A New Method for Identifying and Delineating Spatial Agglomerations with Application to Clusters of Venture-Backed Startups

Edward J. Egan, Stowe, Vermont

James A. Brander, University of British Columbia

December 2020

Please send comments to ed@edegan.com or james.brander@sauder.ubc.ca

Acknowledgment: We wish to thank Brian Ayash for constructive comments and Anne Dayton, former Researcher Manager at the McNair Center for Entrepreneurship and Innovation at Rice University's Baker Institute, for her invaluable support. We are also grateful for the hard work of Kryan Adams, Oliver Chang, Peter Jalbert, Christy Walden, and the other McNair Center students who contributed to this research. We gratefully acknowledge financial support from SSHRC grant 435-2017-0627.

A New Method for Identifying and Delineating Spatial Agglomerations
with Application to Clusters of Venture-Backed Startups

Abstract:

This paper advances a new approach based on hierarchical cluster analysis (HCA) for identifying, delineating, and analyzing agglomerations of economic activities. We apply the approach to clustering of U.S. venture-backed startup firms. HCA identifies nested clusters at varying aggregation levels. We propose a method for selecting the most economically meaningful aggregation level and associated clusters. We also describe a related method that includes a larger periphery in estimated clusters and use heat maps to illustrate the evolution of clusters over time. We indicate how our methods can aid in evaluating cluster support policies and also suggest other potential applications.

Keywords: Agglomeration, Startup, Hierarchical Cluster Analysis, Venture Capital, Entrepreneurship

JEL codes: R12, G24, L26

A New Method for Identifying and Delineating Spatial Agglomerations with Application to Clusters of Venture-Backed Startups

1. Introduction

Understanding the nature of spatial agglomeration in economic activity is fundamental to economic geography, regional science, and related areas. In recent years, this understanding has become particularly important for public policy, as many governments have sought to promote geographic agglomerations or clusters[1] of business activity in emerging industries. In the United States, such policies are carried out by U.S. federal, state, and local governments. These policies are also common in other countries, as illustrated by Canada's Supercluster Initiative and the European Cluster Excellence program.

Proponents of cluster support policies often use "Silicon Valley" as the canonical example of a successful cluster (or network of clusters) and seek to develop a critical mass of interconnected entrepreneurs, early-stage investors, and skilled workers in a specific location. Such policies are often justified using the agglomeration benefits identified in Marshall (1920), particularly labor market externalities, informal information flows, and integrated networks of suppliers and customers.[2] As these factors are not fully internalized by individual entrepreneurs and investors, market failure is likely. Subsidies, tax concessions, infrastructure development, and other policies are often used to encourage cluster development on this basis.

One challenge with such policies is determining the appropriate geographic boundaries for a targeted cluster. Existing policies vary dramatically in their geographic scope. The U.S. Regional Innovation Cluster program supports clusters that are broad in geographic scope, such as the Great Plains Technology and Manufacturing Cluster (in Kansas and Missouri). At the other end of the spectrum, publicly supported clusters such as the Cortex Innovation District in St Louis and the Oklahoma City Innovation District are much smaller in geographic scope and are supported primarily by local governments. If agglomeration economies decay rapidly with distance, then target locations should be small, and a policy that spreads out resources over a large geographic area will be inefficient. But excessively small geographic domains may also be inefficient.

The primary objective of this paper is to advance a new approach for identifying economically meaningful clusters and estimating their geographic scope. Our approach is based on hierarchical cluster analysis (HCA), which is commonly used as a method for putting objects of interest in hierarchical categories such as in biological taxonomy or in classifying historical

[1] We use the terms "agglomeration" and "cluster" interchangeably, as is consistent with most dictionaries and what we take to be common practice.

[2] Faggio et al. (2020) provide a recent study of the Marshallian microfoundations of agglomeration.

documents. We consider two related variants of this approach. We call our main variant the "HCA-regression method" and we believe it to be wholly original as a method for identifying clusters. Our other variant, the "elbow method", has been used in other clustering applications. We show how it can be applied to geographic clustering. We also demonstrate how to assess cluster stability and other aspects of cluster dynamics using "heat maps".

Our second objective is to illustrate our methods using an example of particular relevance for cluster support policies. Specifically, we focus on U.S. entrepreneurial ventures that receive venture capital funding. Most such ventures are in the high-tech sectors and have growth characteristics that are the focus of policy attention. We investigate clustering of venture-backed startups, identify their location and estimate their geographic scope.

Our methods contribute to the basic task in regional science of assessing the amount of clustering exhibited by some activity, as well as identifying the location and boundaries of clusters. We believe that these methods offer significant value-added relative to established methods of measuring and identifying economically meaningful geographic agglomeration, as described in more detail in our literature review. At a minimum, our methods provide a complementary approach to cluster identification. We propose that our methods can be used as an input to public policy analysis and to decision-making by entrepreneurs, venture capitalists, and other market participants.

Section 2 provides a literature review and Section 3 describes our data. Section 4 outlines the HCA approach and Sections 5 and 6 present the elbow method and the HCA-regression method respectively, along with illustrative examples. Section 7 summarizes the overall pattern of agglomeration of startup firms in the United States and Section 8 contains concluding remarks.

2. Literature Review

The concentration of economic activity in agglomerations or clusters has a long history of study in economics, including the classic work of von Thünen (1826) and Marshall (1920). However, the identification and measurement of agglomeration using formal statistical methods is relatively recent.

One important recent contribution is Krugman (1991) who uses a Gini coefficient approach to study the agglomeration of manufacturing industries. Given some variable of interest, such as employment, we can define a Gini coefficient for industry *j* as $G_j = \sum_i (x_i - s_{ij})^2$ where x_i is the share of aggregate employment in location *i* and s_{ij} is the share of industry *j*'s employment in location *i*. This Gini coefficient is 0 if employment in industry *j* has the same pattern as aggregate employment, while clustering of a particular industry in a particular location will generate a positive coefficient. This allows identification of locations where clustering of a particular industry occurs. Ellison and Glaeser (1997) provide an adjustment to this approach to

take account of indivisibilities in plant size. Applications of this adjusted Gini coefficient method for identifying clusters include Rosenthal and Strange (2001) and Ellison et al. (2010).

One limitation of the Gini coefficient approach is that the geographic unit is set exogenously. Typically, the investigator decides on some geographic unit of analysis such as states or metropolitan statistical areas (MSAs) and then investigates whether a given state or MSA exhibits clustering of a particular industry or activity. However, if the relevant scope of agglomeration is either larger or smaller than the assumed unit of analysis, the results can be misleading. One advantage of our approach is that we do not impose exogenous geographic units on the analysis. Instead, we allow variation in the size of clusters that is endogenous and (implicitly) continuous.

We are not the first to move beyond exogenous geographic units. Duranton and Overman (2005) provide a method of measuring what they call "localization". They calculate the distance between each pair of plants in some industry and check whether, at any given distance, the density of plants is higher than would be expected based on random allocation. They find that densities are high relative to random allocation at short distances and are lower at greater distances, implying that plants in the same industry tend to cluster together. This method yields a continuous relative density function of the underlying activity rather than identifying boundaries of a discrete cluster.

An interesting related approach is provided by Buzard et al. (2017) who use "point pattern" methods to study the agglomeration pattern of U.S. R&D labs. This method shows the pattern of agglomeration at various different scales (within ½ mile, within 1 mile, within 5 miles, etc.) and demonstrates that clustering seems to be important at most scales, but particularly at very small scales, as is consistent with Rosenthal and Strange (2003) and Verstaten et al. (2020).

