Journal of Economic Geography (2022) **00**, 1–24 https://doi.org/10.1093/jeg/lbac024

A new method for identifying and delineating spatial agglomerations with application to venture-backed startups

Edward J. Egan and James A. Brander* **[†]

*Stowe, VT, USA
**Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, BC, Canada V6T 1Z2
[†]Correspondence to: *james.brander@sauder.ubc.ca*

Abstract

This article advances a new approach using hierarchical cluster analysis (HCA) for identifying and delineating spatial agglomerations and applies it to venture-backed startups. HCA identifies nested clusters at varying aggregation levels. We describe two methods for selecting a particular aggregation level and the associated agglomerations. The 'elbow method' relies entirely on geographic information. Our preferred method, the 'regression method', uses geographic information and venture capital investment data and identifies finer agglomerations, often the size of a small neighborhood. We use heat maps to illustrate how agglomerations evolve and we describe how our methods can help assess agglomeration support policies.

Keywords: Agglomeration, elbow method, hierarchical cluster analysis, venture capital
JEL classifications: R12, G24, L26
Date submitted: 15 December 2020 **Editorial decision:** 5 July 2022 **Date Accepted:** 29 July 2022

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 of business activity in emerging industries, as noted in Chatterji et al. (2014). In the USA, 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.

Such policies seek to develop a critical mass of interconnected entrepreneurs, early-stage investors and skilled workers in a specific location. Policy intervention is often justified using the agglomeration benefits identified in Marshall (1920), particularly labor market externalities, informal information exchanges and integrated networks of suppliers and customers.[1] 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 technology-based localized development on this basis.

1 Faggio et al. (2020) provide a recent study of the Marshallian micro-foundations of agglomeration.

One challenge with such policies is determining the appropriate geographic boundaries for a targeted location. Existing policies vary dramatically in their geographic scope. The Utah Advanced Materials and Manufacturing Initiative (UAMMI) is supported both by federal government programs and by the Utah state government and serves the entire state. Developments such as the Cortex Innovation Community in St Louis and the Oklahoma City Innovation District (OCID) are much smaller in geographic scope and are supported primarily by municipal 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 article is to advance a new approach for identifying economically meaningful startup agglomerations 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 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 localized agglomeration of economic activity. 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 stability and other aspects of agglomeration dynamics using 'heat maps'.

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

Our HCA-regression method allows for the identification of startup agglomerations at relatively small scales ('micro-agglomerations'), often about the size of a neighborhood, which is the relevant scale for many associated support policies. More specifically, using data from 1995 to 2020, our estimated median agglomeration is about 42 hectares, which corresponds to about 42 city blocks in the U.S. Midwest.[2] Agglomerations identified using the HCA-regression method are largely independent of the size of the starting area. The elbow method identifies larger areas, which are dependent on the starting area, that might be relevant for other policy purposes.

Our methods contribute to the basic task in regional science of assessing the agglomeration pattern exhibited by some activity, as well as identifying the location and boundaries of meaningful agglomerations. 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. We propose that our methods can be used as an input to public policy analysis and to decision-making by market participants such as entrepreneurs, venture capitalists and ecosystem support organizations.

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

2 Relative to the mid-west, city blocks are smaller in the east and larger in the south and the west. Using only data from the post-dot-com era leads to a smaller median estimate of around 25 hectares.

HCA-regression method respectively, along with illustrative examples. Section 7 summarizes the overall pattern of agglomeration of startup firms in the USA and Section 8 contains concluding remarks.

2. Literature review

The concentration of economic activity in agglomerations 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.

An important contribution is Krugman (1991) who uses a Gini coefficient approach to study agglomeration of manufacturing industries. 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 agglomerations include Rosenthal and Strange (2003) and Ellison et al., 2010.

The Gini coefficient method uses exogenous geographic units as locations, such as states or metropolitan statistical areas (MSAs). Duranton and Overman (2005) provide an alternative method of measuring agglomeration that 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 using or identifying boundaries of discrete agglomerations.

A related approach is provided by Buzard et al. (2017) who use 'point pattern' methods to study the agglomeration pattern of U.S. R&D laboratories. This method shows the pattern of agglomeration at various different scales (within 0.5 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 Verstraten et al. (2019). Buzard et al. (2017) also identify and delineate discrete agglomerations. In related work, Buzard et al. (2020) show that patent citation behavior among R&D laboratories is strongly clustered within small-scale localized agglomerations.

Our approach differs from both the localization approach and approaches based on exogenous geographic units. Unlike the localization approach, we seek to identify and delineate discrete agglomerations, estimating shapes and boundaries of the agglomerations. However, unlike the exogenous geographic unit approach, the agglomerations we identify are determined endogenously, although we do require our estimated agglomerations to be within some pre-specified starting area.

There are many reasons why businesses might cluster together in space, including physical geography such as natural ports and river confluences. In the case of recent startups, we might observe clustering because the available space is limited. Such factors no doubt play a significant role in the agglomeration patterns we observe. However, our analysis emphasizes the role of Marshallian agglomeration forces—labor market pooling, customer–supplier relationships and information exchanges (knowledge spillovers)—in generating startup agglomerations.

A recent literature has developed seeking to identify the relative importance of these different agglomeration forces for particular industries, especially high innovation industries

and seeking to understand the implications of the different Marshallian forces. One important insight from this literature is that agglomeration based on information exchange tends to produce smaller and denser agglomerations. For example, Arzaghi and Henderson (2008) provide an analysis of advertising agencies in Manhattan. They argue that information exchange is important in advertising and find evidence of large benefits from close interactions (meetings at which information exchanges occur) and of 'extremely rapid spatial decay' of these benefits.

Information exchange is particularly associated with innovation-intensive industries. Audretsch and Feldman (1996) provide results suggesting that, even after controlling for the degree of geographic concentration in production, innovative activity tends to cluster more in industries where knowledge spillovers (information exchanges) are relatively more important.

In addition, Roche (2020) analyzes the effect of neighborhood-level infrastructure on patenting performance. She suggests that more physically connected street networks increase the likelihood of 'serendipitous knowledge exchange' and also have an advantage in reducing the ratio of travel time to meeting time for various person-to-person interactions. Both of these effects give rises to better patenting performance.

