WORKING PAPER NO. 10-33
THE AGGLOMERATION OF R&D LABS

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

Tony E. Smith
University of Pennsylvania

October 2010

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, Jake Carr, and Robert M. Hunt
Federal Reserve Bank of Philadelphia

Tony E. Smith
University of Pennsylvania

October 2010

ABSTRACT

We document the spatial concentration of more than 1,000 research and development (R&D) labs located in the Northeast corridor of the U.S. using point pattern methods. These methods allow systematic examination of clustering at different spatial scales. In particular, Monte Carlo tests based on Ripley's (1976) *K*-functions are used to identify clusters of labs — at varying spatial scales — that represent statistically significant departures from random locations reflecting the underlying distribution of economic activity (employment). Using global *K*-functions, we first identify significant clustering of R&D labs at two different spatial scales. This clustering is by far most significant at very small spatial scales (a quarter of a mile), with significance attenuating rapidly during the first half mile. We also observe statistically significant clustering at distances of about 40 miles. This corresponds roughly to the size of the four major R&D clusters identified in the second stage of our analysis — one each in Boston, New York-Northern New Jersey, Philadelphia-Wilmington, and Virginia (including the District of Columbia). In this 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 scales. Our global finding of clustering at two spatial scales suggests the possibility of two distinct forms of spillovers. First, the rapid attenuation of significant clustering at small spatial scales is consistent with the view that knowledge spillovers are highly localized. Second, the scale at which larger clusters are found is roughly comparable to that of local labor markets, suggesting that such markets may be the source of additional spillovers (e.g., input sharing or labor market matching externalities).

[*] 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

In models of endogenous growth, knowledge, rather than tangible assets, plays a central role in the economic growth of nations. The model of Romer (1990) assumes that economic agents everywhere have free access to the stock of knowledge. Agrawal, Kapur, and McHale (2008), among many others, point out that immediate accessibility to knowledge is likely to depend on the geographic proximity of agents. Until recently, the vast literature on agglomeration economies has treated these externalities as if they were internalized at a broad geographic aggregate, such as a county, a metro area, or even a state. Recent literature has shown, however, that the benefits associated with knowledge spillovers attenuate rapidly with distance from the source of the externality (Audretsch and Feldman, 1996; Rosenthal and Strange, 2001 and 2008; Keller, 2002; Fu, 2007; Agrawal, Kapur, and McHale, 2008; Arzaghi and Henderson, 2009; and Elvery and Sveikauskas, 2010).

It is well known that economic activity is geographically concentrated. This is also true of research and development (R&D) activity, as is immediately evident from examining a national map of the locations of private R&D establishments, as shown in Figure 1. Notice the very high concentration of R&D labs in the Northeast corridor — stretching from northern Virginia to Massachusetts — which is the focus of this paper. Other concentrations appear around the Great Lakes, Southern California, and California's Bay Area. What is not immediately clear from the map is that spatial concentration of R&D is significantly greater than manufacturing activity in general, a fact established in Buzard and Carlino (2009).

In this paper, we use distance-based techniques to analyze the spatial concentration of the locations of over 1,000 R&D labs in a nine-state area of the Northeast corridor of the United States, a region that accounts for one-third of all R&D labs in the United States. A number of previous papers have used the Ellison and Glaeser (1997) concentration index to measure the clustering of manufacturing employment at the zip code, county, MSA, and state levels of geography. We attempt to delineate the spatial structure of such concentrations in more detail by employing 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) 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 is by far most significant (based on *p*-values) 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.

We also observe a secondary mode of significance at a scale of about 40 miles. This will be seen below 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.

While there is a well-developed theoretical literature describing a variety of microfoundations for agglomeration economies, empirical tests of the various mechanisms are plagued with the problem of observational equivalence. The global findings reported in this paper suggest that examining the concentration of economic activity over a variety of spatial scales may help to address this issue. One possibility is that there is a single mechanism at work operating simultaneously at different spatial scales. For example, in their study of manufacturing activity at the zip code, county, and state levels, Rosenthal and Strange (2001) find that labor market pooling appears to operate at all three levels of geography. Another possibility suggested by our global clustering results is that there are two different mechanisms of spillovers that account for patterns of clustering at two distinct spatial scales (locally in the form of knowledge spillovers and at the scale of labor markets in the form of pooling and matching externalities).[1] This suggests that empirical work should allow for the possibility of multiple mechanisms, both for the purposes of identification and to avoid confounding results.

The second phase of our analysis employs local *K*-function statistics to identify explicit spatial clusters of R&D labs. For each scale, *d*, we identify those labs with strongly significant clustering in their *d*-neighborhoods as *core points* of agglomeration at scale *d*. Those core points that are sufficiently close to share *d*-neighbors are then grouped into clusters designated as *core clusters*. These core clusters yield a natural 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 essentially replicated using the significance-maximizing procedures developed by Besag and Newell (1991) and Kulldorff (1997).

Finally, by applying this hierarchy of core clusters, we are able to refine the internal spatial structure of the four major agglomerations identified. In particular, this spatial structure can be related to key local geographic features such as the presence of freeways and proximity to university centers.

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 both global and local analyses of spatial clustering are then developed in Section 4. Finally, these results are summarized and discussed in Section 5.

[1] As an example, Rosenthal and Strange (2001) find evidence that knowledge spillovers operate at a spatial scale corresponding to zip codes, while labor market pooling appears at all the spatial scales they test.

2. LITERATURE REVIEW

A number of previous papers have used the Ellison and Glaeser (1997) concentration index (the EG index) to measure the clustering of manufacturing employment at the zip code, county, MSA, and state levels of geography (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).

Duranton and Overman (2005) — hereafter referred to as DO — use micro data to identify the postal codes for each manufacturing plant in the UK, allowing them to geocode the data. Geo-coding is an important since DO are not bound by a fixed geographical classification (such as state, MSA, or county definitions) 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 (i.e., kernel density methods). Essentially, DO construct frequency distributions of the pair-wise distances between plants in a given industry. When the mass of the distribution is concentrated on the left of the distribution, this represents a spatial concentration of plants in the industry. Alternatively, if the mass of the distribution is concentrated on the right of the distribution, 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 location were randomly chosen. In addition to considering a discrete measure of coagglomeration (measured at the state, MSA, and county levels of geography), 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.

Much of the previous literature has focused on the spatial concentration of manufacturing employment. Fewer papers have looked at the concentration of innovative activity. Audretsch and Feldman (1996) use the United States Small Business Administration's Innovation Data Base that consists of innovations compiled from the new product announcements sections in manufacturing trade journals. Using a predecessor to the EG index, they found that innovation tends to be relatively more concentrated in industries in which knowledge spillovers tend to be important. A recent paper by Buzard and Carlino (2009) looks at the spatial concentration of R&D labs at the county level of geography. They use a variant of the EG index developed by Guimarães, Figueiredo, and Woodward (2007) to examine the spatial distribution of R&D labs. The purpose of the Buzard and Carlino (2009) paper is to identify R&D clusters that are significantly different from spatial randomness. They show that while economic activity tends to be geographically concentrated, spatial concentration is even more pronounced among establishments doing R&D.

The theoretical literature on urban agglomeration economies has focused on externalities in the production of goods and services rather than in inventive activity itself. Nevertheless, the three most prominent mechanisms explored in this literature — input sharing, matching, and

knowledge spillovers — are also relevant for the location choices for labs engaged in R&D activity. Of these microfoundations, empirical evidence on knowledge spillovers is rather sparse. What the limited research suggests is that there is an extremely rapid distance decay associated with knowledge spillovers (Audretsch and Feldman, 1996; Rosenthal and Strange, 2001 and 2008; Keller, 2002; Fu, 2007; Agrawal, Kapur, and McHale, 2008; Arzaghi and Henderson, 2008; and Elvery and Sveikauskas, 2010). For example, Jaffe, Trajtenberg, and Henderson (1993) find that nearby inventors have a much higher propensity to cite each others' patents, suggesting that knowledge spillovers are indeed localized.

Arzaghi and Henderson (2008) look at the location pattern of firms in the advertising industry in Manhattan. They report that Manhattan accounts for 20 percent of the total national employment in the ad industry, 24 percent of all advertising agency receipts, and 31 percent of media billings. They show that for an ad agency, knowledge spillovers and the benefits of networking with other nearby agencies are large, but the benefits dissipate very quickly with distance from other ad agencies and are gone after roughly one-half of a mile.

Our work differs from past studies in three ways. First, rather than looking at the geographic concentration of firms engaged in the production of goods (such as manufacturing) and services (such as advertising), we consider the spatial concentration of private R&D establishments. The paper by Buzard and Carlino (2009) is a predecessor to this study, but their study describes the location of R&D labs on a single spatial scale (counties). There is also a small but growing literature on research/technology/science parks that represent the clustering of R&D and technology-based institutions that develop on or near a university campus in order to benefit from knowledge flows from the university and among the institutions in the cluster (see Link, 2009 for a recent survey of this literature). This literature takes the cluster as a given construct — the park is the cluster. In our paper, we allow the data to identify clusters, some of which will correspond to science parks. Second, rather than focusing on the concentration of *employment* in a given lab, we look at the clustering of individual R&D *labs*.[2] Third, following DO, we look for geographic clusters of labs that represent statistically significant departures from spatial randomness using simulation techniques.

The paper by Lychagin et al. (2010) is related to our study. Lacking data on geographic distribution of R&D activity, Lychagin et al. (2010) use the address of the first named inventor on patent applications matched with Compustat data on firms to create a spatial distribution of the locations of firms' R&D activities. In a production function analysis, they find that geographic spillovers have a large and economically significant effect on sales growth, even after controlling for product-market proximity and proximity in technology. As in our study, they find evidence that geographic spillovers are highly localized. They also find that the location of a company's inventors is more important for a company's sales than the locations of a company's headquarters. Our study differs in several ways from that of Lychagin et al. (2010). First, our paper is based on the actual locations of R&D activity. Second, we do not take clustering as given, but rather we derive patterns of clustering that are consistent with spillovers mediated by physical distance.

[2] 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.

3. DATA

We hand coded the addresses and other information about private R&D labs contained in the 1998 vintage of the *Directory of American Research and Technology.* Since the directory lists the complete address for each establishment, we were able to assign a geographic identifier (using geocoding techniques) to 3,134 R&D labs in the U.S. in 1998. In this paper, we limit our analysis to 1,035 labs in a modified set of 6,043 zip codes areas in nine states comprising the Northeast corridor of the United States (Connecticut, Delaware, Maryland, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Virginia, including the District of Columbia — the Washington, DC cluster). The (geocoded) locations of these 1,035 labs are depicted by the enlarged dots in Figure 2.[3]

R&D activity is more concentrated in these nine states than is either employment or population. In 1998, 33 percent of all labs were located within this region, 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 innovations counts in their data include Massachusetts, New Jersey, and New York.

In our formal analysis, 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.[4] This is a good benchmark since most of our R&D labs are associated with manufacturing firms.[5] Manufacturing is represented spatially by total manufacturing employment in each of the 6,043 zip code areas in these nine states (shown in Figure 3).

4. POINT PATTERN ANALYSIS

Clustering is identified statistically by statistical tests involving both global and local *K*-function analyses, which have the advantage of identifying the spatial scales at which such clustering is most significant. In the next section, we begin with a global analysis of clustering involving all R&D locations to identify spatial scales at which clustering appears to be most significant. This is followed in Section 4.2 by a more local analysis in which both the location and extent of specific R&D clusters are identified. Here we begin with the significance-maximizing procedure of Besag and Newell (1991) and Kulldorff (1997) in particular and, in Section 4.3, report the clustering results obtained from the SATSCAN procedure of Kulldorff. Next, we develop our multiscale core-cluster procedure based on local *K*-function tests in Section 4.4 and report the clustering results for this procedure as well. Finally, in Section 4.5 we refine our global results to

[3] 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. Alternatively, when a company reported multiple labs at the same address, we treated it as a single lab, producing a set of 951 "company" labs. We repeated all analyses using company labs and found the results to be highly consistent to what is presented below.

[4] Our techniques implicitly assume that manufacturing employment is uniformly distributed across points within a given zip code.