Each of the methods described above has proven valuable and has been applied in subsequent work. Our work offers value-added in several areas. Like Buzard et al. (2017), we can analyze clustering at different scales. However, our scales are endogenous and allow for clusters of very different sizes. Our approach also provides specific boundaries for each cluster. Also, in addition to using locational data, we use economic data that helps identify policy-relevant clustering patterns, and we examine cluster dynamics over time.

We focus on venture-backed startups as our illustrative example partly because of the literature showing the importance of entrepreneurship in emerging clusters, including Delgado et. al (2010), Chatterji et. al (2014), and Andersson and Larsson (2016). Also, we note the line of research started by Acs and Audretsch (1990) showing that venture-backed startups are associated with high levels of innovation. Therefore, the clusters we identify are likely to have high levels of innovation, making our analysis complementary to Carlino and Kerr (2015), who

review agglomeration and innovation, and to Buzard et al. (2017) who investigate agglomeration of R&D labs.

3. Data

Our data on venture-backed startups comes from Thomson-Reuters' VentureXpert database, which is described in Da Rin et al. (2010), Kaplan et al. (2016), and elsewhere. A potential alternative would be to use business registration data, as in Guzman and Stern (2015), which includes both venture-backed and non-venture-backed startups. However, for our purposes, the Thomson-Reuters data has the advantage that it includes a performance measure in the form of venture capital funding, which is essential for our analysis. And we would expect clustering of non-venture-backed high-tech startups to be correlated with venture-backed startup activity in any case.

We use U.S. startup data from 1995 through 2018, giving us 23 years of annual data. For this period, the VentureXpert database has nearly complete population coverage of U.S. venture-backed startups. To be in our data set, a firm must have received a "growth" (i.e., seed stage, early stage, or later stage) venture capital investment between 1995 and 2018. It continues in the data until it has an "exit" – an initial public offering or is acquired—or until it goes 5 years without a new venture capital investment.

The total number of distinct venture-backed startups is 32,106, each of which is in the data for multiple years. We also need the startup's address in usable (geocodable) form, which eliminates a small fraction of the firms, leaving 31,543 startups. We match these startups with Census Places (CPs), often called "cities and towns". We use the full list of CPs from the 2010 Census, which includes 29,573 CPs. While almost all startups are located in CPs, a few are not, and therefore are dropped from the data.

In addition, many CPs are small towns or unincorporated areas with few if any startups. We drop such CPs from the analysis as they do not contain meaningful startup agglomerations. Specifically, for inclusion, we require a CP to have at least one year in which 10 or more startups received venture capital investment. Also, seven CPs are geographically intertwined with other CPs. We simply combine the CPs in those cases, leaving us with 198 geographic units, all of which are cities. We use the term "city" rather than CP from now on.

We also drop specific city-year combinations for which there are fewer than six active startup firms. We are left with 22,979 distinct startups in 198 cities. The number of distinct startup-year combinations is 151,940 and the number of city-year combinations is 3,797.

4. Hierarchical Cluster Analysis

4.1 Methodological Overview

Hierarchical cluster analysis (HCA) is a long-established technique with many applications. Versions of HCA are available in most major statistics packages and scientific computing languages. Standard references on HCA include Ch. 4 in Everitt et al. (2011) and Contreras and Murtagh (2015).

HCA identifies nested categories or groups of some underlying population. The output of HCA is a full set of nested categories, including both high and low levels of aggregation. For example, we might classify plants based on their genetic characteristics (as in Peeters and Martinelli, 1989). At a high level of aggregation, we might have groups such as "grains" and "fruit". At a lower level of aggregation, within the grains group, we might have "wheat" and "corn" and within the wheat group at a still lower level of aggregation we might have No. 2 winter wheat, etc.

HCA can be applied to any collection of objects for which there is a metric that quantifies differences between objects. In some applications, such as identifying plants based on gene sequences, or identifying the authorship of documents based on word choice (as in Aljumily, 2015), the appropriate metric may not be obvious. In our case, however, the natural metric is Euclidean distance as in Cesario et al. (2020). Clusters are groups of startups that are relatively close to each other.

We use top-down or "divisive" cluster analysis—starting with the largest possible cluster in the area under study. Successive layers are created by subdividing clusters in previous layers, providing finer and finer partitions of the data. The process continues until, in the final layer, each startup location is an isolated point. The alternative to divisive clustering is agglomerative or bottom-up clustering, which starts with isolated points and then combines them in a step-by-step process.

If we plot U.S. venture-backed startups on a map of the United States, we would observe the San Francisco Bay area as a geographic cluster of activity. At a lower level, each of San Jose, Berkeley and various other cities in the Bay area might be clusters. And, at a still finer level of aggregation, neighborhoods within San Jose or Berkeley (or elsewhere) might be clusters.

We wish to do more than just generate a nested classification system. We also wish to identify the level of aggregation that best identifies economically meaningful clusters. In our HCA-regression method, we do this using the dynamic pattern of venture capital investment. In this analysis, clusters are groups of venture-backed startups whose evolution best explains the temporal pattern of venture capital investment.

We could, in principle, start with the entire United States as the largest possible cluster in the HCA process. In practice this is not possible for computational reasons. The calculations required to subdivide a cluster increase exponentially in the number of items in the cluster, so there is a huge computational saving if we start out with smaller initial units. We use U.S. Census Places (CPs) as the largest geographic unit. Examples of CPs include Berkeley, San Francisco, and San Jose. This might seem like a low level of aggregation to start with. It is feasible to start with a higher level, such as MSAs, or possibly even states. However, we can tell from our results that there would be little gain in starting at the state level or MSA level as the clusters we identify are generally much smaller than the CP level, as is consistent with Buzard et al. (2017) and Rosenthal and Strange (2003).

4.2 Clustering Algorithms

There are several widely-used clustering algorithms. They differ in how they use distance in the clustering process. They may use absolute distance, squared distance, or some other transformation of distance, and they may focus on within-cluster distances or between-cluster distances. The different algorithms may yield different nested structures, so it is important to use an algorithm that is suitable for the task at hand. Most can be used for either agglomerative or divisive clustering, yielding the same nested structure in both cases.

Investigations of different algorithms as reviewed in Everitt (2011, Ch. 4) indicate that no method is universally preferred. However, Ward's algorithm, which is based on squared Euclidean distance, generally performs well both computationally and in terms of generating an accurate and meaningful nested structure of clusters. It is particularly appealing for geographic contexts.

Ward's approach was developed for agglomerative clustering. At each step, the agglomerative algorithm combines clusters so as to minimize the sum of squared distances between each element in the new cluster and the midpoint of the cluster. Ward's approach can be applied (in reverse) to divisive clustering: At each step, the divisive version of Ward's algorithm subdivides clusters so as to minimize the sum of squared residuals between the cluster elements and the cluster midpoints of the two new sub-clusters.

We use a computationally efficient algorithm that produces the same results as Ward's algorithm for divisive HCA. Specifically, we use a divisive nearest-neighbor chain algorithm written in Python and available from scikit-learn.[3]

[3]Scikit-learn is an open-source resource that supports machine learning using Python. It is available at https://scikit-learn.org. The module we use is available through Scikit-learn's AgglomerativeClustering module. Divisive clustering using Ward's algorithm is available using the appropriate selection of parameters in the module.