Kerr and Kominers (2015) provide a theory of agglomeration in which bilateral agglomeration benefits in the form of information spillovers operate at very small scales, but overlapping bilateral relationships generate larger agglomerations. Thus, Firms A and C might be too far apart for meaningful interactions, but both might interact with Firm B located between them, so all three firms might be in the same agglomeration. That paper shows how different values for the 'radius' of interaction translate into different overall sizes, shapes and densities of agglomerations and applies the theory to patenting.

Davis and Dingel (2019) extend the analysis of agglomeration based on information exchange ('idea exchange' in their terminology) to develop a 'system of cities' model than can explain city–size-based wage differentials and other important comparative properties of cities.

Relative to this literature, our article has several areas of value-added. First, our approach is suitable for the policy objective of identifying areas over which agglomeration support policies, such as tax concessions, could reasonably be applied. Methods based on localization are not immediately suitable for this purpose (although they could likely be extended in some way for such uses). Methods based on exogenous jurisdictions are also not ideal for this purpose.

Assessments of agglomeration and associated support policies could be applied at the ZIP code level, city level, county level, MSA level, Census tract level, etc. While analyzing data organized at the level of each of these units is valuable for some purposes, none of these exogenously determined areas are likely to provide the most efficient areas over which to apply support policies.

Second, even apart from policy purposes, we suggest that it is worthwhile to develop a data-driven method for identifying agglomerations that is independent of pre-existing jurisdictional boundaries. It is also potentially of interest to see how estimated agglomerations correspond to existing geographic classification systems.

We focus on venture-backed startups as our illustrative example partly because of the literature showing the importance of high-growth high-technology entrepreneurship in localized economic development, 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 agglomerations 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 laboratories.

3. Data

Our data on venture-backed startups come from the Thomson Reuters VentureXpert database, which is described in Da Rin et al. (2013), Kaplan and Lerner (2016) and elsewhere. A potential alternative would be to use business registration data, as in Guzman and Stern (2015), which include both venture-backed and non-venture-backed startups. However, for our purposes, the Thomson Reuters data have the advantage that it include a performance measure in the form of venture capital funding, which is essential for our analysis. And we would expect non-venture-backed high-tech startups to be geographically correlated with venture-backed startup activity in any case.

We use U.S. startup data from 1995 to 2020, giving us 26 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 startup must have received a growth venture capital investment (as defined in Egan, 2021) between 1995 and 2020. 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 33,018, almost all of which are in the data for multiple years.

We wish to use HCA to subdivide these startups into clusters that reflect meaningful agglomerations. We could, in principle, start with entire USA as the largest possible cluster in the HCA process. In practice, this is not possible for computational reasons. Using standard methods, the calculations required for HCA increase exponentially in the maximum number of items in a cluster, so there is a very large computational saving if we limit the maximum possible size of clusters. We use MSAs as the largest geographic unit.

For the HCA-regression method, the estimated agglomerations are much smaller than the size of an MSA, as is consistent with Buzard et al. (2017) and Rosenthal and Strange (2003). Furthermore, we compared the results of starting with MSAs with the results starting with the much smaller 'Census designated places' (i.e., towns and cities) and found very similar results for the HCA-regression method. We therefore believe that there would be little gain in allowing for a larger initial size, such as a state.

For the HCA-elbow method, however, the results do depend on the size of the initial unit. In essence, the elbow method, as we implement it, identifies the best way of subdividing any starting unit into just a few smaller areas. The primary value of the elbow method is in identifying the location and shape of clusters at a given level of aggregation.

To use MSAs as the underlying geographical unit, we match startups with MSAs by geocoding startup addresses as longitudes and latitudes, and then determining whether each location falls within an MSA boundary. We use the 2020 U.S. Census TIGER/Line Shapefiles to define the boundaries of MSAs.[3]

3 MSA names and boundaries sometimes change over time and MSAs may enter or drop out of the data as their populations change. However, the boundaries of the larger MSAs, which contain the overwhelming majority of startups, are very stable over time. Using the MSA definitions and boundaries from earlier years, such as 2010, has virtually no effect on the analysis.

Of the 33,018 startups in the VentureXpert database over the sample period, a small fraction is not in MSAs and some do not have usable addresses. After dropping those two groups of startups (about 6% of the original population) we are left with 30,980 startups in the 392 MSAs in the 2020 Census.

However, many MSAs have relatively few startups. We drop such MSAs from the analysis as they do not contain meaningful startup agglomerations. Specifically, for inclusion, we require an MSA to have at least 1 year with 10 or more active startups. Only 88 out of the 392 MSAs satisfy this requirement and these 88 MSAs account for 29,741 startups (about 96% of the startups in MSAs with usable addresses). Thus, a modest share of the MSAs account for the overwhelming majority of venture capital investments.

Even within this group of 88 MSAs, the distribution of startups is highly concentrated and strongly skewed. Online Appendix Table A1 shows the top 20 MSAs with their total number of startups from 1995 to 2020.

4. Hierarchical cluster analysis

4.1. Methodological overview

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 Chap. 4 in Everitt et al. (2011) and Contreras and Murtagh (2015).

HCA identifies nested categories or clusters 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 categories or clusters such as 'fruit' and 'grains'. At a lower level of aggregation, within the fruit category, we might have 'apples' and 'oranges' and within the apple category at a still lower level of aggregation we might have Fuji apples, Granny Smith apples, etc. An example of current relevance is the use of HCA to categorize respiratory virus variants (as in Wentzensen et al., 2010). There are alternative approaches to clustering, such as the '*k*-means' approach which, for any prespecified number of clusters, *k*, minimizes within-cluster variance. Al Kindhi et al. (2019) describe these methods, suggest a hybrid method and apply it to clustering viruses.

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 geographic distance as in Cesario et al. (2020).

We calculate distances between startups using their longitudes and latitudes. This task is complicated by the fact that the earth's surface is curved. The true distance between two points is the distance along the surface of the three-dimensional globe. Treating the two points as if they were on a flat two-dimensional plane, as with standard two-dimensional maps, is inaccurate. While the degree of inaccuracy is small if the startups are close together, it is large enough to affect some of our calculations. We therefore use (open source) PostGIS geographic database software combined with the World Geodetic System 1984 three-dimensional reference ellipsoid model to calculate distances between start-ups. This approach yields very precise and accurate distances.

