Strategic Management Journal
Strat. Mgmt. J., **37**: 10–21 (2016)
Published online EarlyView 15 December 2015 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2451
Received 30 August 2013; *Final revision received 9 November 2014*

ZOOMING IN: A PRACTICAL MANUAL FOR IDENTIFYING GEOGRAPHIC CLUSTERS

JUAN ALCÁCER[1] and MINYUAN ZHAO[2]*
[1] *Strategy Unit, Harvard Business School, Boston, Massachusetts, U.S.A.*
[2] *Management Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.*

Research summary*: This paper advances strategic management research by taking a close look at the reasons, procedures, and results of cluster identification methods, focusing on a density-based algorithm that organically define clusters from actual locations of economic activities. Despite being a popular research topic and analytical tool, geographic clusters are often studied with little consideration given to the underlying economic activities, the unique cluster boundaries, or the appropriate benchmark of economic concentration. Our goal is to increase awareness of the complexities behind cluster identification, and to provide concrete insights and methodologies applicable to various empirical settings. The method we propose is especially useful when researchers work in global settings, where data available at different geographic units complicates comparisons across countries.*

Managerial summary*: Geographic proximity has been recognized as a fundamental factor driving firm performance, especially in knowledge-intensive industries. However, despite increasing interest in the study of geographic clusters—locations with a high concentration of economic activity—we as researchers have not given sufficient consideration to the underlying economic activity, the unique cluster boundaries, or even the definition of economic concentration. In this paper, we carefully examined the existing methodologies for cluster identification and proposed a method that defines clusters based on the actual location of economic activity. This new method is applicable to various empirical settings beyond geographic clusters. In addition, because clusters are defined by actual economic activity rather than administrative boundaries, it allows for meaningful comparison across countries.*

INTRODUCTION

Two trends underpin the recent revival of interest in the role of clusters—geographic units with a high concentration of economic activity. The first trend springs from research proving that geographic proximity can be a fundamental factor driving firm performance. For example, clusters of economic activity may impact firms' location choices (Alcácer and Chung, 2014; Shaver and Flyer, 2000), value creation (Almeida and Kogut, 1999; Delgado, Porter, and Stern, 2014), and value appropriation (Alcácer and Zhao, 2012; Giarratana and Mariani, 2014; Zhao, 2006). The second trend follows the increasing availability of location micro-data, which makes it possible to explore new and important questions. For example, a directory of R&D labs allows researchers to study innovation by examining the spatial distribution of innovative activities within and across firms, or the addresses of plants that adopted the ISO 9000 standard can be used to study the diffusion of firm practices. It is surprising, then, that with a few notable exceptions

Keywords: cluster; geography; agglomeration; location patent data

*Correspondence to: Minyuan Zhao, 3620 Locust Walk, Philadelphia, PA 19104. E-mail: myzhao@wharton.upenn.edu

(Alcácer and Zhao, 2012; Delgado, Porter, and Stern, 2015; Ellison and Glaeser, 1997; Kerr and Kominers, 2015), few papers look carefully at the issue of cluster identification.

This paper highlights the complexities of identifying clusters and then discusses and compares different methods for defining them. In particular, we address three related questions: (1) What economic activity should we measure to assess if a cluster exists? (2) What is the appropriate geographic unit over which that economic activity should be measured? and (3) How much economic concentration is necessary to label a geographic unit a cluster?[1]

We answer these questions with a combination of literature review, theoretical discussion, and illustrations using various algorithms. We emphasize methodologies that employ micro-data—an approach pioneered by geographers (e.g., Duranton and Overman, 2005)—to identify a cluster's contours more precisely. This approach is especially useful when researchers study global settings, where data are available at diverse political and geographic units that complicate cross-country comparisons.

While we use a specific empirical setting for illustrative purposes (clusters in the global semiconductor industry based on patent density), the paper's insights and methodologies are general enough to apply in other contexts. For example, the input data could be firms, individuals, or economic activities, and the output could be used to define geographic units in studies exploring competition and cooperation among hotels in Baum and Mezias (1992), fast-food franchisees in Kalnins and Lafontaine (2004), or electronics stores in Ren, Hu, and Hausman (2011).

HOW TO IDENTIFY CLUSTERS?

Here we explicate the three related questions researchers should consider when defining clusters in a specific context.

What type of economic activity?

A researcher studying agglomeration is inherently interested in an underlying economic activity. For example, to understand how agglomeration influences firms' competitive advantage, the researcher will likely be interested in firms' technical capabilities or knowledge stocks—in which case the underlying economic activity is knowledge creation and dispersion. Marshall (1920) takes firms as the unit of analysis in his seminal work on agglomeration economies. More recently, most agglomeration literature has explored employment concentration rather than the number of establishments (Glaeser *et al.*, 1992). Measuring agglomeration levels by employment is popular because employment data is readily available and because many papers have a public policy orientation.

Whether employment is an appropriate measure of economic activity will depend on the research question. For example, while employment is a plausible measure for studying manufacturing clusters, anecdotal and empirical evidence suggests there is a weaker link between employment and innovation than between employment and manufacturing plants. Audretsch and Feldman (1996) found that R&D activities tend to be more concentrated than production activities. Similarly, Alcácer (2006) found that, in the wireless industry, the distribution of R&D labs is more concentrated than any other activity in the value chain, and that the locations of manufacturing and innovation differ. Hence, publications, patents, or product development datasets would be better sources to identify technology clusters.[2]

A related decision is whether to collect data for a specific industry or for a set of related industries. The literature represented by Marshall (1920), Arrow (1962), and Romer (1986) takes an intra-industry perspective, arguing that proximate firms specializing in the same activity create agglomeration economies by encouraging skilled labor and input providers to make industry-specific investments, and by increasing the amount of industry-specific knowledge in the region. Meanwhile, Jacobs (1969) and Porter (1990) focus on interactions across industries. Input–output linkages attract related industries to locate next

[1] We focus on the characteristics of a cluster rather than on how it emerges. The latter topic has been thoughtfully explored in other research; for example, Ellison and Glaeser (1997) tried to determine if a cluster forms in response to agglomeration benefits or to location endowments.