[5] The two notable exceptions are electronics wholesaling (which includes firms such as Apple computers) and software. Comparable results are obtained when we use total rather than manufacturing employment.

the industry level by grouping R&D labs in terms of their primary industrial research areas at the two-digit standard industrial classification SIC level.

4.1 Global Cluster Analysis

The key question of interest is whether R&D locations are more clustered than would be expected from the spatial concentration of manufacturing alone. To address this question, a natural null hypothesis is that R&D locations are determined entirely by the distribution of manufacturing employment. More formally, our null hypothesis, H_0, is that *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 is deemed to be sufficiently disaggregated to provide reasonable approximations as unions of zip code areas.

A simple two-stage Monte Carlo procedure for generating locations approximately consistent with our null hypothesis is to choose a zip code with probability proportional to manufacturing employment in that zip code, 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 roughly consistent with H_0, where r_i, s_i represent the latitude and longitude coordinates (in decimal degrees) at point *i*. In this context, the question of interest is whether the *observed 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., from the manufacturing employment distribution).

4.1.1 *K-Functions*

The most popular measure of clustering for point processes is Ripley's (1976) *K*-function, $K(d)$,[6] 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 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 [7]

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

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

[7] 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.

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

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

yields 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 there is estimated to be only a one-in-a-hundred chance of observing a value as large as $\hat{K}^0(d)$ under H_0. Thus it may be inferred that at scale *d* there is significant clustering of R&D locations at the 0.01 level. Finally, it should be noted that since the observed value itself is always included in the count, $N^0(d)$, the smallest possible *p*-value given *N* simulations is $1/(N+1)$. For $N = 999$ this yields a minimal *p*-value of 0.001.

4.1.2 *Test Results*

This 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). However, clustering in the present case is so strong relative to manufacturing employment that the estimated *p*-values, $P(d)$, were uniformly 0.001 for all $d = 1,..,100$. In principle it is possible to perform larger numbers of simulations than $N = 999$; but the choice of *N* was deemed to be 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] showed that they were well approximated by the normal distribution for essentially all distances used. This suggested that sharper discrimination could be achieved by calculating the *z*-scores for each observed estimate, $\hat{K}^0(d)$, as given by

$$z(d) \;=\; \frac{\hat{K}^0(d) - \overline{K}_d}{s_d} \;\;,\;\; d = 1,..,100 \tag{3}$$

where $\overline{K}_d$ and s_d are the corresponding sample means and standard deviations for the $N+1$ sample *K*-values. These *z*-scores turned out to yield much sharper (and more interesting) results,

as shown in Figure 4. Notice first that the lowest *z*-score is already more than seven standard deviations above the mean, which explains the constancy of *p*-values above.[8] Next observe that in relative terms, clustering is by far most significant at very small scales of around one-quarter mile. This will be examined in more detail in subsequent sections. Notice also that there is a secondary mode of significance at a scale of about 40 miles (which, in terms of standard deviations, is seen to be about half as pronounced as the primary mode). This will be seen to correspond roughly to the scales of the four major R&D clusters identified in the next section.

4.2 Local Cluster Analysis

While the global analysis above serves to indicate that there is very significant clustering of R&D locations relative to manufacturing employment, it provides little more information other than the spatial scale (in this case distances) at which clustering appears to be most significant. In this section, we focus on the size and location of specific clusters.

4.2.1 *Local K-Functions*

The main tool for accomplishing this is the *local* version, $\hat{K}_i(d)$, of sample *K*-functions for individual pattern points, i, (first introduced by Getis, 1984),[9] which in our present notation is simply the count of additional pattern points within distance d of i, i.e.,[10]

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

Here our basic null hypothesis, H_0, is the same as above. So for any given point i in the observed pattern X^0, the question is whether the associated local *K*-value, $\hat{K}_i^0(d)$, is significantly larger than would be expected under H_0. If so, then one may infer the presence of *local clustering* within a circle of radius d about location x_i. To test for such clustering, the only substantive difference from the procedure above is that location x_i is now held fixed. So 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. Here the appropriate *p*-values for a one-sided test of H_0 with respect to point i now take the form,

---

[8] Based on a *p*-value of approximately 10^{-19} for a z-score of 9.0, one would require a simulation sample size of $N \geq 10^{19}$ before observing any difference in clustering significance between scales.

[9] 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).

[10] 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.

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

where $N_i^0(d)$ is again the number of these $N+1$ at least as large as $\hat{K}_i^0(d)$. Perhaps the single most attractive feature of such local tests is that the resulting *p*-values for each point i in the observed pattern can be *mapped*. This allows one to check for point clustering visually. In particular, groupings of very low *p*-values serve to indicate not only the location but also the size of possible clusters. While such groupings based on *p*-values necessarily suffer from "multiple testing" problems (discussed below), they nonetheless exhibit a qualitative structure that is very informative.

4.2.2 *Test Results*

In this local setting, simulations were again 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 all local tests was $D = \{0.25, 0.5, 0.75, 1, 2, 5, 10, ..., 100\}$ (with 1-mile increments between 10 and 100). Unlike the global results above, where the *p*-values in (2) provided no discriminatory power between different scales, the local results are very different. In particular, it is not surprising to find many isolated R&D locations that exhibit no local clustering whatsoever so that wide variations in significance levels are possible at any given spatial scale. Moreover, it is also natural to expect variations between different scales. Roughly speaking, at very small scales (say less than one-quarter 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. Hence from a visual perspective, the most interesting scales are those intermediate scales at which one begins to see more "coherent" clusters.

In the present case, a visual inspection of *p*-value maps shows that the clearest patterns of distinct clustering can be captured by the smaller set of distances $\{1,5,10\}$. Of these three, the single best distance for revealing the overall clustering pattern in the entire data set appears to be 5 miles. (The other two radial distances are useful for more local structural comparisons and will be applied below.) The *p*-value map at the 5-mile scale is shown in Figure 5. 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.[11] But while these results are visually compelling, it is important to ask whether they can be established by more analytical methods.

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

4.3 Cluster Identification

The global cluster analysis in Section 4.1 was able to identify the *scales* at which clustering is most significant (relative to manufacturing employment). In addition, the local cluster analysis in Section 4.2 added further information about *where* clustering is most significant at each scale. But neither of these methods actually identifies "clusters" explicitly. Moreover, both of these methods suffer from multiple-testing problems, in that many closely related tests are performed. In this section, we consider the well-known "significance-maximizing" approach to clustering that addresses both of these shortcomings.

4.3.1 *Multiple-Testing Problem*

First we must consider the problem of multiple testing in more detail. While global cluster analysis necessarily involves simultaneous tests over a range of spatial scales, $d \in D$, here we focus on local cluster analysis, where the problem of multiple testing is most severe. In this setting, even if it were true that (1) there were no local clustering at all (so that the observed pattern X^0 of R&D locations was in fact consistent with H_0), and (2) all local *K*-function tests were statistically independent, one would still expect by definition that 5 percent of such tests would be significant at the .05 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 discovery rate" can in principle be corrected by reducing the *p*-value threshold level deemed to be "significant" (and indeed this is one reason for focusing only on *p*-values no greater than 0.005 in Figure 5).

But in the spatial setting of local *K*-function analysis, the situation is even more complex. For, as mentioned above in the extreme case of large radial distances, d, the physical overlap between d-neighborhoods of points i and j within distance $2d$ of each other guarantees that their estimated *p*-values, $P_i(d)$ and $P_j(d)$, will be *statistically dependent*. Hence 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.[12]

4.3.2 *The Significance-Maximizing Approach*

These problems are by no means new, and indeed, a number of approaches have been developed for resolving them. Perhaps the best known are the original work of Besag and Newell (1991) and the more recent work of Kulldorff (1997). Both of these procedures seek to 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 by which 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

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

same null hypothesis, H_0, above.[13] 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 .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.

4.3.3 *Results of the SATSCAN Procedure*

We have applied both the Besag-Newell procedure and Kulldorff's SATSCAN procedure to our present data and found them to be in remarkably good agreement with each other. So here we shall only present the results of the (more popular) SATSCAN procedure. In this setting, we ran the maximum of 10 iterations allowed by the SATSCAN software[14] and have displayed the union of the resulting 10 clusters in Figure 6. The purpose of this summary display is to show by comparison with Figure 5 that essentially the same four main areas of significant concentration are obtained by both local *K*-function and significance-maximizing methods.

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 7 (where the slightly darker zip code in the center is the starting seed). This cluster is essentially the "Boston cluster," referred to in Figure 5 above. For purposes of comparison, the Boston area of Figure 5 has been superimposed on Figure 7 to show that the two most significant groupings of R&Ds in the Boston area are essentially contained in this cluster.

This leads naturally to the question of why these two distinct groupings should constitute a *single* cluster. This is essentially due to the approximately *circular* shapes of candidate clusters defined by this particular SATSCAN procedure.[15] In particular, no circular approximation to either of these two groupings is more significant than the single circular cluster shown.[16] An even more dramatic example is provided by the single largest cluster in the New York area, just north of

---

[13] 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 in Section 4.1 above to simulate the sampling distribution of this statistic under H_0.

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

[15] 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).

[16] Note that the cluster in Figure 5 appears to be somewhat flatter than a circle. This is a consequence of the decimal-degree coordinate system used in the map displays. While it is possible to eliminate this distortion by taking separate state-plane projections of each cluster area, we have chosen to maintain a single coordinate system for all graphic representations (other than the Lambert Conformal Conic projection used for the entire US in Figure 1).

New York City in Figure 6,[17] which is shown enlarged in Figure 8 (where the adjacent cluster to the northeast has been removed for sake of clarity and where again the relevant portion of Figure 5 has been superimposed). Here it is evident that *all* significant concentration of R&D labs (at scale $d = 5$ miles) lies along the southern edge of this cluster. Hence, while there is a smaller concentration of labs in the east central portion, it is clear that attempts to capture these concentrations by circular clusters have led to a gross exaggeration of the actual pattern. These two examples should be enough to indicate that the restriction to circular candidate clusters can be very limiting indeed.[18]

In addition to this shape limitation, the sequential nature of cluster identification in these procedures also 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.[19]

4.4 A Multiscale Core-Cluster Approach

Given these limitations of the significance-maximizing approach, it is useful to consider an alternative approach to cluster identification that explicitly uses the *multiscale* nature of local *K*-functions. This clustering procedure starts with the results of the local point-wise clustering procedure in Section 4.2 above and seeks to identify subsets of points that can serve as "core" cluster points at a given selection of relevant scales, d . As in Section 4.2, we focus here on the three scales, $d \in \{1,5,10\}$, that appear to capture the essential substructure of the four main clusters in Figure 5.[20] In most of the procedural 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 required to have maximal significance relative to the present Monte Carlo simulation, i.e., a *p*-value of .001, indicating stronger observed clustering within a *d*-mile radius than in each of the 999 simulations.[21] In

[17] It should be noted that this is only the fourth most significant SATSCAN cluster. The second and third most significant clusters correspond, respectively, to the cluster due west of New York City and the Washington cluster to the south.

[18] Here it should be emphasized that the SATSCAN testing procedure itself can in principle be applied to any set of candidate clusters. In fact, more general "elliptical" clusters are already available in certain SATSCAN modeling options other than the "continuous Poisson" option used here.

[19] Here again it should be noted that such path dependencies can in principle be removed by broadening the class of candidate clusters to include "cluster ensembles," thereby reducing the problem to finding the single most significant cluster ensemble. However, global optimization quickly becomes intractable in these large configuration spaces, and in fact, is precisely the motivation for the sequential approximations used in this significance-maximizing approach. For a local-optimization approach using such cluster ensembles, see Mori and Smith (2009).

[20] One might also include the extremely local scales of $d = .25$ and $d = .50$ miles, since core points at these scales contribute substantially to the global z-score peak near zero in Figure 4. But since all but four of these core points are already core points at scale $d = 1$ mile (by our definition below), they are well represented by the core points at this slightly larger scale. Moreover, these four exceptional core points are all in the same "Central Park" cluster, as illustrated in Figure 10 below.

[21] The use of 999 simulations was designed to maintain comparability with the SATSCAN results, where 999 is the maximum allowable number of simulations. However, to ensure stability of the present choice of core points based on the extreme *p*-value of .001, the local cluster simulations in Section 4.2 were rerun using 9,999 simulations. The core points obtained from this larger simulation were then defined by a *p*-value threshold of .0014 (i.e., all values

order to exclude "isolated" points that simply happen to be in areas with little or no manufacturing, it is also required 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*.[22] The advantages of this core cluster approach are best illustrated by examples, as we now show.

