WORKING PAPER NO. 12-22
THE AGGLOMERATION OF R&D LABS

Gerald A. Carlino
Federal Reserve Bank of Philadelphia

Robert M. Hunt
Federal Reserve Bank of Philadelphia

Jake K. Carr
Ohio State University

Tony E. Smith
University of Pennsylvania

September 2012

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

THE AGGLOMERATION OF R&D LABS[*]

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

Jake K. Carr
Ohio State University

Tony E. Smith
University of Pennsylvania

September 2012

ABSTRACT

We study the location of more than 1,000 research and development (R&D) labs located in the Northeast corridor of the U.S. Using a variety of spatial econometric techniques, we find that these labs are substantially more concentrated in space than the underlying distribution of manufacturing activity. Ripley's *K*-function tests over a variety of spatial scales reveal that the strongest evidence of concentration occurs at two discrete distances: one at about one-quarter of a mile and another at about 40 miles. We also find that R&D labs in some industries (e.g., chemicals, including drugs) are substantially more spatially concentrated than are R&D labs as a whole.

Tests using *local K*-functions reveal several concentrations of R&D labs that appear to represent research clusters. We verify this conjecture using significance maximizing techniques (e.g., SATSCAN) that also address econometric issues related to "multiple testing" and spatial autocorrelation.

We develop a new procedure for identifying clusters – the *multiscale core-cluster* approach, to identify labs that appear to be clustered at a variety of spatial scales. Locations in these clusters are often related to basic infrastructure such as access to major roads. There is significant variation in the industrial composition of labs across these clusters.

The clusters we identify appear related to knowledge spillovers: Citations to patents previously obtained by inventors residing in clustered areas are significantly more localized than one would predict from a (control) sample of otherwise similar patents.

JEL Codes: 031, R12Keywords: Spatial clustering of R&D labs, measures of geographic concentration, localized knowledge spillovers, patent citations

---

[*] We thank Kristian Behrens, Jim Bessen, Satyajit Chatterjee, Giles Duranton, Vernon Henderson, Andy Haughwout, Jim.Hirabayashi, Tom Holmes, Will Strange, Isabel Tecu, and Elisabet Vilasecans-Marsal for comments and suggestions. The R&D labs data used in this paper were painstakingly assembled and checked by Cristine McCollum, Elif Sen, Annette Swahla, and especially Kristy Buzard. 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

Researchers have established a strong empirical relationship between a country's investments in research and development (R&D), the resulting innovations, and productivity growth. In theoretical models of endogenous growth developed by Romer (1990), Grossman and Helpman (1991), and Aghion and Howitt (1992) investment in R&D is critical for economic growth. Empirical studies have presented evidence of the importance of R&D investment for technological progress, productivity, and growth (see Akcay, 2011, for a recent survey of this literature). The fertility of R&D depends on many factors, but surely an important one is the exchange of ideas. Even though we live in an era of global commerce and ubiquitous electronic communications networks, as we demonstrate in this paper, it remains the case that physical proximity is a key ingredient in the innovative process.

Steve Jobs understood this when he helped to design the layout of Pixar Animation Studios. The original plan called for three buildings, with separate offices for animators, scientists, and executives. Jobs chose instead a single building, with a vast atrium at its core. To ensure that animators, scientists, and executives frequently interacted and exchanged ideas, Jobs moved the mailboxes, the cafeteria, and the meeting rooms to the center of the building. And Mervin Kelly, who, for a time, ran Bell Labs (AT&T's R&D lab), was "convinced that physical proximity was everything."[1] Kelly participated in the design of a building that opened in 1941 "where everyone would interact with one another." Hallways were designed to be so long that when walking a hall's length one would encounter "a number of acquaintances, problems, diversions and ideas. A physicist on his way to lunch in the cafeteria was like a magnet rolling past iron filings." Within this unique culture, Bell Labs' employees developed some of the most important inventions of the 20th century, including the transistor, the laser, and the solar cell.

These are two examples of individual companies taking conscious steps to maximize spatially mediated knowledge spillovers within their own organizations. But there is also highly suggestive evidence that these spillovers also occur across organizations: All else equal, companies tend to locate R&D establishments near the R&D establishments of other firms. This pattern of "clustering," illustrated for about 1,000 private R&D labs located in the Northeast corridor of the United States in Figure 1, is established rigorously in this paper.[2] Earlier research (e.g., Jaffe, Trajtenberg, and Henderson (1993) — hereafter JTH) documents a pattern of spatial concentration (often described as localization) in patent citations: all else equal, patents are more likely to cite earlier patents from inventors located nearby than ones obtained by inventors living farther away. Such patterns are consistent with the hypothesis that there are important knowledge spillovers that tend to decrease with distance. In this paper we establish that, all else equal, the spatial concentration of patent citations is higher among patents obtained inside a research cluster than similar patents obtained by inventors living outside such a cluster.

A number of previous papers have used the Ellison and Glaeser (1997) — hereafter EG — concentration index to measure the clustering of manufacturing employment at the zip code, county, MSA, and state levels of geography. Rather than using fixed geographic units, such as

[1] Jon Gertner, "True Innovation," *New York Times*, February 25, 2012.

[2] While it is not immediately clear from the figure that the spatial concentration of R&D is significantly greater than manufacturing activity in general, this fact is established in Buzard and Carlino (2011).

counties or metropolitan areas, we use continuous measures to delineate the spatial structure of the concentrations of R&D labs. Specifically, we use Ripley's (1976) *K*-function methods to analyze locational patterns over a range of selected spatial scales (e.g., within a quarter mile, 1 mile, 5 miles, etc.). This approach allows us to consider the spatial extent of the agglomeration of R&D labs as well as how rapidly the clustering of labs attenuates with distance. Following Duranton and Overman (2005) — hereafter DO — and Ellison, Glaeser, and Kerr (2010), we look for geographic clusters of labs that represent statistically significant departures from spatial randomness using simulation techniques. Specifically, "randomness" in this case is not taken to mean a uniform distribution of R&D activity. Rather, since we are primarily interested in R&D concentration not explainable by manufacturing alone, we focus on departures from the distribution of manufacturing employment.

In the first phase of the analysis, we employ global *K*-function statistics to test for the presence of significant clustering over a range of scales. There are two important findings from this global analysis. First, the clustering of labs exhibits a significant peak at very small spatial scales, such as distances of about one-quarter of a mile. Second, we find that the significance of clustering dissipates rapidly with distance. This rapid attenuation of significant clustering at small spatial scales is consistent with the view that knowledge spillovers are highly localized. The finding of rapid attenuation of significant clustering is particularly important since, among the Marshallian externalities that have motivated the literature on agglomeration economies, knowledge spillovers have proven to be the hardest to verify empirically. Rosenthal and Strange (2001) find that proxies for knowledge spillovers positively affect EG concentration measures but only at zip code levels. Rosenthal and Strange (2008) introduce spatial decay into the estimation of agglomeration externalities, but they assume no attenuation within the first mile. Arzaghi and Henderson (2008) show that for ad agencies in New York City, information spillovers attenuate very rapidly, within several blocks.

If knowledge spillovers operate, we would expect them to be important in location decisions of knowledge-based activities such as R&D. Importantly, our finding that the most significant localization of R&D labs occurs within a two- to three-block radius and attenuates rapidly thereafter is consistent with the mounting evidence for the attenuation of human capital spillovers at small spatial scales.

We also observe a secondary mode of significance at a scale of about 40 miles. This will be seen to correspond roughly to the scales of the four major R&D agglomerations identified in the second phase of our analysis — one each in Boston, New York-Northern New Jersey, Philadelphia-Wilmington, and Virginia, including the District of Columbia (hereafter referred to as Washington, DC). The scale of this clustering is roughly comparable to that of labor markets and hence is consistent with the view that agglomeration economies at the level of labor markets (e.g., externalities associated with pooling and matching) are important for innovative activity (see, for example, Carlino et al., 2007).

Given the strong clustering found at small scales, the question remains as to where this clustering occurs. In the second stage of the analysis, explicit clusters are identified by a new procedure based on local *K*-functions, which we designate as the *multiscale core-cluster* approach. This new approach yields a natural nesting of clusters at different spatial scales. In particular, *core clusters* are identified at each scale containing those points involved in the most significant clustering at that scale. By construction, core clusters at smaller scales tend to be nested in those at larger scales. Such core clusters thus yield a hierarchy that can serve to reveal the relative

spatial 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 four major agglomerations mentioned above. As a consistency check, these results are replicated using the significance-maximizing procedures developed by Besag and Newell (1991) and Kulldorff (1997).

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 is both a higher bar and avoids a potential measurement issue at very small spatial scales that may occur when we use manufacturing employment as our baseline. We find at small spatial scales (such as within a two- to three-block area) that 37 percent of the industries studied are significantly more concentrated compared with overall R&D labs, and none are significantly more dispersed. The rapid attenuation of significant clustering of labs for many individual industries bolsters our view that at least one important component of knowledge spillovers must be highly localized.

Finally, using patent data, we are able to provide evidence consistent with the hypothesis that knowledge spillovers are highly localized within the clusters of R&D labs we identify. Patents contain information about the location of inventors as well as citations to prior patents on which they are built. As with citations to academic articles, we interpret patent citations as tangible evidence of knowledge spillovers.[3] If the clustering patterns we identify are motivated, at least in part, by spillovers that attenuate with distance, we should expect to find a comparable clustering of citations. In other words, we should expect to see that citations of patents generated within a cluster should come disproportionately from previous patents generated in that same cluster. We find that citations are a little over four times more likely to come from the same cluster as earlier patents than one would predict using a (control) sample of otherwise similar patents.

To place these results in perspective, we begin in the next section with a review of the relevant literature. This is followed in Section 3 with a brief discussion of data sources. The statistical methodology and test results for the global analyses of spatial clustering are developed in Section 4, while the local analyses of clustering are discussed in Section 5. In Section 6 we introduce a new approach (multiscale core-cluster approach) for indentifying explicit R&D clusters. In Section 7, we provide a detailed discussion of the internal spatial structure of the four major R&D agglomerations identified by our analysis. In Section 8, we show that citations of patents generated within a cluster come disproportionately from within the same cluster as previous patents. We conclude in Section 9.

2. LITERATURE REVIEW

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. EG extended the spatial Gini coefficient to condition not only on the location of manufacturing employment but to also on industrial structure.

[3] The number of citations a patent receives is also correlated with the estimated value of the patent. See Harhoff et al. (1999).

A number of recent studies have used the EG index to measure the 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, Glaeser, and Kerr, 2010). Holmes and Stevens (2004) take a broader approach and use employment data for all U.S. industries, not just manufacturing, and find that among the 15 most concentrated industries, six are in mining and seven are in manufacturing; only two industries fall outside mining and manufacturing (casino hotels and motion picture and video distribution).

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

To address these issues, DO used micro data to identify the postal codes for each manufacturing plant in the UK, thus allowing these data to be geocoded. Geocoding is important, since DO are not bound by a fixed geographical classification but base their approach on the actual distance between firms. Additionally, rather than using a specific index to measure geographic concentration, such as the EG index, DO take a nonparametric approach (based on kernel densities). Essentially, DO construct frequency distributions of the pair-wise distances between plants in a given industry. When the mass of the distribution is concentrated at short distances, this represents a spatial concentration of plants in the industry. Alternatively, if the mass of the distribution is concentrated at longer distances, this represents a more dispersed spatial pattern. Importantly, DO consider whether the number of plants at a given distance is *significantly* different from the number that would have been found if their locations were randomly chosen.

A few other studies have used continuous measures of concentration. In addition to considering a discrete measure of coagglomeration (measured at the state, MSA, and county levels), Ellison, Glaeser, and Kerr (2010) follow DO and also consider more spatially continuous measures of coagglomeration. Marcon and Puech (2003) use distance-based methods to evaluate the spatial concentration of French manufacturing firms and find that some industries are concentrated, while other industries are dispersed. Arbia, Espa, and Quah (2008) use a *K*-function approach to study the spatial distribution of patents in Italy during the 1990s. Kerr and Kominers (2010) develop a model where the costs of interaction among agents define the distance over which forces for agglomeration of activity operate. In one application, Kerr and Kominers (2010) use data on patent citations and show that technologies with short distances over which agents interact are characterized by smaller and denser concentrations relative to technologies allowing for interactions over longer distances. In another application, Murata, et al. (2011) apply a continuous approach to test for the localization of knowledge spillovers using U.S. patent data. Using tests introduced by JTH, Murata, et al. (2011) find evidence supporting the localization of knowledge spillovers.

Our work differs from past studies in a number of ways. Rather than looking at the geographic concentration of firms engaged in the production of goods (such as manufacturing), we use a

new location-based data set that allows us to consider the spatial concentration of private R&D establishments. Rather than focusing on the overall concentration of R&D *employment*, we analyze the clustering of individual R&D *labs*.[4] 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. 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. One advantage of *K*-functions is that they can easily be disaggregated to yield information about the *spatial locations* of clusters at various scales. Our tests for the localization of R&D labs also control for industrial concentration and, in particular, the concentration of manufacturing employment.[5] Finally, in addition to these cluster-identification results, we show that patents from these clusters generate citations that are more localized than are patent citations in general.

3. DATA

Our primary data source is the 1998 vintage of the *Directory of American Research and Technology*. Using the complete address information for each R&D establishment, we were able to geocode the locations of more than 3000 labs. For this paper, we limited the analysis to 1,035 R&D labs in ten states comprising the Northeast corridor of the United States (Connecticut, Delaware, Maryland, Massachusetts, New Hampshire, New York, New Jersey, Pennsylvania, Rhode Island, and Virginia, including the District of Columbia — the Washington, DC cluster). These labs are plotted in Figure 2.[6] Since there are approximately 6,043 zip codes in these states, there is on average one R&D facility for every six zip codes in this part of the country.

Even at the most aggregate level, it is easy to establish that R&D activity is relatively concentrated in these ten states. For example, in 1998 one-third of private R&D labs (and 32 percent of private R&D expenditures) were located within this region, as compared with 22 percent of total employment (21 percent of manufacturing employment) and 23 percent of the population. This concentration is consistent with Audretsch and Feldman (1996), who report that three of the top four states in terms of innovation in their data include Massachusetts, New Jersey, and New York.