4.3 Layers of Clustering

The first step is to identify the hierarchical structure of clustering for each city-year in the data. A key innovation in our approach is that we convert groups of locations from the HCA to clearly defined geographic clusters. To do this, it is necessary to specify the acceptable shapes for clusters. We assume two standard conditions for the shape. It should be convex: If two startups are in the same cluster, then any startup located on a straight line between those two startups should also be in the cluster. And the shape should be no larger than necessary to contain all startups in the cluster. These two properties imply that the shape should be a "convex hull" – the convex polygon with the smallest perimeter containing all startups in the cluster. A convex hull might be a triangle, a diamond shape, or a more complex polygon. A convex hull may also be degenerate in that it may be a point, as when there is just a single isolated startup, or it may be a line segment, as when there is an isolated pair of startups in nearby locations.

Isolated single startups or pairs of startups cannot reasonably be called be called clusters in the economic sense. From now on, we refer only to non-degenerate convex hulls created by our HCA algorithm as clusters, and we refer separately to lines and points. We sometimes refer to the full set of clusters, line segments, and points as "groups", which is common terminology when using statistical methods for classification.

We describe the different levels of aggregation in the clustering structure as "layers".[4] The first layer puts all startups in a given city in a single cluster. In the second layer, the initial cluster is subdivided into two. If, for example, an initial single cluster of 20 startups in a city consists of 5 startups in close proximity to each other near one edge of the city and 15 startups on the other side of the city, the algorithm would convert those two groupings to two distinct clusters in the second layer.

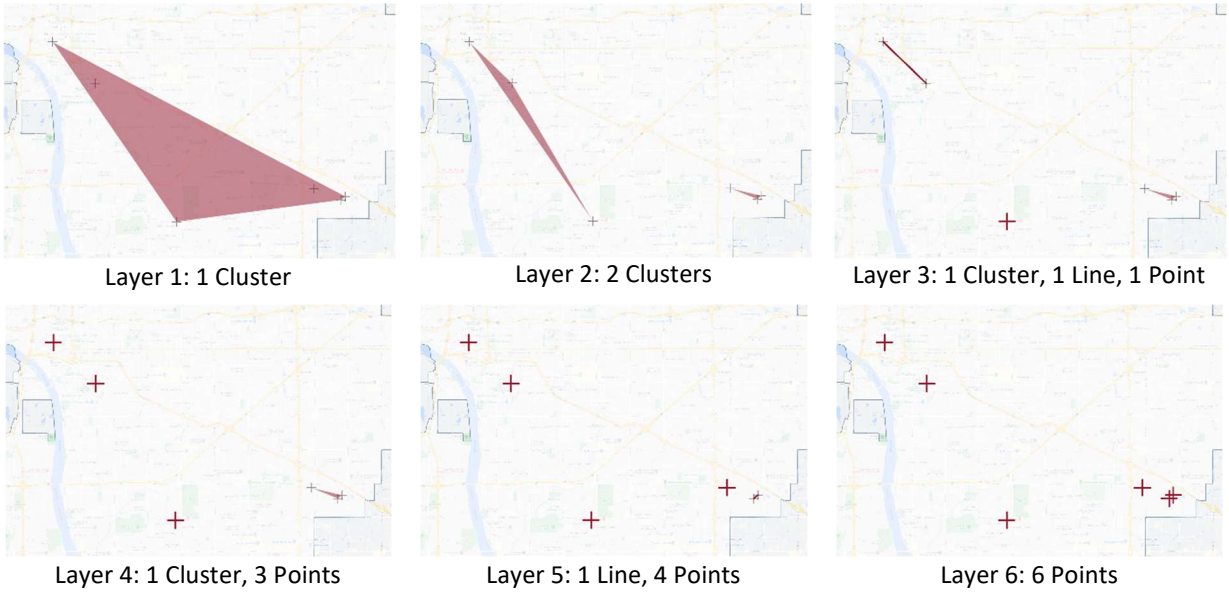
The third layer of clustering subdivides one of the two clusters in the previous layer—making whichever partition provides the biggest reduction in the sum of squared distances of startups from the cluster midpoints. This process continues, doing one new subdivision at each step until eventually each startup location is just one point on the map. As the process proceeds, single startup locations may be split off into points, or pairs of locations may be split off into line segments. It is possible to have two or more startups in the same building and therefore at a single point, and it is possible to have three or more startups on a line segment.

Figure 1 illustrates this process for Tulsa in 2003, which is among the simplest possible examples as it has only 6 startups—the minimum number allowed for inclusion in our data set.

[4] The results of HCA can be illustrated in a tree diagram or "dendrogram" and different levels of aggregation are sometimes called "steps" in the “agglomeration schedule”.

This process is shown in a series of cluster maps for the corresponding layers in the HCA process.

Figure 1: Decomposition of Tulsa, OK, 2003 into 6 Layers (shown at 1:130,000 scale)



Layer 1: 1 Cluster | Layer 2: 2 Clusters | Layer 3: 1 Cluster, 1 Line, 1 Point

Layer 4: 1 Cluster, 3 Points | Layer 5: 1 Line, 4 Points | Layer 6: 6 Points

The number of layers for a given city-year is always equal to the number of startup locations in that city-year. Most city-year combinations have more startup locations than Tulsa and therefore have more layers. San Francisco and New York each have over a thousand startup locations for most years and therefore more than a thousand layers.

For a given city-year, each layer will have a number of clusters, a number of lines, and a number of points, which sum to the number of groups. Layer 1 always has one cluster and no lines or points. Layer 2 normally has two clusters. The number of clusters tends to increase as layering proceeds until clusters are gradually replaced by points and line segments. As shown for Tulsa in 2003 in Figure 1, Layer 1 has one cluster, Layer 2 has two clusters, and Layer 3 reverts to only one cluster (along with one line segment and one point). Layer 4 has one cluster and three points, etc. The number of clusters goes, in order, 1, 2, 1, 1, 0, 0 in this case. However, in general, neither the increase nor the decrease is guaranteed to be monotonic.

Identifying clusters for a city-year amounts to choosing the appropriate layer. For Tulsa, if Layer 2 is the appropriate layer for 2003, that implies that Tulsa had two clusters in 2003. But if Layer 1 was the appropriate layer, then we would conclude that Tulsa had only one cluster in 2003 and if Layer 6 were correct, we would say there were no clusters, just six seemingly unrelated startups. The cluster map for the chosen layer shows the precise location and delineation of any clusters for the city-year under consideration.

Therefore, the next step is to select the appropriate layer and associated cluster map for each city-year. We propose two methods, the elbow method and the HCA-regression method. These methods can be used in conjunction to provide a reliable basis for cluster identification.

5. Choosing a Clustering Layer Using the Elbow Method

With HCA, it is possible to apply heuristic methods based entirely on the underlying metric (Euclidean distance in our case) to identify the best layer and therefore delineate clusters. The literature in this area is small, but the best-known such heuristic is probably the "elbow" method, first suggested by Thorndike (1953) and described in, for example, Bholowalia and Kumar (2014) and Ketchen and Shook (1996).

We are using Ward's algorithm for divisive HCA, which efficiently partitions observations into groups such that new step in the hierarchical structure (i.e. each new layer) maximizes the increase in explained variation. The elbow method is a natural extension of Ward's algorithm as it selects a particular layer (and therefore a particular structure of clusters) based on marginal changes in this explained variation.