[2] Obviously the usefulness of patents as a data source for cluster identification depends on whether patents are good indicators of innovation, which seems to be the case in industries such as semiconductors (Macher, Mowery, and Di Minin, 2008), pharmaceuticals (Cohen *et al.*, 2000), and chemicals (Ahuja and Katila, 2001), among others. For areas such as biotechnology and pharmaceuticals, publications are an alternative data source for local innovative activities (Furman *et al.*, 2005).

DOI: 10.1002/smj

to one another (Ellison, Glaeser, and Kerr, 2010). For example, new semiconductor technology may come from and be used by firms in diverse industries: aircrafts, automobiles, electronics, medical devices, etc. These firms compete in different product markets while learning from each other in the same technology field (Alcácer and Zhao, 2012). In such circumstances, firms and individuals from different industries are not only drawn by the common factors in the region, but also benefit from interacting with one another (Arbia, Espa, and Quah, 2008). Therefore, the economic activity chosen to identify clusters should reflect the importance of cross-industry interactions.

What geographic unit?

Whatever the underlying economic activity is, the geographic units should be defined based upon the economic activity of interest. Different economic activities have different geographic ranges. For instance, knowledge is sticky, suggesting a limited geographic range, such as metropolitan areas. In contrast, moving intermediate goods across distances is easier, suggesting a broader geographic range, such as states or countries. This consideration is often absent in the extant literature.

There are two broad approaches to the definition of geographic units. One is to use predetermined administrative units, such as countries, states, metropolitan areas, or economic areas. The other is to generate geographic units for the analysis organically based on the density of the economic activity under investigation.

Predetermined units

Predetermined geographic units are common in the strategy literature, mainly because most U.S. employment data is generated at the county level and then aggregated into economic areas or states. These administrative boundaries are adequate in some cases. For example, for research exploring the institutional environment's effect on firms' location decisions, defining the geographic unit by state would capture variations in state legislation. Indeed, it has been shown that patent citations are sensitive to state borders (Alcácer and Gittelman, 2006; Jaffe, Trajtenberg, and Henderson, 1993; Singh and Marx, 2013).

Unfortunately, actual economic activity does not always follow the neat borders of predetermined geographic units, which were often created for reasons other than studying the underlying economic activity.[3] Figure 1 helps to illustrate the potential problems caused by predetermined definitions of geographic units. Each point in Figure 1 represents an innovation and each rectangular shape represents a predetermined geographic unit (A, B, C, D, E). Assuming that clusters are defined as areas containing more than five dots, the data would reveal the existence of two clusters: cluster 1 in unit A and cluster 2 between units B and E.

This example illustrates three basic problems with using predetermined geographic units to identify clusters. First, the number of clusters may increase by aggregating numerous low-density locations within the same geographic unit. For example, geographic unit C would be labeled a cluster even when there is not a single location within the area that satisfies the density requirement of a cluster. The larger the area of the geographic unit, the more likely it will capture false positives—units identified as clusters when they are not. Having large areas of sparsely populated activities is against the ideas that agglomeration economies deteriorate fast over distance (Rosenthal and Strange, 2003) and that clusters are "places where the occurrence of certain events is more pronounced" (Czamanski and de Ablas, 1979).

Second, a cluster may be perceived as larger than it actually is. For example, within geographic unit A, the three points to the right would be added to cluster 1. As a result, the size of the cluster—the level of economic activity within it—would be artificially high. The concept of density also varies across locations. For example, in densely populated areas such as Japan and Western Europe, some traditional clustering methods tend to identify a large area as one cluster, even if they are divided by clear boundaries (e.g., mountains).

Third, a cluster's borders may extend beyond a geographic unit. For instance, cluster 2 is in both geographic units B and E. Guo, Peuquet, and Gahegan (2002) used the concept of "density-

[3] Economic areas may be the only exception, although they are still based on predetermined country borders. Each economic area consists of at least one node (a metropolitan or densely populated area that serves as a center of economic activity) and the surrounding counties that are economically related to the node(s). Commuting patterns are the main factor used to determine economic relationships among counties.

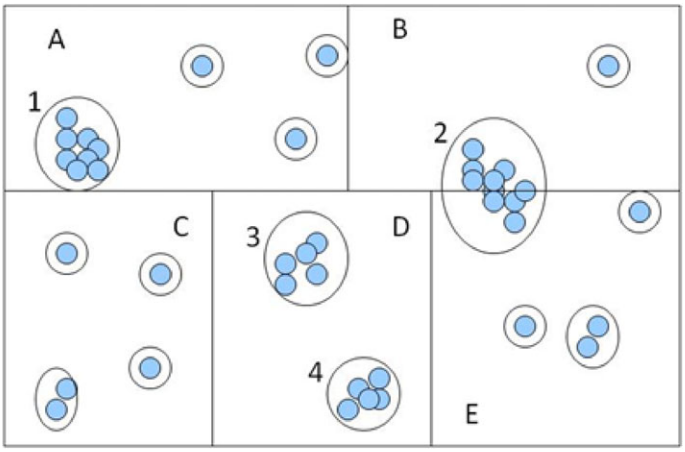

Figure 1. Geographic distribution of economic activity

connected" to describe a continuous space of high-density economic activity. Delgado, Porter, and Stern (2010) addressed this by including a measure of cluster strength within a region of related clusters—a measure that captures the strength of similar clusters in neighboring regions. Similarly, Kerr and Kominers (2015) showed that the size and shape of clusters vary across industry/technology fields.