In our formal analysis below, the concentration of R&D establishments is measured relative to a baseline of economic activity as reflected by the amount of manufacturing employment in the zip code, as reported in the 1998 vintage of Zip Code Business Patterns. These data are plotted in Figure 3. Since our main objective is to describe the localization of total R&D labs, manufacturing employment represents a good benchmark, since the vast majority of our R&D

[4] The study by Guimarães, Figueiredo, and Woodward (2007) is the only other study we are aware of that looks at spatial clustering at the establishment level. Specifically, they look at the geographic concentration of over 45,000 plants in 1999 for concelhos (counties) in Portugal.

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

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

labs are owned by manufacturing firms.[7] Since R&D labs may choose locations for different reasons than those of manufacturing establishments, later in the paper, we also examine the concentration of labs conducting R&D in *specific* industries, as compared to the locations of *all* R&D labs.

For the analysis in Section 8 of this paper, we use patent and citation data obtained from the NBER Patent Data Project.[8] We use data for patents granted in the years 1996-2006. In particular, we are interested in the geographic distribution of citations to patents obtained by inventors living within one of the R&D clusters we identify in Section 6 of the paper. As with journal articles, patent documents often include citations to earlier patents that are somehow related to the current invention. We follow the previous literature in using the home address of the first inventor on the patent to locate the patent in space. We obtained the specific coordinates for the patents we used from the Patent Dataverse.[9]

4. GLOBAL CLUSTER ANALYSIS

The key question of interest is whether the overall pattern of R&D locations in the ten states we examine exhibits more clustering than would be expected from the spatial concentration of manufacturing in those states. To address this question statistically, our null hypothesis is that R&D locations are determined entirely by the distribution of manufacturing employment:

H_0: *The probability of finding a randomly selected R&D lab in any given area is proportional to manufacturing employment in that area*.

Although we do not have employment data for arbitrary areas, our zip code geography for the Northeast corridor should be sufficiently disaggregated to provide reasonable approximations for the purposes of our global cluster analysis (as unions of zip code areas).[10]

A simple two-stage Monte Carlo procedure for generating locations consistent with our null hypothesis is to randomly draw a zip code with a probability that is proportional to manufacturing employment in that zip code, relative to manufacturing jobs in all zip codes in our data, and then to choose a random location within that zip code. By repeating this procedure for a set $n = 1035$ location choices, one generates a pattern, $X = (x_i = (r_i, s_i) : i = 1,..,n)$, of potential R&D locations that is consistent with H_0, where (r_i, s_i) represents the latitude and longitude coordinates (in decimal degrees) at point *i*. This process is repeated many times for each R&D location in the data set. In this way, we can test whether the *observed point*

[7] There are two notable exceptions in our labs data: electronics wholesaling (which includes firms such as Apple computers) and software. As a robustness check, we ran many of our tests using total employment as a back cloth and found comparable results.

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

[9] Specifically, we used the location information contained in the file inventors5s_9608.tab downloaded from http://dvn.iq.harvard.edu/dvn/dv/patent. For details, see Lai, D'Amour, and Fleming (2009).

[10] The median area of zip codes in our data set is 17 square miles, which corresponds roughly to a radius of 2.3 miles.

pattern, $X^0 = (x_i^0 = (r_i^0, s_i^0) : i = 1,..,n)$, of R&D locations is "more clustered" than would be expected if the pattern were generated randomly (i.e., randomly drawn from the manufacturing employment distribution).

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

4.1 *K*-Functions

The most popular measure of clustering for point processes is Ripley's (1976) *K*-function, $K(d)$,[11] which (for any given mean density of points) is essentially the expected number of additional points within distance *d* of any given point. Hence if $K(d)$ is higher than would be expected under H_0, then this may be taken to imply *clustering* of R&D locations relative to manufacturing at a spatial scale *d.*

For testing purposes, it is sufficient to consider sample estimates of $K(d)$. If for any given point *i* in pattern $X = (x_i : i = 1,..,n)$, we denote the number (count) of additional points in *X* within distance *d* of *i* by $C_i(d)$, then the desired *sample estimate*, $\hat{K}(d)$, is given simply by the average of these point counts, i.e., by [12]

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

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

If one simulates a number of point patterns, $X^s = (x_i^s : i = 1,..,n)$, $s = 1,..,N$, by the above procedure, and for a selection of radial distances, $D = (d_1,..,d_k)$, constructs the corresponding sample *K*-functions, $\{\hat{K}^s(d) : d \in D\}$, $s = 1,..,N$, then at each *scale*, $d \in D$, these values yield an approximate sampling distribution of $K(d)$ under H_0. Hence if the corresponding value, $\hat{K}^0(d)$, for the observed point pattern, X^0, of R&D locations is sufficiently large relative to this distribution, then this can be taken to imply significant clustering relative to manufacturing. More precisely if the value $\hat{K}^0(d)$ is treated as one additional sample under H_0, and if the

[11] The term "function" refers to the fact that $K(d)$ is in principle defined for all $d \geq 0$.

[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.

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

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

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

For example, if $N = 999$ and $N^0(d) = 10$, so that $P(d) = 0.01$ then under H_0, there is estimated to be only a one-in-a-hundred chance of observing a value as large as $\hat{K}^0(d)$. Thus, at spatial scale *d* there is significant clustering of R&D locations at the 0.01 percent level of statistical significance.

4.2 Test Results for Global Clustering

Our Monte Carlo test for clustering was carried out with $N = 999$ simulations at radial distances, $d \in D = \{0.25, 0.5, 0.75, 1, 2, ..., 99, 100\}$, i.e., at quarter-mile increments below 1 mile and at 1-mile increments from 1 to 100 miles. We find that clustering is so strong, relative to manufacturing employment, that the estimated *p*-values were 0.001 for all spatial scales we considered. Thus, our conjecture that private R&D activities exhibit significant agglomeration is extremely well supported by the data.

Note that, using our approach, the smallest possible *p*-value that can be generated is $1/(N+1)$. For $N = 999$, the smallest possible *p*-value is then 0.001, which suggests that we may be underestimating the statistical significance of our results. Of course, we could increase the number of draws, but we chose $N = 999$ because it was sufficiently large to obtain reliable estimates of the sampling distributions under H_0. Analysis of these distributions—both in terms of Shapiro-Wilk (1965) normality tests and normal quintile plots (not shown)—indicate that they were well approximated by the normal distribution for all the spatial scales we tested.

4.3 Variation in Global Clustering by Spatial Scale

To obtain a sharper discrimination between results at different spatial scales, we calculate the *z*-scores for each observed estimate, $\hat{K}^0(d)$, as given by

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

where $\bar{K}_d$ and s_d are the corresponding sample means and standard deviations for the $N+1$ sample *K*-values. These *z*-scores are depicted in Figure 4a. Notice first that the lowest *z*-score is already more than seven standard deviations away from the mean, which explains the constancy of *p*-values reported above.

A key finding from the global *K*-function analysis is that the overall clustering of R&D labs is by far most significant (based on *z*-scores) at very small spatial scales, such as distances of one-

quarter mile. While still highly significant, the *z*-scores decline rapidly up to a spatial scale of about 5 miles. We also observe a secondary mode of significant clustering for the totality of all labs at about 40 miles, as shown in Figure 4a. In terms of standard deviations, this is about half as pronounced as the primary mode.

This pattern of *z*-scores is consistent with two strands of empirical research on human capital spillovers and agglomeration economies. For example, there are a number of papers that establish very rapid attenuation of effects with distance in studies of the concentration of manufacturing employment (Rosenthal and Strange, 2001 and 2008, and Elvery and Sveikauskas, 2010), of innovative activity (Audretsch and Feldman, 1996; Keller, 2002; and Agrawal, Kapur, and McHale, 2008); and of locations of advertising firms in New York City (Arzaghi and Henderson, 2008). Other studies find evidence of positive effects of agglomeration at much greater distances (Rosenthal and Strange, 2008, and Elvery and Sveikauskas, 2010). Carlino et al. (2007) establish robust correlations between patent intensity (patents per capita) and job density (jobs per square mile) for 280 U.S. cities in the 1990s. Such patterns are consistent with models of labor market search that exhibit matching externalities (Berliant et al., 2006, and Hunt, 2007).

4.4 Precision at Very Small Spatial Scales

The global *K*-function approach is suitable for detecting non-random concentrations over a wide spectrum of spatial scales. For our particular application of this technique, however, we need to address a potential confounding factor with our data. The concern is that while we have identified the precise locations of R&D labs, our data on employment are a sum of jobs for each zip code in the data set. Thus, we are implicitly assuming that jobs are distributed uniformly across a zip code. This creates the *possibility* of potential bias at spatial scales that are much smaller than a zip code.[13]

For a number of reasons, we don't believe there is much, if any, bias in our significance measures at very small spatial scales. First, as discussed in the next section, we have constructed alternative tests for concentration for which the underlying data do not vary in terms of their spatial precision. In the next section, we test for the concentration of R&D labs in specific industries relative to the locations of all R&D labs in our data set. As we found in Figure 4a, for practically all industries, our measures of statistical significance are higher at very small spatial scales (e.g., a quarter of a mile) than they are at intermediate or even larger scales.

To be conservative, however, we constructed a second counterfactual exercise designed to sweep out any possible effect of differences in spatial precision in our global *K* analysis. In this alternative formulation, rather than using the actual location of R&D labs, we assign each lab randomly to a point in the zip code where they are located. Thus, the pseudo location of R&D labs has the same spatial precision as our zip-code-level employment data.[14] Figure 4b shows the *z*-scores for this alternative exercise. Comparing Figures 4a and 4b, it is clear the only material differences in *z*-scores occur at distances of about 3 miles or less. The plots are otherwise the same.

[13] This potential source of bias was first pointed out to us by Gilles Duranton and Isabel Tecu.

[14] We thank Gilles Duranton for suggesting the alternative counterfactual exercise.

We conclude first that any differences in the spatial precision of our R&D lab data and the employment data have no effect on the significance measures for spatial scales of 3 miles or more. Second, even using this alternative specification, it is clear that the global *K* statistic remains highly significant at very small spatial scales. Thus, the evidence for clustering at even very small distances is not an artifact of measurement error. Finally, it is almost certainly the case that the *z*-scores reported for smaller spatial scales depicted in Figure 4b are biased downward, since we have deliberately introduced additional noise into our information on the location of R&D labs.[15] This appears to be confirmed by the pattern of *z*-scores for measures of the relative concentration of R&D labs described in the next section.

4.5 Relative Clustering of R&D Labs by Industry

We believe that the distribution of manufacturing jobs is a reasonable, relatively 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. One example might include partial public funding of technology parks.

In this section, we modify our null hypothesis: we assume 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. The limitation of this approach is that we cannot say anything about the clustering of R&D labs in general. But the benefit is two-fold. First, we can incorporate into our null hypothesis factors that are likely to influence the location of R&D in general. Second, we can assess whether specific industries exhibit more spatial concentration of their R&D than for all R&D labs taken together. Note that, here, we are constructing a test of *relative* spatial concentration, since we have already established that R&D facilities are significantly more concentrated in space than manufacturing activity in general.

To accomplish this, we grouped labs in terms of their primary industrial research areas at the two-digit SIC level.[16] We apply a variant of the global *K*-function procedure by taking random draws of the count of R&D labs from the full population of 1,035 labs.[17] Table 1 reports the *p*-values for each of the 19 two-digit SIC industries for selected distances. We find that at a distance of a quarter-mile, seven of these 19 industries (37 percent) are significantly more

[15] To put it another way, in order to eliminate the possibility of a false positive result, this alternative approach increases the likelihood of false negative results.

[16] This assignment is based on information contained in the Bowker Directory. The two-digit level is used to achieve sufficient sample sizes for testing purposes. This yields 19 industrial groups with corresponding SIC designations: 10, 13, 20-23, 26-30, 32-39, and 73. It also reduces the likelihood that the presence of outliers, in terms of industry specialization, might also lead to false positives. Consider, for example, a company developing advanced, large caliber cannon, which may require a proving ground isolated from other activities.

[17] In particular, this identification procedure is carried out (in a manner similar to DO) in terms of standard random-permutation tests based on global *K*-function statistics.

localized (at the 0.05 percent level) than are R&D labs in general[18] None are significantly more dispersed.[19]

The *z*-scores for the seven industries with the most significant patterns of clustering are displayed graphically in Figure 5. Because we are especially interested in the attenuation of *z*-scores at small scales, these *z*-scores are given in increments of 0.25 miles up to 5 miles. For all but one of these industries, the clustering of R&D labs is by far most significant at very small spatial scales — a quarter mile or less. The lone exception is Miscellaneous Manufacturing Industries (SIC 39), where the highest *z*-score occurs at a distance of just under 2 miles.

In addition, Figure 5 reveals a very rapid distance decay of the *z*-scores for each of the seven industries. The rapid spatial attenuation of *z*-scores supports our view that at least one important component of knowledge spillovers in these industries is highly localized. For most of these industries there is nearly a monotonic decline in *z*-scores as spatial scale increases. In four instances, at distances above 3 miles, the industry's R&D labs are no more concentrated spatially than R&D labs in general. Two exceptions do stand out—Chemicals and Business Services: labs in these industries are also spatially concentrated, relative to all R&D labs, at much larger spatial scales. Note that, in our data, all but one of the R&D labs in the Business Services category are associated with firms engaged in computer programming or data processing.

The results for the chemical and allied products industry (SIC 28) merit some additional discussion, if for no other reason than that 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, a share more than twice as large as any other two-digit SIC industry. Thus, at least within the geographic area we study in this paper, this industry is a major contributor to the overall clustering pattern of R&D shown in Figure 4a. Nevertheless, as Figure 5 demonstrates, evidence of clustering occurs in many other industries. In other words, the clustering of R&D labs is not a phenomenon specific to drugs and chemicals.

5. LOCAL CLUSTER ANALYSES

The global analysis documents that R&D facilities in these ten states are indeed clustered at a variety of spatial scales. In this section we a use variation of our techniques to identify specific R&D clusters and the labs that belong to them. The main tool for accomplishing these tasks is the *local* version of sample *K*-functions for individual pattern points (first introduced by Getis, 1984).[20] Basically, this local version at i is simply the count of all additional pattern points

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

[19] With respect to dispersion, two of the 19 industries are found to be significantly more disbursed starting at a distance of 5 miles, while an additional industry exhibits some degree of relative dispersion at 50 miles.