It is important to be precise about terminology. In HCA, the term 'cluster' has a specific technical meaning that does not always coincide with the how the term is used in

economics. At the highest level of aggregation in the HCA process, a 'cluster' consists of all elements in the population under study—all startup locations in an MSA in our case. At the lowest level of aggregation, clusters consist of individual members of the population—individual startup locations in our case. Neither the 'cluster of the whole' nor individual startup locations would be called 'clusters' in normal economic usage.

From now on we use the term 'cluster' only in its technical HCA sense. Our estimated agglomerations are a particular set of clusters identified in the HCA process. Other clusters are not taken to be agglomerations. For example, clusters that consist of individual startups or pairs of startups are, for obvious reasons, not taken to be agglomerations.

4.2. Clustering algorithms

There are several widely used hierarchical 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 results, so it is important to use an algorithm that is suitable for the task at hand.

Investigations of different algorithms as reviewed in Everitt (2011, Chap. 4) indicate that no method is universally preferred. We use Ward's algorithm, which is based on squared distance. It is widely used and generally performs well both computationally and in terms of generating an accurate and meaningful nested structure of clusters. Using squared distance is appealing for geographic contexts given work such as Arzaghi and Henderson (2008), who find a large nonlinear spatial decay in the benefits of proximity that can be approximated by a quadratic formulation and Rosenthal and Strange (2003), who find similarly rapid (non-linear) spatial decay of proximity effects.

4.3. Applying Ward's algorithm to startup locations

We use Ward's algorithm to identify the hierarchical structure of clustering for each MSA-year in the data. To identify specific cluster boundaries, it is necessary to specify the acceptable cluster shapes. In our main analysis, we assume two standard conditions for the shape. A cluster 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 that 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. We also briefly consider relaxing the shape assumptions by 'buffering' the proposed agglomerations.

A convex hull might be a triangle, a diamond shape or a more complex polygon. It may also be degenerate in that it may be a point, as when there is just a single isolated startup location or it may be a line segment, as when there is an isolated pair of neighboring startup locations. All convex hulls are 'clusters' in the technical sense used in HCA even if they are degenerate and therefore have no area. We use the term 'areas' to refer to non-degenerate convex hulls—those with positive area. This terminology—areas, line segments, points, degenerate shapes and convex hulls—is standard in geometry. We require our estimated agglomerations to be areas.

Ward's algorithm can be implemented in either a top-down ('divisive') sequence of steps or a bottom-up ('agglomerative') sequence. Divisive clustering starts with the cluster of the whole then generates finer and finer partitions of the data in successive steps until

each location is a separate cluster. Agglomerative clustering does the reverse, starting with each individual location as a cluster and gradually building toward the cluster of the whole. Both approaches yield exactly the same hierarchical clustering structure. We use a version of Ward's algorithm for agglomerative HCA written in Python and available from scikit-learn. We modified the algorithm slightly to input pre-computed distances calculated using PostGIS rather than allowing the algorithm to compute distances internally, and recoded the clustering structure so that the results were divisive.[4]

For any given MSA-year, each startup location, y_i, is a point on a surface. Denoting the mean location of all startups in an MSA as y^{**}, we can calculate SS_{tot}, the total sum of squared distances of startups in the MSA from y^{**}.

$$SS_{tot} = \sum_i (y_i - y^{**})^2. \tag{1}$$

Ward's algorithm proceeds by choosing, for each step and the associated number of clusters, the particular clustering pattern that 'explains' as much of the total sum of squares as possible.

We refer to each step in Ward's algorithm as a 'layer'. Layer 1 is the layer in which there is only one cluster and all startup locations are in it (i.e., Layer 1 contains the cluster of the whole). In Layer 2, the startup locations are divided into two clusters. In Layer 3, one of the clusters from Layer 2 is divided into two, yielding three clusters in total and so on. The total number of layers equals the number of distinct startup locations.

For any given layer, if statistical cluster j contains n_j startups and has mean location y_j^*, the explained sum of squares, SS_{exp}, is

$$SS_{exp} = \sum_j n_j (y_j^* - y^{**})^2. \tag{2}$$

This explained sum of squares is sometimes called the 'between' sum of squares as it is based on the distances between clusters. The unexplained sum of squares SS_{unexp}, sometimes called the 'within' sum of squares, is the sum of the within-group sums of squares. It is the variation in location not explained by division into clusters. The explained and unexplained sums of squares must sum to the total sum of squares.

$$SS_{tot} = SS_{exp} + SS_{unexp}. \tag{3}$$

In Layer 1, the cluster structure does not 'explain' any of the locational differences between startups. The total sum of squares is 'within-group' and is therefore unexplained. For Layer 2, the algorithm divides the cluster of the whole into two sub-clusters so as to maximize the explained sum of squares. To take an extreme example, if half the startups in an MSA-year were in a building (at one street address) in one part of the MSA and the other half were in a building in another part of the MSA, the division of the cluster of the whole into two sub-clusters (one for each building) would explain all the locational variation in the data and the explained sum of squares would equal the total sum of squares.

4 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. Our modified version is available in the online supplementary material.

More generally, for each layer, as Ward's algorithm chooses clusters so as to maximize the explained sum of squares, it also maximizes the standard R^2-statistic—the ratio of the explained sum of squares to the total sum of squares.

$$R^2 = \frac{SS_{exp}}{SS_{tot}}. \tag{4}$$

Online Appendix Section A2 applies this method to the Allentown–Bethlehem–Easton MSA in Pennsylvania and New Jersey in 1999, showing the full sequence of steps in the HCA process. Each step identifies a layer in the clustering process.

5. Choosing a clustering layer using the elbow method

The next task is to identify the specific layer that corresponds to our estimated agglomerations. With HCA, it is possible to apply heuristic methods based on distance to identify such a layer and therefore delineate the estimated agglomerations. 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). Our implementation of 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 the explained variation in location as measured by the R^2 statistic.

A hypothetical example illustrating the elbow method is provided in Online Appendix Section A3. The 'elbow' is the point in the graph of R^2 versus the layer number where the slope of graph falls sharply, resembling a bent elbow. This cannot occur at Layer 1 and we also rule out Layer 2.[5] The rule we use is to choose the subsequent layer after which the marginal increase (first difference) in R^2 drops most sharply. This is where the second difference in R^2 is minimized.

