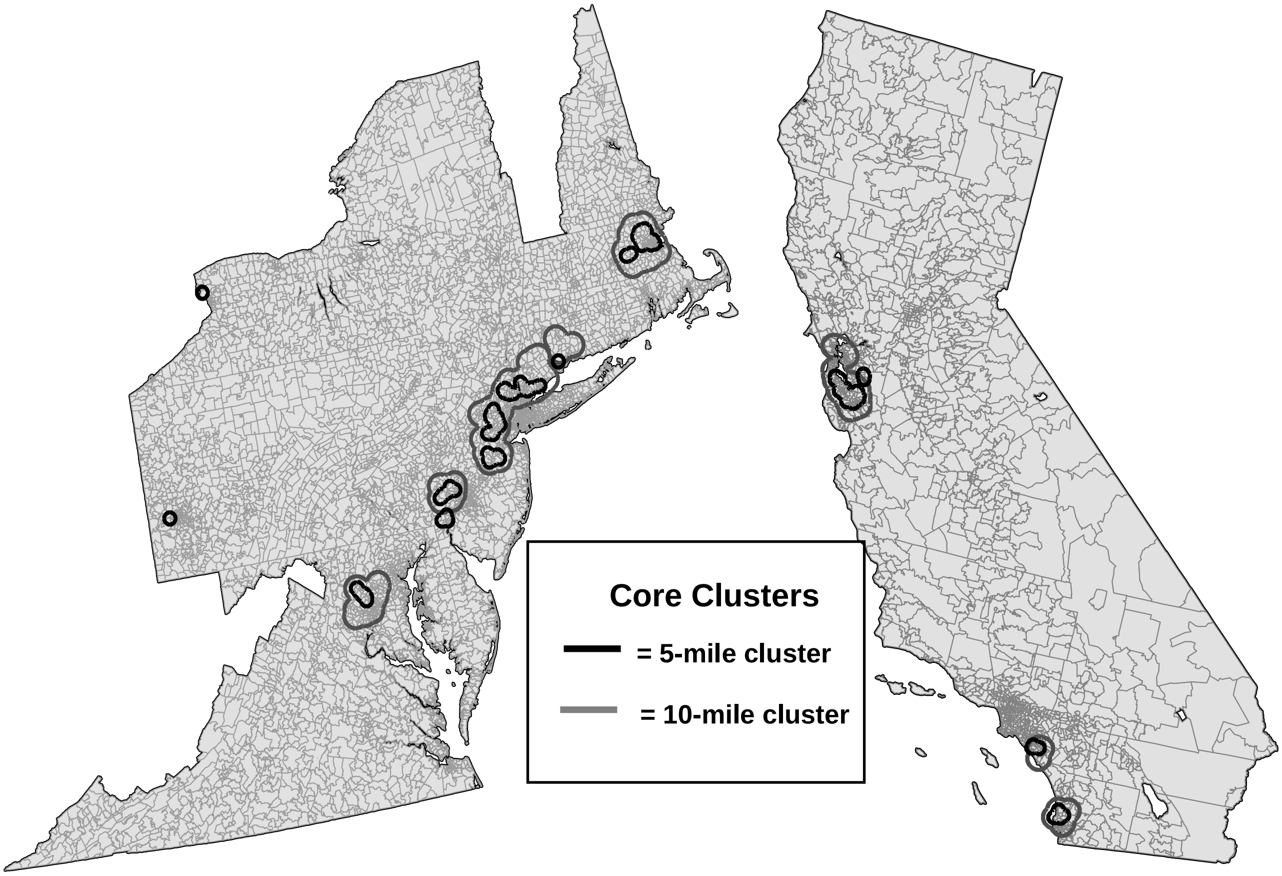

Figure 1a: Northeast Corridor Core Clusters
***d* = 5, 10**

Figure 1b: California Core Clusters
***d* = 5, 10**

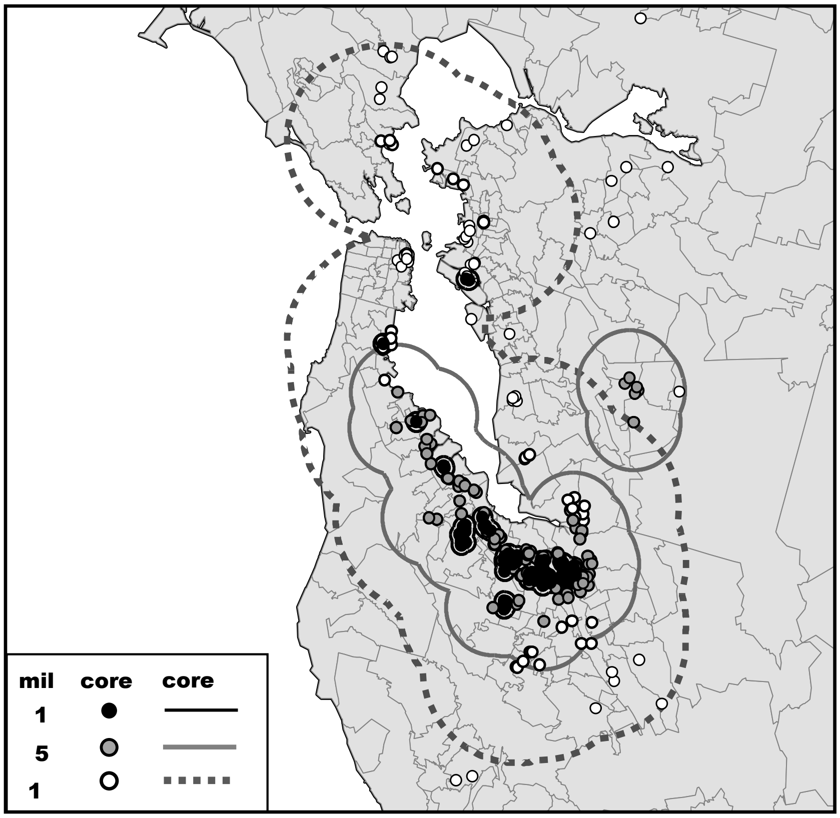

