1-2016

Zooming in: A Practical Manual for Identifying Geographic Clusters

Juan Alcacer
Harvard Business School

Minyuan Zhao
University of Pennsylvania

Recommended Citation

Zooming in: A Practical Manual for Identifying Geographic Clusters

Abstract

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.

Disciplines
Management Sciences and Quantitative Methods

Zooming In: A Practical Manual for Identifying Geographic Clusters

Juan Alcácer

jalcacer@hbs.edu

Harvard Business School

Minyuan Zhao

myzhao@umich.edu

Ross School of Business – University of Michigan

August 2013

Zooming In: A Practical Manual for Identifying Geographic Clusters

Abstract

This paper takes a close look at the reasons, procedures, and results of cluster identification methods. Despite being a popular research topic in strategy, economics, and sociology, 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 organic cluster identification methodology we propose is especially useful when researchers work in global settings, where data available at different geographic units complicates comparisons across countries.

1. INTRODUCTION

The concept of clusters—high concentrations of economic activity in a specific geographic unit—is at the core of a vast amount of research in economics, management, urban planning, sociology, and public policy. It is surprising, then, that few papers look carefully at the issue of cluster identification. With a few notable exceptions (Ellison and Glaeser, 1997; Porter, 1990), most papers overlook the nuances embedded in this seemingly intuitive concept.

In this paper, we detail the reasons, procedures, data, and results of our effort to identify geographic clusters. Our goal is to increase awareness of the complexities behind cluster identification, and to provide a concrete method, used in Alcacer (2006) and Alcacer and Zhao (2012), that can help researchers define clusters more accurately. In particular, we address three

related questions in cluster identification: (1) What economic activity should we measure to determine clustering? (2) What is the appropriate geographic unit over which economic activity should be measured? and (3) What levels of economic concentration are high enough for the geographic unit to be labeled a cluster?[1]

We answer these questions with a combination of literature review, theoretical discussion, and illustrations with various algorithms. While we use a specific empirical context (the global semiconductor industry) for illustrative purposes, the insights and methodologies are general enough for other contexts. The organic cluster identification methodology we propose is especially useful when researchers work in global settings, where data available at different geographic units complicates comparisons across countries.

2. HOW TO IDENTIFY CLUSTERS?

In this section, we discuss the three questions that researchers should consider when they define clusters in a specific context: the measure of economic activity in a location, the geographic unit over which economic activity should be measured, and the concentration threshold required to classify a location as a cluster.

2.1 What type of economic activity?

When studying agglomeration, a researcher is inherently interested in some underlying economic activity. For example, to understand how agglomeration influences firms' competitive advantage, the researcher is likely interested in firms' technical capabilities or knowledge stocks. In this 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

---

[1] There are other decisions that need to be addressed when defining clusters. For example, Ellison and Glaeser (1997) try to determine if a cluster is the result of agglomeration benefits or endowments. We focus on the characteristics of clusters regardless of how those clusters emerged.

recently, most of the literature studying agglomeration in economic activity has used employment concentration instead of the number of establishments (Glaeser *et al.*, 1992). The popularity of employment-based agglomeration level measures is driven by the availability of employment data and by the public policy orientation of most papers examining economic development.

Employment appears to be a plausible measure for studies of manufacturing clusters. However, there is anecdotal and empirical evidence that the link between employment and innovation (as opposed to the link between employment and manufacturing plants) is weaker. Audretsch and Feldman (1996) find that R&D activities tend to be more concentrated than production activities. Similarly Alcácer (2006) finds the distribution of R&D labs in the wireless industry is more concentrated than any other activity in the value chain, and that locations of manufacturing and innovation in that industry differ.

For technology clusters, patents are a better data source for cluster identification because patents are associated with the locations in which inventors innovate. 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 *et al.*, 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).