One consequence of using this R^2 rule is that the chosen layer always occurs relatively early, selecting only a few (relatively large) areas. The reason is that the large changes in R^2 always occur early and, correspondingly, the large changes in the first difference also occur early. Layer 3 is commonly chosen but Layers 4–6 emerge with reasonable frequency.

The areas in the elbow layer are our estimated agglomerations. In some cases, we estimate only one agglomeration in an MSA. This occurs if one or more startups are sufficiently far from the main group that the initial subdivisions of the cluster of the whole simply break off the isolated startups. The elbow layer will then have just one area (i.e. one agglomeration) along with one or more isolated points or line segments. It is also possible to have two estimated agglomerations. This occurs when we get two areas and one or more lines or points in the elbow layer.

The elbow method as we implement it does a good job of estimating a small number of fairly large agglomerations in each MSA. One contribution of the method is in estimating a specific size, shape and location for each agglomeration, which has value-added over and above estimating the number of agglomerations. We see this method as potentially

5 Because layer numbers start at Layer 1, the initial first difference is the entire R^2 and the initial second difference (Layer 2) is also very large. This 'end point effect' biases the elbow selection toward Layer 2, so we drop Layer 2 as a possibility.

useful for dividing an MSA into different administrative districts, possibly for administration of start-up support policies or for other underlying populations such as hospitals or police stations. In addition, it yields a sharp comparison with the HCA-regression method discussed in Section 6.

Figures 1 and 2 illustrate the selection of the elbow and estimated agglomerations for two MSAs. One is the San Francisco–Oakland–Berkeley ('San Francisco') MSA, which has the largest number of startups in our data (5725 over the 1995–2020 period) and the other is for Madison, WI, which is approximately at the median among the 88 MSAs used in our analysis (95 startups over the 1995–2020 period). Figure 1 shows, for the year 2006, how the R^2 changes with the layer number and where the elbow appears for each MSA. Figure 2 shows the associated estimated agglomerations.[6]

For San Francisco, the elbow occurs at Layer 3, which identifies three clusters. Therefore, these three clusters are our estimated agglomerations using the elbow method. Each of these clusters has positive area. As noted by a referee, the (northern) cluster containing parts of San Francisco, Berkeley and Oakland might not seem like an obvious cluster but there is a lot of connectivity, including convenient bridges, between its areas. For Madison, the elbow occurs at Layer 4, which identifies four clusters. However, only three of these clusters have positive area. The fourth cluster is a point (an isolated startup) at the very top of the map close to the eastern edge and marked with a plus (+) sign. Two of the Madison areas are fairly large and one, in the south-central area, is quite small. These three areas are the agglomerations estimated by the elbow method for Madison for 2006.

If the agglomerations are meaningful, we would expect them to persist over time. Figure 3 uses heat maps to illustrate the stability of agglomerations over time. 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 MSA, we take partially transparent cluster maps for each year and lay them on top of each other. For each year, the areas are lightly shaded. As an area or part of an area is identified in more and more years, it takes on an increasingly dark hue, as illustrated in Figure 3 for the San Francisco and Madison MSAs.

The dark core areas of the heat maps coincide closely with the agglomerations identified in the single year of 2006. More broadly, the heat maps indicate that it is possible to identify relatively stable agglomeration cores. For San Francisco, the three estimated agglomerations are relatively stable. One agglomeration corresponds to the cities of San Francisco, Berkeley and Oakland, along with Marin County. The other two agglomerations are northern Silicon Valley and the East Bay south of Oakland. For Madison, the heat map arguably suggests only two agglomeration cores.

6. Identifying agglomerations using the HCA-regression method

We view an economic agglomeration of startups as a collection of startups in geographic proximity with some positive economic interaction effect. While agglomerations estimated using the elbow method likely exist at least partly because of Marshallian agglomeration economies, the method does not explicitly use agglomeration-related economic information.

6 Our agglomeration maps use the EPSG:3857 Pseudo-Mercator projection with a Google Maps base layer.

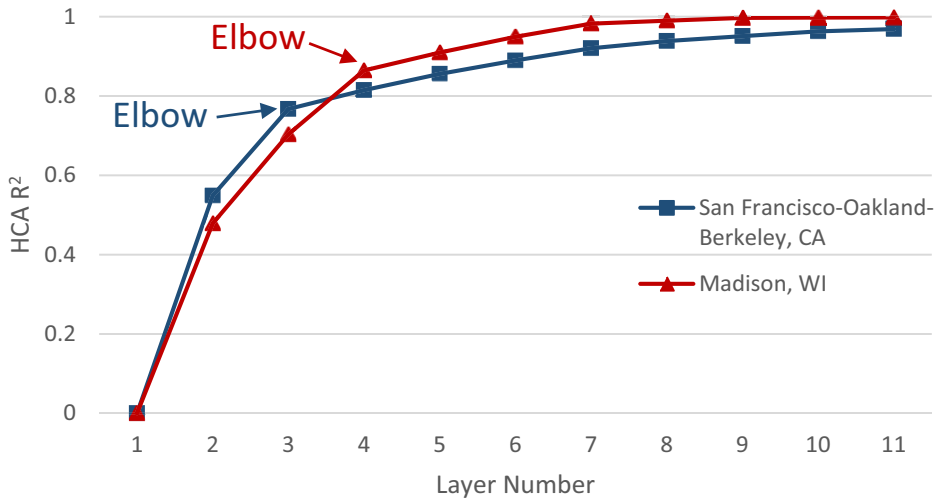

Figure 1. Identifying the elbow for San Francisco and Madison, 2006.