[20] 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 *i.* Hence $\hat{K}_i(d)$ is basically a (maximum likelihood) estimate of size one for $K_i(d)$. For a range of alternative measures of local spatial association, see Anselin (1995).

within distance *d* of i. In terms of the notation in expression (1) above, the *local K-function,* $\hat{K}_i$, at location i is given for each distance, *d*, by,[21]

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

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

5.1 Local Testing Procedure

For the remainder of the paper, we use the same null hypothesis employed in Section 4.1 (R&D labs are distributed in a manor proportional to the distribution of manufacturing employment). The only substantive difference from the procedure used in that section is that the actual point pattern associated with location *i*, x_i, is held fixed. The appropriate simulated values, $\hat{K}_i^s(d),\ s = 1,..,N$, under H_0 are obtained by generating point patterns, $X^s = (x_j^s : j = 1,..,n-1)$, of size $n-1$, representing all points other than i. The resulting *p*-values for a one-sided test of H_0 with respect to point i take the form,

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

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

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

5.2 Test Results for Local Clustering

For our local cluster analysis, simulations were performed using $N = 999$ test patterns of size $n-1$ for each of the $n\ (=1035)$ R&D locations in observed pattern X^0. The set of radial distances (in miles) used for the local tests was $D = \{0.5, 0.75, 1, 2, 5, 10, 11, 12.., 100\}$.

In our global analysis, the associated *p*-values were essentially the same for nearly all spatial scales. That is not the case for the local analysis. It is not surprising to find that many isolated R&D locations exhibit no local clustering whatsoever, so that wide variations in significance levels are possible at any given spatial scale. It is also natural to expect variations in tests of statistical significance at different spatial scales. At very small scales (say, less than one-quarter

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

of a mile), one expects to find a wide scattering of very small clusters, such as industrial parks that include more than one R&D lab. At the other extreme (say, 100 miles) one expects to find very large clusters, based mostly on the strong overlap of *K*-function areas around each location. From a visual perspective, at least, the most interesting scales are those intermediate scales at which one begins to see more "coherent" clusters.

A visual inspection of *p*-value maps for our tests shows that the clearest patterns of distinct clustering can be captured by the smaller set of distances {0.5,1,5,10}. Of these four, the single best distance for revealing the overall clustering pattern in the entire data set appears to be 5 miles, as illustrated in Figure 6.[22] For clarity, we have shown only three levels of *p*-values. 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, which can be roughly described (from north to south) as the "Boston," "New York City," "Philadelphia," and "Washington DC" agglomerations.[23] But while these patterns are visually compelling, it is important to establish the results more formally.

6. IDENTIFYING CLUSTERS USING ROBUST METHODS

The global cluster analysis in Section 4 identified the *scales* at which clustering is most significant (relative to manufacturing employment). The local cluster analysis in Section 5.1 provided information about *where* clustering is most significant at each spatial scale. But neither of these methods formally identifies or defines "clusters," the combinations of specific labs that belong in a set of labs subject to mutual influence by other members of the set. In this section, we apply some additional techniques to identify clusters. In the process of doing so, we will address some econometric issues that could potentially contaminate these and our earlier results.

6.1 The Multiple-Testing Problem

Our method of identifying clusters is, by construction, a local cluster analysis. Because we are testing over multiple locations (some nearby) and spatial scales (some quite large), we must address two aspects of a "multiple testing" problem.[24]

Suppose there was in fact no local clustering of R&D labs (so that the observed pattern X^0 of R&D locations could not be distinguished statistically from the patterns generated under our null hypothesis). Suppose also that all local *K*-function tests were statistically independent from each other. Then, by construction, we should still expect 5 percent of our resulting test statistics to be statistically significant at the 0.05 percent level. So when many such tests are involved (in our case, 1,035 tests at each scale, $d \in D$), one is bound to find some degree of "significant clustering" using standard testing procedures. As is well known, this type of "false positive rate"

[22] We use the results for the entire set of distances in the robustness sections that follow.

[23] The one exception here is a small but significant agglomeration in Pittsburgh.

[24] A global cluster analysis, conducted over many spatial scales, may also suffer from this problem, but the problem is made worse for the local cluster technique, which we address in the text.

can be mitigated by reducing the *p*-value threshold level deemed to be "significant." That is one reason why we focus only on *p*-values no greater than 0.005 in Figure 5.

This adjustment alone is not sufficient in instances where the assumption of statistical independence of the tests is also violated. This is a likely possibility when our statistics for detecting local clustering are calculated over radial distances that are larger than half the distance between any two points for which the statistic is being calculated. The resulting *p*-value map must necessarily exhibit some degree of (positive) spatial autocorrelation, much in the same way that kernel smoothing of spatial data induces autocorrelation.[25]

6.2 The Significance-Maximizing Approach

A number of econometric approaches have been developed for resolving multiple testing problems in spatial applications. Perhaps the best known are the original work of Besag and Newell (1991) and the more recent work of Kulldorff (1997). Both approaches resolve the multiple-testing problem by conducting only a *single* test.

In the present setting, one focuses on zip code areas (cells) and replaces individual locations with counts of R&D labs in each area (cell counts). Using *centroid* distance between cells, candidate clusters are then defined as unions of *m*-nearest neighbors to given "seed" cells, and a test statistic is constructed to determine the single *most significant* cluster. In both of these *significance-maximizing* procedures, the notion of "significance" is essentially defined with respect to tests based on the same null hypothesis, H_0, above.[26] To determine a *second* most significant cluster, the zip code areas in the most significant cluster are removed, and the same procedure is then applied to the remaining zip code areas. This procedure is typically repeated until some significance threshold (such as a *p*-value exceeding 0.05) is reached.

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

We have applied both the Besag-Newell procedure and Kulldorff's SATSCAN procedure to our data and found them to be in remarkably good agreement with each other. Thus, we present only the results of the (more popular) SATSCAN procedure. In this setting, we ran the maximum of 10 iterations allowed by the SATSCAN software,[27] and the results from the union of these

---

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

[26] In our present setting, the Besag-Newell (1991) procedure directly uses H_0 to define a nonhomogeneous Poisson process of R&D frequency counts in each zip code area. The appropriate test statistic is then simply the observed total count in each candidate cluster. The SATSCAN procedure of Kulldorff (1997) uses a more complex likelihood-ratio statistic (under H_0) for each candidate cluster and then employs essentially the same simulation procedure as in Section 4.1 above to simulate the sampling distribution of this statistic under H_0.

[27] This software is available online at http://www.satscan.org/download_satscan.html.

iterations are plotted in Figure 7. Comparing this figure to Figure 6 (derived using our local *K*-function), it is evident that both procedures are identifying essentially the same areas.

Turning next to the specific clusters identified by SATSCAN, we start with the single most significant cluster found in "stage 1" of the procedure, as shown by the darkened set of zip code areas in Figure 8 (where the slightly darker zip code in the center is the starting seed). This cluster is essentially the "Boston cluster," referred to in Figure 6 above. For purposes of comparison, the Boston area of Figure 6 has been superimposed on Figure 8 to show that the two most statistically significant groupings of R&D Labs (based on the local *K*-Function analysis) in the Boston area are essentially contained in this cluster. Again, there appears to be a reasonable correspondence between the results reported in Section 5 and those found here.

Still, the patterns presented in Figure 6 naturally raise the question as to why two distinct groupings of labs identified in the local *K*-function analysis should constitute a *single* cluster as identified by the SATSCAN procedure. It is due to the approximately *circular* shapes of candidate clusters defined by this particular implementation of the procedure.[28] In particular, no circular approximation to either of these two groupings is more significant than the single circular cluster shown.

An even more dramatic example is provided by the single largest cluster in the New York area, just north of New York City in Figure 7, which is shown enlarged in Figure 9 (again the relevant portion of Figure 6 has been superimposed). Here it is evident that *all* significant concentrations of R&D labs (at scale $d = 5$ miles) lie along the southern edge of this cluster. While there is a smaller concentration of labs in the east central portion, it is clear that attempts to capture these concentrations by circular shapes may have distorted the identification of the actual cluster. [29]

In addition to this shape limitation, the sequential nature of cluster identification in these procedures introduces other types of "path-dependence" problems. In particular, the removal of clusters identified at each stage necessarily modifies the neighborhood relations among the remaining zip codes at later stages. So at a minimum, these modifications require careful "conditional" interpretations of all clusters beyond the first cluster.[30]

To conclude, tests using both the Newell procedure (not shown) and Kulldorff's SATSCAN procedure are generally consistent with the results found in our local *K*-function analysis. This suggests that the results reported in Section 5 are not attributable to the kinds of multiple testing problems outlined above. Nevertheless, we can improve upon our implementation of the SATSCAN procedure for the purposes of formally identifying clusters. We accomplish this in the next section.

6.3 A Multiscale Core-Cluster Approach

[28] Our particular model corresponds to the "continuous Poisson" option in the SATSCAN software, for which neighborhoods are required to be "circular" (as defined by a seed area together with its first *m*-neighbors).

[29] It should be noted that the option of using more general "elliptical" clusters is available in certain SATSCAN modeling options other than the "continuous Poisson" option used here.

[30] Methods for addressing such path dependencies have been developed, but they require global optimization, which can be intractable for some applications. See Mori and Smith (2009).

It is useful to consider an alternative approach to cluster identification that explicitly uses the *multiscale* nature of local *K*-functions. This procedure starts with the results of the local point-wise clustering procedure in Section 5.2 and seeks to identify subsets of points that can serve as "core" cluster points at a given selection of spatial scales, d. Here we focus on the three scales, $d \in \{1,5,10\}$, that appear to capture the essential substructure of the four main clusters in Figure 6. In most of the discussion below, we focus on the 5-mile scale for purposes of illustration and consider scales 1 and 10 only when substantive comparisons between the scales are made.

At each scale, d, a *core point* is an R&D lab with an associated *p*-value of 0.001 or lower, derived in the local *K*-function analysis using the 999 simulations described in Section 5.1.[31] 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 created a *d*-mile-radius buffer around each core point (in ArcMap) and identified the sets of points in each *connected component* of these buffer zones as a *core cluster* of points at level *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 level *d*.[32]

The advantages of this core-cluster approach are best illustrated by examples. We begin with the single most significant cluster identified by SATSCAN—the Boston cluster shown in Figure 8. We noted earlier that the local *K*-function analysis produced two distinct concentrations of R&D labs within a single cluster as identified using the SATSCAN procedure. The corresponding results for the multiscale approach are shown in Figure 10. The core points for the spatial scales $d = 1,5,10$ are plotted along with their corresponding core clusters.

For example, at the 5-mile scale we see that there are indeed two core clusters, defined by all of the labs inside each of the dark gray buffer zones (with corresponding core points also shown in dark gray). However, when the scale is expanded to 10 miles, these two clusters merge into a single core cluster that is roughly comparable to the SATSCAN cluster in Figure 6, but which now contains precisely those labs that contribute to the significance of at least one core point at this scale.

Conversely, when the scale is reduced to 1 mile, a richer picture of local concentration emerges. Here, the largest core cluster at the 5-mile scale is now seen to contain six individual 1-mile core clusters, while the smaller core cluster at 5 miles contains only a single 1-mile core cluster. Note finally that while such clusters tend to be nested by scale, this is not always the case. In particular, there is a conspicuous 1-mile core cluster near the bottom of the figure that is not contained in any 5-mile core cluster. There happens to be a concentration of five R&D labs in close proximity that are relatively isolated from the other labs. So while this concentration is

[31] The use of 999 simulations was designed to maintain comparability with the SATSCAN results, where 999 is the maximum allowable number of simulations. As a check, we also ran the local cluster simulations in Section 5.1 with 9,999 simulations. The core points identified from this exercise were, with a few minor exceptions, the same as those obtained from the original 999 simulations.

[32] The present definition of "core cluster" is designed to ensure that individual clusters are disjoint sets. Topologically, this requires that each such cluster be generated by sets of core points that are 2*d-path connected,* where a 2*d-path* is a sequence of points in the set with adjacent points no more than a distance of 2*d* apart. In other words, "adjacent" core points on such paths should be capable of sharing at least one *d*-neighbor.

picked up at the 1-mile scale (and in fact at the half-mile scale as well), it is too small by itself to be picked up at the 5-mile scale.

Our second example illustrates one of the strong local concentrations of R&D labs that contribute to the peak of significance for the smallest spatial scales described in Section 4.3 above. Figure 11 plots a cluster of 17 labs just south of Central Park in New York City. The figure shows core points at the quarter-mile and half-mile scale as well as the 1-mile scale. The quarter-mile core cluster of five labs is denoted by the darkest buffer containing four black points (where the lowest of these points contains two labs). This is a particularly strong cluster since *all* labs are within one-quarter mile of each other, and hence all are core points at the quarter-mile scale. The larger 1-mile core cluster is indicated by the dashed buffer. The 1-mile core points are more difficult to show, since they are also half-mile or even quarter-mile core points. To distinguish these, a larger circle has been placed around each of the eight 1-mile core points. All points other than the five white points (labeled "Other Labs") are half-mile core points, with the associated core cluster shown in dark gray. The only one of these that is not either a 1-mile or quarter-mile core point is shown by the single dark gray point (which also contains two labs).

To gain further insight into the differences between these core clusters, the zip codes shown in Figure 11 are shaded to depict the relative number of manufacturing jobs. The darkest one of these has more manufacturing jobs (22,000) than any other zip code in our data set. Notice that the 1-mile core cluster overlaps part of this dense manufacturing area, while the quarter-mile and half-mile core clusters do not. This explains why the half-mile core point closest to this area (the two labs at the dark gray point) as well as the quarter-mile core point closest to this area (the two labs at the lowest black point) are not also core points at the 1-mile scale. It is also of interest to note that this strong concentration of labs was not among the 10 most significant clusters identified by SATSCAN (although it might very well be close to the top 10).[33]

These examples serve to illustrate some of the attractive features of this multiscale core-cluster approach. First and foremost, such representations add a scale dimension not present in other clustering methods. In essence, this approach extends the multiscale feature of local K-functions from individual points to clusters of points. Moreover, individual core-cluster shapes are seen to be more sensitive to the actual configuration of points than those found in the significance-maximizing method. Finally, since all core clusters are determined simultaneously, the problems of "path dependencies" discussed above do not arise.

Still, this multiscale approach is not a substitute for more standard approaches such as significance-maximizing. We can establish statistically significant clusters, but we cannot necessarily rank order clusters in terms of statistical significance. In particular, this method 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). While individual core points can be said to reflect relative (threshold) significance levels, there is no way to assign precise statistical significance to the *core clusters* they generate. Moreover, such representational schemes 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). Hence, the main

[33]This also shows that at micro scales such maximal-significance procedures can be very sensitive to the particular shapes of zip code areas (cells). In this case, the two adjacent zip code areas containing most of these labs happen to be closer to other neighbors (in centroid distance) than they are to each other.

objective of this procedure is to yield *visual* representations of clusters that capture both their relative shapes and concentrations in a natural way. Since there is no universally accepted definition of clusters, it seems prudent to analyze this problem from many viewpoints and look for areas of substantial agreement among them.