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) (the MAR school) 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) (the Jacobs-Porter school) focus on interactions across industries. Input-output linkages attract related industries to locate next to one another (Ellison *et al.*, 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 and interact with each other in the same technology field (Alcácer and Zhao, 2012). Depending on the importance of such cross-industry interactions, the measures of economic activity should reflect them accordingly.

2.2 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. In contrast, moving intermediate goods across distances is easier, suggesting a broader geographic range. The activity's geographic range then helps define the appropriate geographic unit. Economic activity with a more limited geographic range should be examined using smaller geographic units, such as counties or metropolitan areas; while activity with a greater range can be examined using larger geographic units, such as states or even countries. This consideration is often absent in the extant literature: a particular definition for geographic units is used with little consideration of the underlying economic activity, or based on citations of other papers that use the same geographic units for the analysis of totally different economic activities.

There are two approaches to the definition of geographic units. One is to use predetermined administrative or statistical units, such as countries, states, metropolitan areas, or economic area. This is a one-size-fits-all approach. The other is to use the data on the economic activity under

investigation, and organically generate geographic units for the analysis. This is an approach customized for a specific industry (i.e. semiconductors or biotechnology), location (i.e. Japan or Jamaica), or phenomenon (i.e. R&D or manufacturing).

Predetermined units

The common practice in the literature is to use pre-determined geographic units, such as metropolitan areas, county, or states in the U.S., and countries for locations outside of the U.S. Data availability generally drives that decision: most U.S. employment data is generated at the county or metropolitan-area levels and then aggregated to economic areas or states. In some cases, such administrative boundaries are adequate. For example, if the research focus is institutional environment's effect on firms' location decisions, then variations across state legislation would suggest that state is the right geographic unit for the study.

However, in the case of knowledge spillover, we need a unit that reflects the pooling of knowledge and resources. In an effort to identify geographic areas that mimic economic activity, the Bureau of Economic Analysis (BEA) defined 179 economic areas spanning the continental U.S. Each economic area consists of at least one node (a metropolitan or densely populated area that serves as 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. Each economic area includes, as far as possible, both the work site and residences of its labor force. The BEA's definition of economic areas not only captures the boundaries of labor pools as described by commuting patterns between work and home, but also the bounds of supplier pools because supplying firms draw from the same labor pool, and one of the main conduits of knowledge flow is employee turnover in the local labor pool.

Unfortunately, in many cases, actual economic activity does not follow the neat borders of predetermined geographic units, which were often created for other reasons than studying the specific underlying economic activity such as agglomeration economies. Figure 1 helps to illustrate the potential problems caused by inappropriate, 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.

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. Delgado *et al.* (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. Failing to consider contiguous clusters would require empirical models that address spatial correlations. Granted, the prevalence and consequences of these problems will vary depending on theoretical arguments for clustering and the specific empirical context.

Organically defined units

Instead of following pre-determined geographic units, the borders of clusters can be identified organically to reflect as much as possible the actual spatial distribution of the data. 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. Any pre-determined geographic unit is not going to be common across countries, rendering impossible a meaningful comparison across clusters.

There are several existing methods for cluster generation, mainly partition clustering and hierarchical clustering. In a partition clustering process—*k-means* and *k-medians* being the most frequently used—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 are the most compact ones allowed by the data. Depending on the specification, the cluster centers may or may not be one of the original points in the sample.

While this clustering process is convenient, there are several drawbacks. First, the optimization process is approximate, and the result depends on the number k predetermined by the researcher. With large datasets, the selection of k is often arbitrary. Second, this process minimizes the sum of distances or squared distances from cluster centers, so the result tends to

be k clusters of similar sizes, which does not always square with reality. Finally, this method is 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 cut off arbitrarily when the locations are too far away from the cluster center.

Hierarchical clustering does not require a predetermined k. In fact, it does not even generate a specific number of clusters. 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. The researchers decide on the appropriate distance function for the optimization process, and then decide which level of the hierarchy (or equivalently, the number of clusters) to use for the analysis. While hierarchical clustering is more flexible than partition clustering, they share the same limitation with regard to geographic areas of irregular shapes.