To begin with, recall that the single most significant cluster identified by SATSCAN is the Boston cluster shown in Figure 7 above. Here it was noted that there are two distinct concentrations of R&D labs within this cluster.[23] This observation can now be made sharper by our multiscale representation, as seen in Figure 9. Here the core points at each scale $d = 1, 5, 10$ are shown, along with their corresponding core clusters. For example, at the 5-mile scale we now 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 5, 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. Here 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 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 near zero in the global *z*-scores of Figure 4. In particular, there is a highly concentrated cluster of 17 labs just south of Central Park in New York City, as depicted in Figure 10. Here we have shown 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

---

rounding down to .001 or less). With only minor exceptions, these final core points agreed with those obtained from the original 999 simulations.

[22] 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.

[23] The specific nature of these labs will be discussed more fully in Sections 5.1 below. For the moment, we focus only on their spatial configuration.

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 shaded background shows the relative levels of manufacturing in each zip code. The darkest of these is actually the single *highest* manufacturing zip code (more than 22,000 employees) in all of the 6,043 zip codes in this study area. 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).[24]

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.

But it is equally important to emphasize that this multiscale approach should not be regarded as a substitute for more standard approaches such as significance-maximizing. 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 parameters values by which they are defined (here 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).[25] Hence the main objective of this procedure is to yield *visual* representations of clusters that capture both their relative shapes and concentrations in a natural way.[26] Since there is no universally accepted definition of "clusters," it seems most prudent to analyze this problem from many viewpoints and look for areas of substantial agreement among them.

[24] 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.

[25] Of course, one can carry out informal "sensitivity analyses" with respect to alternative choices of these parameters. Our present choices are in fact the result of such investigations and were deemed to be best representatives on this basis.

[26] The situation here is somewhat analogous to choropleth map representations of surfaces, where choices of scale divisions (such as by "natural breaks") are intended simply to capture overall variations in a visually informative way.

5. SUMMARY AND DISCUSSION OF AGGLOMERATION RESULTS

To summarize the overall structure of R&D agglomerations identified within our 10-region study area, we begin in the following section with a discussion of our main findings at the global level. This is followed in Section 5.2 with a more detailed discussion of the internal spatial structure of the four major agglomerations found at the metropolitan level. In particular, we identify the primary research areas associated with individual core clusters of labs. In Section 5.3 we relate these spatial structures to key local geographic features such as proximity to freeways and the presence of university centers. Finally in Section 5.4, we briefly compare the spatial structures of those R&D labs with primary research areas in specific industries.

5.1 Forces for agglomeration of R&D labs

Our most important finding from the global *K*-function analysis in Section 4.1 above 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 (as in Figure 4). Moreover, there is a rapid decline in significance up to about 5 miles. This is consistent with the mounting evidence for attenuation of human capital spillovers at small spatial scales found in the studies of the concentration of manufacturing employment (Rosenthal and Strange, 2001 and 2008, and Elvery and Sveikauskas, 2010); in studies of innovative activity (Audretsch and Feldman, 1996; Keller, 2002; and Agrawal, Kapur, and McHale, 2008); and in a study of the concentration of the advertising industry in New York City (Arzaghi and Henderson, 2008).

We also observe a secondary mode of significant clustering for all labs taken together at about 40 miles, or roughly at a scale corresponding approximately to a local labor market (metropolitan area). This secondary cluster is consistent with the view that matching and input sharing externalities are likely relevant at the spatial scales of those markets. This finding corroborates the positive effects of agglomeration economies at greater distances found by Rosenthal and Strange (2008) and Elvery and Sveikauskas (2010).

The pattern of clustering observed for R&D labs is also evident in the clustering patterns of labs in individual industries. Figure 11 presents graphs of the *z*-scores for the seven three-digit SIC industries (inorganic chemicals, plastics, drugs, organic chemicals, miscellaneous chemicals, electronic components, and computer programming and data processing) that account for the largest share of labs in our sample.[27] Because we are especially interested in the attenuation of *z*-scores at small scales, Figure 11 shows graphs of *z*-scores for the seven industries in increments of 0.25 miles up to 5 miles. [28] In each of the seven industries, clustering of R&D labs is by far most significant at very small spatial scales — a half mile or less. Figure 11 also reveals a very rapid distance decay of the *z*-scores for each of the seven industries in that the *z*-scores fall very dramatically during the first 5 miles of distance between labs. The rapid spatial attenuation of *z*-scores for labs in individual industries is consistent with the view that at least one important component of knowledge spillovers must be highly localized.

[27] Here we grouped labs by the industries listed as their primary research areas.

[28] The dashed line shown in Figure 11, as well as in Figure 12, represents a value of 1.96. Thus, *z*-scores at or above the dashed line are taken as significant in a statistical sense.

The graphs in Figure 12 show the clustering of labs between 1 mile and 100 miles in increments of 1 mile. As found for all labs taken together (Figure 4), we also find a secondary mode of a significant clustering of labs for each of the seven industries. As found for all R&D labs (Figure 4), this secondary mode of significance occurs at scales of about 40 miles for labs doing R&D in four industries (plastics, drugs, industrial organic chemicals, and electronic components). For miscellaneous chemical products and industrial inorganic chemicals, the secondary mode of significance occurs at scales of around 70 miles. In the case of computer programming, the secondary node occurs at a distance of about 13 miles. This finding of a secondary mode of significance at scales roughly associated with metropolitan areas is consistent with agglomeration economies associated with sharing and matching of workers among labs, which can be afforded in broad local labor markets.

5.2 Major Areas of Agglomeration

To summarize our results on the spatial location and extent of R&D agglomerations, we focus on our multiscale representation in terms of core clusters developed in Section 4.4 above. As a parallel to the graphical summary of SATSCAN results in Figure 6, a composite of all core clusters at scales of $d = 1, 5, 10$ miles are shown in Figure 13, where for example the outer gray contours correspond to core clusters at scale $d = 10$. A comparison with Figure 6 shows that areas of significant agglomeration are in general agreement with SATSCAN, but that core clusters are by construction closer in shape to the patterns seen in the local *K*-function results for $d = 5$ in Figure 5. Here again Boston, New York, Philadelphia, and Washington, DC, are seen to be the major areas of agglomeration. Hence we now proceed to discuss each of these areas in more detail (for convenience, core clusters shall be referred to simply as clusters).

5.2.1 *The Boston Agglomeration*

There are 187 R&D labs within Boston's single 10-mile cluster, as shown in Figure 9 above.[29] 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 in Figure 9 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.

5.2.2 *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 14) 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 thirty-seven (16 percent) do research on industrial chemicals. Within this highly elongated 10-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

[29] The map legend in Figure 9 applies to all map figures in this section.

"Central Park" cluster shown in Figure 10 above. About two-thirds of the 17 labs in this cluster are conducting research on drugs, perfumes and cosmetics, or computer programming and data processing.

5.2.3 *The Philadelphia Agglomeration*

As seen in Figure 15, there is a large 10-mile cluster to the west of Philadelphia (where the city of Philadelphia is shown in darker gray). Here there are a total of 49 labs, with 16 labs conducting research on drugs, and another 16 labs doing 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.

5.2.4 *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 16 (with the city of Washington, DC, in darker gray). Here, 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. [30]

5.2.5 *The Pittsburgh Area*

In addition to these four major areas of agglomeration, notice from Figure 13 above 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 17 (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). Here 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.

[30] 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.

5.3 The Importance of Highways and Universities

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

5.3.1 *Boston Area*

A prime example is provided by the locations of R&D labs in the Boston area. As seen in Figure 9, 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 18, 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 9 are seen in Figure 18 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 9 is seen to be centered precisely on the intersection of Interstate 90, with the outer circumferential highway being Interstate 495.

5.3.2 *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 14 is seen from Figure 19 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.

5.3.3 *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 of Figure 15 with the major routes shown in Figure 20. 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 15 (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.

5.3.4 *Washington, DC, Area*

Finally, in the metropolitan area of Washington, DC, we see from a comparison of Figure 16 and Figure 21 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 exhibit close proximity 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). [31]

5.4 Relative Clustering of R&D Labs by Industry

Turning finally to a consideration of R&D concentrations by industry, we begin by summarizing our observations in Section 5.2 above. We found that within our study area, the main areas of R&D activity are in drugs, chemicals, plastics, and computer programming and data processing. Spatially, the labs that conduct R&D in drugs tend to be concentrated in Boston, New York City, and Philadelphia. Those labs that are conducting R&D in chemicals are mostly concentrated in New York City and Pittsburgh. Those labs that are involved with plastics are mostly concentrated in New York City, Philadelphia, and Washington, DC. Finally, those labs that are doing R&D in computer programming and data processing are mostly concentrated in Boston, Washington, DC, and New York City.

Recall also from Section 5.1 that we examined global clustering patterns for labs in the seven three-digit SIC industries that accounted for the largest share of R&D labs in our sample. In addition to the results for individual industries, we also carried out a global analysis of R&D clustering within industries *relative* to R&D as a whole. Here we grouped labs in terms of their primary industrial research areas at the two-digit SIC level.[32] A variant of the global *K*-function procedure above was then used to identify scales at which various industries were more clustered than would be expected for comparable random samples of R&D labs from the full population of 1,035 labs.[33] Here our main finding was that the chemical and allied products industry (SIC 28) is not only the single largest area of R&D [with more than twice as many labs (463) as any other two-digit SIC], but it also exhibits the most significant concentration at both small scales (under a mile) and large scales (around 40 miles) relative to R&D labs as a whole. Hence, at least within our study area, this single industry appears to be a major contributor to the overall clustering pattern of R&D shown in Figure 4.

[31] As Figure 22 reveals, the 5-mile core cluster just west of the city of Pittsburgh (as seen in Figure 17) is almost precisely at the intersection of two major routes (Interstate 279 and Interstate 79) .

[32] The two-digit level was used here to achieve sufficient sample sizes for testing purposes. This yielded 23 industrial groups with corresponding SIC designations: 10,13,16,20-23,26-30,32-39,50,73, and 87.

[33] In particular, this identification procedure was carried out in terms of standard random-permutation tests based on global *K*-function statistics.

6. CONCLUDING REMARKS

In this article, we use distance-based techniques to analyze the spatial concentration of the locations of over 1,000 R&D labs in a nine-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).

There are two important findings that 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. Here 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 designated as the multiscale core-cluster approach. 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.

There is a rich theoretical literature describing a variety of mechanisms for agglomeration economies as well as knowledge spillovers that may explain the spatial distribution of both production and innovation. Empirical tests of these theories are beset with the problem of observational equivalence. How can we identify the precise mechanisms at work? The results of this paper suggest that one approach to this problem is to examine locational configurations of firms at a variety of spatial scales. In particular, our results suggest two possibilities. Perhaps there are two mechanisms of spillovers that explain patterns of clustering at two distinct scales (next door and at the level of local input markets). Or there could be a single mechanism which, for reasons not well understood, operates simultaneously at two spatial scales. The former suggests that empirical work must be sufficiently flexible to allow for multiple mechanisms, both for the purposes of identification and to avoid confounding results. The latter suggests that a very particular form of spatial spillover is at work that could help to narrow the range of candidate theories to be tested.

An important next step in our investigations is to determine the extent to which the spatial concentrations of R&D labs identified in this paper contribute to the productivity of innovative activity. To address this issue, we are in the process of matching patent activity to individual R&D labs. In a future version of this paper, we hope to shed light on the role that localized R&D spillovers play on the patent productivity of labs.

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*, Vol. 64 (2008), pp. 258-69.

Anselin, L. "Local indicators of spatial association – LISA," *Geographical Analysis*, Vol. 27 (1995), pp. 93-115.

Arzaghi, Mohammad, and J. Vernon Henderson. "Networking Off Madison Avenue," *Review of Economic Studies*, Vol. 75 (2008), pp. 1011-1038.

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

Besag, J., and J. Newell. "The Detection of Clusters in Rare Diseases," *Journal of the Royal Statistical Society,* Vol. 154 (1991), pp. 327-333.