Organically identified clusters

Instead of following predetermined geographic units, scholars in geography suggest identifying the borders of clusters organically to more accurately reflect the actual spatial distribution of the data (Ester *et al.*, 1996). With this approach, areas 1 and 2 in Figure 1 would be identified as clusters regardless of whether they belong to specific geographic units. Such organically defined clusters are even more relevant in global settings, where any predetermined geographic unit is not going to be common or comparable across countries.

There are several main methods for organic identification of clusters: partition clustering, hierarchical clustering, and density-based clustering (Ester *et al.*, 1996; Guo *et al.*, 2002). In a partition-clustering process—*k-means* and *k-medians* being used most frequently—researchers first fix the number of clusters to *k*. Then the algorithm will look for *k* cluster centers so that, when all the *N* data points in the sample are assigned to the nearest centers, the sum of distances (in the case of *k-medians*) or squared distances (in the case of *k-means*) from each data point to its respective cluster center is minimized. That is, the resulting *k* clusters will be as compact as the data allows. Depending on the specification, a cluster's center may or may not be one of the original points in the sample.

Hierarchical clustering does not require a predetermined *k*; nor does it generate a specific number of clusters. Instead, as indicated by the name, it generates a hierarchy of clusters with one big cluster including all *N* data points at the top and *N* clusters, each comprising one data point, at the bottom of the tree. Researchers decide on the appropriate distance function for the optimization process and on the level of the hierarchy (or equivalently, the number of clusters) to use for the analysis (e.g., Alcácer, 2006).

While the partition and hierarchical clustering methods are convenient, they have several drawbacks. First, their results hinge on the number *k* (in the case of partition clustering) or on the level in the tree (in the case of hierarchical clustering) determined by the researcher. With large datasets such decisions are often arbitrary. Second, these methods tend to generate clusters of similar sizes, an outcome that may not square with reality. Finally, these methods are not effective in handling clusters with irregular shapes. Long stretches of concentrated economic activity, such as in the northeast corridor of the United States, tend to be truncated when portions are located too far from the cluster center.[4]

[4] To correct this problem, Carvalho *et al.* (2009) extended the traditional hierarchical clustering approaches by imposing geographic contiguity.

Conscious of these limitations, geographers have proposed a third approach: density-based clustering. For example, Ester *et al.* (1996) challenged the existing partitioning and hierarchical clustering methods and demonstrated how a density-based algorithm can better capture clusters with irregular shapes. Guo *et al.* (2002) presented a more interactive and visual tool for hierarchical subspace clustering. Specifically, users can adjust the space definition and density threshold to identify high-density clusters. Consistent with this literature, Alcácer and Zhao (2012) developed a density-based clustering algorithm, explained in Density-based cluster identification: an example, in an effort to bring the geography research to the strategy audience.

How concentrated should the economic activity be?

Not all geographic concentrations constitute a cluster. By definition a cluster requires a large concentration of economic activity; but how much concentration is really required? Again taking Figure 1 as an example, if the cutoff value is 4 instead of 5 (i.e., geographic units are labeled clusters if they contain more than 4 points), concentrations 3 and 4 would be designated as clusters (along with concentrations 1 and 2).

The idea of activity exceeding a benchmark concentration of activities was forwarded by Ellison and Glaeser (1997) in their examination of U.S. manufacturing. Using U.S. states as the unit of analysis, they used a "dartboard" approach to define the benchmark: without agglomeration, a state's number of manufacturing establishments should be determined by random throws at a dartboard, with each state's size equal to its square miles of area. States with establishments in excess of this dartboard threshold are considered "agglomerated."

A related question is the appropriate number of clusters in a study. Approaches for determining the number of clusters have been diverse in the literature. Alcácer and Chung (2007) looked at a continuum of industry-employment concentration (i.e., the levels of geographic agglomeration), while Alcácer and Chung (2014) defined clusters as those locations in which agglomeration levels are above the mean. Ellison and Glaeser (1997) plotted their agglomeration measurement and set an arbitrary cutoff value based on their understanding of the results. Delgado *et al.* (2014) defined strong clusters as the top 20 percent of economic areas in terms of the magnitude of cluster specialization.

The appropriate cutoff point should offer a good balance between coverage (a fair representation of the industry under study) and selection (true clusters with agglomeration economies). On the one hand, including an adequate number of clusters in the sample is important to minimize sample bias and to provide enough variation to isolate the effects of specific variables. For instance, because universities are a common feature in large clusters, the effect of educational institutions on innovation may be underestimated if clusters without universities are excluded from a sample. On the other hand, too large a selection may include locations that are not truly clusters, introducing unnecessary noise into the analysis. Alcácer and Zhao (2012) chose the top 25 clusters because those 25 accounted for 84 percent of innovations in the semiconductor industry while offering a sufficient variety of competitive local environments. When the cutoff criteria are unclear, robustness checks are useful to ensure that the results are not dependent on the number of clusters considered.

DENSITY-BASED CLUSTER IDENTIFICATION: AN EXAMPLE

This section gives an example of the organically-defined cluster identification approach based on the density-based algorithms in the geography literature (Ester *et al.*, 1996) and implemented in Alcácer and Zhao (2012). To further explore the pros and cons of the available methods, we also compare the density-based algorithm with predetermined geographic units and the hierarchical clustering algorithm. While the empirical context for this example is innovation activities in the global semiconductor industry, proxied by patent applications, the identification method can be applied to a wide range of location and non-location data, such as sales data from restaurant chains, establishment locations provided by the U.S. census, subsidiary locations documented by Dun & Bradstreet, or relational distances in social networks.