To overcome the limitations of existing clustering algorithms, we developed an organic cluster identification algorithm that takes full advantage of information from the data. Each cluster will start from a high-density point and organically expand in all directions until the density tapers off or the distance between the neighboring points gets too large. That way, we are not constrained by any specific number of clusters or by the distance from an arbitrary cluster center. The resulting clusters can take any shapes or sizes, reflecting what we see in reality. In Section 3, we will describe the algorithm in detail and compare it with the existing algorithms.

2.3 How much activity concentration is enough to label a location as a cluster?

Not all geographic concentrations constitute a cluster. The definition of a cluster implicitly requires a large concentration of economic activity; the question is precisely how much concentration is required for an agglomeration to become a cluster. Again with Figure 1 as an

example, if the cut-off 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 become clusters (along with concentrations 1 and 2).

The idea of activity exceeding a benchmark is forwarded by Ellison and Glaeser (1997) when examining agglomeration of U.S. manufacturing. Using a geographic unit of the 50 states, 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". In cluster-related researches, each researcher needs to define a benchmark threshold appropriate to the setting.

A related question is the appropriate number of clusters in a study. Approaches to determine 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 (2013) 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 the intuition from their results. Delgado *et al.* (2012) 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 field or 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 may be a common feature in large

clusters, the effect of educational institutions on innovation may be underestimated if clusters without universities are excluded from the sample. On the other hand, too large a selection may include locations that are not truly clusters, introducing unnecessary noise to the analysis. Alcácer and Zhao (2012) chose the top 25 clusters because they account 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. In Alcácer and Zhao (2012), the results remain robust with the selection of top 10 or top 50 clusters.

3. Defining organic clusters

This section addresses the issue of selecting the geographic unit size over which economic activity should be measured, using the global semiconductor industry as the context to introduce an organic cluster identification algorithm. We also compare this algorithm with predetermined geographic units as well as with the hierarchical clustering algorithm.

3.1 Preparing the location data

Semiconductor firms routinely patent their innovations, leaving a trail of the geographic distribution of R&D activities in the industry. Thus, our algorithm primarily uses the density of patents in a given location to identify the contours of a cluster. Unfortunately the inventor addresses provided by the USPTO were noisy. Typos were common, especially for foreign locations. Therefore, our first non-trivial task was to manually clean all names for which an electronic match with the latitude and longitude data was not found.

We found three types of typos. The first type was inconsistencies in abbreviations and acronyms for city names, mainly, but not exclusively, in American locations. For example, Los Angeles could be spelled fully or as LA, L.A., etc.; Salt Lake City appeared often as SLC; words

like West, East, North and South could be abbreviated or spelled fully. The second type was mismatches between city and state or city and country. For example, Ann Arbor, Michigan (state code MI) was sometimes assigned to Minnesota (state code MN), Mississippi (state code MS) or Missouri (state code MO). Israeli cities such as Tel-Aviv that were sometimes assigned to the U.S. state of Illinois because the code for both Illinois and Israel is IL. The third type was incomplete addresses mainly associated with foreign inventors. For example, many addresses included a postal code but missed either the province or the city. In most cases, we were able to recover the city information using the directory of each country's postal codes.

This name cleaning process required many rounds of improvement and revision. A program to fuzzy-match geographic names offered potential matching options to the research assistants and authors. We then used lists of provinces, major cities, and zip codes for most countries in our dataset to resolve ambiguous matches, aided by internet searches. In the case of international addresses, we asked individuals knowledgeable about specific countries or languages to perform the manual checks. The authors were able to cover Latin America, Spain, France, Italy, Portugal, Australia, the U.K., the U.S. and the Greater China region. We also asked native speakers from Korea, Singapore, Germany, Russia and Slavic countries, Scandinavia, and Japan to clean up the data of these countries.

Next, we obtained latitude and longitude information from two sources. 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).