Buzard, Kristy, and Gerald A. Carlino. "The Geography of Research and Development Activity in the U.S.," Working Paper No. 09-16, Federal Reserve Bank of Philadelphia (2009).

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*, Vol. 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*, Vol. 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*, Vol. 105 (1997), pp. 889-927.

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

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

Fu, Shihe. "Smart Café Cities: Testing Human Capital Expenditure in the Boston Metropolitan Area," *Journal of Urban Economics*, Vol. 61 (2007), pp. 89-111.

Getis, A., "Interaction Modeling Using Second-Order Analysis", *Environment and Planning*, Vol. 16 (1984), pp. 173-183.

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

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).

Jaffe, Adam, M. M. Trajtenberg, R. Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 108 (1993), pp. 577-598.

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

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

Link, Albert, N. "Research, Science, and Technology Parks: An Overview of the Academic Literature," in Charles W. Wessner, Ed., *Understanding Research, Science and Technology Parks: Global Best Practice: Report of a Symposium*, Washington: The National Academic Press (2009), pp. 127-139.

Lychagin, Sergey, Joris Pinkse, Margaret E. Slade, and John Van Reenen. "Spillovers in Space: Does Geography Matter?" Working Paper 16188, NBER Working Paper Series (2010).

Marcon, E., and Puech, F. "Evaluating the Geographic Concentration of Industries Using Distance-Based Methods," *Journal of Economic Geography*, Vol. 3 (2003) pp. 409-428.

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).

Ripley, B.D. "The Second-Order Analysis of Stationary Point Patterns," *Journal of Applied Probability* Vol. 13 (1976), pp. 255–266.

Romer, Paul, 1990, "Endogenous technological change, *Journal of Political Economy* Vol. 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*, Vol. 64 (2008), pp. 373-389.