Geocoding location data

Our proposed algorithm to identify clusters uses geocoded location data as input. Thus the first step in the process is to associate latitude and longitude coordinates to each locational observation. Although the datasets for strategy research are seldom geocoded, the process of geocoding has been simplified tremendously in the past few years with the availability of comprehensive geocode location data that can be merged to any datasets containing addresses.

In our particular example, in which we analyze clustering of R&D activities in the semiconductor industry, we use location data associated to patents. Semiconductor firms routinely patent their innovations, leaving a trail of the geographic distribution of R&D activities. Thus, in our example, the proposed algorithm primarily uses the density of patents in a given location to identify the contours of a cluster. Researchers who are interested in interfirm interactions may want to use firm or establishment density as input.

We relied on the technological classification from Derwent World Patents Index (DWPI) to obtain the universe of semiconductor patents applied for between 1998 and 2001 and granted between 2001 and 2004. After removing duplicates in patent families (Gittelman and Kogut, 2003), our semiconductor patent sample consisted of 23,675 patent families. From these patents we collected all the inventor locations[5] and manually cleaned all location names to remove typos and omissions.

Next, we obtained latitude and longitude information by matching the inventor locations to two official repositories of place names approved by the U.S. Board on Geographic Names (US BGN), a Federal interagency body chartered by public law to maintain uniform feature name usage. For U.S. locations, we used the Geographic Names Information System (GNIS) of the U.S. Geological Survey. For foreign locations, we used the Geonet Names Server (GNS) of the National Geospatial Intelligence Agency. Besides its wide coverage of 5.5 million location names worldwide, the GNS dataset uses phonetic variations to capture spellings from different alphabets (as in Asian countries) and from alphabets with extra characters (as in Scandinavian and Slavic countries). Since both datasets are managed by US BGN, the definitions of U.S. and foreign locations are comparable.[6]

Through the matching process we were able to identify 38,621 unique locations in the U.S. and 61,385 unique locations outside of the U.S., as defined in GNIS and GNS, respectively, with a success rate of almost 100 percent.

Identifying the cluster contours

With the location data prepared, we generated the contours of each cluster based on the density of patents in a given location. Duranton and Overman (2005) suggested that spatial clustering methods should look at maps of points on a continuous space rather than discrete grids. As in some of the geography models (Ester *et al.*, 1996; Guo *et al.*, 2002), the algorithm we propose takes full advantage of information from the data. Basically, each cluster starts from a high-density point and organically expands in all directions until the density tapers off or the distance between the neighboring points becomes too large. In this way we are not constrained by any specific number of clusters or by the distance from an arbitrary cluster center. Clusters generated by this method can—like actual clusters—take on any shape or size.

The main steps in the algorithm are the following. First, we loaded the data described in Geocoding location data. Each observation denotes the location of an inventor by his or her latitude and longitude. Where a patent was developed by several inventors at different locations, we assign a fraction of the patent to each location depending on the percentage of inventors from that location. We then calculated the density for each location by adding all patents with the same latitude and longitude, using the maximum/minimum values of latitude and longitude to establish the map's border.

Second, we identified the highest density location (i.e., the location with the largest number of inventors) and assigned it a "cluster ID." We identified each location within its Neighborhood Radium (NR) and evaluated its respective patent densities. NR determines the most basic geographic unit in miles, or how big each "dot" on the map should be. For example, if we set the NR for

[5] Inventor locations are mailing addresses, so they can be either work or home addresses. Therefore, the identified clusters will potentially capture commuting patterns as well as the density of innovative work.

[6] Our approach to geocoding locations uses free, reliable datasets that are easy to download and use. Researchers can also use proprietary datasets; for example, the one used by Google in its mapping applications (https://developers.google.com/maps/documentation/geocoding/). Regardless of the method, it is important to note that the geocoding step is independent from the proposed algorithm that utilizes the geocoded data as input.

American locations at 20 miles, the algorithm identifies all observations within a 20-mile radius of the focal location and checks patent density at these locations.

We used a different NR value for each country based on average commuting distances that we obtained from various sources.[7] Because the inventor address can be his/her home or work address, the identified clusters will capture the commute patterns as well as the density of innovative work, which is in line with the way economic areas are defined. In countries where we could not obtain commuting distances, we used those of neighboring countries or the regional average. Similarly, researchers may decide to use a different NR for different regions in the same country (e.g., California vs. New York), with the understanding that running the algorithm by regions may artificially segregate otherwise continuous clusters.

Another crucial parameter in the algorithm is the Contour Threshold (CT), which is the minimum density value for a location to be considered part of a cluster. Locations with a density value above the CT are added to the cluster and used as focal locations from which new NRs could be drawn. That is, the cluster continues to expand from there. If any location in the new NR has already obtained a cluster ID, the two clusters are merged and every member of the new cluster assumes the ID of the existing one. Therefore, this algorithm can capture bipolar or multipolar clusters such as the San Francisco–San Jose cluster in California (Kerr and Kominers, 2015) and the Denver–Boulder–Greeley cluster in Colorado (Billings and Johnson, 2014).[8]

Locations with a density value below the CT indicate the start of a low-density area and thus become the border of the cluster contour. When any new location considered within the NR of an already clustered location has fewer patents than the CT, or when any location is farther away than the number of miles in the NR, the cluster is "closed" and the contours are drawn. The location with the next highest density then receives a new ID and the above steps are repeated. This continues until all locations are associated with a cluster ID.