We first applied this process to global patents from the semiconductor industry. We relied on the technological classification from the Derwent World Patent Index (DWPI) to obtain the universe of semiconductor patents applied for between 1998 and 2001, and granted between 2001 and 2004.[2] After removing duplicates in patent families (Gittelman and Kogut, 2003), our semiconductor patent sample consisted of 23,675 patent families, from which we were able to identify 38,261 unique locations in the U.S. and 61,385 unique locations outside of the U.S., with an almost 100 percent success rate. Most of the locations for which we could not attach latitudes and longitudes are in China, where a large number of cities share identical spellings. Without any province information, we were unable to find a unique match.

This effort also generated a list of common typos in location names and a set of programs that help to clean and geocode locations. We applied these tools to the entire UPSTO dataset, creating a comprehensive dataset with latitudes and longitudes for most patents granted since 1975. This dataset will be available to the public together with this note.

3.2   Identifying the cluster contours

With the location data prepared, we can generate the contours of each cluster based on the density of patents in a given location. The main steps in the algorithm are the following:

First, we load the data described above. Each observation denotes a particular patent with the latitude and longitude of the inventor. If a patent is developed by a group of inventors residing at different locations, there are as many observations as unique locations of the inventors, with each location accounting for one patent.[3] Then we calculate the density for each location by adding up

---

[2] For details of the data source, see Alcácer and Zhao (2012).

[3] Depending on the research question, one may also count this as a fraction of a patent. For example, if we are interested in the participation of local inventors, each patent represents a project that the local talents are involved in, so counting it as one patent is reasonable. On the other hand, if we are interested in relative contribution of various locations to the firms' R&D, calculating the fractions is more appropriate.

all the patents with the same latitude and longitude. The maximum and minimum latitude and longitude values in the dataset are used to set up the border of the map.

Second, we identify the location with the highest density (i.e. the location with the largest number of patents) and assign it a cluster ID. Then we identify all locations within its Neighborhood Radium (NR) and evaluate their respective patent densities. NR determines the most basic geographic unit, in miles. Put simply, this is to define how big each "dot" on the map should be. For example, we set the NR for American locations at 20 miles, meaning that the algorithm will identify all observations within a 20-mile radius from the focal location and check the patent density at these locations.[4]

Another crucial parameter in our 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 as members of the cluster and used as new focal locations from which new NRs could be drawn for evaluation. That is, the cluster will continue to expand from there. If any location in the new NR has already obtained an existing cluster ID, the two clusters are merged and every member of the new cluster assumes the ID of the existing one. 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 next location with the highest density is then given a new ID and the above steps are repeated. This continues until all locations are associated with a cluster ID.

---

[4] We used a different NR value for each country based on average commuting distances, which we obtain from various sources. In countries where we could not obtain commuting distances, we used those of neighboring countries or the regional average.

Figure 2 shows this process visually. Figure 2a 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 2b) and checks that the densities of those locations are not below the CT. All the points in Figure 2b 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 2c and each location with a higher-than-CT density triggered a new NR. The process continues in Figure 2d, when lower peaks—location with lower densities—emerged. In Figure 2e, 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 locations with just one patent (stand-alone points in Figure 2f), locations with low and medium patent densities, and areas with high patent densities. With our global semiconductor data, our algorithm generated 5,234 units, among which the singletons are not used in the analysis.

We can also reproduce the output on a map to have a visual assessment of the algorithm, where adjacent clusters are mapped with different colors depending on the cluster IDs they assume. Figure 3 gave the examples of the U.S., Japan and Western Europe. As shown in the maps, both density and distance matter in cluster identification. The clusters generated from our algorithm are not constrained by specific shapes or sizes, but follow the actual observations of innovation activities.

Although our organic cluster identification algorithm has a number of advantages, it also has some drawbacks. It demands latitude and longitude data for each location. 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 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.