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

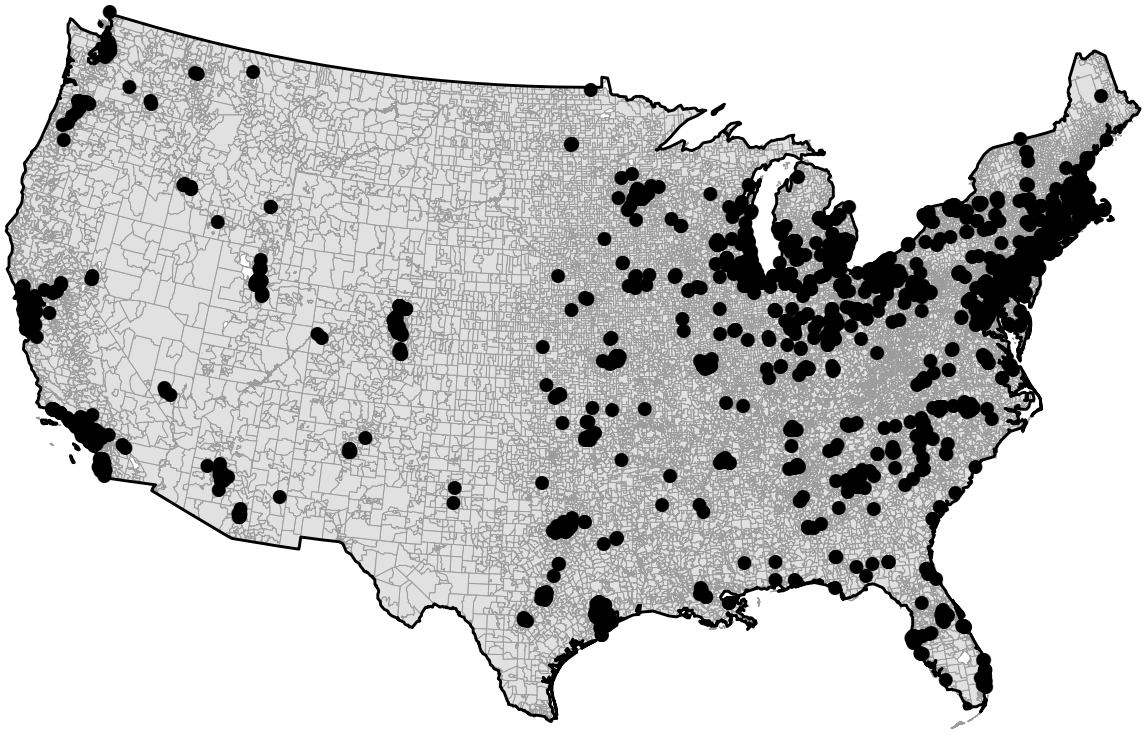

Figure 1. Location of R&D Labs

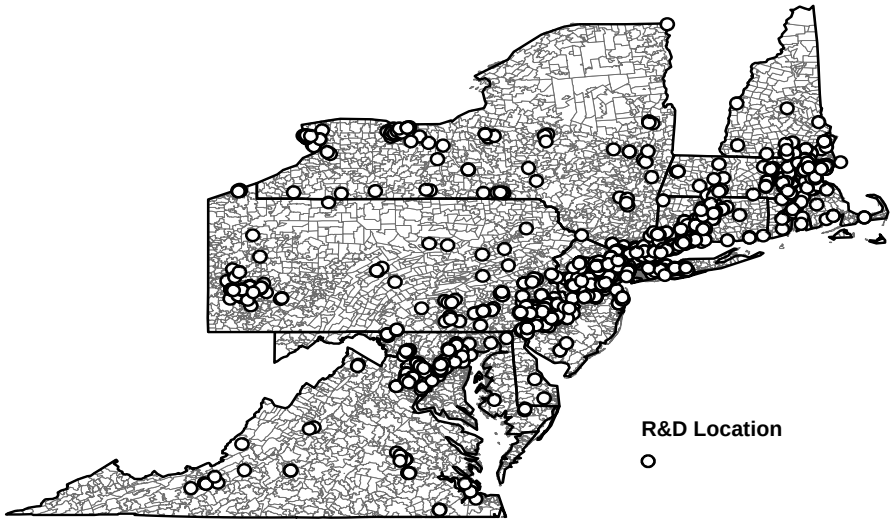

Figure 2. R&D Locations

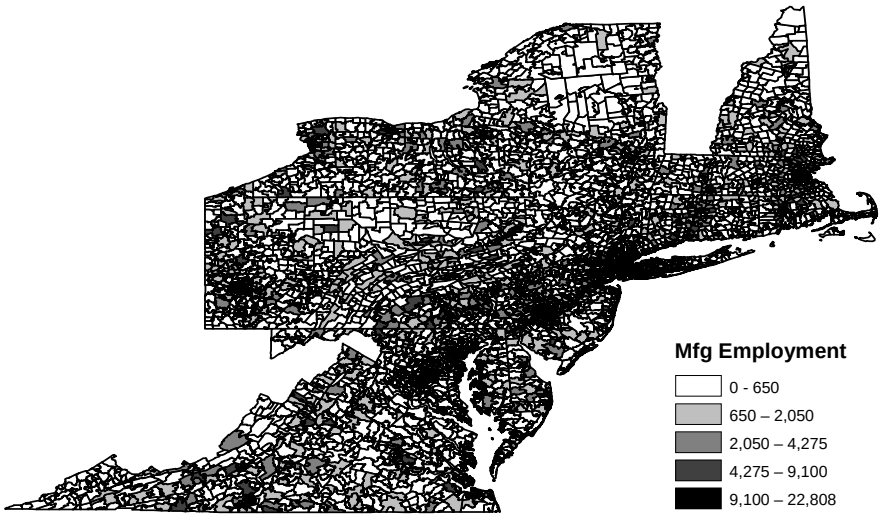

Figure 3. Manufacturing Employment

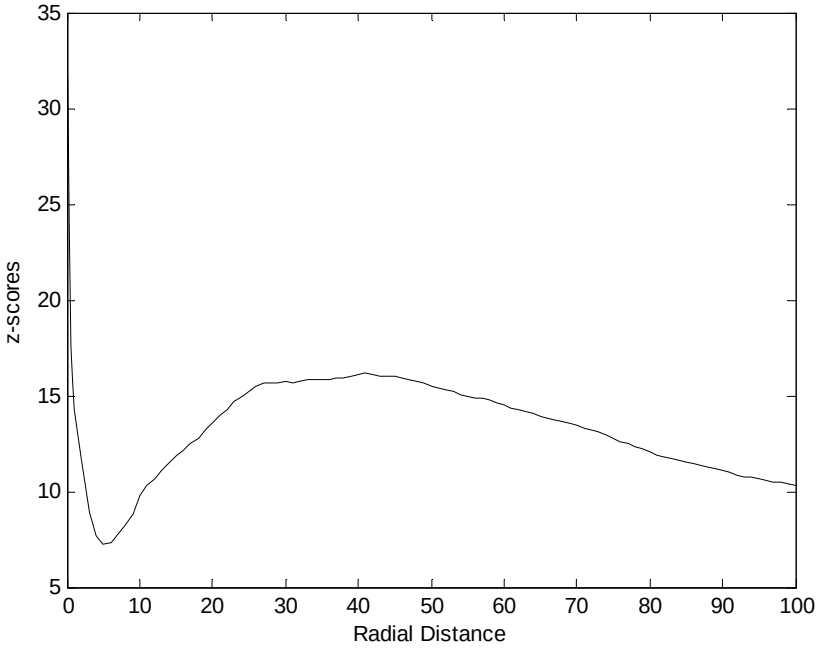

Figure 4. Clustering Z-Scores

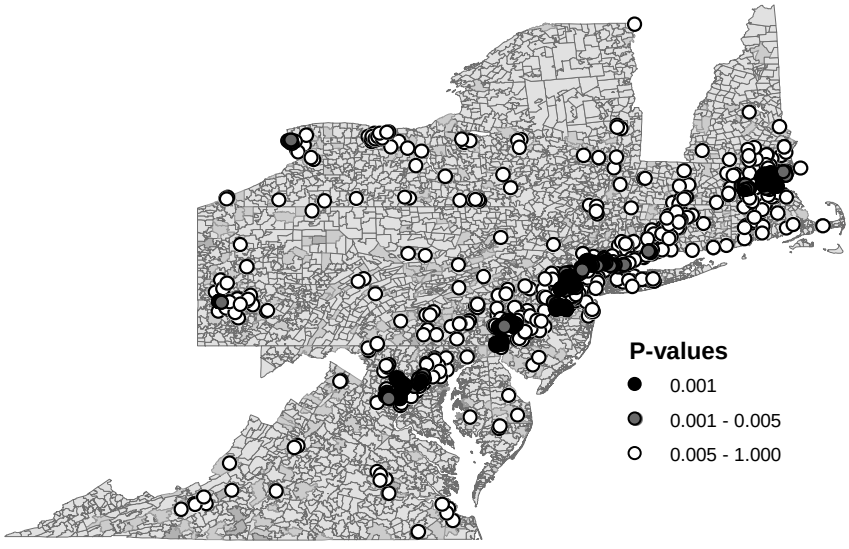

Figure 5. Local *K*-Function *P*-values at $d = 5$ Miles

Figure 6. Union of the Top 10 SATSCAN Clusters

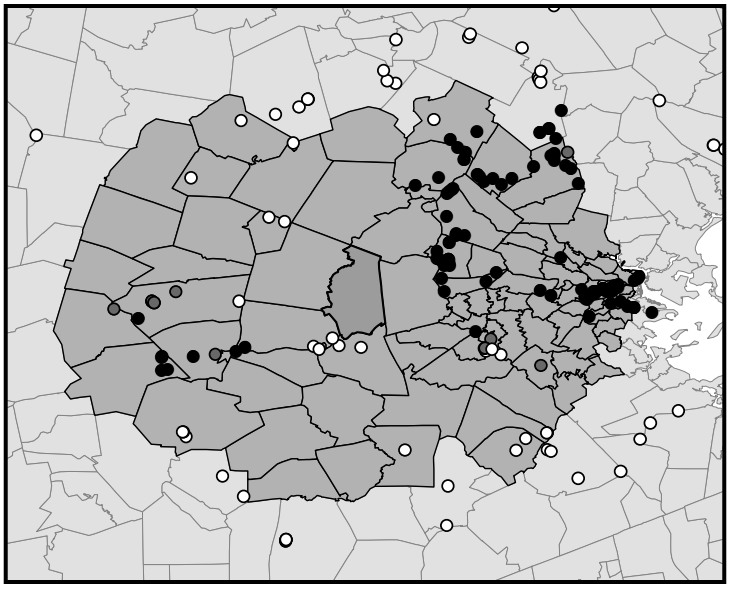

Figure 7. Boston Cluster in SATSCAN