Figure 2 shows this process visually. Figure 2(a) shows the highest density locations in the sample in a two-dimensional universe. The algorithm looks for locations that are within the NR of each focal point (Figure 2(b)) and checks that the densities of those locations are not below the CT. All the points in Figure 2(b) passed this test and were added to their respective cluster IDs. Note that two high-density points were connected within the NR and were joined as one cluster. New locations were evaluated in Figure 2(c), and each location with a higher-than-CT density triggered a new NR. The process continues in Figure 2(d), when lower peaks—locations with lower densities—emerged. In Figure 2(e), the new locations added for analysis were either below the CT or far away from the NR. Thus, the final contours of the clusters were assigned. Note that the algorithm's output includes geographic units with just one patent (singletons) in Figure 2(f), units with low and medium patent densities, and units with high patent densities.

With our global semiconductor data, the proposed algorithm generated 5,234 units, excluding singletons. The clustering results are shown in Figure 3.

Comparing methods of cluster identification

An ideal geographic unit to capture the concept of clusters should increase the chance that two neighboring locations are assigned to the same cluster and reduce the chance that two distant points are assigned to the same cluster. In other words, measuring Type I errors (in which a location is added to a cluster where it does not belong) and Type II errors (in which a location is not added to a cluster where it belongs) allows us to evaluate how well a given geographic unit can capture the concept of a cluster. Therefore, we compared proxies of Type I and Type II errors generated using the density-based algorithm and using the predetermined areas commonly found in the literature: state, economic area, county, metropolitan areas (for American inventors only), country, and hierarchical clusters (for all inventors).

Specifically, we took all inventors in our semiconductor data (246,620 inventors at 100,006 unique locations worldwide, among which 104,742 inventors and 38,621 unique locations were in the U.S.) and explored whether a given pair of inventors would be classified as being in the same geographic unit (same cluster) using various unit definitions.

[7] The OECD compiles data on commuting patterns from different national sources and makes them available as part of its family database. For more information, see http://www.oecd.org/els/family/database.htm

[8] By merging the two neighboring clusters into one, we are unable to identify overlapping clusters caused by, for example, directional commute flows. This is a limitation of the algorithm.

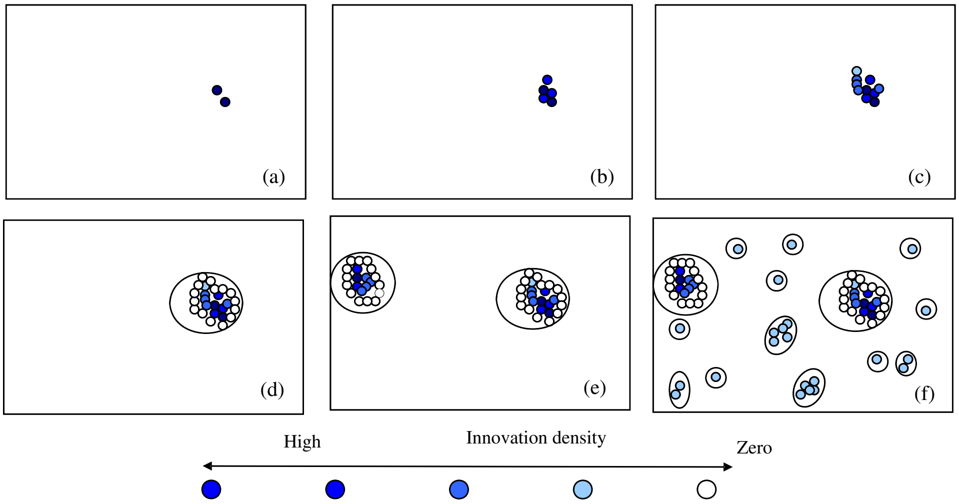

Figure 2. Density-based cluster identification algorithm

The pairs were generated as follows: For each inventor in our sample (focal inventor), we randomly selected another inventor who was not associated with the same patent. We know the latitudes and longitudes of each pair and can calculate the actual distance between them. We assumed that if two inventors were within 20 miles of each other, they were likely part of the same geographic unit, regardless of how that unit was defined. A good unit definition would therefore recognize pairs as being in the same cluster when they were less than 20 miles apart (minimizing Type II errors) and would not group two inventors in the same cluster when they were farther than 20 miles apart (minimizing Type I errors). As with any measurement, minimizing both error types is practically impossible.[9]

Table 1 shows the results of this exercise with predetermined definitions of clusters, using state, economic area, metropolitan area, and county for the U.S. inventors, and country for global inventors. Each panel corresponds to a different definition of geographic units and has two rows, the first for pairs in which the two inventors are less than 20 miles apart, the second one for pairs that are more than 20 miles apart. From 104,742 randomly generated pairs, 71,095 have both inventors within 20 miles of each other and 33,647 are separated by more than 20 miles. The column labeled *Not classified* indicates the number of pairs for which one of the inventors didn't belong to any geographic unit. Columns *Different units* and *Same unit* indicate how many pairs were classified as members of the same unit or of different units.

We start the exercise in Table 1 using states. Of the 71,095 pairs within 20 miles of each other, 69,985 pairs (98%) were classified into a cluster. The remaining 1,110 pairs corresponded to inventors who were near one another but living in different states (typically in the Northeast). The number of distant inventors that would have been classified as belonging to the same cluster (when they probably do not) was also high at 17,773 (53%). Most of these happened in states like California and Texas, where several clusters exist within the same state boundary.

Defining clusters by economic areas offers an improvement by reducing Type II errors; i.e., most pairs within 20 miles (70,818) were recognized as co-located. Economic areas were also better at separating pairs that were more than 20 miles apart into different clusters; i.e., 48 percent of pairs were classified as belonging to different clusters.