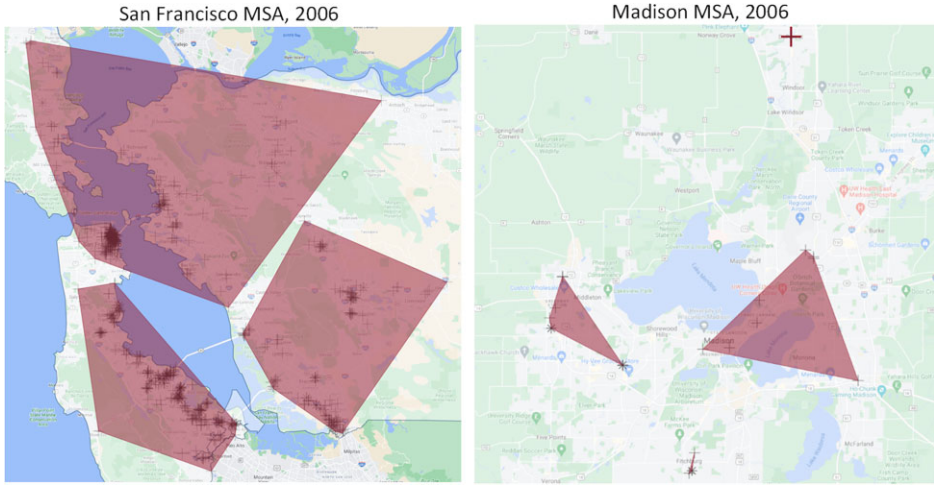

Figure 2. Elbow-based agglomerations.

In this section, we propose and illustrate the HCA-regression method for identifying agglomerations. This method does explicitly use economic information in the form of year-by-year amounts of venture capital investment received by each MSA.

If agglomeration economies are important, the agglomeration structure in an MSA would affect the total amount of venture capital investment received. For example, if two ventures are close together, a venture capitalist might be able to visit both in the same trip, possibly making both more attractive investment targets than if they are far apart. Or possibly exchanges of information between the two ventures in close proximity are helpful to both and make them both more attractive investment targets.

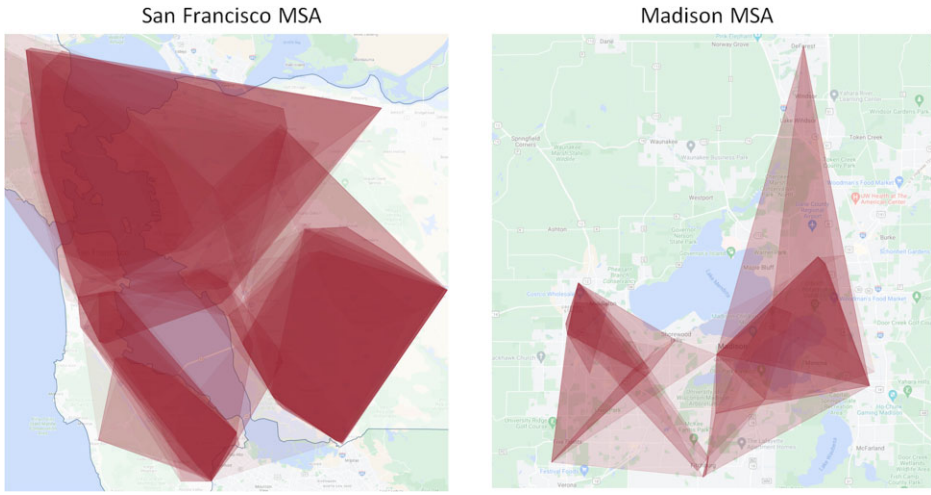

Figure 3. Elbow-based heat maps 1995–2020.

We can specify the dependence of venture capital investment on the agglomeration structure as

$$Y_{it} = f(n_i; X_{it}) + e_{it}, \tag{4}$$

where Y_{it} is the venture capital investment received by MSA i at time t, n_i is the number of agglomerations in MSA i, X_{it} is a vector of variables characterizing the agglomeration structure and e_{it} is an error term that includes other influences on investment normalized to have expected value zero.

We do not know in advance the number of agglomerations in a given MSA and must estimate it. This estimation can be done using Equation (4). For MSA i at time t, we observe Y_{it} and we can observe the X_{it} variables associated with any possible value of n_i. We determine our estimated value of n_i according to an optimization criterion based on R^2 statistics. Specifically, we choose the value of n_i so that the R^2 value for the regression implied by Equation (4) is high. Using the R^2 statistic in this way is a natural extension of Ward's algorithm, which is also based on R^2 statistics. Using a related goodness-of-fit statistic such as an F-test or an adjusted R^2 or using a maximum-likelihood approach would yield very similar results.

We assess Equation (4) for all admissible possible values of n_i. Each possible value is associated with a clustering structure (using the term clustering in the technical sense) and has a specific number of areas. Only one of those clustering structures is the estimated agglomeration structure and the areas in that structure are the estimated agglomerations.

We include five variables in X_{it} to characterize a clustering structure: the number of areas, the average size of the areas, the average startup density in the areas, the average separation of area midpoints and the share of the MSA's startups in areas (rather than as isolated points or in line segments). These five variables have substantial explanatory power and adding additional clustering structure variables adds very little additional explanatory power.

As can be seen in Online Appendix Figure A1 for Allentown, a given number of areas is not uniquely associated with a particular structure. In Allentown, in 1999, there were

three different possible clustering structures (Layers 1, 3 and 4) with one area (i.e. with one proposed agglomeration), each of which has a different cluster structure and therefore a different set of values for X_{it}. For each candidate value of n_i, we must identify one particular layer with n_i areas as the 'representative layer' of the clustering hierarchy. We then use the values of X_{it} obtained from the associated layer of the clustering process.

In the Allentown 1999 example, Layer 2 is the only layer with exactly two areas and therefore two potential agglomerations. It follows that Layer 2 is the representative layer for $n = 2$ (the two-area layer). However, it is not immediately obvious which layer provides the best one-area representation as Layers 1, 3 and 4 all have one area.

We suggest that Layer 3 is the most representative one-area layer. Layer 1 is not ideal because the area subdivides naturally into two areas at the following step. This area is 'unstable' in that it does not maintain its approximate size and shape beyond one layer. Layer 3 is suitable because, beyond Layer 3, all we are doing is forcing the HCA process to drop individual startups or pairs of startups from clusters. In general, even for large MSAs, the cores of identified areas do not change much as we proceed beyond the representative layer and the general location of the areas remains stable. We therefore provide the following definition.

Definition: The *representative n-area layer* is the first layer with n startup areas that has the property that no later (higher numbered) layer has more startup areas.

If only one layer has n areas, that layer must be the representative n-area layer. If multiple layers have n areas, this definition chooses the most representative of those layers. Therefore, for each MSA-year, we have a representative n-area layer for each value of n from 1 up to the maximum number of startup areas for that MSA-year.