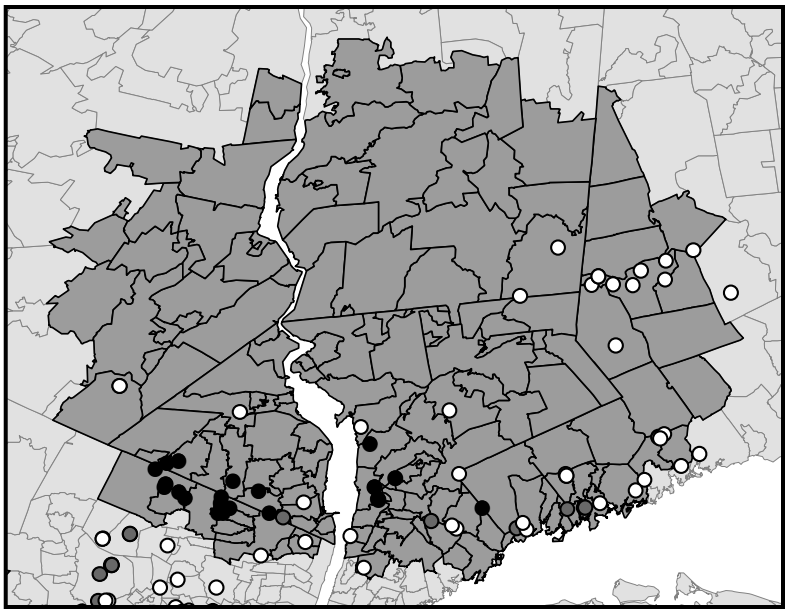

Figure 8. Largest New York Cluster in SATSCAN

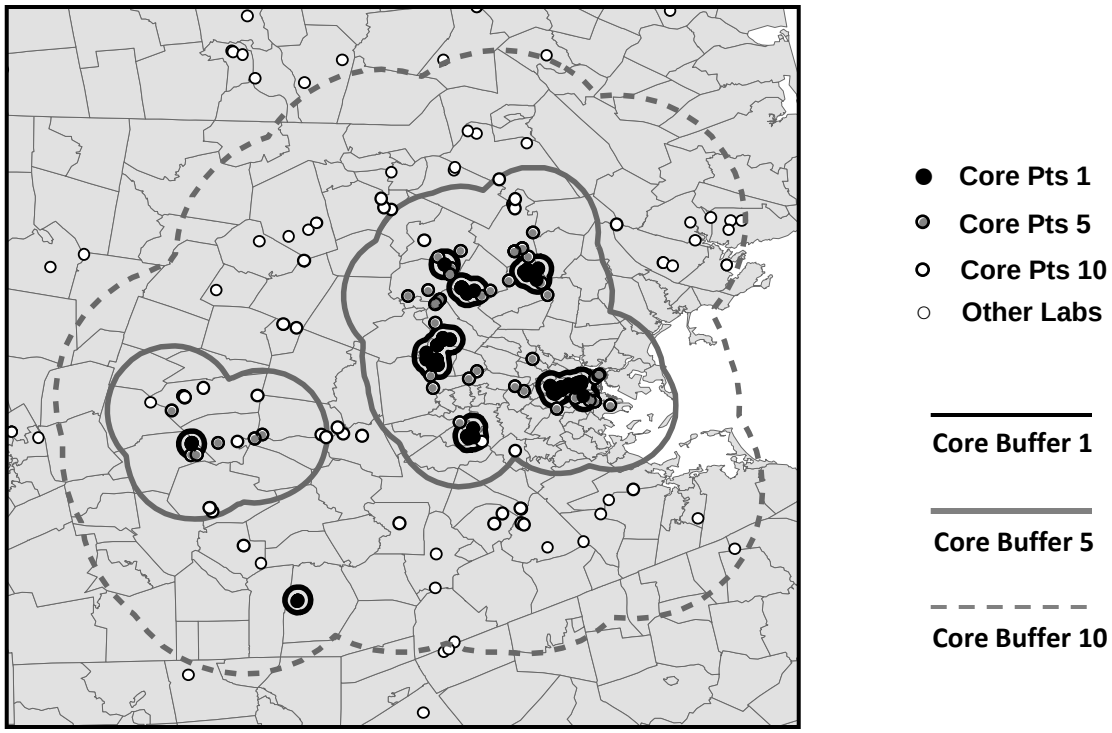

Figure 9. Boston Core Clusters

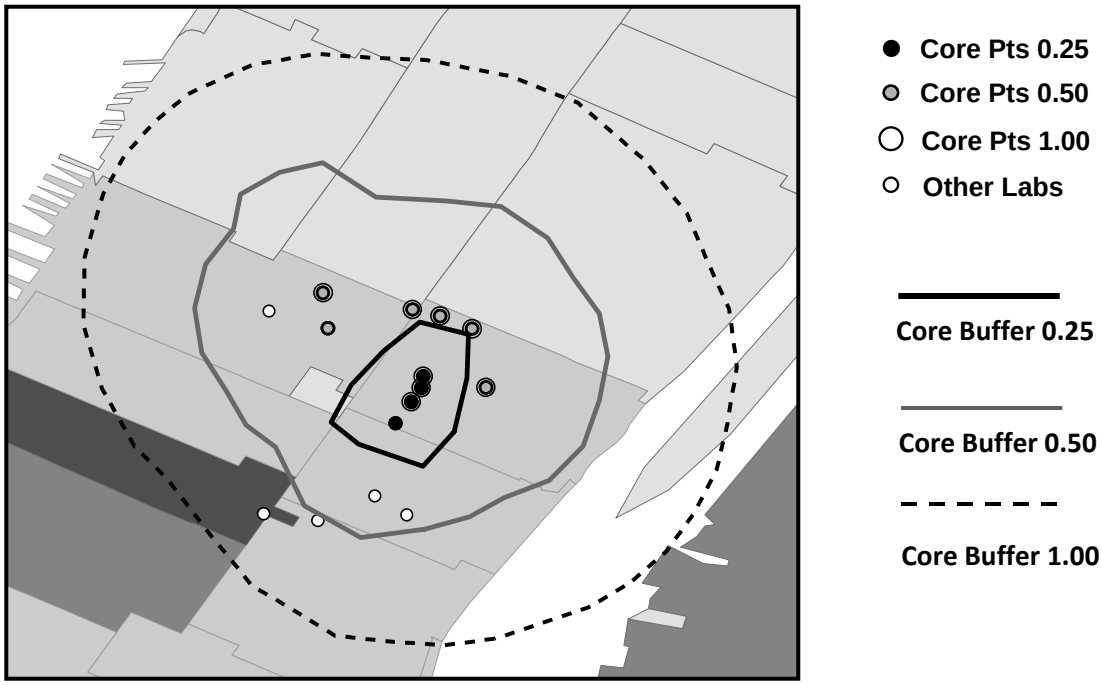

Figure 10. Central Park Core Clusters

Figure 11. Industry Z-scores

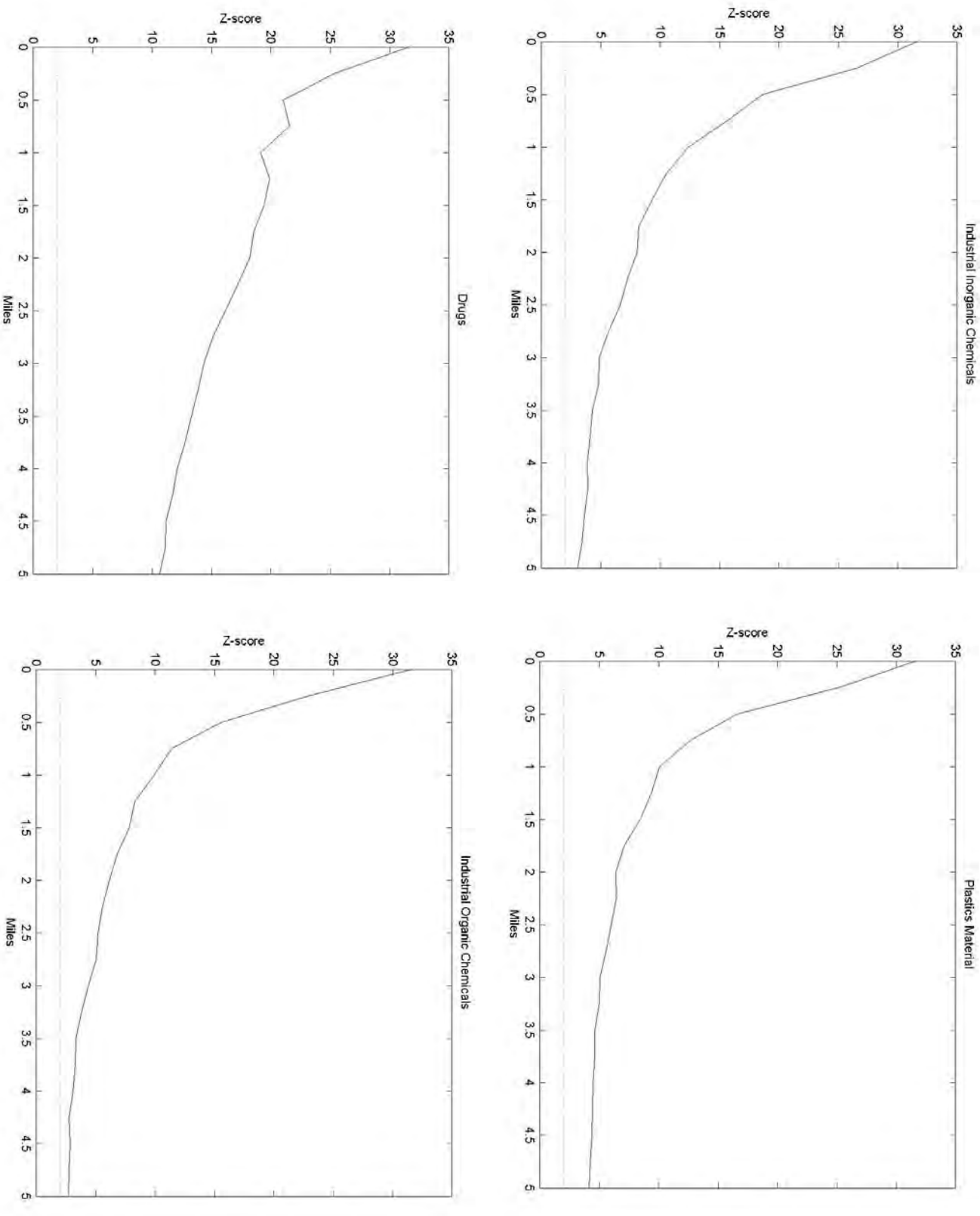

Figure 11. - Continued

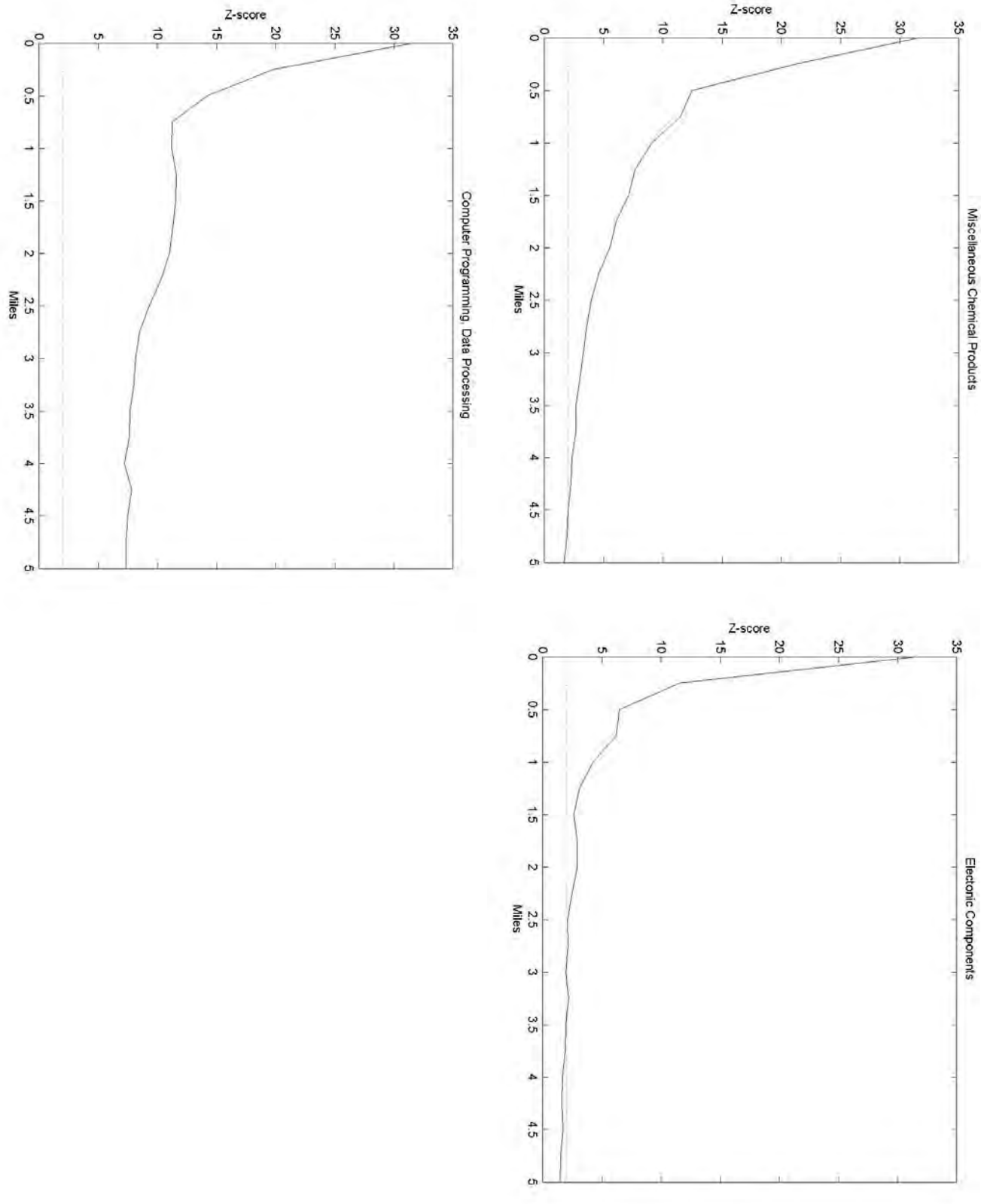

Figure 12. Industry Z-scores

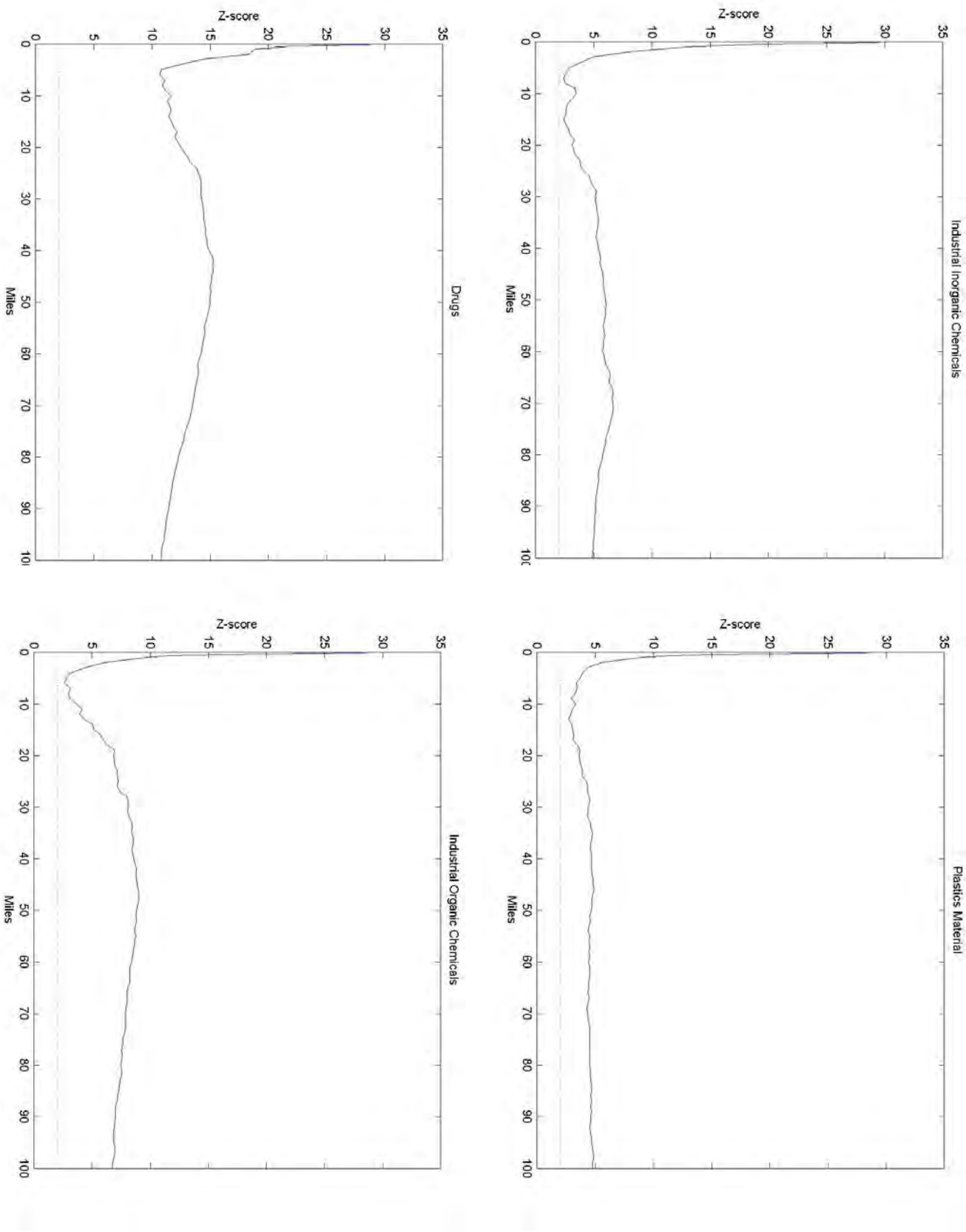

Figure 12. - Continued

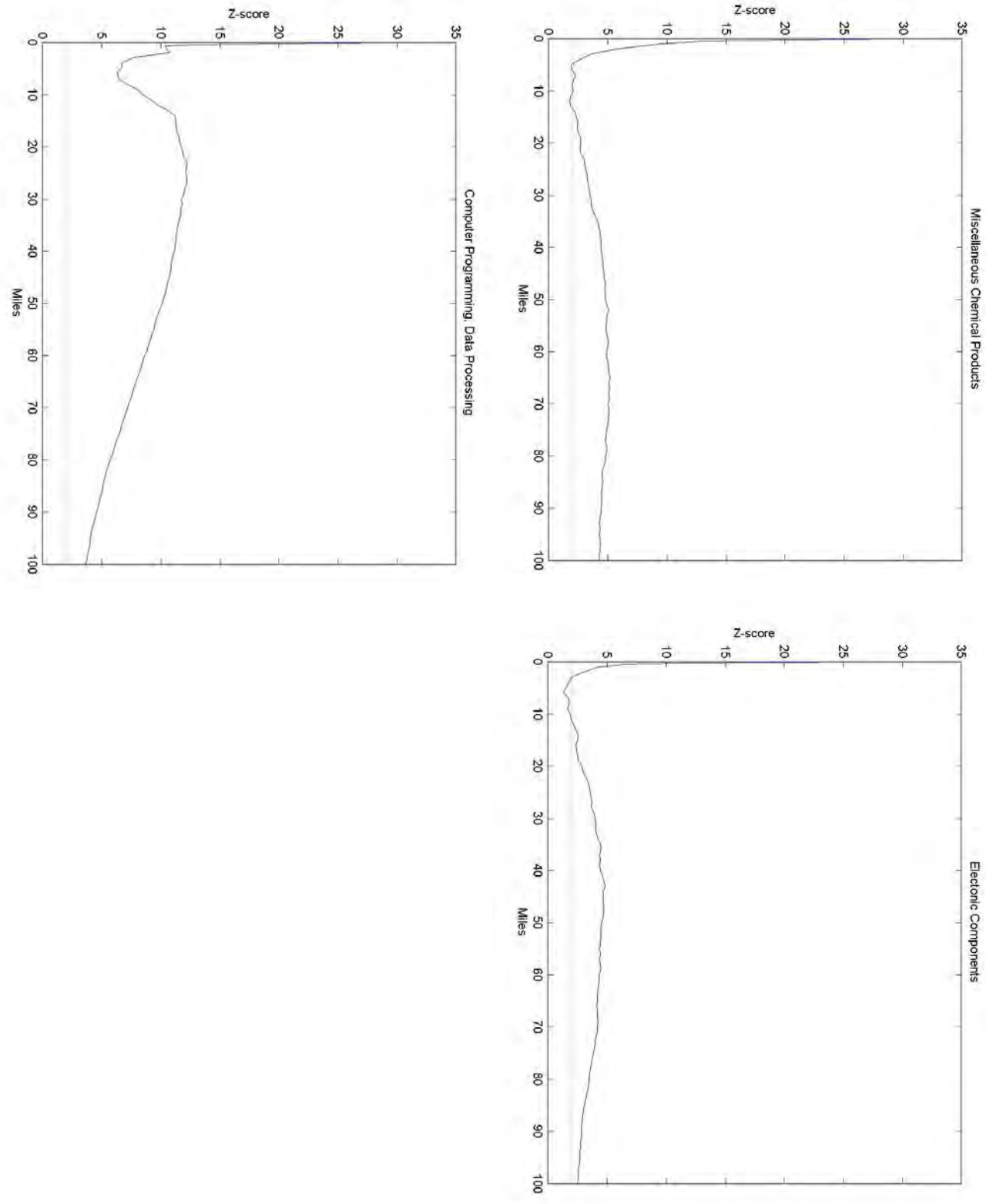

Figure 13. Multiscale Core Clusters (d = 1,5,10)