Among all of the geographic units considered, counties performed best in terms of avoiding potential Type I errors. Most pairs 20 miles apart would have been classified as belonging to different clusters if counties defined the cluster boundaries. However, counties were also the unit with the lowest number of neighboring pairs recognized as belonging to the same cluster (74%).

[9] Note that we believe it is not accurate to presume that pairs more than 20 miles apart are not in the same cluster. For example, imagine two inventors working in Silicon Valley and living in opposite ends of the city's commuting zone: San Francisco and Mountain View. These inventors should be part of the same cluster despite living many miles apart. In other words, expecting Type I errors close to 0 is not realistic.

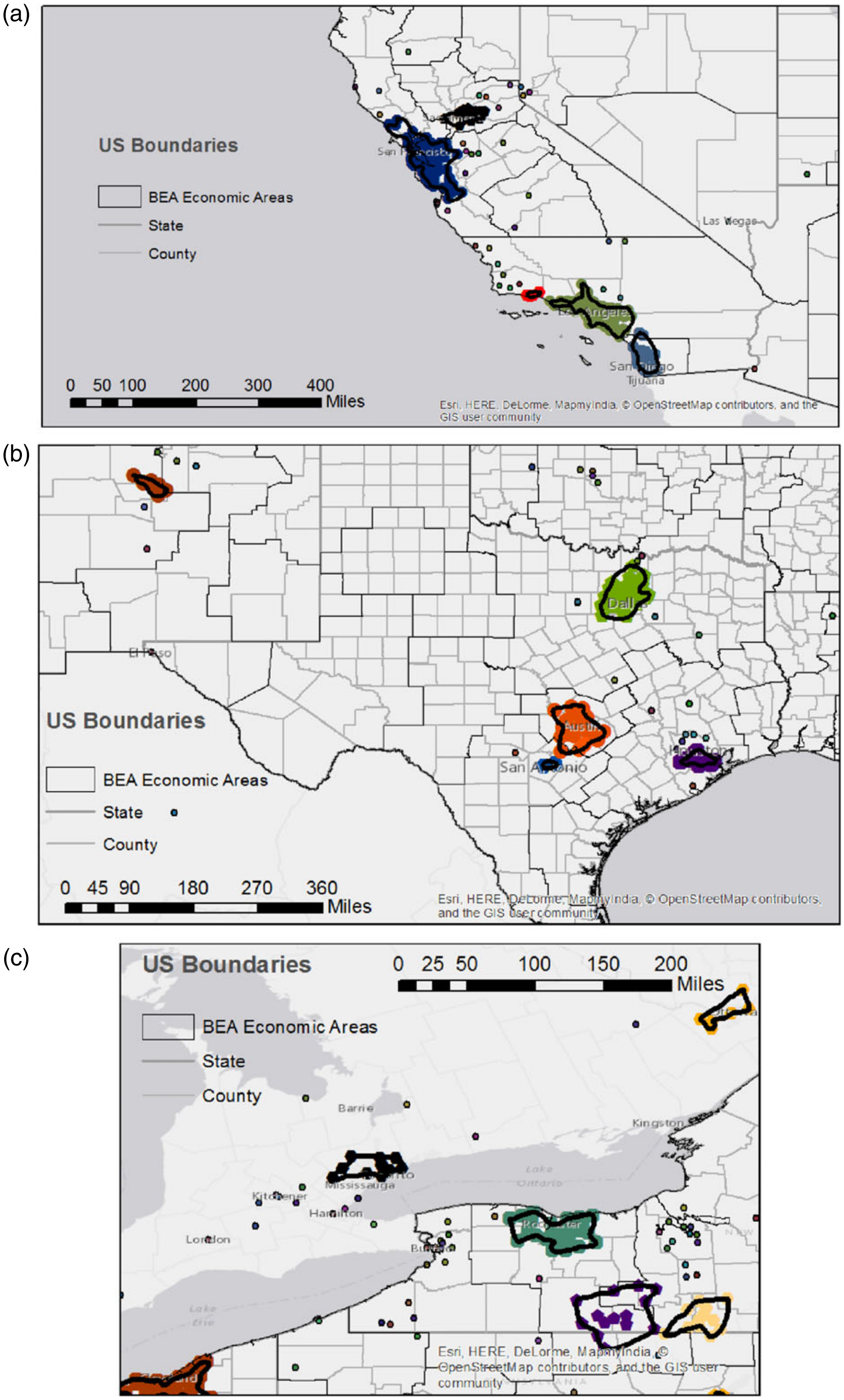

Figure 3. Density-based clusters—(a) California, (b) South U.S., and (c) North New York State. These maps show the clusters for semiconductor patents between 1998 and 2001 generated using algorithm parameters of radius 15 and threshold 10. Boundaries are shown for individual clusters. BEA Economic Areas define the relevant regional markets surrounding metropolitan and micropolitan statistical areas

Table 1. Comparing predetermined geographic units

	Distance (miles)	Obs.	Different units	Same unit	Not classified	Correct classification (%)
State	[10,20]	71,095	1,110	69,985		98
	(20, …)	33,647	15,874	17,773		47
Economic area	[10,20]	71,095	277	70,818		100
	(20, …)	33,647	16,281	17,366		48
MSA	[10,20]	71,095	537	69,256	1,302	97
	(20, …)	33,647	15,509	15,748	2,390	46
County	[10,20]	71,095	18,210	52,885		74
	(20, …)	33,647	30,238	3,409		90
Country (excluding U.S.)	[10,20]	85,209	34	85,175		100
	(20, …)	56,669	10,140	46,529		22

Metropolitan Statistical Areas (MSAs) are a commonly used unit of analysis. Unlike economic areas, though, MSAs do not span all U.S. locations. As a consequence, compared to economic areas, MSAs performed slightly worse in identifying neighboring as well as distant pairs. Note that MSAs are very accurate at identifying neighboring pairs when both inventors fall into an economic area. This suggests that the problem with MSAs is not their definition but the fact that they do not cover all U.S. territories.