For any city-year, each startup location is a point in two-dimensional space. Denoting the mean location of all startups as y^{**}, the total sum of squares, $\mathrm{SS}_{\mathrm{tot}}$, is the sum of squared distances of startups from the overall mean location.

$$\mathrm{SS}_{\mathrm{tot}} = \sum_i (y_i - y^{**})^2 \tag{2}$$

For any specific layer of the HCA, and its corresponding groups, we can divide this total sum of squares into the explained sum of squares and the unexplained sum of squares, just as when doing analysis of variance (ANOVA). If statistical cluster j contains n_j startups and has mean location y_j^*, the explained sum of squares, $\mathrm{SS}_{\mathrm{exp}}$, is:

$$\mathrm{SS}_{\mathrm{exp}} = \sum_j n_j \left(y_j^* - y^{**}\right)^2 \tag{3}$$

When using ANOVA this explained sum of squares is sometimes called the "between" sum of squares as it is based on the distances between groups. The unexplained sum of squares $\mathrm{SS}_{\mathrm{unexp}}$, sometimes called the "within" sum of squares, is the sum of the within-group sums of squares. The explained and unexplained sums of squares must sum to the total sum of squares.

$$\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{\mathrm{exp}} + \mathrm{SS}_{\mathrm{unexp}} \tag{4}$$

For each layer, Ward's algorithm chooses groups so as to maximize the ratio, R^2, of the explained sum of squares to the total sum of squares. This is the standard meaning of the R^2 statistic – the ratio of the explained sum of squares to the total sum of squares.

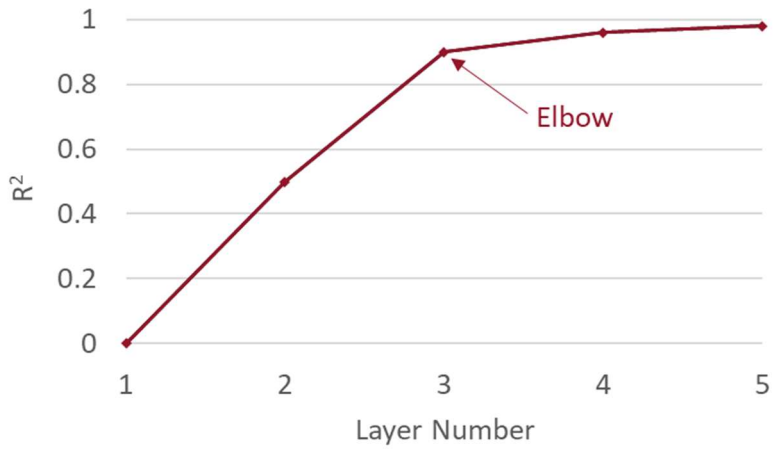
$$R^2 = \frac{\mathrm{SS}_{\mathrm{exp}}}{\mathrm{SS}_{\mathrm{tot}}} \tag{5}$$

The elbow method focuses on how this sum of squares ratio, R^2, varies from layer to layer. In layer 1, all startups are in a single cluster and so the explained sum of squares is zero: $R^2 = 0$. As we proceed through the layers, this R^2 value increases until the final layer, in which each location is in a separate group so $R^2 = 1$.

It is helpful to consider an extreme case. Suppose that a triangular city has a concentration of startups in each corner and none elsewhere. For Layer 1, there is a single cluster including all startups in the entire city. The midpoint of that cluster is the same as the overall midpoint, so $y_j^* = y^{**}$ and $\mathrm{SS}_{\mathrm{exp}} = 0$ from (3) and $R^2 = 0$ from (5). For Layer 2, one corner will become a separate cluster. The within-group or unexplained sum of squares for the other cluster is still substantial, but the unexplained (within-cluster) sum of squares for the separated cluster is now very low. The overall unexplained sum of squares drops sharply and the R^2 rises correspondingly.

For the third layer, all three corners are separated as distinct clusters and the overall explained sum of squares and the R^2 rise sharply again. However, further layers will provide only slight increases in the R^2 by forcing the clusters in corners to subdivide further. A plot of R^2 as a function of the layer number would look like Figure 2.

Figure 2: Hypothetical Elbow Method Example



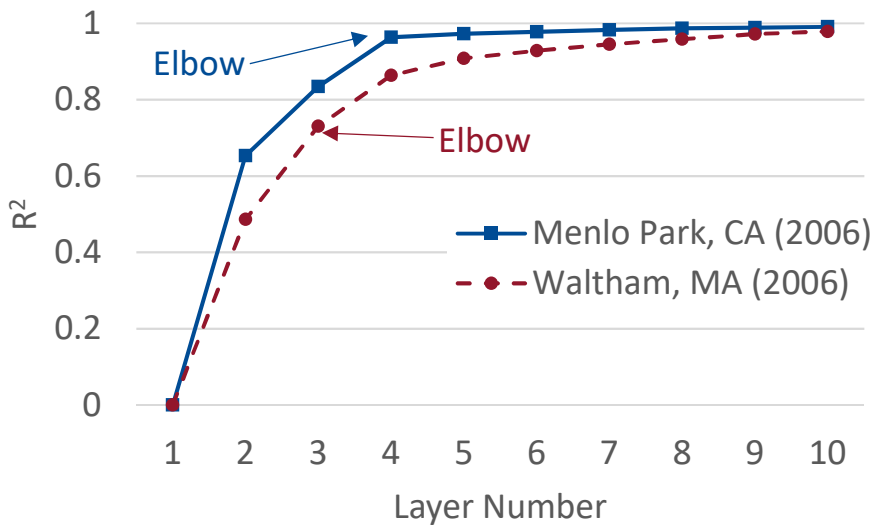
The elbow method selects the layer after which the marginal increase in R^2 drops sharply. It appears as an "elbow" in the diagram. In this example, the elbow appears at layer 3, in which there are 3 clusters. This is, by construction, the correct number of clusters, as least from a purely locational point of view, with one in each corner of the city.

The elbow in Figure 2 is easy to identify by visual inspection. More generally, it is necessary to have a specific rule for identifying the elbow. There are several reasonable possible rules. The rule we use is to choose the layer after which the marginal increase in R^2 drops most sharply, as

in Figure 2. This is where the second difference in R^2 is minimized. This is a natural choice for a stopping rule but reasonable alternatives would be to use the degree of concavity (the ratio of the second difference to the first difference) or the change in elasticity. One feature of our choice is that increases in R^2 tend to be large for the first few layers then decline sharply. Therefore, declines in the improvement rate also tend to be large in the first few layers, resulting in an elbow selection that also arises early on. We adjust for this by ruling out the first second difference as a candidate selection.

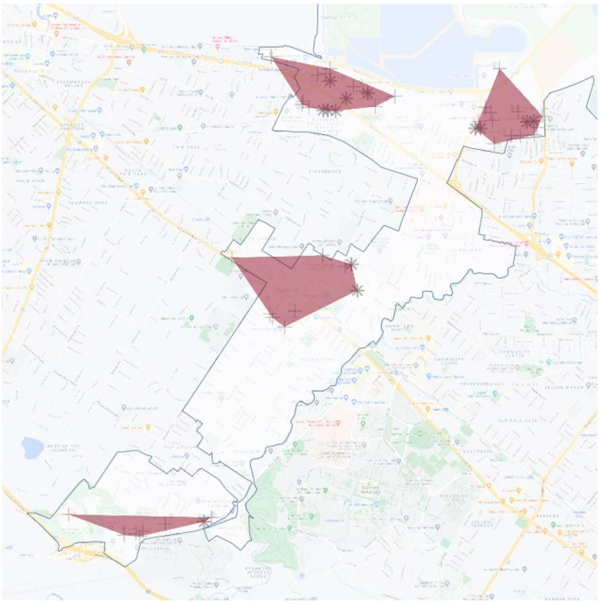
Figure 3 illustrates the elbow method for two relatively small cities in 2006, located in major high-tech areas. It shows Menlo Park, CA which is next to Stanford University in the heart of Silicon Valley, and Waltham, MA which is inside Route 128, the inner beltway of the Boston-Cambridge-Newton MSA.