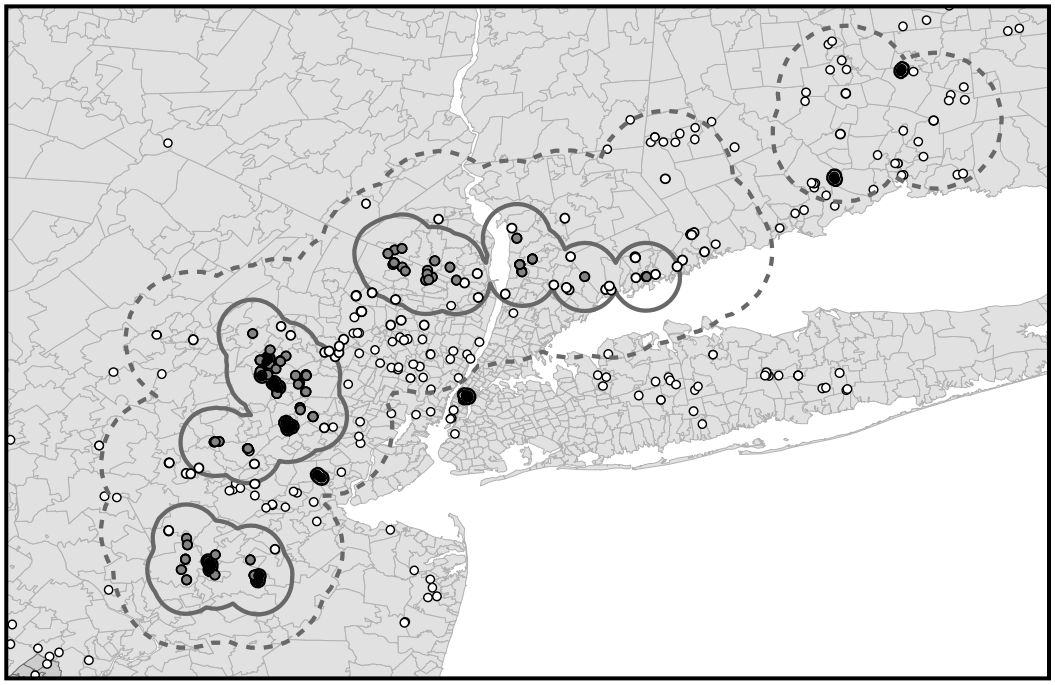

Figure 14. New York Core Clusters

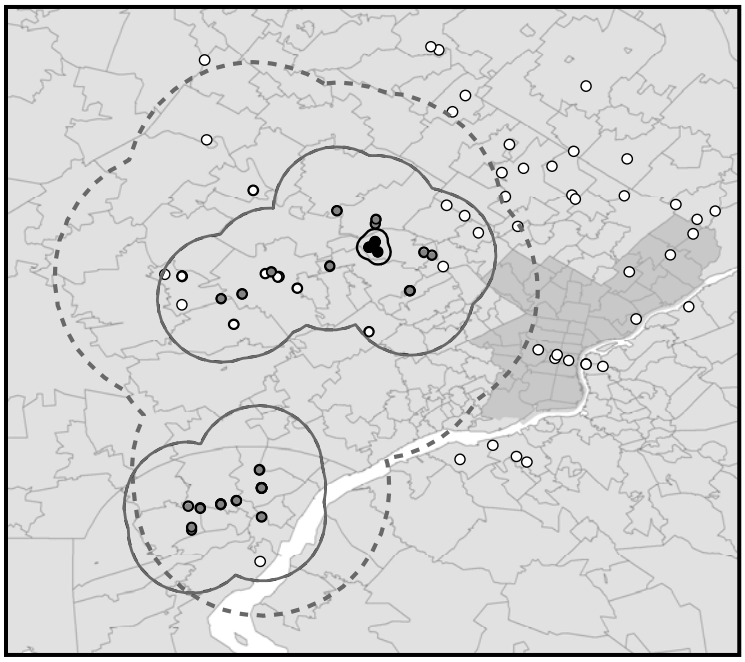

Figure 15. Philadelphia Core Clusters

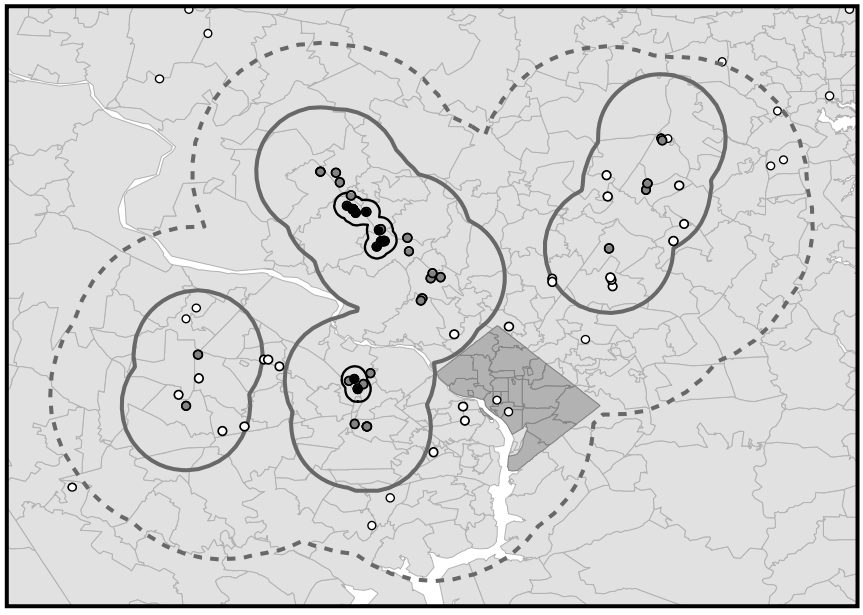

Figure 16. Washington, DC, Core Clusters

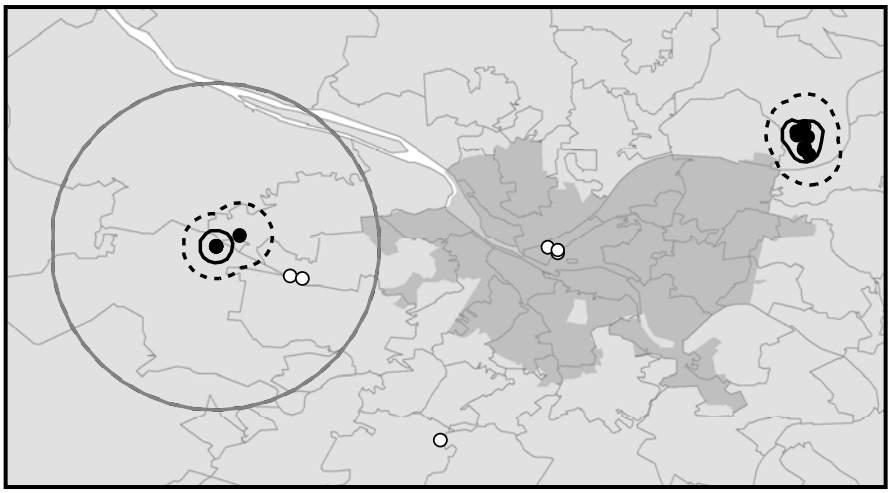

Figure 17. Pittsburgh Core Clusters

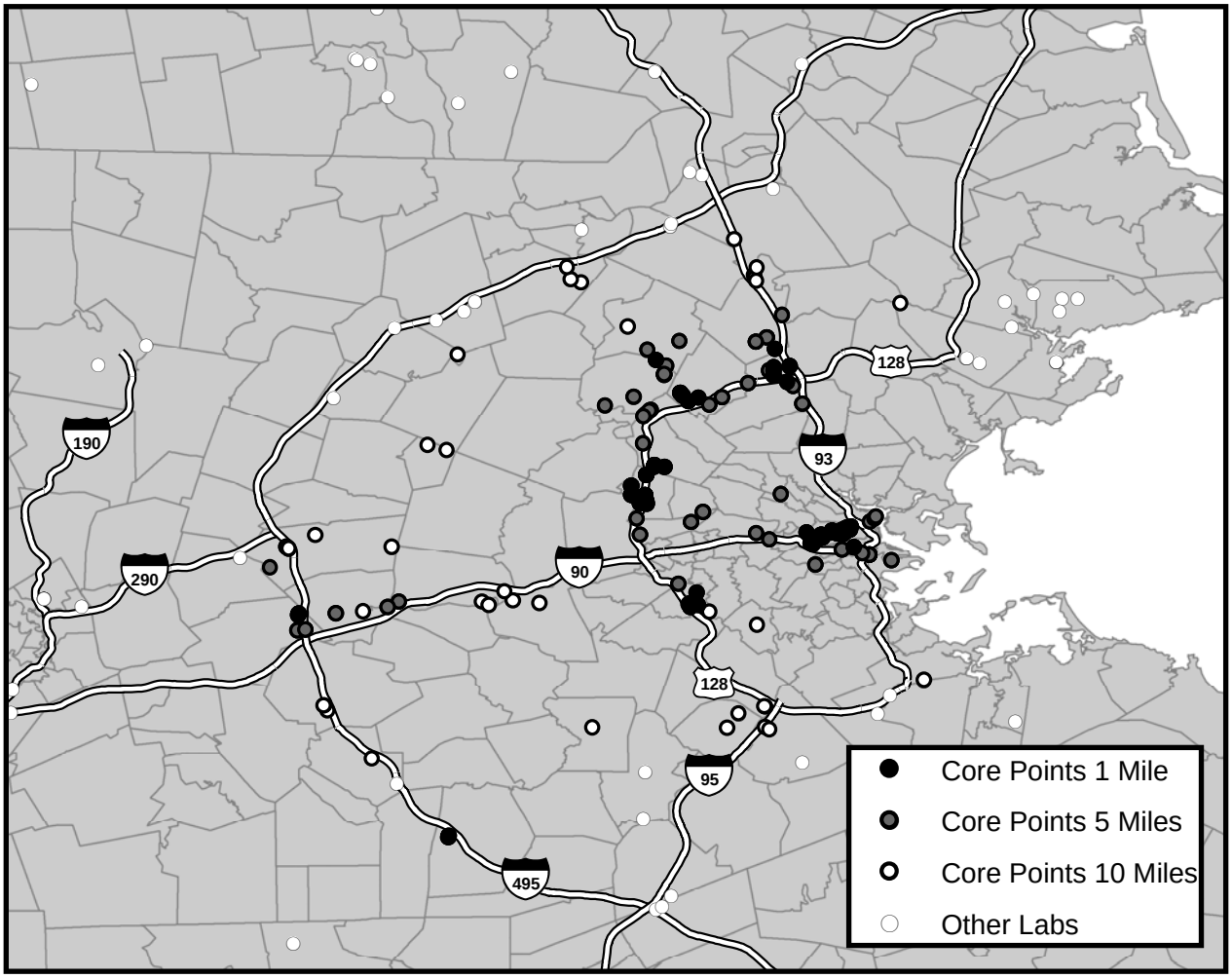

Figure 18. Boston Core Points and Major Routes