3.3 Testing the results 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 it should not belong to) and Type II errors (in which a location is not added to a cluster it belongs to) 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 for different geographic units, including clusters generated with our algorithm described in the previous subsection, and those used 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 worldwide at 100,006 unique locations, with 104,742 of those at 38,621 unique U.S. locations) and explored whether a given pair of inventors would be classified as being in the same geographic unit (same cluster) using various unit definitions.

The pairs were generated as follows: For each inventor in our sample (focal inventor) we drew randomly another inventor that 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 are likely to be part of the same geographic unit, regardless of how that unit is defined. A good unit definition would recognize pairs as being in the same cluster when they are less than 20 miles apart (minimizing

Type II errors) and would not group two inventors in the same cluster when they are farther than 20 miles apart (minimizing Type I errors). As with any measurement, minimizing both error types is practically impossible.[5]

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 foreign 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 different units.

We start the exercise in Table 1 using states. In this case, of the 71,095 pairs within 20 miles of each other, 69,985 pairs (98%) were classified as belonging to a cluster. The remaining 1,110 pairs corresponded to inventors that were close to each other but living in different states. Most of these cases occurred 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 boundaries.

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

---

[5] 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 miles apart. In other words, expecting Type I errors close to 0 is not realistic.

better at separating pairs that were more than 20 miles 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. Unfortunately, counties were also the unit with the lowest number of neighboring pairs recognized as belonging to the same cluster (74%).

Metropolitan Statistical Areas (MSAs) are a commonly used unit of analysis. According to the U.S. Office of Management and Budget (OMB), an MSA "consists of one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core." Unlike economic areas, though, MSAs do not span the whole territory of the U.S. Therefore, some locations in our sample did not fall into any MSAs. As a consequence, compared to economic areas, MSAs performed slightly worse in identifying neighboring pairs as well as distant pairs.[6]

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. For example, Singh (2008) studied knowledge flows within multinationals using countries as the unit of analysis. Note that the number of pairs is different: we have 141,878 foreign inventors in our sample for which we created random pairs. Among those pairs, 85,209 were within 20 miles of each other and 56,669 were more distant. Not surprisingly, using countries as geographic units

---

[6] 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.

performed well at capturing neighboring pairs but poorly at identifying distant pairs. Only 22% of the distant pairs were identified as 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 1 shows the results of our own cluster identification algorithm for U.S. locations. Compared with all the predetermined geographic units in Table 1, our 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 our algorithm also works well at identifying isolated inventors as singletons.

For foreign locations, our 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%). Taken together, the performance results for both U.S. and non-U.S. locations suggest that our cluster identification algorithm outperforms most commonly used geographic units.

Table 2 also compares our algorithm with the hierarchical clustering algorithm with centroid linkage, for U.S., non-U.S., and all locations. In each region, we picked the number of clusters that closely mimics the number from our algorithm, so the outcomes are comparable in sizes but different in their contours. Our algorithm performed better for distant pairs, although both approaches performed similarly well for neighboring pairs.

4. CONCLUSION

Given the large amount of research in firms' location strategies, and in economic geography more generally, we believe that it is important for researchers to understand the factors behind the choice of an appropriate geographic unit for empirical analysis, as well as the implications of such choices. In this paper, we discussed three crucial considerations in the identification of

clusters. First, the measure of economic activity in a location should reflect the specific phenomenon under study. Second, when choosing the geographic unit over which economic activity is measured, one should consider the requirement of the research questions as well as the degree of distortion that each choice may incur. Finally, the balance between coverage and selection determines the concentration threshold required to classify a location as a cluster.

We also provided a new method to identify a cluster organically based on the economic activities in the data. This useful alternative for the definition of geographic units offers unique advantages in precision, flexibility, and applicability to cross-country studies. The process by which the algorithm was developed also sheds light on the various tradeoffs that researchers must address in geography research.

REFERENCES:

Ahuja G, Katila R. 2001. Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic Management J.* **22:** 197-220.

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

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

Alcácer J, Chung W. 2013. Locations strategies for agglomeration economies. *Strategic Management J.* forthcoming.