7. DESCRIPTION OF SPECIFIC R&D CLUSTERS

In this section we provide a more detailed discussion of the internal spatial structure of the four major agglomerations found at the metropolitan level. In particular, in Section 7.1 we identify the primary research areas associated with individual core clusters of labs. In Section 7.2 we relate these spatial structures to key local geographic features such as proximity to freeways and the presence of university centers. Finally, we briefly compare the spatial structures of those R&D labs with primary research areas in specific industries.

7.1 Major Areas of Agglomeration

Figure 12 plots all the core clusters at spatial scales of $d = 1, 5, 10$ miles. The outer gray contours correspond to core clusters at scale $d = 10$, for example. This map can be compared to the K-function results for $d = 5$ in Figure 6 and the results using SATSCAN plotted in Figure 7.

Reviewing these maps, it is clear that each technique reveals Boston, New York, Philadelphia, and Washington, DC, to be areas of significant spatial concentration in R&D, relative to the underlying pattern of manufacturing activity. The clusters identified using the multiscale approach for $d = 10$ correspond reasonably well to the ones identified via SATSCAN, but they are closer in shape to the pattern of the most significant local *p*-values found for labs using the local *K*-function approach. Given the multiplicity of techniques we have employed, these results seem quite robust.

The Boston Agglomeration

There are 187 R&D labs within Boston's single 10-mile cluster, as shown in Figure 10.[34] Most of these labs conduct R&D in five three-digit SIC code industries — computer programming and data processing, drugs, lab apparatus and analytical equipment, communications equipment, and electronic equipment. The largest 5-mile cluster shown in Figure 10 contains 108 labs, which account for 58 percent of all labs in the larger 10-mile cluster. At the 1-mile scale, Boston has eight clusters, six of which are centered in the largest 5-mile cluster. The largest of these 1-mile clusters contains 30 labs, half of which conduct research on drugs.

The New York City Agglomeration

The single largest cluster identified within our 10-state study area is the 10-mile cluster above New York City (shown in Figure 13) that stretches from Connecticut to New Jersey. This cluster contains a total of 235 R&D labs. Sixty-four (27 percent) of these labs conduct research on drugs, and 37 (16 percent) do research on industrial chemicals. Within this highly elongated 10-

[34] The map legend in Figure 10 applies to all map figures in this section.

mile cluster, three distinct 5-mile clusters were identified. Most of the concentration is seen to occur in the two clusters west of New York City, which in particular contain five of the nine 1-mile clusters identified. Among these 1-mile clusters, the largest is the "Central Park" cluster shown in Figure 11. About two-thirds of the 17 labs in this cluster are conducting research on drugs, perfumes and cosmetics, or computer programming and data processing.

The Philadelphia Agglomeration

As seen in Figure 14, there is a large 10-mile cluster to the west of Philadelphia (where the city of Philadelphia is shown in darker gray), where there are a total of 49 labs. Of these 49 labs, 16 conduct research on drugs, and another 16 do research in the plastics materials and synthetic resins industry. This cluster in turn contains two 5-mile clusters. The most prominent of these is centered in the King of Prussia area directly west of Philadelphia and contains 30 labs, with 40 percent doing research on drugs. The second 5-mile cluster is centered in the city of Wilmington to the southwest. Here, about 25 percent of the labs are also engaged in research on drugs, but most (almost 60 percent) are doing research on plastics materials and synthetic resins.

The Washington, DC, Agglomeration

The final area of concentration is the 10-mile cluster around Washington, DC, which contains 76 R&D labs as shown in Figure 15 (with the city of Washington, DC, in darker gray), where three 5-mile clusters can also be seen. The most prominent of these is directly west of Washington, DC, and contains 37 (almost one-half) of the labs in the larger cluster. Thirty percent of the firms in this 5-mile cluster do research in the areas of computer programming and data processing. In turn, this cluster contains two 1-mile clusters, the largest of which (to the north) contains 16 labs with 44 percent conducting research on drugs. [35]

The Pittsburgh Area

In addition to these four major areas of agglomeration, notice from Figure 12 that there is a smaller agglomeration consisting of two 1-mile core clusters in the Pittsburgh area, one of which is contained in a 5-mile cluster. These are shown enlarged in Figure 16 (with the city of Pittsburgh in darker gray). In the 5-mile cluster (dark gray buffer) there are eight labs, six of which are in its 1-mile sub-cluster (dashed black buffer). Five of these are actually at the same location, denoted by the half-mile cluster (solid black buffer), where the three main areas of research are in plastics materials and synthetic resins, chemicals, and paints and allied products. The 1-mile cluster on the eastern edge of Pittsburgh contains seven labs, with the center three defining the half-mile cluster shown. All but one of these seven labs is conducting research in the areas of laboratory apparatus and analytical, optical, measuring, and control equipment.

7.2 The Importance of Highways and Universities

[35] It is also worth noting that the 5-mile cluster containing these two 1-mile clusters appears to be somewhat questionable in this case. Here, a scale choice of, say, around 4 miles would have produced two distinct clusters that might provide a more appropriate representation of this particular configuration. However, for the sake of comparability across the study area, we have chosen to use a common set of scales throughout.

It is likely that access to both major highways and major research universities is an important determinant of the location and development of innovative activity. This is clearly evident in the four major agglomerations identified here.

Boston Area

A prime example is provided by the locations of R&D labs in the Boston area. As seen in Figure 10, the largest 1-mile cluster (just west of Boston) is centered in Cambridge, home to both Harvard and MIT. The strength of the Boston area's R&D activity has been especially supported by the strength of MIT in electrical engineering, a core discipline for R&D in the computer and electronics industries.

Turning next to Figure 17, observe that Cambridge also has good access to both Interstate 93 (running north to south) and Interstate 90 (running east to west). Similarly, many of the labs in the major 5-mile Boston cluster of Figure 10 are seen in Figure 17 to be located along Route 128 (Interstate 95), which is the inner ring highway around the city. In particular, four of the six 1-mile clusters in this grouping are located along the Route 128 corridor. This corridor also has junctions with Interstate 93 and Interstate 90. Further to the west of Route 128, the smaller 5-mile cluster in Figure 10 is seen to be centered precisely on the intersection of Interstate 90, with the outer circumferential highway being Interstate 495.

New York Area

Given its size, the New York area is by far the most complex. But here again, both the shapes and locations of core clusters are heavily influenced by major highways. In particular, the main 5-mile cluster west of New York City shown in Figure 13 is seen from Figure 18 to be nested within the triangle of Interstates 78, 287, and 80 (also 280) and is most concentrated in Morristown, just south of the 287-80 intersection. Even more dramatic is the elongated shape of the northern 5-mile cluster stretching along Interstate 87. As for universities, the 5-mile cluster southwest of New York City is clearly concentrated around Princeton University, which is active in all areas of research. Finally, the strong "Central Park" cluster in Manhattan is, of course, in close proximity to a host of research universities, including both Columbia and New York University.

Philadelphia Area

Another example of the importance of highways, and especially locations close to the junction of two major highways, is seen by comparing the Philadelphia core clusters in Figure 14 with the major routes shown in Figure 19. Notice first that the major 5-mile cluster (west of Philadelphia) essentially follows the confluence of both the Pennsylvania Turnpike (Interstate 76) and Route 202. In fact, the only significant 1-mile sub-cluster (located in King of Prussia, PA) is almost precisely at the intersection of these two major routes. Further south, Route 202 basically runs through the middle of the second 5-mile cluster in Figure 14 (located in Wilmington, DE). The labs in the Philadelphia cluster are also in close proximity to a number of high-quality engineering and medical schools, including the University of Pennsylvania, Drexel University, Temple University, and Lehigh University.

Washington, DC, Area

Finally, in the metropolitan area of Washington, DC, we see from a comparison of Figure 15 and Figure 20 that essentially all core R&D points of the main 5-mile cluster (including its two 1-mile sub-clusters) are stretched along Interstate 270 to the north of Washington, together with the "Washington Beltway" (Interstate 495) to the west. In addition, the smaller 5-mile clusters to the east and west of the main cluster are close to Interstate 95 and Interstate 66, respectively. In terms of universities, the University of Maryland is just north of Washington, DC, inside the Beltway. In particular, the 5-mile cluster to the east along Interstate 95 is between the University of Maryland (to the south) and Johns Hopkins University in Baltimore (to the north). [36]

8. CLUSTERS AND KNOWLEDGE SPILLOVERS FROM PATENT CITATIONS

So far we have established a body of evidence demonstrating that R&D labs are indeed clustered, and we have posited a method for identifying specific clusters. In this section, we test whether there is any additional evidence that these clusters are potentially associated with knowledge spillovers that are attenuated by distance.

To do that, we follow in the tradition of JTH, who developed a method for studying the geographic extent of knowledge spillovers using patent citations. These are citations to earlier patents that are included in subsequent ones, much like references in an academic journal article.[37] These citations are a concrete indication of the transmission of information from one generation of innovation to another.

JTH test for the "localization" of 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 patent from an inventor living in a particular geographic area (in their study either a state or a consolidated metropolitan statistical area (CMSA)). The control group is a set of patents that are similar to citing patents in the treatment group, but that do not cite the specific patent in that geographic area. In this instance "similar" means that the control patents are selected so that they come from the same 3-digit (technology) patent classification and were issued at about the same date. In this way, the statistical test should control for the pre-existing spatial concentration of technologically related activities, which may be driven by a host of factors other than the specific spillovers they seek to measure.

JTH construct two proportions, one for the treatment group and one for the control group. The proportion is the number of citing patents that are from the same geographic area as the patent they cite divided by the total number of patents that cite a patent from that area. A statistically significant positive difference in these ratios for treatment and control groups is then taken to be a potential indication of localized spillovers. In this setting, JTH find that patent citations are two times more likely to come from the same state and about six times more likely to come from the

[36] As Figure 22 reveals, the 5-mile core cluster just west of the city of Pittsburgh (as seen in Figure 16) is almost precisely at the intersection of two major routes (Interstate 279 and Interstate 79) .

[37] This analogy should not be taken too literally. Referencing a prior patent may implicitly limit the scope of the current one. Thus, there is a disincentive to include gratuitous citations, which may occur in some journal articles.

same metropolitan area as earlier patents than one would expect based on the sample of control patents.

Here, we construct a comparable test statistic but substitute the R&D clusters identified in Section 6.3 for the state and CMSA geography used by JTH. This provides us with an alternative way to test for possible localized knowledge spillovers at smaller spatial scales than found in much of the preceding literature.[38] Recall that the boundaries of our clusters are determined by interrelationships among the R&D labs in our sample and therefore should more accurately reflect the appropriate boundaries in which knowledge spillovers are most likely to be at work. In that sense, the geography of our clusters should be better suited for studying knowledge spillovers than are states, metropolitan areas, or other political boundaries.

For out tests, we use the boundaries identified by our 5-mile and 10-mile buffer clusters.[39] Recall that we identified nine 5-mile clusters, of which two are in Boston, three in New York, two in Philadelphia, and two in Washington, DC. We identified four 10-mile clusters, one each in Boston, New York, Philadelphia, and Washington.

The patent citation counts we use are constructed from the NBER U.S. Patent Citations Data File and refer to citations from subsequent patents granted in the U.S. As is customary in much of the literature, we assign patents to locations according to the residential address of the first inventor named on the patent.[40] We begin with a set of $n_o = 9105$ *originating* patents, $\{o_i : i = 1,..,n_o\}$, that were granted to inventors living in one of our 5-mile clusters in the years 1996-97 (see Table 2).[41] If the *forward citations* for patent o_i are denoted by $\{c_{ij} : j = 1,..,n_i\}$, then (after removing self-citations), these originating patents received a total of $\sum_{i=1}^{n_o} n_i = 90{,}159$ forward citations over the years 1996-2006.[42] For each of these citing patents c_{ij}, we attempted to identify a unique control patent, $\tilde{c}_{ij}$, issued in the same year and 3-digit technology class, but which did not cite the originating patent, o_i. We were successful about 60 percent of the time. More specifically, if $\tilde{n}_i (\leq n_i)$ denotes the number of o_i-citing patents, c_{ij}, for which a control, $\tilde{c}_{ij}$, was found, then the total number of forward citations actually used for testing was $\sum_{i=1}^{n_o} \tilde{n}_i = n = 54{,}532$.[43] Given

[38] One exception is Murata et al. (2011), who test for (and find) evidence of localized knowledge spillovers using patent citations mapped to the level of "census places." in the U.S., which are somewhat larger than zip codes.

[39] Ideally, we would also like to conduct the analysis using the boundaries determined for the one-half mile and 1-mile buffer clusters. Unfortunately, we were unable to find a sufficient number of control patents to confidently conduct the analysis for the clusters defined at those spatial scales.

[40] Note that we are not conditioning our test on whether an inventor works at one of the R&D labs in our sample. At the same time, not all inventors who do work at one of those labs live close enough to work to fall within our 5- or 10-mile buffer clusters. This should not affect the validity of our statistical test, since we are using the same method of classifying patents for both the treatment and the control groups.

[41] The following formulation of the proportions used for testing purposes is based largely on Murata et al. (2011).

[42] Since self-citations may not result from knowledge spillovers, we not only removed inventor self-citations, but we also excluded citing patents owned by the same organizations as the originating patent.