Figure 3: Using the Minimum Second Difference to Locate the Elbow

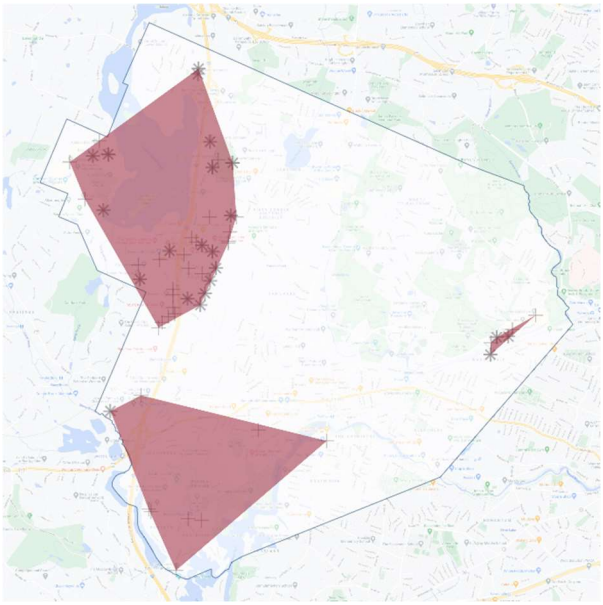


For Menlo Park the graph is every bit as definitive as the example in Figure 2, and the elbow occurs at Layer 4. For Waltham, the elbow is less obvious as it looks like either Layer 3 or Layer 4 might be best. It is close, but Layer 3 is the preferred layer. Having selected the layer using the elbow method, we can draw the associated cluster maps, showing that Menlo Park has three clusters and Waltham has four.

Figure 4: Elbow-based Cluster Maps at 1:65,000 scale



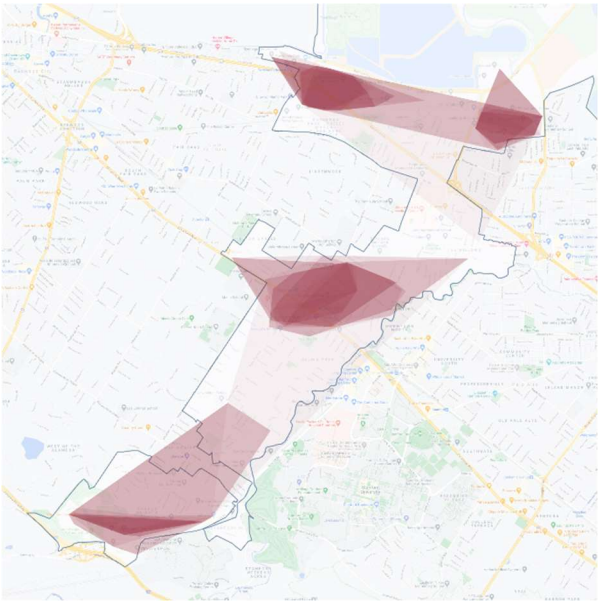
Menlo Park, CA, in 2006



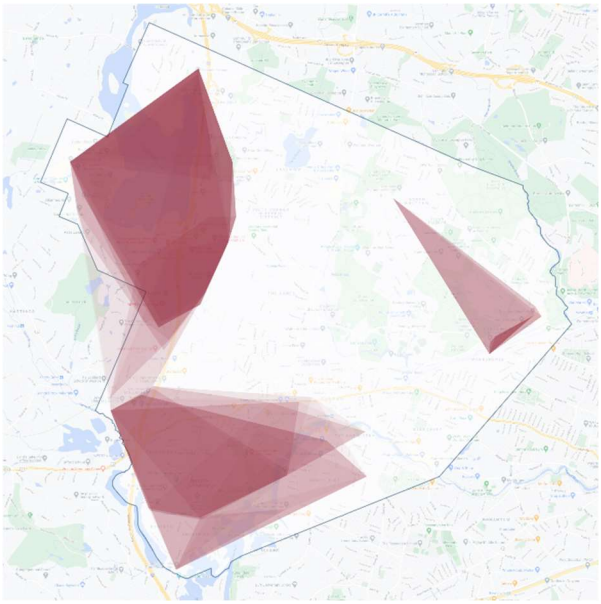
Waltham, MA, in 2006

Figures 3 and 4 apply only to 2006. If the clusters are meaningful we would expect them to persist over time. We can illustrate the stability of clusters over time by using heat maps. A heat map is a data visualization technique that uses color or shading intensity to illustrate how some phenomenon varies over some dimension of interest (time in our case). For a given city, we take the cluster maps for each year and lay them on top of each other. For each year, the clusters are lightly shaded. As a cluster or section of a cluster is identified in more and more years, it takes on an increasingly dark hue, as illustrated in Figure 5 for Menlo Park and Waltham.

Figure 5: Elbow-based Heat Maps at 1:65,000 scale



Menlo Park, CA, 1995-2018



Waltham, MA, 1995-2018

The elbow method is a step up in rigor from just looking at map of startups and using subjective visual inspection to identify clusters. However, like visual inspection, it relies entirely on spatial relationships between startups. It is fairly "conservative" in that it does not have much sensitivity to pruning outliers. For example, in the lower left (southwest) cluster in Menlo Park, the core of the cluster is fairly small, but in a few years a couple of outliers north and east of the main cluster cause the estimated cluster to be much larger than the core (dark shaded) area. Nevertheless, these heat maps show a fairly clear picture. The dark areas in both cities reflect stable venture-backed startup communities.

6. Choosing a Clustering Layer Using Regression

In this section we illustrate our HCA-regression method for estimating and delineating clusters. One difference from the elbow method is that the regression method uses economic information in the form of venture capital investment data to augment locational data. We suggest that the resulting estimated clusters have more economic and policy relevance than estimates derived from purely locational information.

The first step in our HCA-regression method is to define a representative layer for each possible number of clusters in a given city-year. If we look at Figure 1, showing the series of HCA layers for Tulsa in 2003, we see that Layer 2 is the only layer with exactly two clusters. Therefore, Layer 2 is the representative two-cluster layer. However, it is not immediately obvious which layer provides the best one-cluster representation as Layers 1, 3, and 4 all have one cluster.

We suggest that Layer 3 is the most representative one-cluster layer. Layer 1 is not ideal because it subdivides naturally into two sub-clusters. This cluster is "unstable" in that it does not maintain its approximate size and shape beyond one layer. Layer 4 is arguably less suitable than Layer 3 because, beyond Layer 3, all we are doing is forcing the HCA process to drop individual startups or pairs of startups from estimated groups. In general, the cores of identified clusters do not change much as we proceed beyond the representative layer and the general location of the clusters remains stable. This property is implied by following formal definition.

Definition: The *representative m-cluster layer* is the first layer with *m* clusters that has the property that no later (higher-numbered) layer has more clusters.

If only one layer has *m* clusters, that layer must be the representative *m*-cluster layer (as with Layer 2 for Tulsa in 2003). If multiple layers have m clusters, this definition chooses the most representative of those layers. Therefore, for each city-year, we have a representative m-cluster layer for each value of *m* from 1 to up to the maximum number of clusters for that city-year.

The next step for a given city is to run a series of regressions, using annual observations, for that city. One regression is based on the representative one-cluster layers, one is based on the representative two-cluster layers, and so on. The dependent variable in each regression is next year's venture capital investment in the city, and is the same for each regression. The explanatory variables reflect various aspects of the cluster structure implied by the representative layer for that regression and therefore differ across regressions. Our objective is to choose the "best" of these regressions. That regression identifies the estimated number of clusters and the associated cluster structure for the city.