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

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. *Amer. Econom. Rev.* **86:** 630–640.

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.

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

Delgado M, Porter M, Stern S. 2012. Clusters, convergence, and economic performance. *NBER Working Paper* 18250.

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

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.

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

Glaeser EL, Kallal HD, Scheinkman JA, Shleifer A. 1992. Growth in cities, *Journal of Political Economy* 100: 1126–1152

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

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

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

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

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

Singh J. 2008. Distributed R&D, cross-regional knowledge integration, and quality of innovative output. *Research Policy* **37**(1): 77-96.
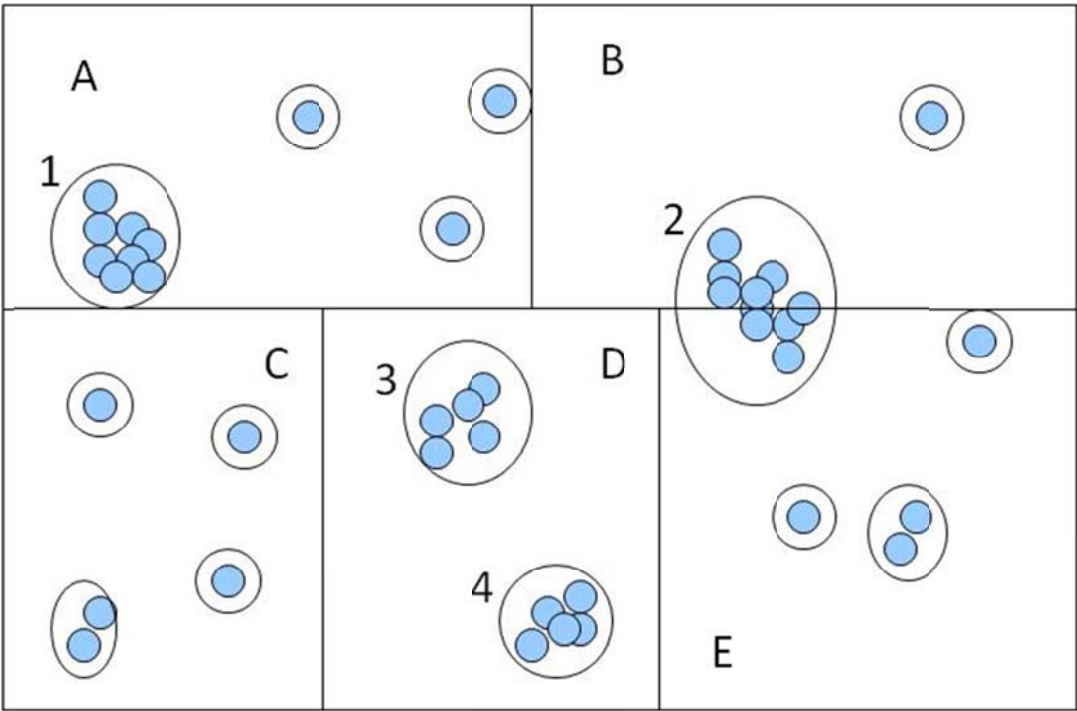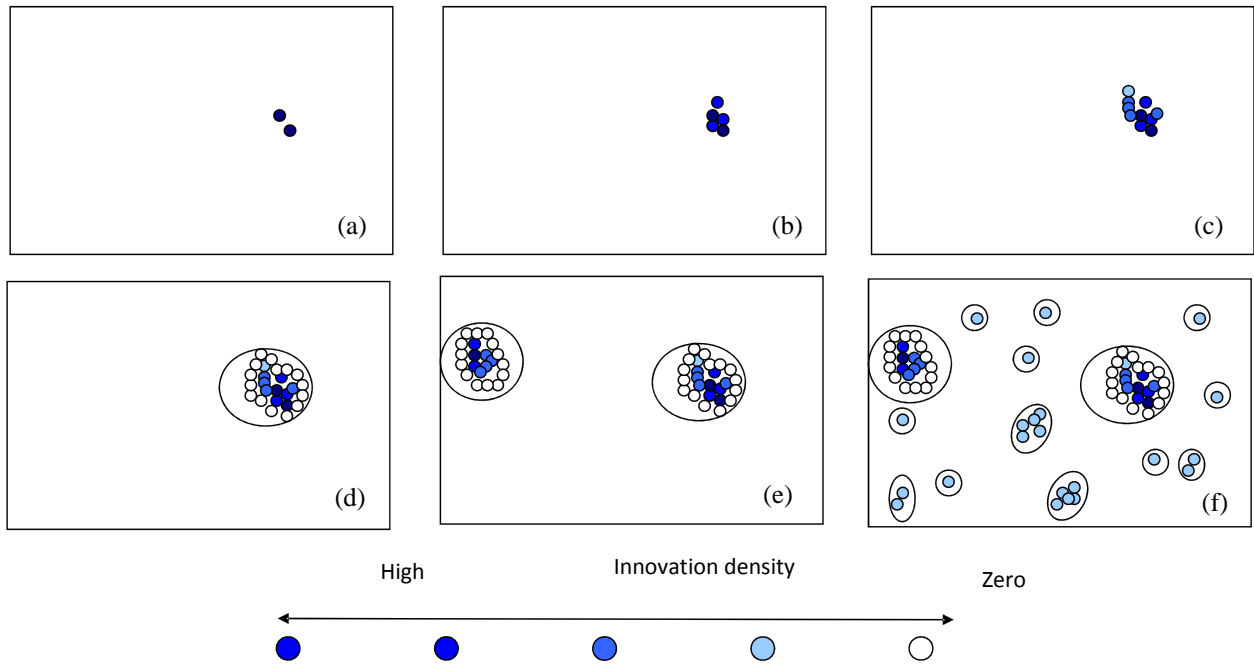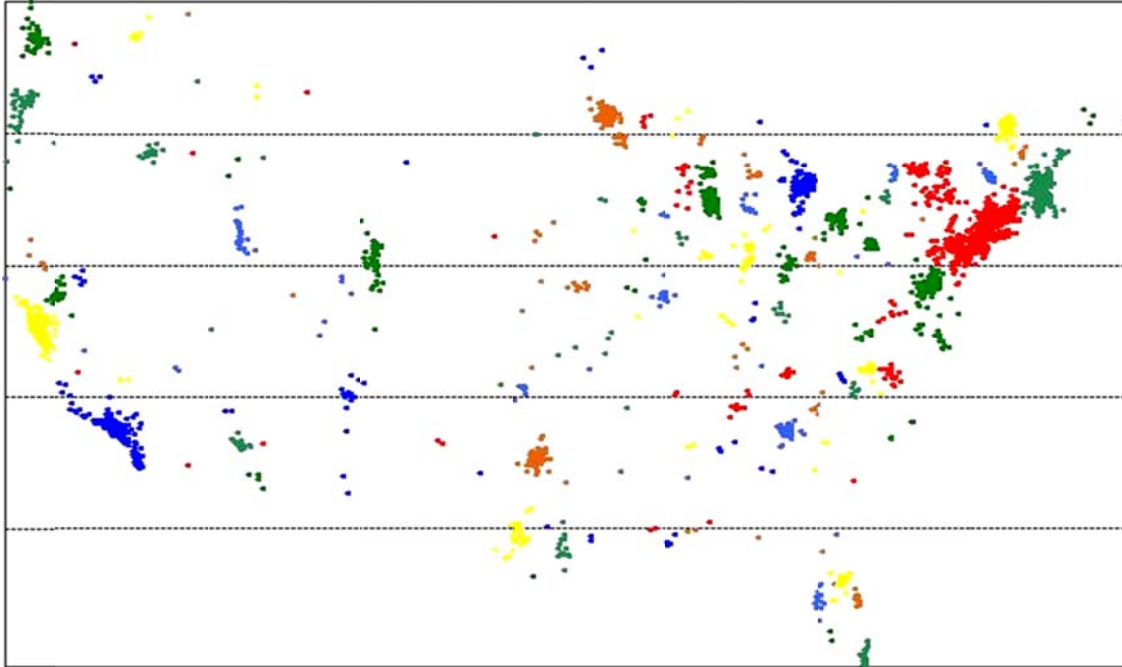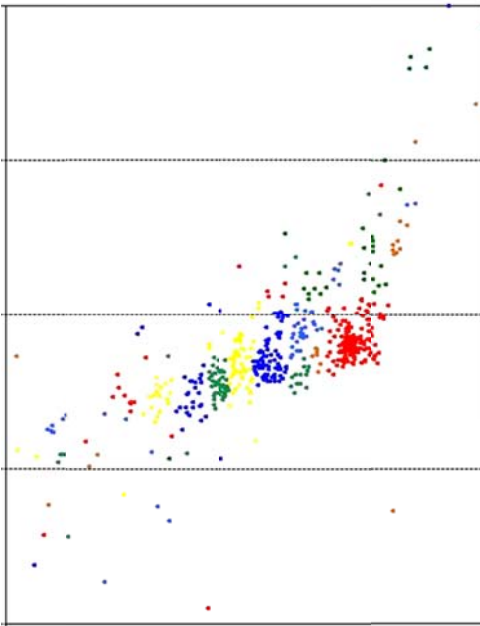
Figure 1: Geographic distribution of economic activity