The next step is to run a series of regressions motivated by Equation (4) for each MSA using annual observations for that MSA. One regression is based on the representative one-area layer, one is based on the representative two-area layers and so on. Each of these regressions is conditional on a particular candidate value of n (i.e., on a particular number of areas).

For relatively large values of n, there may not be an observation for each year. We impose a requirement of at least 15 annual n-area observations for area number n to be considered. In other words, we require at least 15 years with n or more areas.

The regression specification for each number n for each MSA is as follows:

$$y_t = a_n + b_n X_{nt} + e_{nt}, \tag{5}$$

where y_t is the amount of venture capital invested in the MSA in year t, a_n is a constant term for the n-area representative layer, b_n is a vector of slope parameters, X_{nt} is a vector of explanatory variables for the n-area representative layer at time t and e_{nt} is a random error for the n-area representative layer at time t.

Each regression and therefore each value of n generates an R^2 statistic, giving us a set of combinations of n and R^2. We fit a cubic function to those points and find the integer value of n that maximizes that function. We use a cubic function instead of a quadratic because most of the figures are skewed to the right.

Taking the value of n that maximizes the estimated cubic function instead of simply taking the value of n with the highest R^2 makes little difference for most MSAs, especially the smaller MSAs, but it has the advantage of smoothing the effect of random errors due to unobserved factors and provides more reliable estimates. Figure 4 shows the estimated

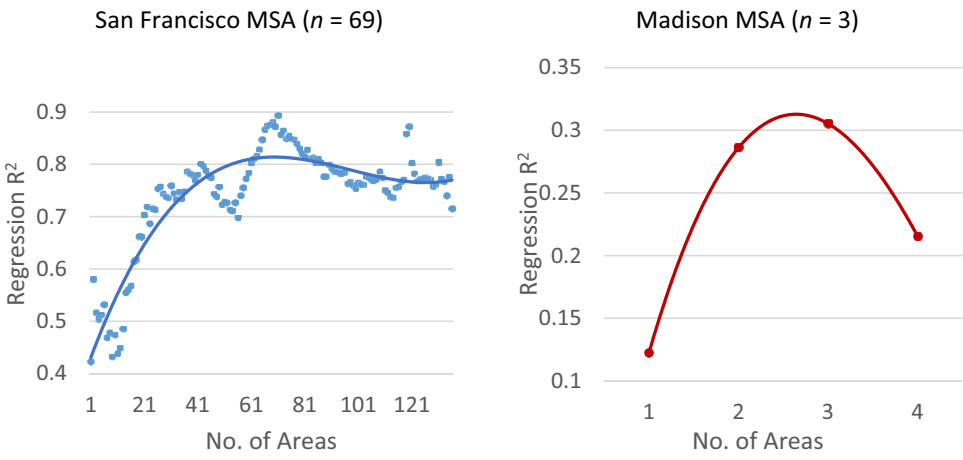

Figure 4. Using regression R^2 statistics to select the number of clusters.

cubic functions and underlying combinations of R^2 and n for the San Francisco and Madison MSAs.

The regression method selects just one agglomeration count for each MSA. That count applies to every year for which the requisite number of areas exists. It is possible for an MSA to have less than this count of areas for some years, although that happens only rarely and, if it does, typically arises early in the period before significant later entry. In such cases, we assume that the number of agglomerations equals the number of areas in those years.

Figure 5 shows the HCA-regression-based cluster map for the San Francisco MSA in 2006 and the multi-year heat map. We use a larger scale than in Figures 2 and 3 because the estimated agglomerations for the San Francisco MSA are much smaller and hard to see if we include the entire MSA. The maps show the part of the MSA with the densest collection of startups, which is the northeast part of the city of San Francisco.

The diagrams make two main points. First, the estimated agglomerations are small. We estimate 69 agglomerations in the San Francisco MSA and some are only a few blocks in size. If we took the number of areas associated with the highest actual R^2 value instead of the maximum of the fitted cubic function in Figure 4, the estimated number of agglomerations would be 71 and the resulting maps would be virtually indistinguishable at the scale shown. As the maps shown cover only a small part of the area of the MSA, they do not contain all 69 agglomerations. However, they do contain close to half the estimated agglomerations in the MSA, indicating a high level of concentration of agglomerations within the MSA.

The heat map has more shaded areas and somewhat larger shaded areas than the 2006 map as it includes estimated agglomerations for all years, not for just 2006. The core areas in the heat map correspond closely to the agglomerations in 2006, indicating that the agglomerations are relatively stable over time.

Figure 6 shows the same diagrams for the Madison MSA, which is a much smaller MSA. In this case, we show three maps, one for 2006, a heat map for the full period and a 'buffered' heat map. The buffered map relaxes our shape requirements for the agglomerations. Our base approach requires agglomerations to be areas shaped as polygons (i.e.,

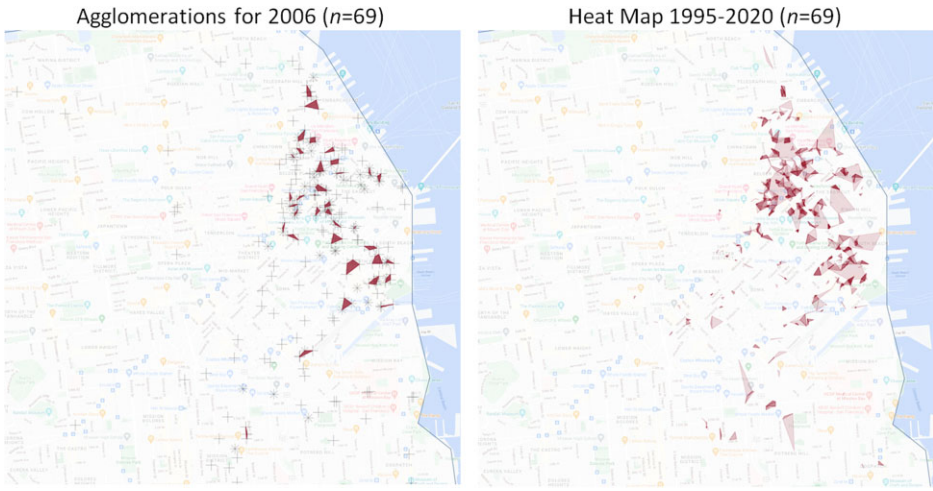

Figure 5. HCA-regression agglomerations for San Fransico MSA.