The last panel in Table 1 shows the results for international locations. Most previous studies examining innovation across countries used country as the definition of locations. Among the 141,878 random pairs of foreign inventors in our sample, 85,209 pairs were within 20 miles of each other and 56,669 pairs were more distant. Not surprisingly, using countries as geographic units performed well at capturing neighboring pairs but poorly at identifying distant pairs. Only 22 percent of the distant pairs were identified as being in different clusters.

Table 2 shows results similar to those presented in Table 1, this time for organically generated clusters. The first panel in Table 2 shows the results of the density-based cluster identification algorithm for U.S. locations. Compared with the predetermined geographic units in Table 1, the density-based algorithm generated the fewest Type I and Type II errors: it recognized more neighboring pairs as belonging to the same cluster and separated more distant pairs into different clusters. A more detailed analysis of Type II errors suggests that the density-based algorithm also works well at identifying isolated inventors as singletons. For foreign locations, the density-based algorithm performed slightly poorer for neighboring pairs (99.65% vs. 99.96%) but significantly better for distant pairs than the country definition (61% vs. 22%).

Table 2 also compares the density-based algorithm with the hierarchical clustering algorithm with centroid linkages among U.S., non-U.S., and all locations. In each region, we picked a number of clusters that closely mimicked the number the algorithm proposed, so that the outcomes were comparable in number but had different contours. The density-based algorithm performed better for distant pairs, but the two approaches performed similarly for neighboring pairs.

Finally, Figure 3(a–c) show the clusters defined using the density-based algorithm for California, the southwestern US and northern New York state, respectively. These are real data examples of the stylized facts captured in Figure 1. For all maps we use the density-defined clusters as well as county, economic area (EA) and state boundaries. The maps show that county boundaries are not the optimal choice to capture clusters since in many cases clusters span multiple counties. Economic areas do a better job, but sometimes they are too wide and encompass multiple clusters (e.g., the Rochester economic area has three distinctive clusters: Rochester, Ithaca, and Syracuse) or they aggregate isolated locations (e.g., the economic area of Los Angeles includes the LA cluster plus disperse locations in the area). Conversely, sometimes economic areas are too narrow (e.g., the Austin cluster includes Austin and San Antonio economic areas).

Taken together, the performance results for both U.S. and non-U.S. locations suggest that density-based cluster-identification algorithms outperform most commonly used geographic units. That said, it will continue to fall on researchers to determine the most appropriate method for a specific context. Although density-based

Table 2. Comparing organically identified geographic units

	Distance (in miles)	Obs.	Different units	Same unit	Correct classification (%)
Density-based clustering (U.S.)	[10,20]	71,095	37	71,058	100
	(20, …)	33,647	19,392	14,255	58
Density-based clustering (non-U.S.)	[10,20]	85,209	297	84,912	100
	(20, …)	56,669	34,335	22,334	61
Density-based clustering (all locations)	[10,20]	156,304	334	155,970	100
	(20, …)	90,316	53,727	36,589	59
Hierarchical clustering (U.S.)	[10,20]	71,095	16	71,079	100
	(20, …)	33,647	15,780	17,867	47
Hierarchical clustering (non-U.S.)	[10,20]	85,209	207	85,002	100
	(20, …)	56,669	29,505	27,164	52
Hierarchical clustering (all locations)	[10,20]	156,304	223	156,081	100
	(20, …)	90,316	45,285	45,031	50

cluster-identification algorithms like the one we proposed have a number of advantages, they also have some drawbacks. They demand latitude and longitude data for each location. In the case of our density-based algorithm, it also requires obtaining realistic parameter values for both NR and CT, which entails some manual adjustments to ensure accuracy. For example, inappropriate values of NR or CT would result in very long clusters in the densely populated areas around cities such as Tokyo and New York City. Repeatedly checking establishment locations as reported in Dun & Bradstreet helped us gain a better understanding of the clustered activities in those locations and adjust the parameters accordingly.

Once the clusters are defined, they can be analyzed along many different dimensions: the density of competitors in a cluster, the concentration of economic activities among local entities, and the presence of complementary activities there.

CONCLUSION

Given the large amount of research on the geographic dimensions of strategy, we believe it is important for researchers to understand the factors at play when choosing a geographic unit for empirical analysis, and also to understand the implications of those choices. This paper discussed three crucial considerations for identifying clusters. First, the measure of economic activity in a location should reflect the specific phenomenon under study. Second, when choosing the geographic unit across which economic activity is measured, one should consider the research question's requirement(s) as well as the degree of distortion each option might incur. Finally, the concentration threshold at which a location is classified a cluster should balance coverage (a fair representation of the industry under study) and selection (true clusters with agglomeration economies).

We also provided a new, density-based method to identify organically a cluster that offers unique advantages in precision, flexibility, and applicability to cross-country studies. As an algorithm to help researchers sort observations into relevant groups, it works not only with physical distance but also for distance in social networks, technological fields, and business relationships. Therefore, its potential application is much broader than the literature on geographic locations; for example, it can be used to identify clusters of friends, clusters of related advances in the technological landscape, and clusters of suppliers/customers. The process by which the algorithm was developed also sheds light on the various tradeoffs researchers must address in geography research. Of course, such methods also present higher requirements for data and computing power, and they do not apply to clusters with special features such as directional linkages. With that in mind, researchers seeking the optimal method for a given study must carefully consider that study's theoretical question and its specific empirical context.