For a city's *m*-cluster regression, we find the representative m-cluster layer for each year. The maximum number of clusters may vary from year to year. Therefore, for relatively large values of *m*, there may be not be an observation for each year. We impose a requirement of at least 12 annual *m*-cluster observations for cluster number *m* to be considered.

The explanatory variables we use for these regressions are based on eight underlying variables: the number of points, the number of line segments, the number of clusters, the number of startups in points, the number of startups on line segments, the number of startups in clusters, the total length of line segments, and the total area of clusters. These variables describe the cluster structure. For a given city, this information varies slightly from year to year, as does the amount of venture capital investment the following year.

We form principal components of these eight variables. We use principal components because there is a high degree of multicollinearity among the eight underlying variables. In addition, given the modest number of observations in each regression, we wish to reduce the

dimensionality of the data set. In addition, almost all the relevant variation in the data is contained in three principal components. Using principal components makes all the regressions comparable and efficiently uses the information in the data.

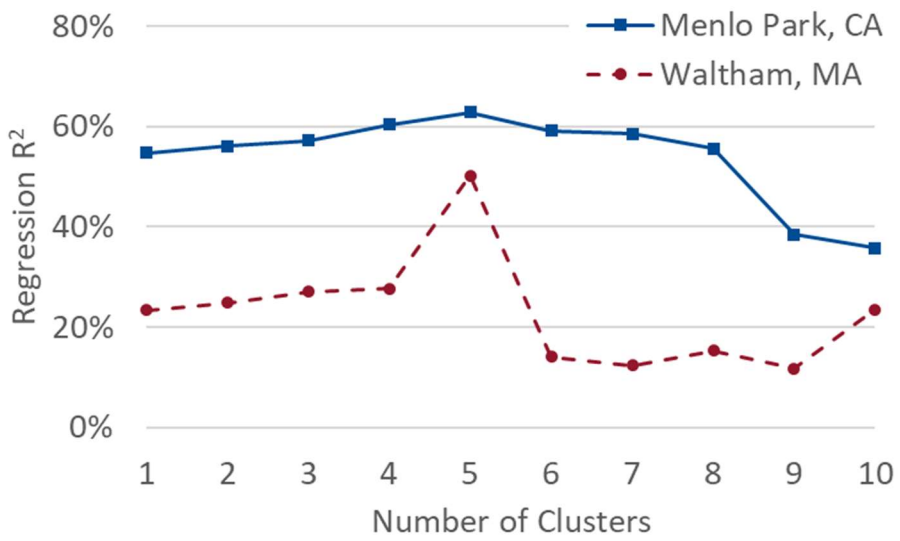
The regression specification for each cluster number m for each city is as follows:

$$y_{t+1} = a_m + b_m C_{mt} + e_{mt} \quad (5)$$

where y_{t+1} is the amount of venture capital invested in the city in year $t + 1$, a_m is a constant term for representative layer m, b_m is a vector of slope parameters, C_{mt} is a vector of principal components for representative layer m at time t, and e_{mt} is a random error for representative layer m at time t. We run this regression for each admissible cluster number m.

The criterion we use to select the "best" regression is to choose the regression with the highest R^2 statistic. This is conceptually consistent with the approach used by Ward's algorithm for HCA and with the elbow method. Also, as we use principal components, this R^2 criterion coincides with other commonly used goodness of fit criteria such as the Akaike Information Criterion, the Bayesian Information Criterion, or the adjusted R^2. Figure 2 illustrates how the R^2 varies with the number of clusters for Menlo Park and Waltham. For both Waltham and Menlo Park, the estimated number of clusters is 5.

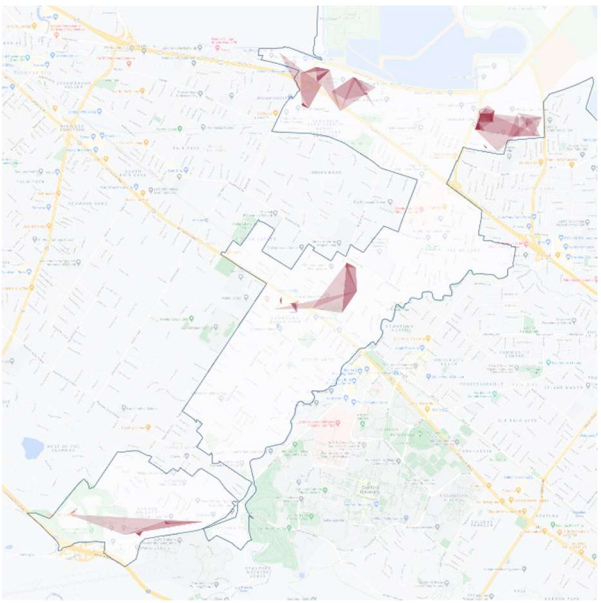
Figure 6: Using Regression R^2 Statistics to Select the Number of Clusters



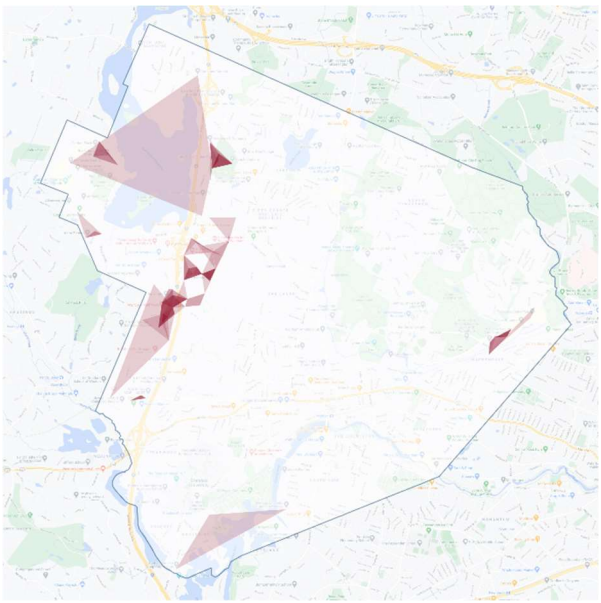
The regression method selects just one cluster count for each city. That cluster count applies to every year for which the requisite number of clusters can be identified. It is possible for a city to have less than this count of clusters for some years. This typically happens early in the period before significant later entry. If that happens, then we conclude that the city had the next-highest number of clusters for those years. Menlo Park has 5 clusters in every year so this issue

does not arise. However, Waltham has only 4 in 1996 and 1997 (and 5 in every other year). In any case, if our method is estimating meaningful clusters, the cluster map should not change dramatically from year to year. Variation over time can be visualized using heat maps, as in Figure 7 for Menlo Park and Waltham.

Figure 7: HCA-Regression-based Heat Maps at 1:65,000 scale



Menlo Park, CA, 1995-2018



Waltham, MA, 1995-2018

Comparing these heat maps based on the regression method with the elbow method heat maps shows that both methods are identifying the similar core clusters in the same areas, but the areas identified by the regression method are smaller and more tightly focused around the cluster cores. More broadly, compared with the elbow method, the regression method provides more refinement, finding five clusters in any given year rather than four (Menlo Park) or three (Waltham). And the cluster boundaries identified by the regression method are tighter, resulting in smaller areas for the estimated clusters and for the associated heat maps.

For example, in Waltham, Figure 5 (using the elbow method) shows one fairly large cluster in the northwest part of the state. In Figure 7 (using the regression method), there are several distinct clusters in that same area, with tight boundaries, along with some light-colored areas that sometimes have enough startups to be grouped with one of the core areas in a cluster.