[43] We could increase this match rate by relaxing the matching criteria (e.g., looking for patents in adjacent years or related patent classes), but it would be at the expense of reducing the similarity between the treatment and control groups. We chose not to do that.

these basic data, the desired proportion for the remaining "treatment" citations can be constructed by counting for each originating patent, o_i, the number, m_i, of citations, c_{ij}, with residential addresses in the same cluster as o_i. The fraction of all citations that are in the same cluster as their originating patent is then given by the *treatment proportion*,

$$p = \frac{\sum_{i=1}^{n_o} m_i}{\sum_{i=1}^{n_o} \tilde{n}_i} = \frac{1}{n}\sum_{i=1}^{n_o} m_i \qquad (6)$$

Similarly, if $\tilde{m}_i$ denotes the number of "control" citations, $\tilde{c}_{ij}$, with residential locations in the same cluster as o_i, then the desired fraction of these citations among all control citations is given by the *control proportion*,

$$\tilde{p} = \frac{\sum_{i=1}^{n_o} \tilde{m}_i}{\sum_{i=1}^{n_o} \tilde{n}_i} = \frac{1}{n}\sum_{i=1}^{n_o} \tilde{m}_i \qquad (7)$$

Finally, the desired test statistic is simply the difference between these proportions, i.e., $p - \tilde{p}$. Under the null hypothesis of "no relation," this difference of independent proportions is well known to be asymptotically normal with mean zero and thus provides a well-defined test statistic.[44]

As shown in Table 2, across all of the 5-mile clusters, 3.9 percent of the treatment patents were obtained by an inventor living in the same cluster as an inventor of the cited patent. For the control group, this occurs among less than 1 percent of citing patents. Thus, any patent citing an earlier patent in one of our 5-mile clusters is on average 4.3 times more likely to be in that cluster than would be expected by chance alone. This differential is statistically significant, as are the differentials for each of the clusters (the smallest *z* statistic is over 10). The range of differentials varies from as little as 2.8 times for one of the Boston clusters to as high as 12.3 times for one of the Washington clusters.

Table 3 presents the comparable analysis for our four 10-mile buffer clusters. For these geographies we identified 16,424 originating patents granted in 1996-97 that received a total of 160,224 subsequent citations. We successfully matched 62 percent (99,255) of those citing patents to a control patent. At this spatial scale, any patent citing an earlier patent in one of our 10-mile clusters is on average 1.7 times more likely to be in that cluster than would be expected by chance alone. This differential is statistically significant, as are the differentials for each of the clusters (the smallest *z* statistic is over 6).

While hardly conclusive, the smaller differential identified for the 10-mile buffer clusters, when compared to the differential for the 5-mile clusters, is consistent with the attenuation in *z*-scores

[44] 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 valid here. But for the present large sample size, $n > 50,000$, this is of little consequence since the *t* and standard normal distributions are virtually identical.

we observed with rising spatial scales in our analysis of the global *K*-function estimates. It is also consistent with the rapid attenuation of knowledge spillovers with distance found in other empirical studies.

9. CONCLUDING REMARKS

In this article, we use several distance-based econometric techniques to analyze the spatial concentration of the locations of over 1,000 R&D labs in a ten-state area in the Northeast corridor of the United States. Rather than using a fixed spatial scale, we attempt to describe the spatial concentration of labs more precisely, by examining spatial structure at different scales using Monte Carlo tests based on Ripley's *K*-function. Geographic clusters at each scale are then identified in terms of statistically significant departures from random locations reflecting the underlying distribution of manufacturing activity (employment).

Two important findings emerged from the global *K*-function analysis. First, the clustering of labs is by far most significant (based on *z*-scores) at very small spatial scales, such as distances of about one-quarter of a mile, with significance attenuating rapidly during the first half-mile. The rapid attenuation of significant clustering at small spatial scales is consistent with the view that knowledge spillovers are highly localized. We also observe a secondary mode of significance at a scale roughly associated with metropolitan areas. This secondary cluster is consistent with the view that agglomeration economies associated with the scale of labor markets (e.g., externalities associated with pooling and matching of skilled workers) is important for innovative activity.

While the global *K*-function analysis indicates that there is very significant clustering of R&D locations relative to manufacturing employment, it provides little more information other than the spatial scale (distances) at which clustering appears to be most significant. Local *K*-function analysis is useful for identifying the location and extent of *specific* concentrations of labs. In this paper, we introduce a novel way to identify clusters, called *the multiscale core-cluster* approach. The local *K*-function analysis identified four major clusters (one each in Boston, New York-Northern New Jersey, Philadelphia-Wilmington, and Washington, DC). These four clusters roughly correspond to the size of the secondary mode of clustering (approximately at a distance of 40 miles) identified by the global *K*-function. We also found that R&D labs tend to concentrate along major highways and often at or near junctions of major highways. Each of these clusters has distinct characteristics, especially in terms of the mix of industries the R&D labs serve.

In the final section of the paper, we apply a familiar test for potentially identifying localized knowledge spillovers using patent citations. Jaffe, Trajtenberg and Henderson (1993) found evidence of localization of patent citations at the level of individual states or CMSAs. We verify that this occurs for much smaller scales using tests based on our 5- and 10-mile buffer clusters. This suggests that our multiscale core-cluster approach is identifying economically significant location patterns that may be related to knowledge spillovers that attenuate with distance.

REFERENCES

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

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

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

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

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

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

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

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

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

Buzard, Kristy, and Gerald A. Carlino. "The Geography of Research and Development Activity in the U.S.," in *Handbook of Economic Geography and Industry Studies*," Frank Giarratani, Geoff Hewings, and Philip McCann (eds.) Cheltenham: Edward Elgar Publishing (2011).

Carlino, Gerald, Satyajit Chatterjee, and Robert Hunt. "Urban Density and the Rate of Invention," *Journal of Urban Economics*, 61 (2007), pp. 389-419.

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

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

Duranton, Gilles, and Henry G. Overman. "Testing for Localization Using Micro-Geographic Data," *Review of Economic Studies*, 72 (2005), pp. 1077-1106.

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

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

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

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

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

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

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

Harhoff, Dietmar, Francis Narin, F. M. Scherer, and Katrin Vopel. “Citation Frequency and the Value of Patented Inventions,” *Review of Economics and Statistics*, 81 (1999), pp. 511–15.

Holmes, Thomas J., and John J. Stevens. “Spatial Distribution of Economic Activities in North America,” in: J.V. Henderson and J.-F Thisse (eds.), *Handbook of Regional and Urban Economics, Vol. IV: Cities and Geography*. North Holland, Amsterdam: Elsevier (2004).

Hunt, Robert. "Matching Externalities and Inventive Productivity," Federal Reserve Bank of Philadelphia Working Paper 07-7 (2007).

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

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

Kerr, William R., and Scott Duke Kominers. “Agglomerative Forces and Cluster Shapes,” Harvard Business School Working Papers 11-061, Harvard Business School, 2010.

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

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

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

Mori, T. and T.E. Smith. “A Probabilistic Modeling Approach to the Detection of Industrial Agglomerations,” *Discussion Paper* 682, Kyoto Institute of Economic Research, Kyoto University, Kyoto, Japan (2009).

Murata, Yasusada, Ryo Nakajima, Ryosuke Okamoto, and Ryuichi Tamura. “Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach,” (2011) *Tsukuba Economics Working Papers* 2010-010, Economics, Graduate School of Humanities and Social Sciences, University of Tsukuba.

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

Romer, P. “Endogenous Technological Change,” *Journal of Political Economy*, 98 (1990), pp. S71-102

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

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

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

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

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

TABLE 2: Localization Test for 5 Mile Buffer 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
Framingham-Marlborough-Westborough, MA	419	5,171	194	3.75	2,517	81	3.22	2,517	7	0.28	**11.6**	**16.7**
Boston-Cambridge-Waltham-Woburn, MA	2,657	29,584	1,228	4.15	16,874	794	4.71	16,874	282	1.67	**2.8**	**14.3**
Princeton, NJ	916	9,756	347	3.56	6,528	270	4.14	6,528	40	0.61	**6.8**	**17.7**
Parsippany-Basking Ridge, NJ	1,714	16,503	591	3.58	10,223	410	4.01	10,223	54	0.53	**7.6**	**17.8**
Greenwich, CT-White Plains, NY-Montvale, NJ	963	8,010	328	4.09	4,958	198	3.99	4,958	18	0.36	**11.0**	**18.5**
Wilmington, DE	513	2,686	60	2.23	2,691	56	2.08	2,691	15	0.56	**3.7**	**10.7**
King of Prussia, PA	726	4,053	105	2.59	3,372	93	2.76	3,372	24	0.71	**3.9**	**12.5**
Chantilly-Sterling, MD	218	3,773	81	2.15	1,455	37	2.54	1,455	3	0.21	**12.3**	**14.8**
Columbia-Rockville-Mc Lean, MD	976	10,623	299	2.81	5,914	199	3.36	5,914	60	1.01	**3.3**	**13.0**
All 5 Mile Clusters	9,105	90,159	3,233	3.59	54,532	2,138	3.92	54,532	503	0.92	**4.3**	**15.4**

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

TABLE 3: Localization Test for 10 Mile Buffer 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
Boston	4,894	54,834	2,269	4.14	31,529	1,364	4.33	31,529	777	2.46	**1.8**	**9.2**
New York	7,727	73,020	2,672	3.66	46,741	1,763	3.77	46,741	1,205	2.58	**1.5**	**6.3**
Philadelphia	1,976	11,288	269	2.38	9,677	241	2.49	9,677	98	1.01	**2.5**	**9.5**
Washington	1,827	21,082	548	2.60	11,308	318	2.81	11,308	126	1.11	**2.5**	**10.3**
All 10 Mile Clusters	16,424	160,224	5,758	3.59	99,255	3,686	3.71	99,255	2,206	2.22	**1.7**	**7.8**