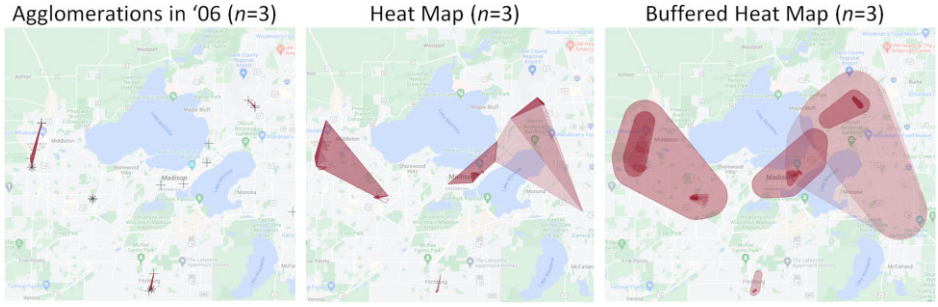

Figure 6. HCA-regression agglomerations for Madison MSA.

convex hulls). However, realistically, the economic reach of an agglomeration would extend outside the boundary of the polygon. Many of the startups in a given area lie on the boundary of the area and might be expected to have an economic sphere of influence that extends some distance in all directions, including outside the area. As noted by a reviewer, we can think of the identified agglomerations as the kernels of VC-backed 'hotspots' that identify general areas of policy interest as the geographic range of agglomeration externalities would extend beyond the area identified by the HCA-regression process.

The buffered heat map extends the boundaries of the areas identified in the HCA-regression process by half the average distance between each corner and the centroid in all directions. Thus, the cross-sectional dimension of the area increases by about 50%. This produces larger areas and makes the boundaries smooth (i.e., no corners).[7]

7 The actual algorithm is as follows: (i) Identify all corners of the initial area (polygon). (ii) Determine the centroid of the area. (iii) Move each boundary point directly away from the centroid by a distance equal to half the average distance between the corners and the centroid.

The use of buffering is potentially important in actual policy applications. Kerr and Kominers (2015) find overlapping regions of interaction for patent citations. In that spirit, it is possible that knowledge flow spillovers generate overlapping startup agglomerations that can be captured by buffering.

The comparison between the elbow method and the regression method for identifying agglomerations is somewhat different for large MSAs such as San Francisco than for small MSAs such as Madison. For San Francisco, the elbow method identifies just three high-level agglomerations in the MSA for 2006 and, correspondingly, three core agglomerations in the heat map for the full 1995–2020 period. This compares with 69 estimated agglomerations in 2006 using the regression method and a similar pattern of core areas in the heat map.

For San Francisco and other large MSAs, the elbow method and the regression method are doing very different things. The elbow method identifies a small number of areas within an MSA. The HCA-regression method identifies much smaller agglomerations based on evidence of agglomeration economies provided by venture capital investment patterns. For Madison, both the elbow method and the regression method identify just three agglomerations and the buffered head map for the regression method has a clear similarity to the elbow method heat map.

For MSAs overall, the smaller estimated agglomerations estimated by the regression method are generally in the interior of elbow method agglomerations. For smaller MSAs that have only a few estimated agglomerations, the number is generally similar to the number of agglomerations obtained using the elbow method.

7. Aggregate agglomeration properties

We now examine the agglomeration patterns of all U.S. venture-backed 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 MSAs and over years. As different MSAs 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 time.

If a given MSA-year has n startup locations, it also has n layers in the HCA process. The layer 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 layer index to go from 0 to 1. We normally use percentages, so the index goes from 0% to 100%. Using the layer index allows us to see how far along the HCA process has progressed when we reach the layer that defines the agglomerations for a given year and MSA. For the HCA regression methods, this typically occurs shortly before we are halfway through the process. For the elbow method, it generally occurs much earlier, as shown in Table 1.

Table 1. MSA-year summary statistics, all MSAs 1995–2020

	HCA-regression ($N = 1761$)			Elbow ($N = 2123$)		
	Median	Mean	SD	Median	Mean	SD
Layer index	45.2%	42.2%	26.1%	10.5%	16.1%	16.5%
Number of agglomerations	2	7.8	16.4	3	2.5	1.0
Frac. in agglomerations	55.8%	58.3%	31.7%	96.4%	85.5%	23.9%
Startups in agglomerations	11	39.0	82.5	23	97.0	225.5
Agglomeration area (ha)	42.5	6323.9	46,236.9	5402.7	21,369.5	43,129.9
Agglomeration density (startups/ha)	0.1	8.4	48.5	0.0	0.6	12.9
Agglomeration separation (km)	4.8	6.7	8.1	11.0	13.1	11.3
Total number of startups in MSA	24	99.1	235.1	24	98.4	225.0
VC invested (2020 $m) in MSA	40.1	373.2	1503.6	43.6	373.6	1422.6

In Table 1, each MSA-year is a distinct observation. There are 2123 MSA-years in our data set with the necessary information to apply the elbow method. However, only 1761 MSA-years are included in the summary statistics for the regression method because of the data requirements for that method.[8]

The minimum number of agglomerations an MSA-year can have is 1. One way this occurs is if all startups in a given MSA-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-MSA clustering. More commonly, a low number of areas occurs later in the HCA process after outlier startups are pruned from the main areas.

The median number of estimated agglomerations is only two for the regression method. (Most MSAs are more like Madison than like San Francisco.) However, for the regression method, the estimated agglomeration structure occurs 50% of the way through the HCA process. For many MSA-years, earlier layers have more than three areas but the HCA process pares the structure down to identify a small number of 'core' areas.

The geographic size of areas is shown in hectares. A hectare has the same area as a square that is 100 m on each side. A typical city block in the U.S. Midwest has an area of about 1 hectare. So, for the regression method, the median estimated agglomeration covers about 42 hectares or 42 city blocks, about the size of small neighborhood.[9] The median agglomeration size for the elbow method is much larger, over 5000 hectares (about 20 square miles).

Using the elbow method, the majority of startups in any MSA are allocated to an agglomeration rather than being pruned out as outliers. Specifically, at the median, an MSA-year has 24 startups and 23 of those are in agglomerations (96%). Using the regression method, just under half the startups (11 out of 24) are in agglomerations at the median.