Our interpretation is that the elbow method and the regression method are consistent with each other but that the regression method provides more precision. The evolution of venture capital investment over time is better explained by smaller and more precise cluster estimates. Such an effect would arise if, for example, venture capitalists can more easily visit different entrepreneurs if they are located within a few blocks of each other rather than being a few

kilometers. apart, or if entrepreneurs are more likely to benefit from informal information flows (and therefore be more attractive to venture capitalists) if they are located very close to each other.

7. Aggregate Cluster Properties

The previous sections specify and illustrate how our HCA-based methods work, using specific examples. We now examine the clustering of all U.S. startups using our full data set.

7.1 Summary Statistics

One set of questions about the aggregate results relates to how far the HCA process progresses before reaching the selected cluster structure. Does this occur after only a few HCA layers, or does it occur further along in the HCA process? Answering this question requires comparing across cities and over years. As different cities have dramatically different numbers of startups and correspondingly different numbers of layers it is helpful to construct a layer index that allows comparisons across cities and over years.

If a given city-year has n startup locations, it also has n layers in the HCA process. The index for layer j, $L(j)$ is

$$L(j) = \frac{j-1}{n-1} \tag{6}$$

As the first layer is layer 1, using this formula allows the index to go from 0 to 1. We normally use percentages, so the index goes from 0% to 100%, as in Table 1, which provides some summary statistics regarding clusters estimated using both methods

Table 1: Cluster City-Year Summary Statistics, All U.S. Startup Cities 1995-2018

	HCA-Regression (N=3,474)			Elbow (N=3,770)		
	Median	Mean	SD	Median	Mean	SD
Layer Index (%)	40	36.1	24.2	18	20.3	14.7
No. of Clusters	2	3.8	7.5	2	2.1	1.0
Startups in Clusters	9	19.5	38.2	13	37.1	84.5
Cluster Area (ha)	24	512	2,097	161	2,085	7,409
Cluster Density (Startups/ha)	0.5	9.8	59.0	0.1	1.1	10.5
Cluster Separation (km)	0.9	2.3	3.4	2.9	4.2	5.0
Total Number of Startups	16	41.3	87.0	15	39.0	83.9
VC Invested (2018 $m)	34.3	142.5	530.1	32.2	133.9	509.8

The middle columns of Table 1 report the characteristics of the cluster structures estimated by the HCA-regression method. Each city-year as a distinct observation. There are 3,797 city-years in our data set, but not every city-year has full information, leaving us with 3,474 observations.[5]

The minimum number of clusters a city-year can have is one. One way this occurs is if all startups in a given city-year are estimated to be in a "cluster of the whole", which occurs in layer 1 of the HCA process. Such outcomes show the absence of within-city clustering. More commonly, a low number of clusters occurs later in the HCA process after outlier startups are pruned from the main clusters. The representative 1-cluster layer can only be layer 1 if the city-year never achieves two clusters, which happens in 432 out of 3,474 city-years.

The median number of clusters is only two, even though the estimated structure occurs 40% of the way through the HCA process. For many city-years, earlier layers have more than two clusters but the HCA process pares down the structure down to identify a small number of "core" areas as clusters and the regression method selects those associated cluster structures.

Areas are shown in hectares. A hectare has the same area as a square that is 100 meters on each side. A typical city block in the U.S. Midwest has an area of about one hectare. So, for the regression method, the median estimated cluster covers about 24 hectares or 24 city blocks.

The median density of startups within clusters is about 0.5 startups per hectare, although there is considerable variation and skewness, with the mean being much larger than the median.

The elbow method results are shown in the last three columns of Table 1. We have results for almost the entire data set, losing less than 1% of the city-years due to data completeness issues. As with the regression method, it is possible to have a single cluster. It happens if the initial few layers remove outliers (isolated startup or pairs) in line segments or points, but do not generate new clusters.

The median number of clusters is only 2 with the elbow method as well. The elbow method generates estimated clusters that are larger in area but less concentrated than the regression method, with correspondingly lower densities. However, the cores of the elbow-method clusters closely match the clusters generated by the regression method.

Using the elbow method, the vast majority of startups in any city are allocated to a cluster rather than being pruned out as outliers. Specifically, at the median, a city-year has 15 startups and 13 of those are in clusters, with only 2 outside of clusters.

[5] The missing city-years occur when a given city does not have the requisite 12 years with a given cluster count that we require in order to use the regression method.

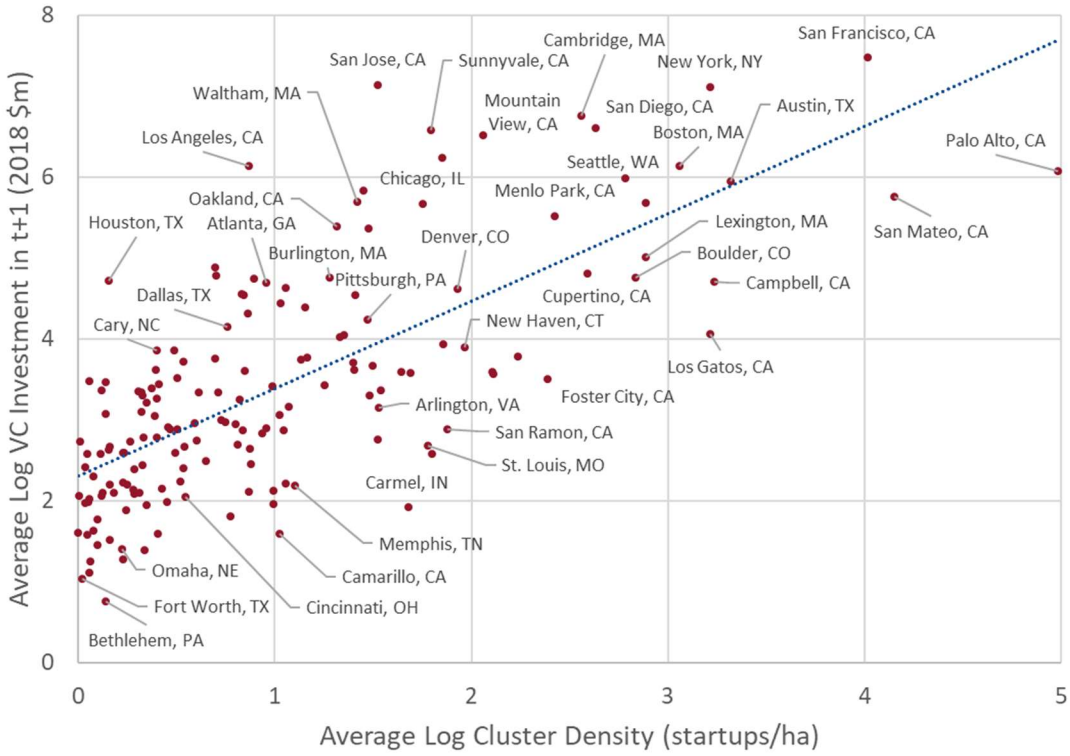
7.2 Cluster Density

One important question for both public policy and decision-making by entrepreneurs and venture capitalists relates to the optimal density of clusters and whether normal market forces would lead to insufficient clustering. The possibility of under-clustering arises from externalities. A profit-maximizing startup considering locating near other startups will consider any benefits it might get from close proximity to other startups. However, it will not internalize benefits that it might confer on other firms. It is therefore quite possible that actual cluster densities in the absence of government policy will be suboptimal.