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

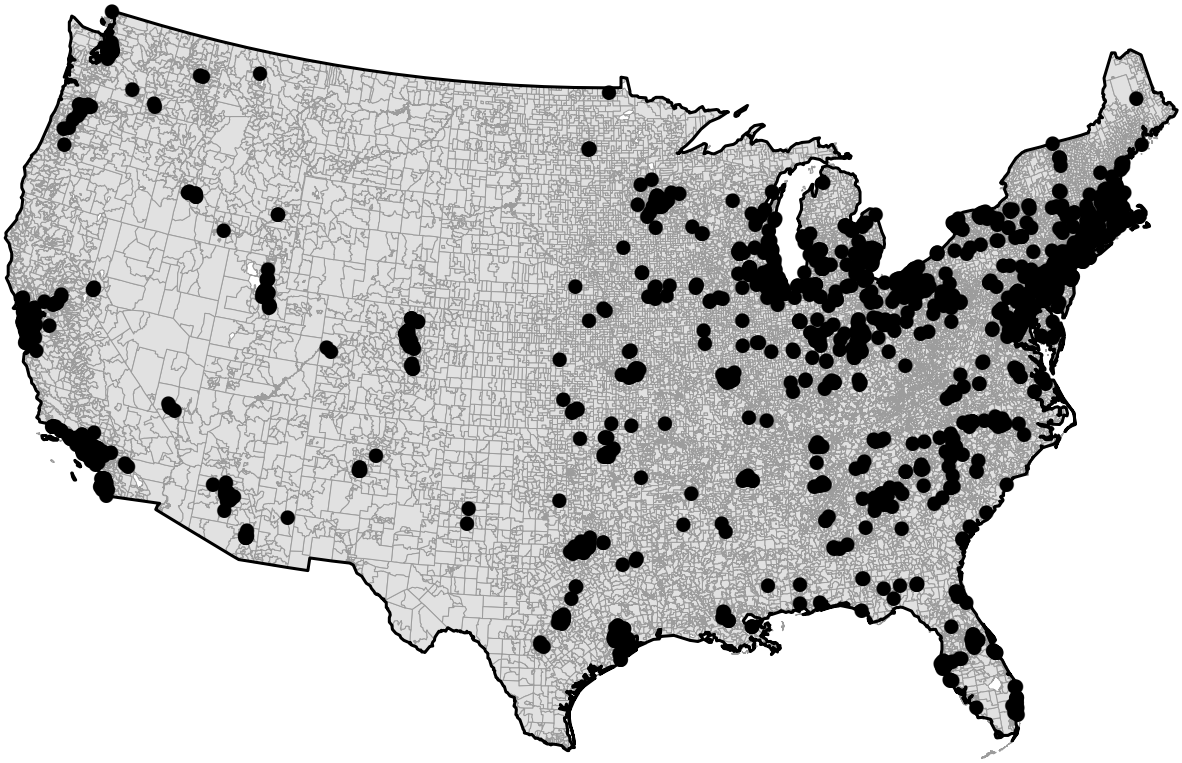

Figure 1. Location of R&D Labs

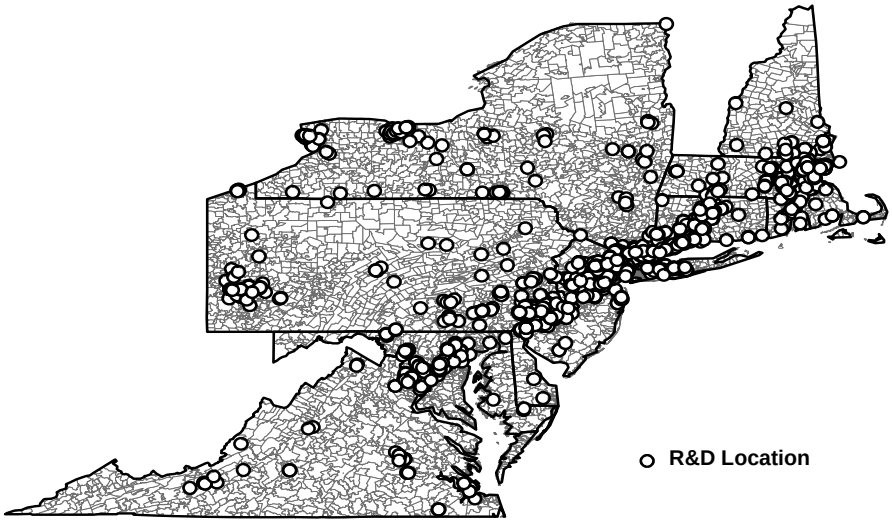

Figure 2. R&D Locations

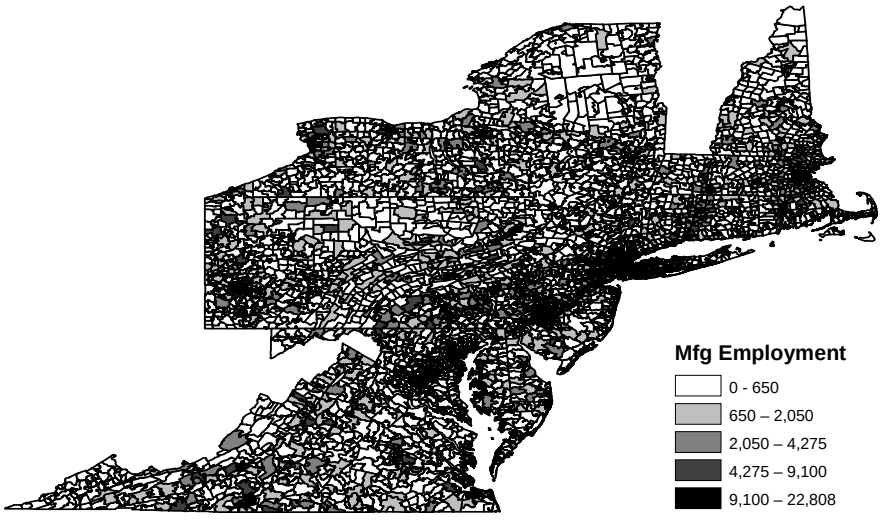

Figure 3. Manufacturing Employment

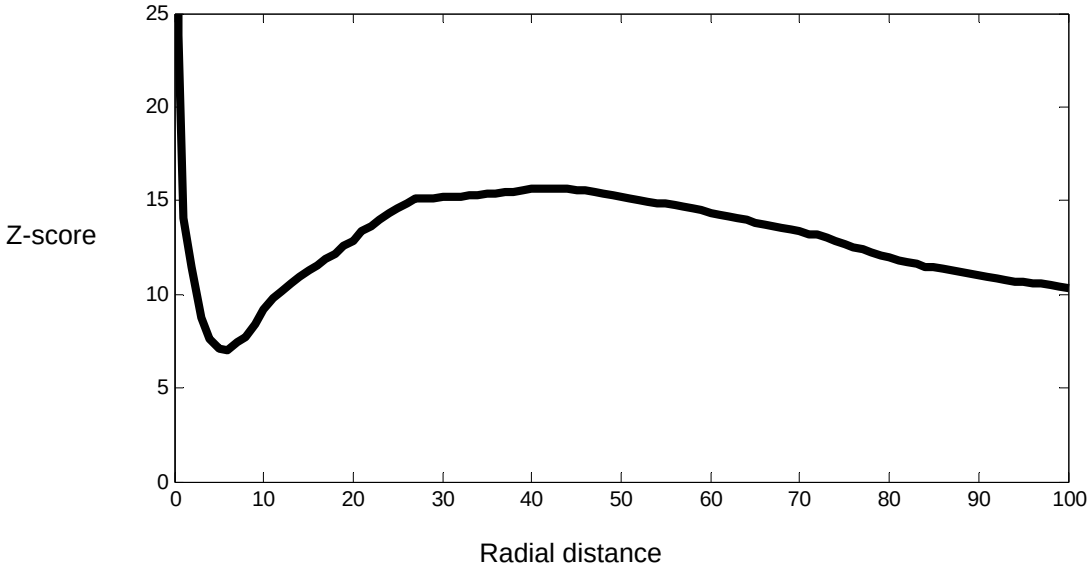

Figure 4a. Z-scores for Global Clustering

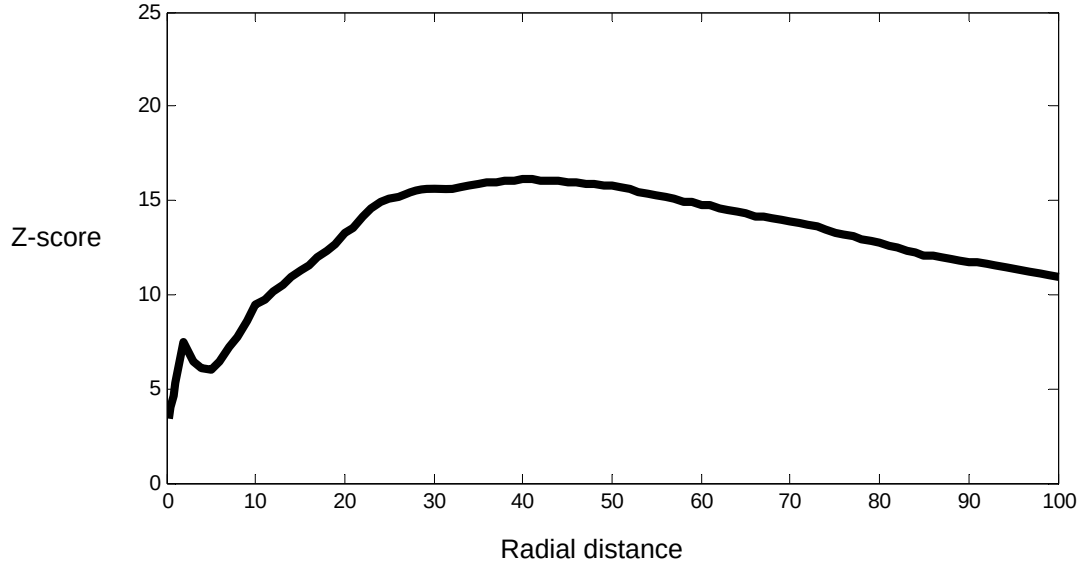

Figure 4b. Alternative Z-scores for Global Clustering

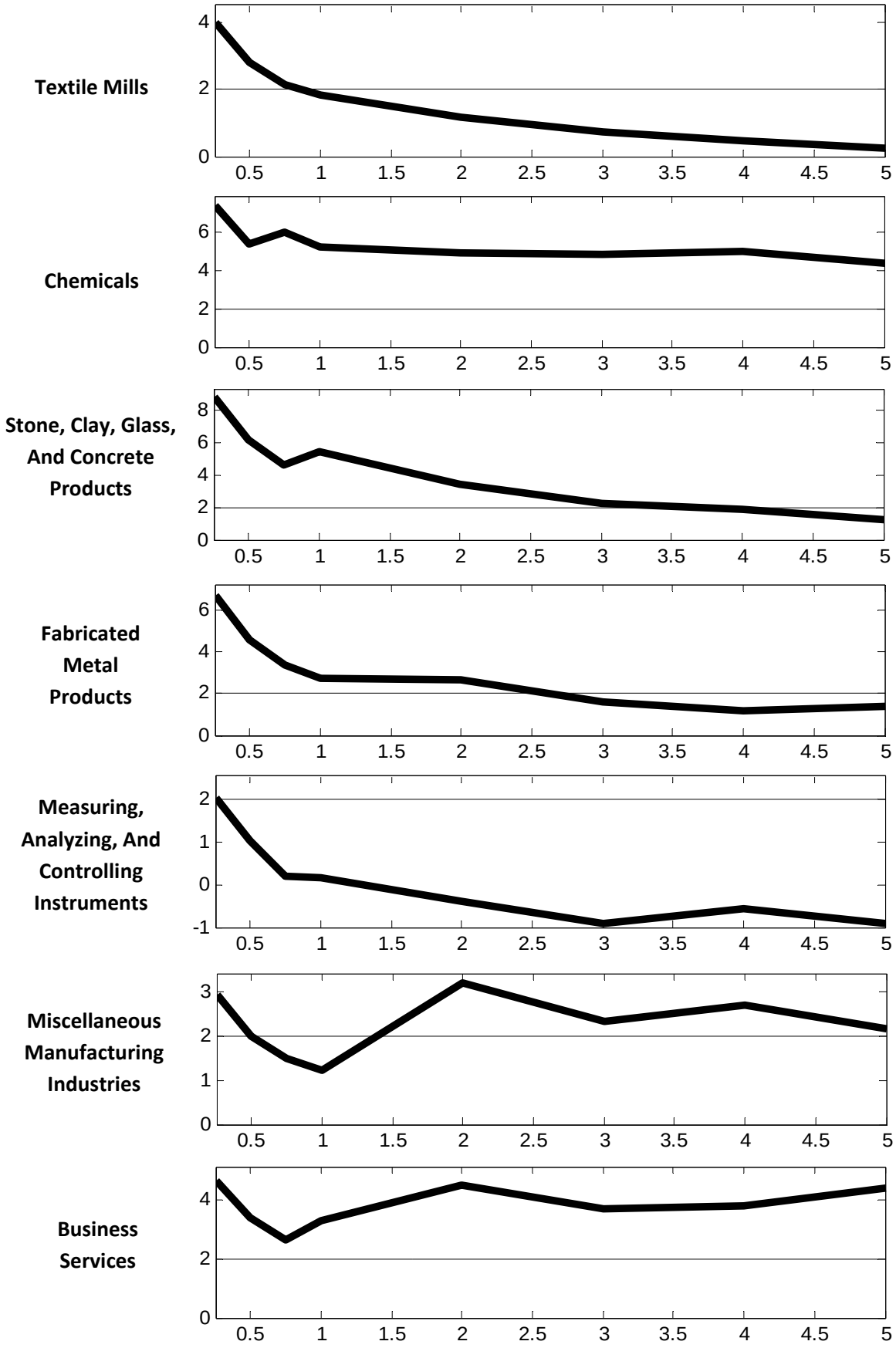

Figure 5. Industry Z-Scores

Figure 6. Local *K*-Function *P*-values at $d = 5$ Miles

Figure 7. Union of the Top 10 SATSCAN Clusters

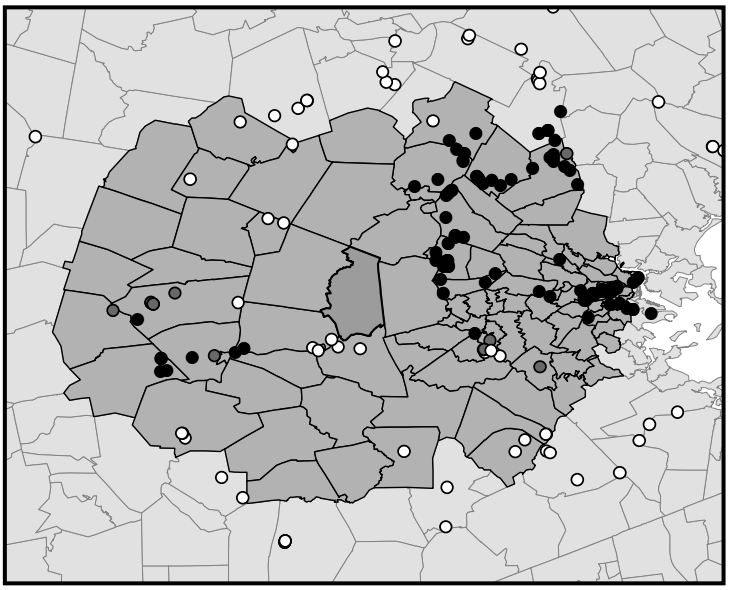

Figure 8. Boston Cluster in SATSCAN

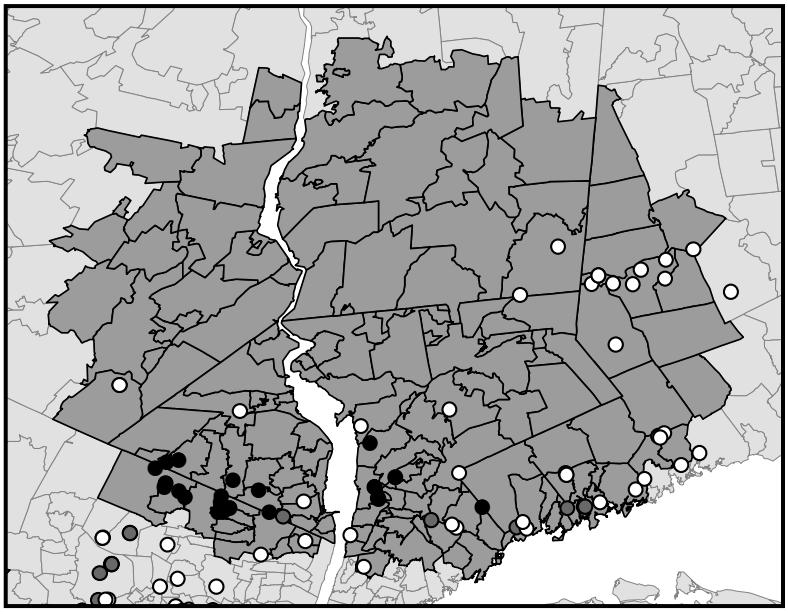

Figure 9. Largest New York Cluster in SATSCAN

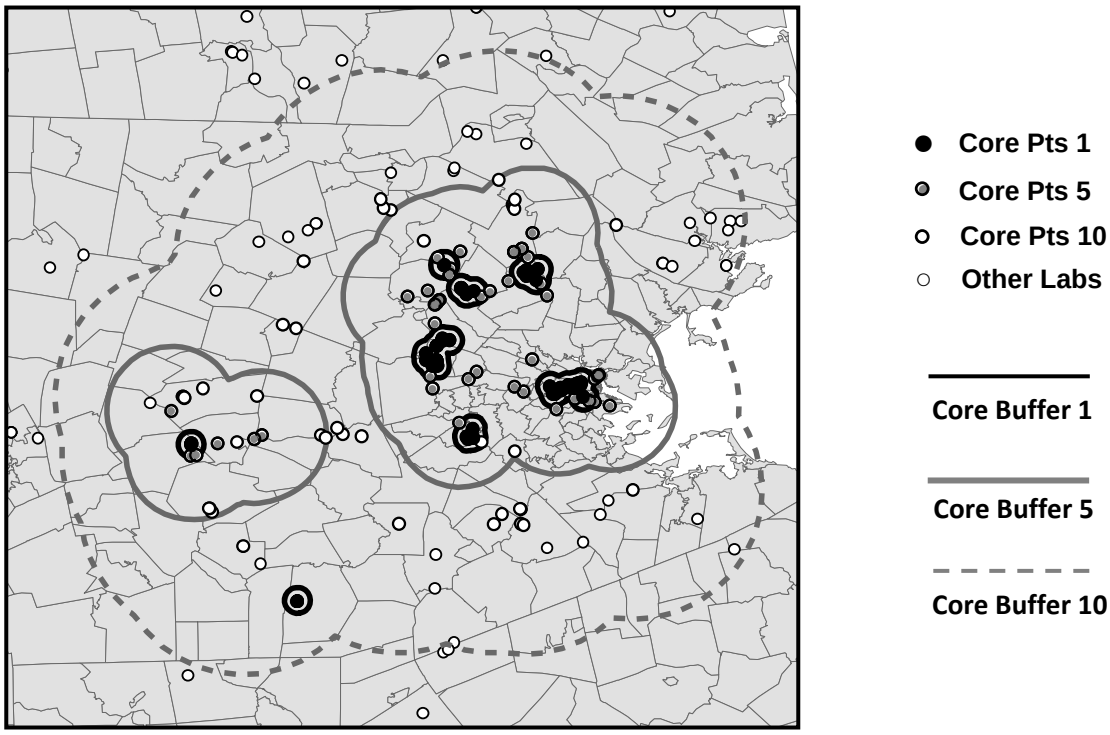

Figure 10. Boston Core Clusters

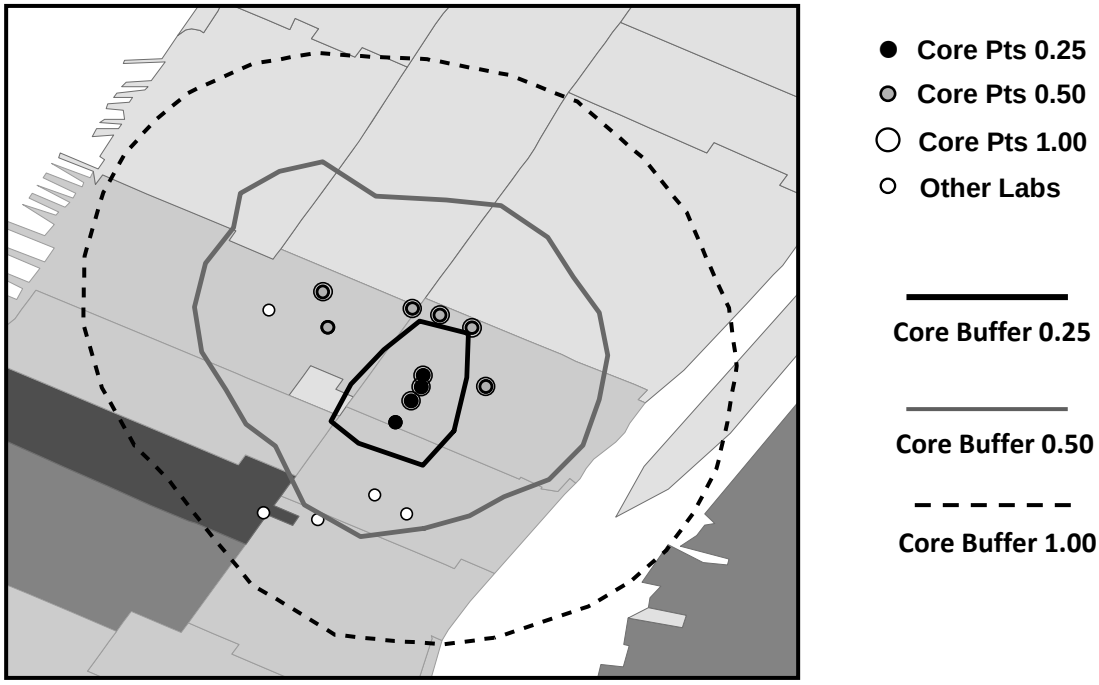

Figure 11. Central Park Core Clusters

Figure 12. Multiscale Core Clusters (d = 1,5,10)