(a)

(b)

(c)

(d)

(e)

(f)

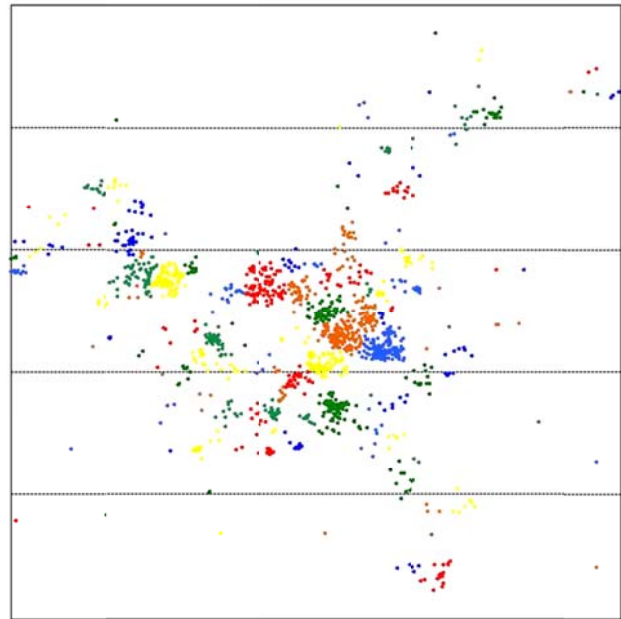High   Innovation density   Zero

Figure 2: Organic Cluster Identification Algorithm

(a) North America

(b) Japan

(c) Western Europe

Figure 3: Geographic distribution of semiconductor inventors and cluster identification

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 US)	[10,20]	85,209	34	85,175		100%
	(20,…)	56,669	10,140	46,529		22%

Table 2: Comparing organically identified geographic units

	Distance (in miles)	Obs	Different units	Same unit	Correct classification
Our organic clustering (US)	[10,20]	71,095	37	71,058	100%
	(20,…)	33,647	19,392	14,255	58%
Our organic clustering (non-US)	[10,20]	85,209	297	84,912	100%
	(20,…)	56,669	34,335	22,334	61%
Our organic clustering (all locations)	[10,20]	156,304	334	155,970	100%
	(20,…)	90,316	53,727	36,589	59%
Hierarchical clustering (US)	[10,20]	71,095	16	71,079	100%
	(20,…)	33,647	15,780	17,867	47%
Hierarchical clustering (non-US)	[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%