The optimal density will not be the same for every type of startup and every location. Some locations and some industries would benefit from closer clustering that others. Even if the benefits of clustering were fully internalized by firms, we would observe different densities in difference clusters. Estimating the optimal cluster density is beyond the scope of this paper. However, our analysis provides some suggestive information, as shown in Figure 8, which shows venture capital investment plotted against cluster density.

Each point in the plot is a city. The x-coordinate for a city shows the average (log) cluster density in that city, based on the regression method, for the years that city is in the data. The *y*-coordinate shows the average (log) venture capital investment received in that city in the following year. Not all city-names are in the figure, due to crowding, but the cities with highest cluster densities and VC investment rates are shown. Palo Alto (next to Stanford University), San Mateo (also close to Stanford in Silicon Valley), and San Francisco are at the high end on both dimensions

Figure 8: The Relationship between Cluster Density and VC Investment using Layers Selected with HCA-Regression



The relationship illustrated in Figure 8 does not arise by construction. Although we expect VC investment in a city to be positively affected by the total number of startups, it would not necessarily be affected by the density of startups within clusters. This point is illustrated in Table 2, which reports regression results associated with Figure 8.

Table 2: Regressions of VC Investment on Cluster Density

The dependent variable is the log of venture capital investment (2018 $m). Robust standard errors, clustered at the city level, in are parentheses. *** indicates p<= 0.01.

	(1)	(2)
Log Hull Density (startups/ha)	1.080***	0.255***
	(0.0897)	(0.0479)
Log Number of Startups		1.227***
		(0.0521)
Constant	2.312***	-0.651***
	(0.0996)	(0.153)
Observations	166	166
R-squared	0.51	0.865

Specification 1 shows the regression in Figure 8. Specification 2 adds the average (log) number of startups in the city as an explanatory variable. The number of startups has a strongly significant positive effect on venture capital investment but, even after adjusting for that effect, cluster density still has a strongly significant effect.

We do not wish to overstate the implications of Table 2 and acknowledge that rigorous demonstration of any causal effects would require attention to potential endogeneity issues. However, Table 2 and Figure 8 provide at least suggestive information regarding the effects of cluster density.

8. Concluding Remarks and Further Applications

The primary objective of this paper is to introduce a new approach, based on hierarchical cluster analysis (HCA), for identifying and delineating clusters of economic activity. We illustrate this approach using venture-backed startups.

We describe two variants of the approach. Our primary variant, the HCA-regression method, combines locational information and economic information to identify and delineate clusters of venture-backed startups at potentially policy-relevant scales. The estimated clusters are relatively small, with a median size of about 24 hectares (i.e. about 24 city blocks), although with substantial variance and skewness. Many cities apply cluster support policies to areas of comparable size. Our other variant, which uses only locational information, is the elbow method. This method estimates larger clusters than the HCA-regression method, with a median cluster size of about 160 hectares.

The clusters estimated by the regression method are almost invariably within the elbow method clusters. In effect, the regression method identifies cluster cores, while the elbow method includes a larger periphery over which the agglomeration economies are not as strong.

The elbow method uses only locational information and confirms that startups are not randomly or uniformly distributed throughout a city. There are many reasons for such clustering in addition to Marshallian agglomeration economies, such as local topography and transportation networks. The elbow method cannot distinguish between clustering that arises due to agglomeration economies and clustering that arises for other reasons. The regression method incorporates economic information that identifies clusters based on the effects of proximity on venture capital investment.

We apply heat maps to cluster maps derived using both methods, showing how the estimated clusters evolve over time. The heat maps show that the clusters we estimate have relatively stable cores that persist for many years.

We suggest that these methods could contribute to decision-making over the investments that many cities and states continue to make in trying to develop and support economic clusters. Depending on policy objectives, the methods could be applied to a population other than venture-backed startups, possibly focusing on a particular industry or some activity, such as

innovation. Correspondingly, economic variables other than venture capital investment might be used to help identify clusters. For example, if the focus is on innovation, performance variables such as patent applications, patents awarded, R&D spending, or scientific and technical employment might be used.

These methods could also be applied much more broadly. The underlying population could be something like criminal activity or instances of a particular disease. The methods could also be applied to metrics other than geographic distance. For example, we could estimate clusters of product varieties using these methods.

In summary, the HCA approach we describe contributes to the longstanding research agenda seeking to identify and summarize the geographic distribution of economic activity. In addition, it provides a potentially powerful technique to identify clustering and its consequences in a wide range of other applications.

References

Acs, Z. J., Audretsch, D. B. (1990). *Innovation and Small Firms*. MIT Press.

Aljumily, R. (2015). Hierarchical and non-hierarchical linear and non-linear clustering methods to “Shakespeare authorship question”. *Social Sciences*, *4*(3), 758-799.

Andersson, M., Larsson, J. P. (2016). Local entrepreneurship clusters in cities. *Journal of Economic Geography*, *16*(1), 39-66.

Bholowalia, P., Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Buzard, K., Carlino, G. A., Hunt, R. M., Carr, J. K., Smith, T. E. (2017). The agglomeration of American R&D labs. *Journal of Urban Economics*, *101*, 14-26.

Carlino, G., Kerr, W. R. (2015). Agglomeration and innovation. In *Handbook of Regional and Urban Economics* (Vol. 5, pp. 349-404). Elsevier.

Cesario E., Vinci A., Zhu X. (2020) Hierarchical Clustering of Spatial Urban Data. In: Sergeyev Y., Kvasov D. (eds) *Numerical Computations: Theory and Algorithms*. NUMTA 2019. Lecture Notes in Computer Science, vol 11973. Springer, Cham.

Chatterji, A., Glaeser, E., Kerr, W. (2014). Clusters of entrepreneurship and innovation. *Innovation Policy and the Economy*, *14*(1), 129-166.

Duranton, G., Overman, H. G. (2005). Testing for localization using micro-geographic data. *Review of Economic Studies*, 72(4), 1077-1106.

Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl (2011) *Cluster Analysis* 5th ed., Wiley.

Contreras, P., Murtagh, F. (2015) "Hierarchical Clustering", Ch. 6 in Hennig, Christian, Marina Meila, Fionn Murtagh, and Roberto Rocci, eds.(2015) *Handbook of Cluster Analysis*. CRC Press.

Da Rin, M., Hellmann, T., Puri, M. (2013). A survey of venture capital research. In *Handbook of the Economics of Finance* (Vol. 2, pp. 573-648). Elsevier.

Ellison, G., Glaeser, E. L. (1997). Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105(5), 889-927.

Faggio, G., Silva, O., Strange, W. C. (2020). Tales of the city: what do agglomeration cases tell us about agglomeration in general? *Journal of Economic Geography*, 20(5), 1117-1143.

Guzman, J., Stern, S. (2015). Where is Silicon Valley? *Science*, 347(6222), 606-609.

Kaplan, S. N., Lerner, J. (2016). Venture capital data: Opportunities and challenges. In *Measuring Entrepreneurial Businesses: current knowledge and challenges* (pp. 413-431). University of Chicago Press.

Ketchen D.J., Shook C.L. (1996) The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* 17(6):441–458.

Krugman, P. R. (1991). *Geography and trade*. MIT press.

Marshall A (1920) *Principles of Economics*. London: MacMillan.

Peeters, J. P., Martinelli, J. A. (1989). Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theoretical and applied genetics*, *78*(1), 42-48.

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

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267-276.

von Thünen, J.H. (1826). Translated into English and reprinted as *Von Thünen's Isolated State*, Pergamon Press, 1966.

Verstraten, P., Verweij, G., Zwaneveld, P. J. (2019). Complexities in the spatial scope of agglomeration economies. *Journal of Regional Science*, *59*(1), 29-55.