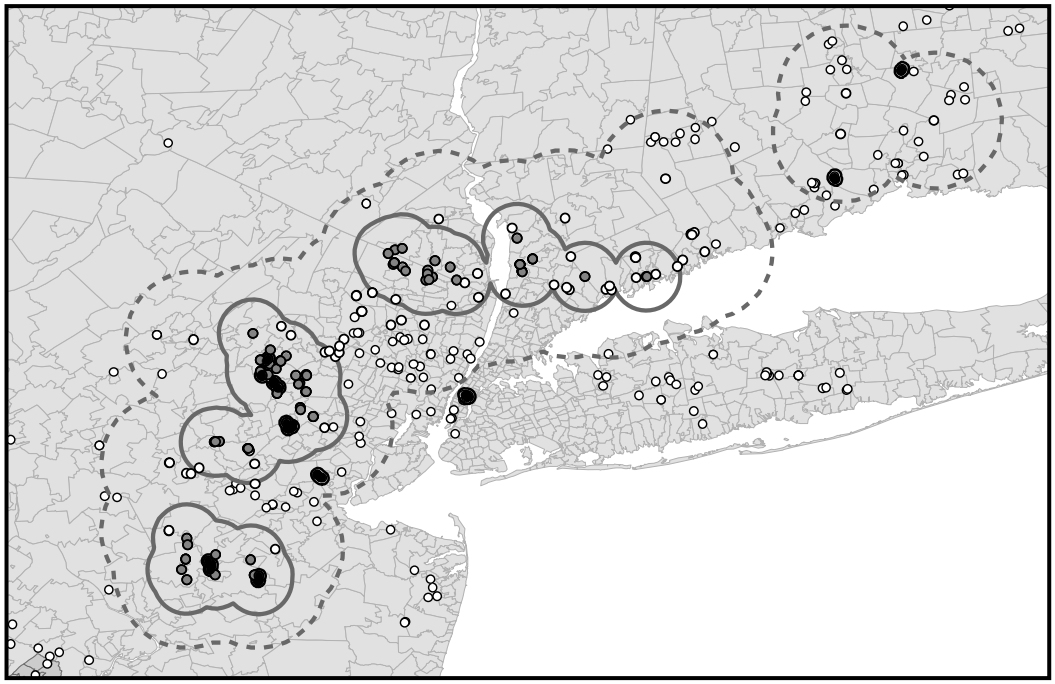

Figure 13. New York Core Clusters

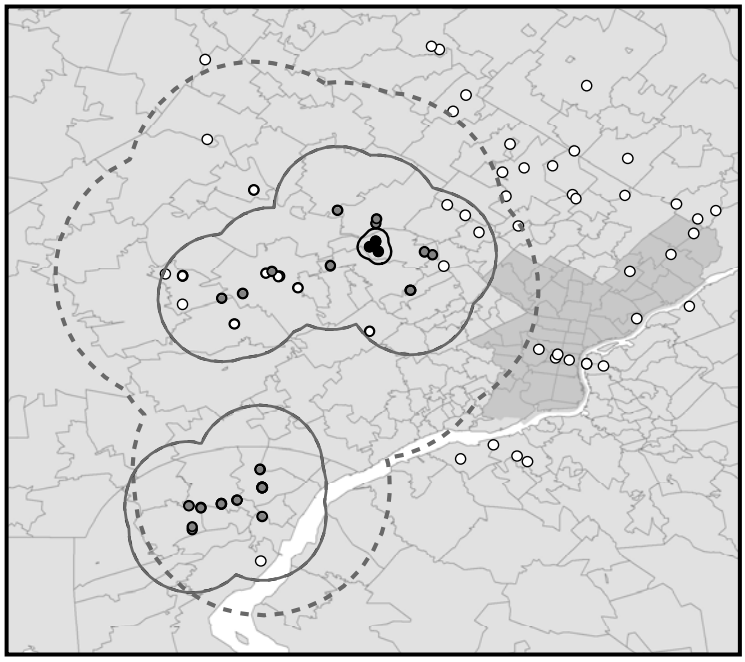

Figure 14. Philadelphia Core Clusters

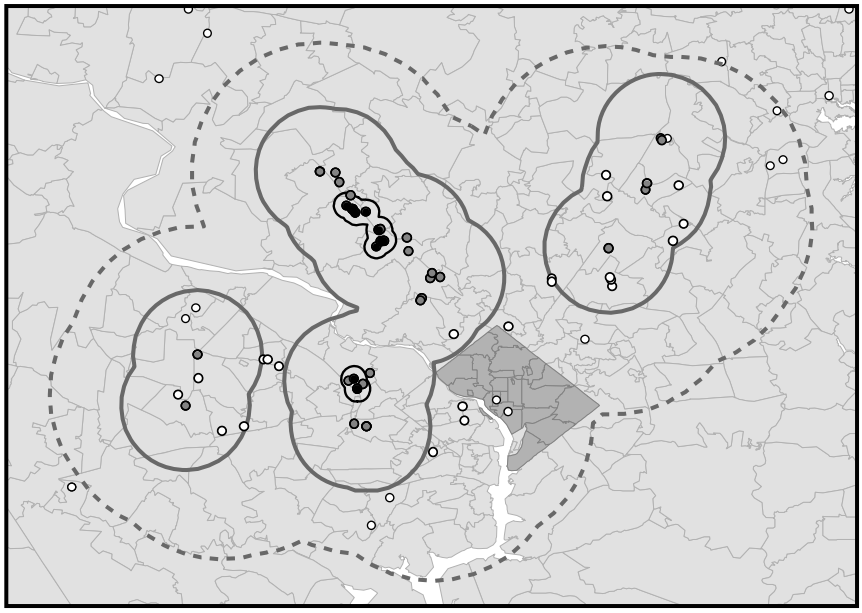

Figure 15. Washington, DC, Core Clusters

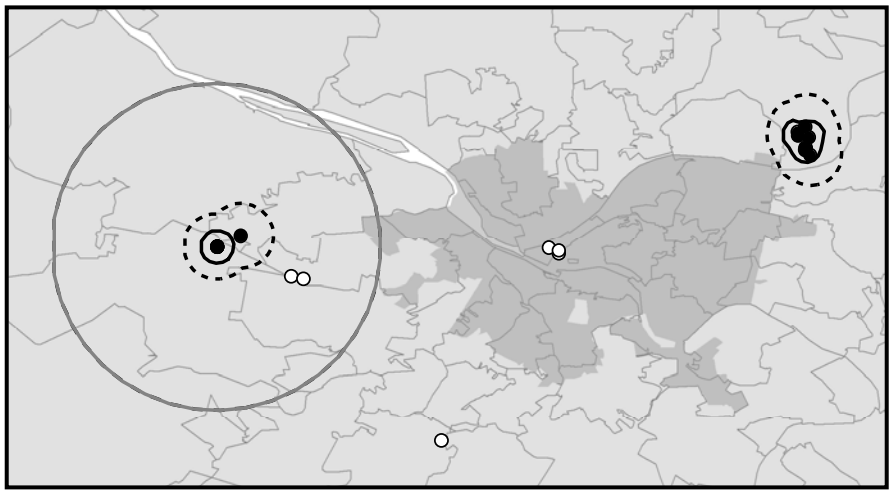

Figure 16. Pittsburgh Core Clusters

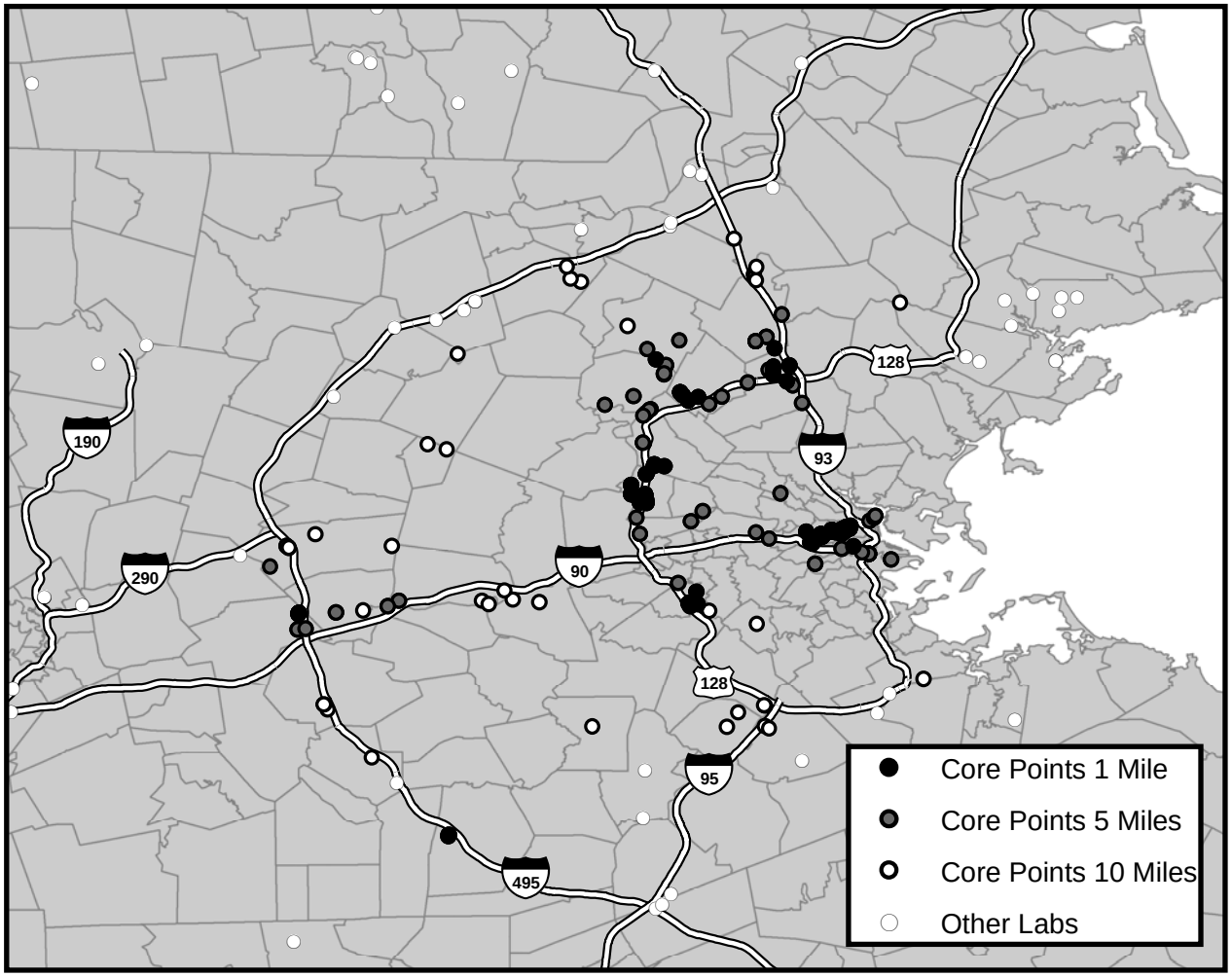

Figure 17. Boston Core Points and Major Routes