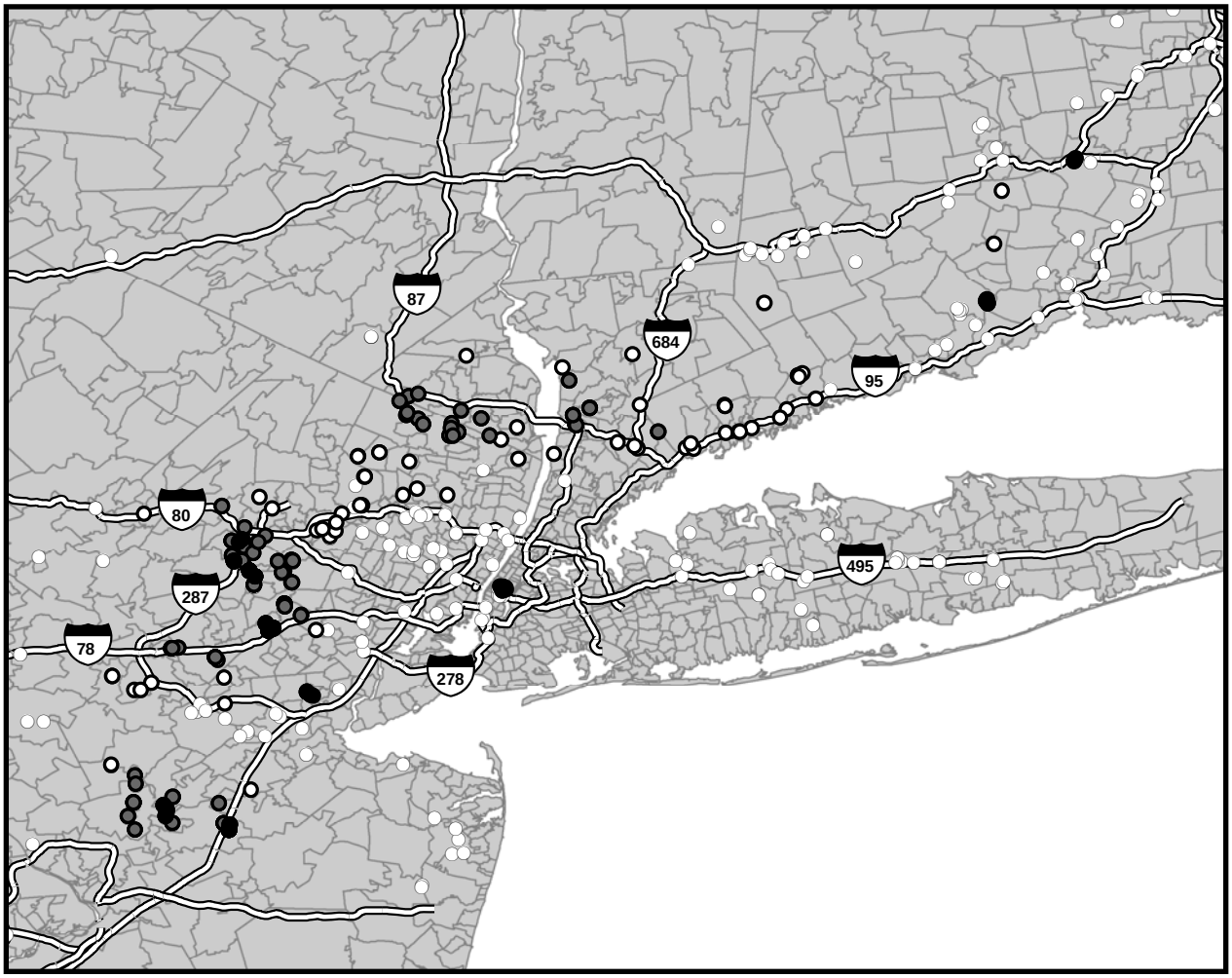

Figure 19. New York City Core Points plus Major Routes

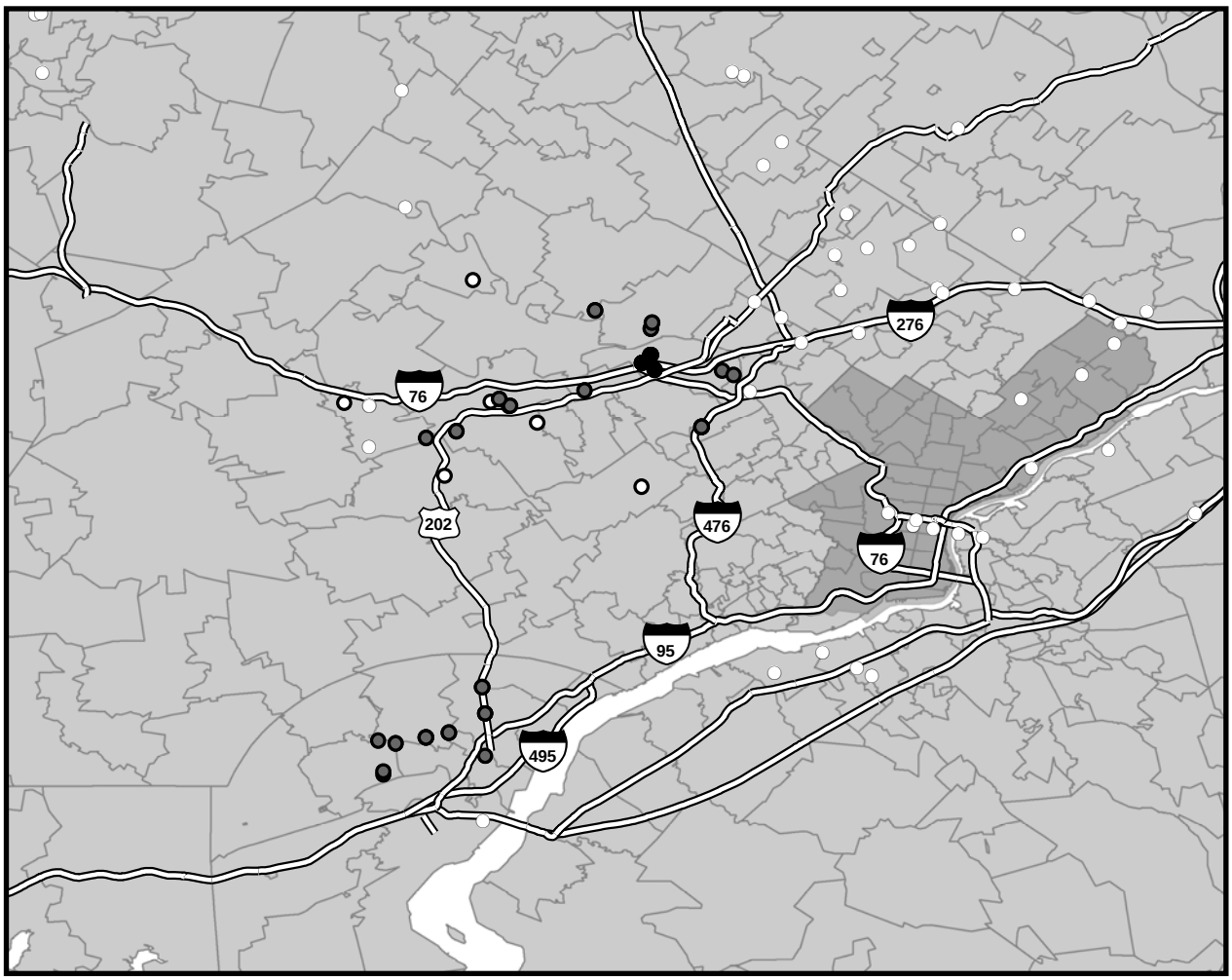

Figure 20. Philadelphia –Wilmington Core Points plus Major Routes

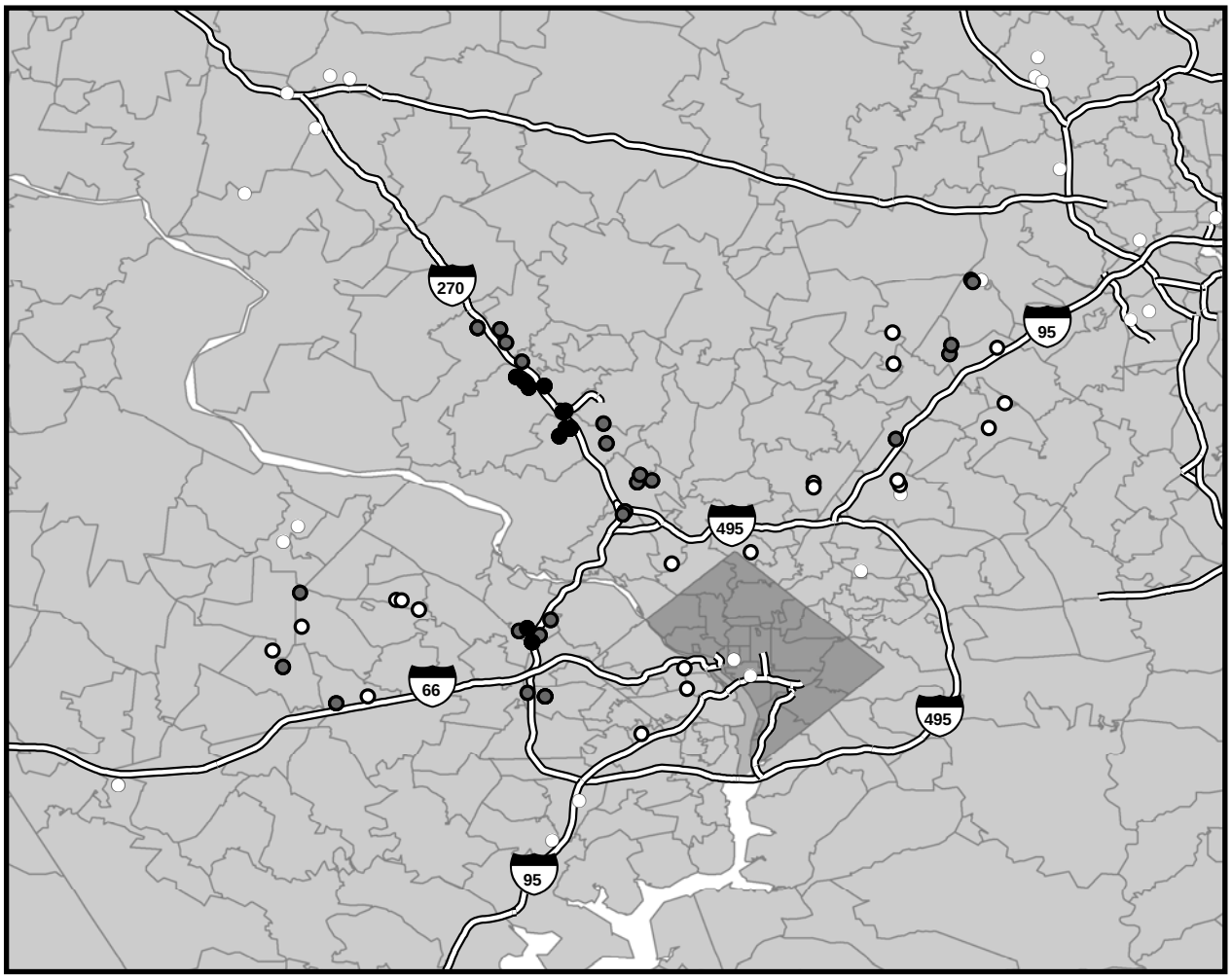

Figure 21. Washington, DC – Northern Virginia Core Points plus Major Routes

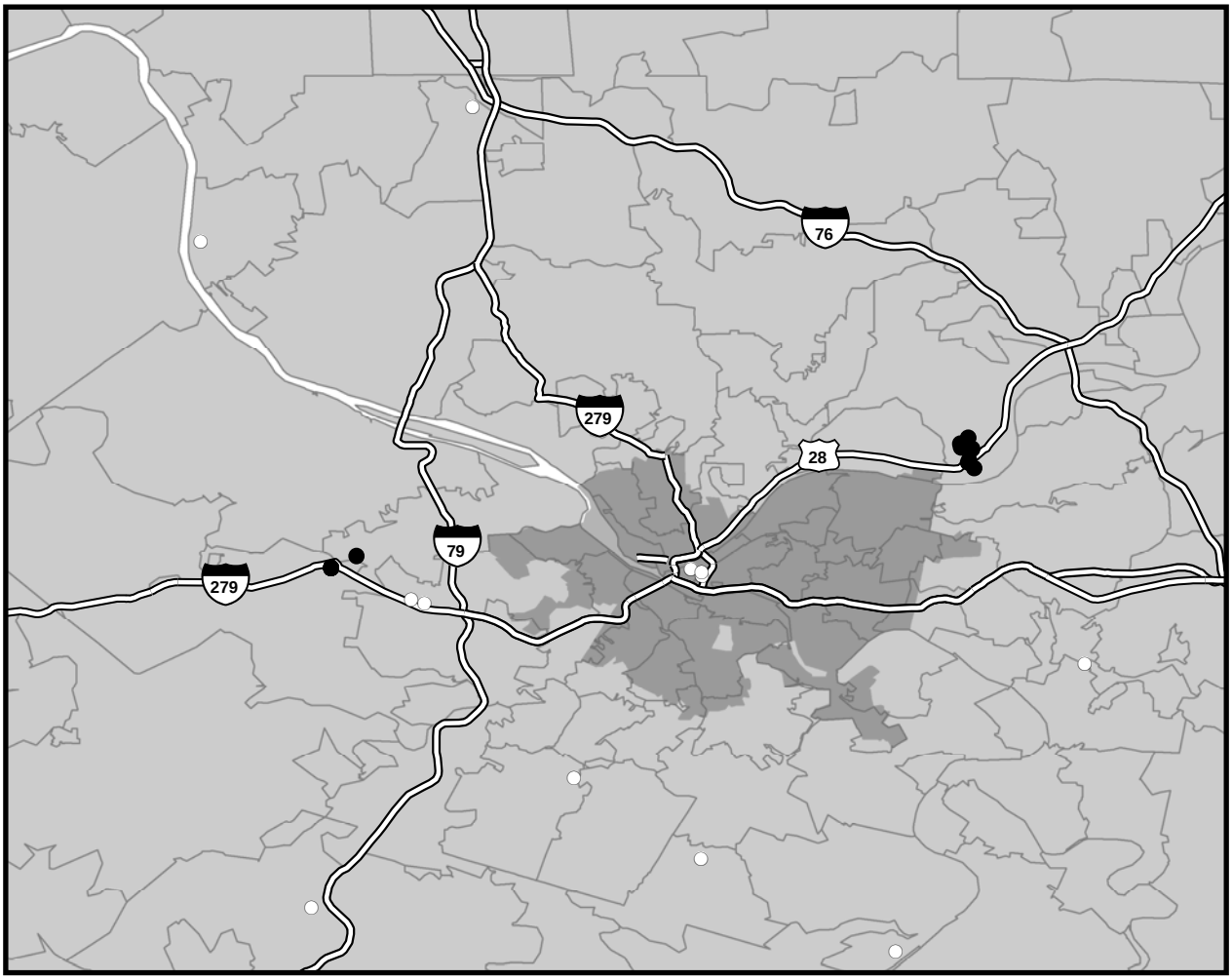

Figure 22. Pittsburgh Core Points plus Major Routes