REFERENCES

Ahuja G, Katila R. 2001. Technological acquisitions and the innovation performance of acquiring firms: a longitudinal study. *Strategic Management Journal* **22**(3): 197–220.

Alcácer J. 2006. Location choices across the value chain: how activity and capability influence co-location. *Management Science* **52**(10): 1457–1471.

Alcácer J, Chung W. 2007. Locations strategies and knowledge spillovers. *Management Science* **53**(5): 760–776.

Alcácer J, Chung W. 2014. Locations strategies for agglomeration economies. *Strategic Management Journal* **35**(12): 1749–1761.

Alcácer J, Gittelman M. 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. *Review of Economics and Statistics* **88**(4): 774–779.

Alcácer J, Zhao M. 2012. Local R&D strategies and multi-location firms: the role of internal linkages. *Management Science* **58**(4): 734–753.

Almeida P, Kogut B. 1999. Localization of knowledge and the mobility of engineers in regional networks. *Management Science* **45**(7): 905–917.

Arbia G, Espa G, Quah D. 2008. A class of spatial econometric methods in the empirical analysis of clusters of firms in the space. *Empirical Economics* **34**(1): 81–103.

Arrow KJ. 1962. The economic implications of learning by doing. *Review of Economic Studies* **29**: 155–172.

Audretsch DB, Feldman MB. 1996. R&D spillovers and the geography of innovation and production. *American Economic Review* **86**: 630–640.

Baum J, Mezias S. 1992. Localized competition and organizational failure in the Manhattan hotel industry, 1898-1990. *Administrative Science Quarterly* **37**: 580–604.

Billings S, Johnson E. 2014. Agglomeration within an Urban Area. Mimeo (University of North Carolina Charlotte).

Carvalho A, Ywata P, Melo Albuquerque G, Rezende de Almeida J, Guimaraes R. 2009. Spatial hierarchical clustering. *Revista Brasileira de Biometria* **27**(3): 411–442.

Cohen WM, Nelson RR, Walsh JP. 2000. Protecting their intellectual assets: appropriability conditions and why U.S. manufacturing firms patent (or not). NBER Working paper 7552. National Bureau of Economic Research, Cambridge, MA.

Czamanski S, de Ablas LAQ. 1979. Identification of industrial clusters and complexes: a comparison of methods and findings. *Urban Studies* **16**(1): 61–80.

Delgado M, Porter M, Stern S. 2010. Clusters and entrepreneurship. *Journal of Economic Geography* **10**(4): 495–518.

Delgado M, Porter M, Stern S. 2014. Clusters, convergence, and economic performance. *Research policy* **43**(10): 1785–1799.

Delgado M, Porter M, Stern S. 2015. Defining Clusters of Related Industries, *Journal of Economic Geography*, forthcoming

Duranton G, Overman HG. 2005. Testing for localisation using micro-geographic data. *Review of Economic Studies* **72**: 1077–1106.

Ellison G, Glaeser EL. 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economy* **105**(5): 889–927.

Ellison G, Glaeser E, Kerr W. 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review* **100**(3): 1195–1213.

Ester M, Kriegel H, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* **96**: 226–231.

Furman JL, Kyle MK, Cockburn I, Henderson RM. 2005. Public and private spillovers, location and the productivity of pharmaceutical research. *Annales d'Economie et de Statistique* **79**(80): 167–190.

Giarratana MS, Mariani M. 2014. The relationship between knowledge sourcing and fear of imitation. *Strategic Management Journal* **35**(8): 1144–1163.

Gittelman M, Kogut B. 2003. Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science* **49**(4): 366–382.

Glaeser E, Kallal H, Scheinkman J, Shleifer A. 1992. Growth in cities. *Journal of Political Economy* **100**: 1126–1152.

Guo D, Peuquet D, Gahegan M. 2002. Opening the black box: interactive hierarchical clustering for multivariate spatial patterns. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems*, New York, NY, 131–136.

Jacobs J. 1969. *The Economy of Cities*. Random House: New York.

Jaffe AB, Trajtenberg M, Henderson R. 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* **108**(3): 577–598.

Kalnins A, Lafontaine F. 2004. Multi-Unit ownership in franchising: evidence from the fast-food industry in Texas. *RAND Journal of Economics* **35**(4): 749–763.

Kerr W, Kominers S. 2015. Agglomerative forces and cluster shapes. Review of Economics and Statistics, forthcoming.

Macher J, Mowery D, Di Minin A. 2008. Semiconductors. In *Innovation in Global Industries: U.S. Firms Competing in a New World (Collected Studies)*, Macher JT, Mowery DC (eds). National Academies Press: Washington, DC.

Marshall A. 1920. *Principles of Economics* (revised 8th edn). MacMillan: London, UK.

Porter ME. 1990. The competitive advantage of nations. *Harvard Business Review* **68**(2).

Ren CR, Hu Y, Hausman J. 2011. Managing product variety and collocation in a competitive environment: an empirical investigation of consumer electronics retailing. *Management Science* **57**(6): 1009–1024.

Romer P. 1986. Increasing returns and long-run growth. *Journal of Political Economy* **94**: 1002–1037.

Rosenthal S, Strange W. 2003. Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* **85**(2): 377–393.

Shaver M, Flyer F. 2000. Agglomeration economies, firm heterogeneity, and foreign direct investment in the United States. *Strategic Management Journal.* **21**(12): 1175–1193.

Singh J, Marx M. 2013. Geographic constraints on knowledge spillovers. *Management Science* **59**(9): 2056–2078.

Zhao M. 2006. Conducting R&D in countries with weak intellectual property rights protection. *Management Science* **56**(7): 1185–1199.