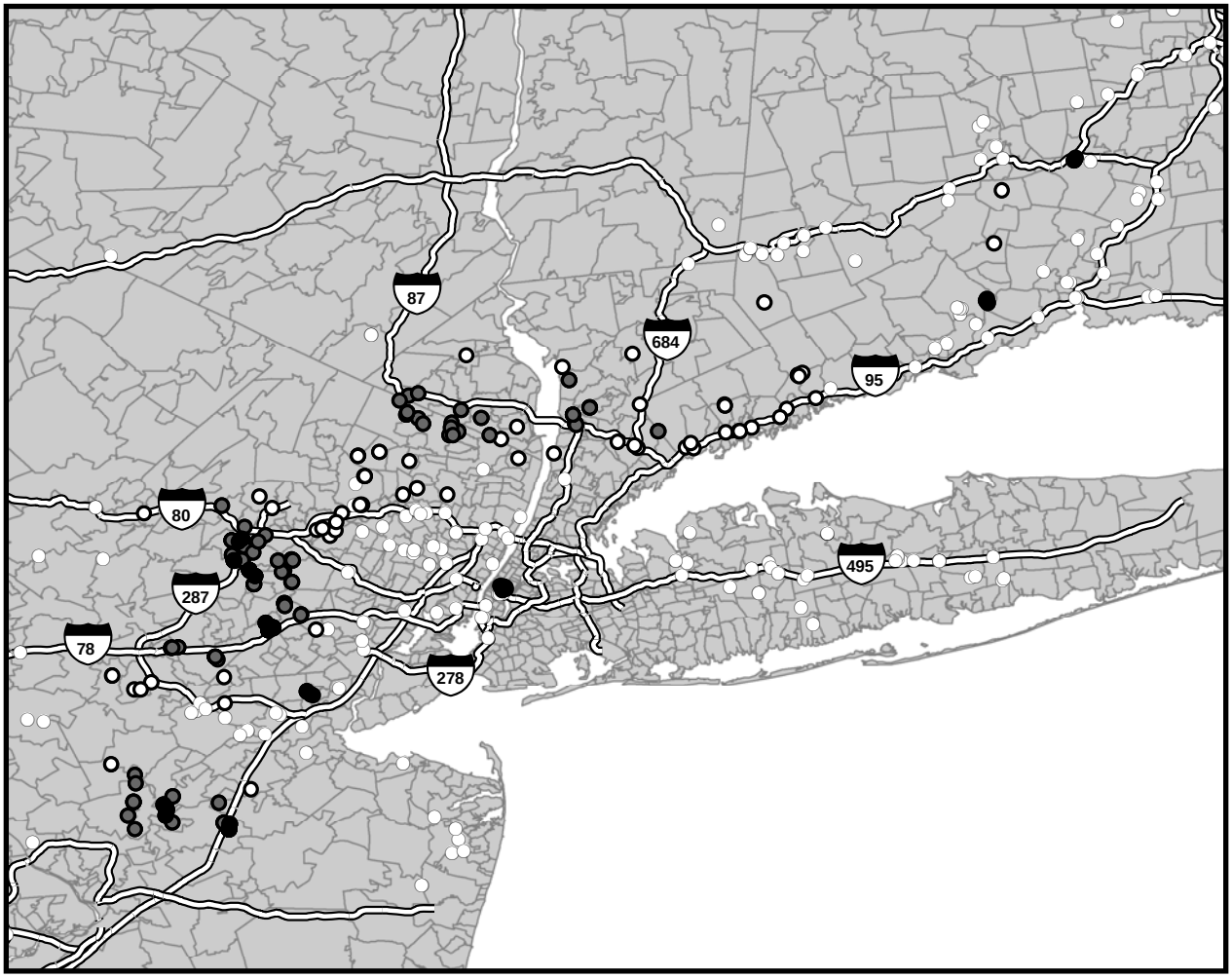

Figure 18. New York City Core Points plus Major Routes

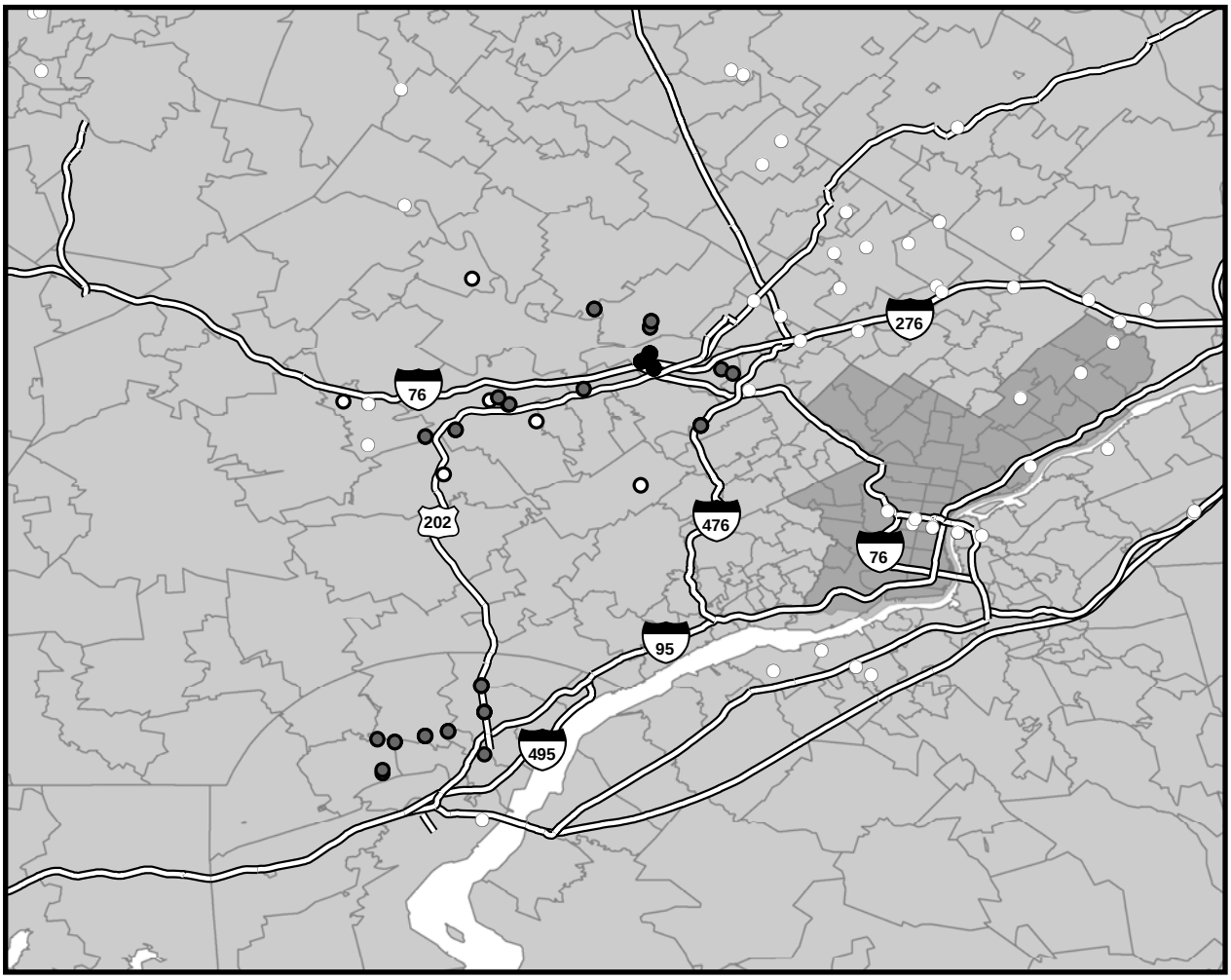

Figure 19. Philadelphia –Wilmington Core Points plus Major Routes

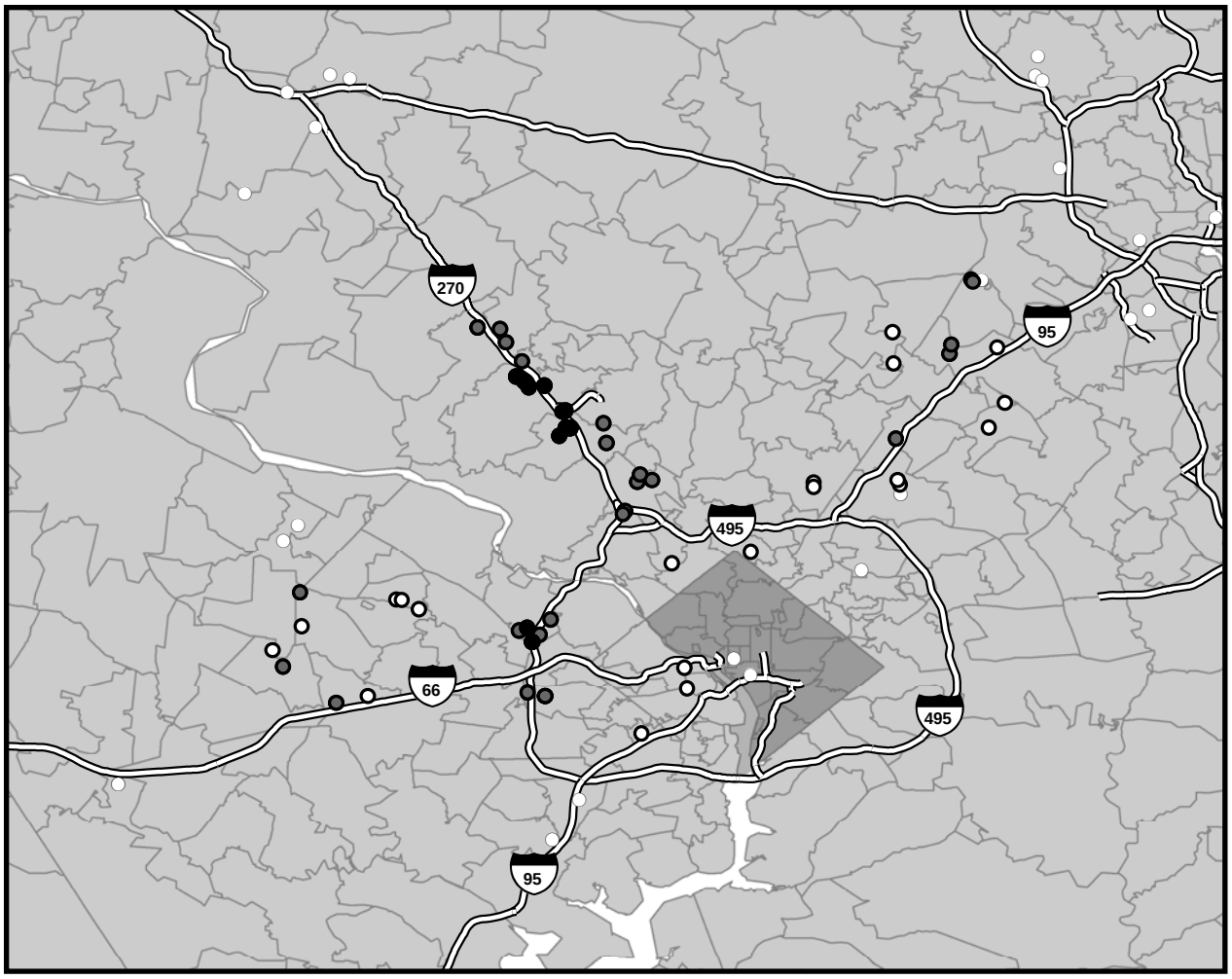

Figure 20. Washington, DC – Northern Virginia Core Points plus Major Routes

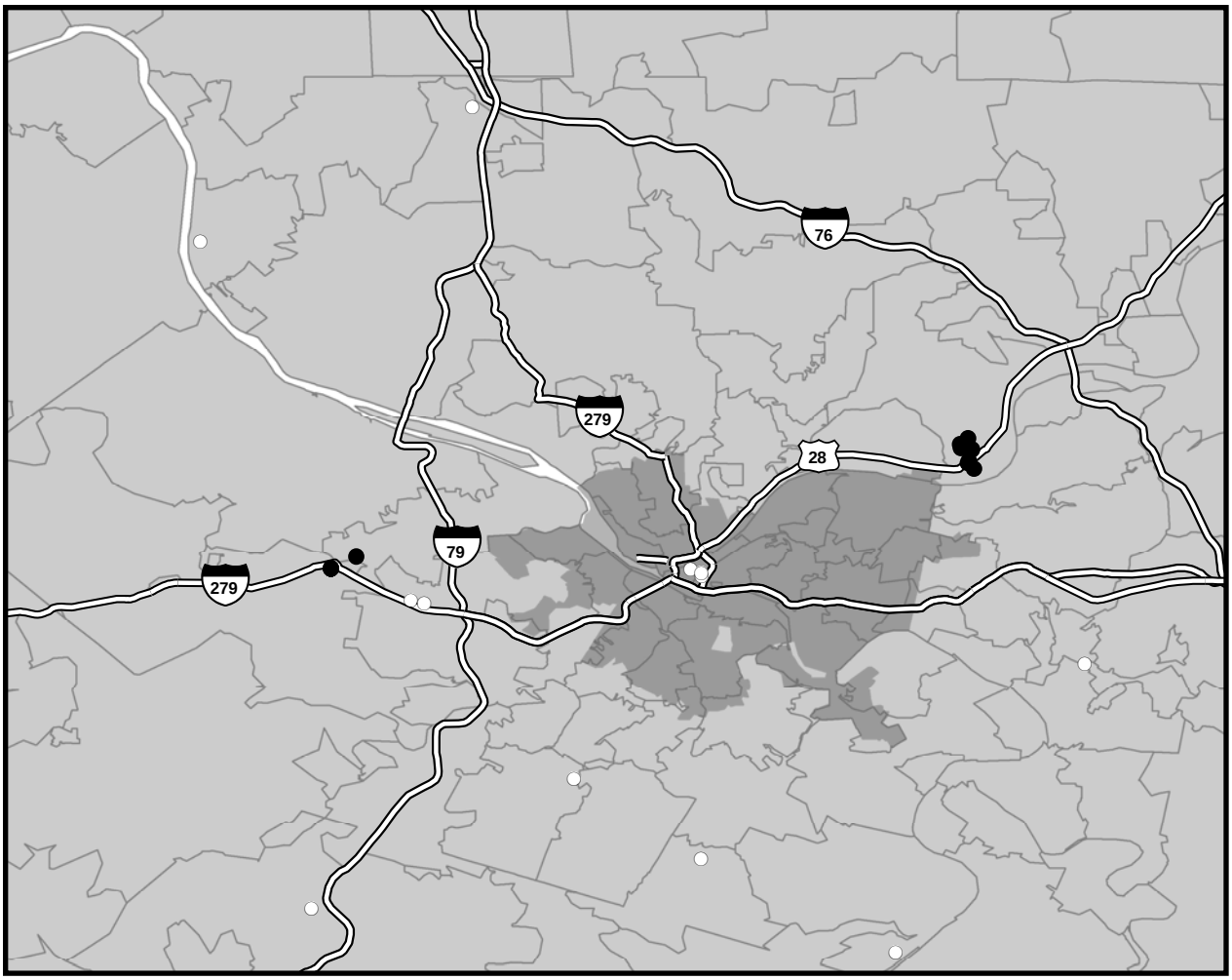

Figure 21. Pittsburgh Core Points plus Major Routes