Figure 2: Multiscale Core Clusters in the San Francisco Bay Area

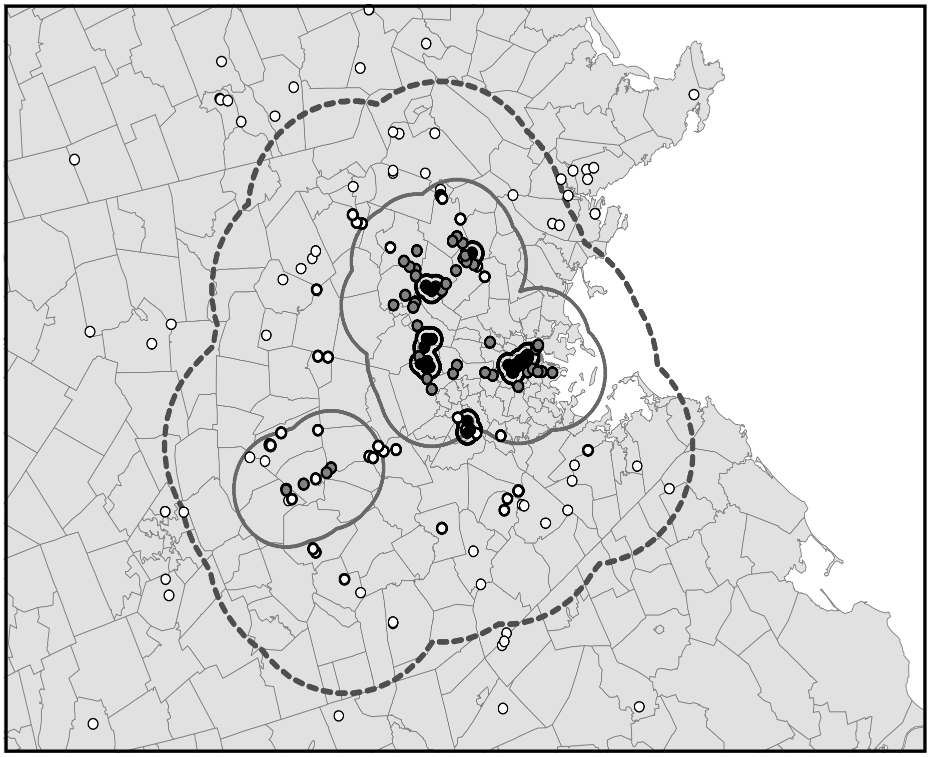

Figure 3: Multiscale Core Clusters in Boston