8 The missing MSA-years occur when a given MSA does not have the requisite 15 years with a given area count that we require in order to use the regression method.

9 Agglomerations identified with the regression method tend to get smaller over time if the number of startups increases over time and startup density therefore increases. We believe that this compression of agglomerations as density rises captures a meaningful effect.

7.2. Agglomeration density

One important question relates to the optimal density of agglomerations and whether normal market forces would lead to insufficient (or possible excessive) agglomeration. The possibility of inefficient agglomeration 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 agglomeration densities in the absence of government policy will be suboptimal. Congestion externalities, on the other hand, could generate excessive density.

The optimal density will not be the same for every type of startup and every location. Some locations and some industries would benefit more from agglomeration than others. Even if the benefits of close proximity were fully internalized by firms, we would observe different densities in different agglomerations. Estimating the optimal density is beyond the scope of this article. However, our analysis provides some suggestive information, as shown in Figure 7, which shows venture capital investment plotted against agglomeration density.

Each point in the plot is an MSA. The *x*-coordinate for an MSA shows the average (log) agglomeration density in that MSA, based on the regression method, for the years that MSA is in the data. The *y*-coordinate shows the average (log) venture capital investment received in that MSA. Not all MSA-names are in the figure, due to crowding, but all the available data points are shown. The San Diego, San Francisco, New York and Boston MSAs have the highest startup density levels in agglomerations and the San Francisco, San Jose, Boston and New York MSAs have the highest level of venture capital investment.

The relationship illustrated in Figure 7 does not arise by construction. The total amount of VC investment in an MSA would not necessarily be affected by the density of startups. This point is illustrated in Table 2, which reports regression results associated with Figure 7.

Specification 1 is the regression in Figure 7, showing the positive relationship between agglomeration density and total venture capital in an MSA. Specification 2 includes density and the other four variables used as explanatory variables for the regression given by Equation (5) that we use to identify the number of agglomerations in an MSA.

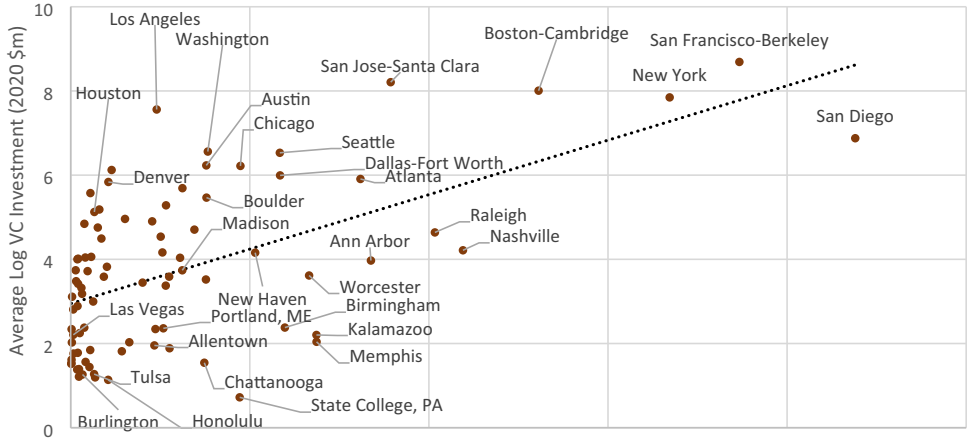

Figure 7. The relationship between VC investment and agglomeration density for agglomerations selected by the HCA-regression method.

Table 2. Regressions of VC investment on cluster density

	(1)	(2)	(3)
Log agglomeration density (startups/ha)	0.369***	1.063***	0.186***
	(0.039)	(0.247)	(0.0435)
Fraction of startups in agglomerations		−1.887***	−1.338***
		(0.329)	(0.338)
Number of agglomerations		0.009***	0.009***
		(0.001)	(0.001)
Log agglomeration area (ha)		0.885***	
		(0.230)	
Agglomeration separation (km)		0.009***	0.009***
		(0.001)	(0.001)
Constant	4.554***	2.878***	3.957***
	(0.207)	(0.319)	(0.187)
Observations	1761	1761	1761
R-squared	0.343	0.637	0.624

Notes: The dependent variable is the log of venture capital investment (2020 $m) in an MSA-year. Robust standard errors, clustered at the MSA level, are in parentheses.
***$p \leq 0.01$.

Specification 3 is the same as Specification 2 except that we drop agglomeration area as a regressor to address any possible concern arising from area being the denominator of the density variable. Even after adjusting for these controls, agglomeration density still has a statistically significant effect.

Table 2 sheds some light on the traditional question as to whether agglomeration economies are more important at the MSA level or the neighborhood level. The MSA level would be more important if such economies come from being able to hire the same workers, whereas neighborhood effects would be more important if the major source of agglomeration economies is in-person information flows. The significance of agglomeration density suggests that neighborhood effects are important as the agglomerations underlying Table 2 are about the size of neighborhoods.

We acknowledge that the relationships shown in Table 2 are just correlations and that we have not demonstrated any causality. However, the results at least suggest that MSAs in the lower left part of Figure 7 have inefficiently low density and could benefit from policies that would encourage denser agglomerations.

More density is not always good, however. While we have emphasized agglomeration economies, particularly of the information exchange type, there also are counter-balancing costs, such as congestion costs. After all, it would not be efficient to have all startups in a single location. An efficient structure would balance the costs and benefits of agglomeration. Nevertheless, Figure 7 and Table 2 provide suggestive evidence that positive agglomeration externalities have not been fully internalized by current agglomeration structures.

7.3. Evolution of MSAs and agglomerations

As noted by a reviewer, our methods can be used to identify agglomerations and MSAs of particular interest due to unusual or exceptional growth patterns at the MSA level or at the

Figure 8. Evolution of downtown Santa Monica's 2000 agglomeration by 2020.

agglomeration level. Online Appendix Figure A3 shows the number of active venture-backed startups for each year for the five MSAs with the most startups. In addition, we use the HCA-regression method to identify persistent and high-growth agglomerations.

Figure 8 shows the evolution of the agglomeration with the most startup growth between 2000 and 2020: downtown Santa Monica in the Los Angeles MSA. The large red polygon shows the estimated 2000 agglomeration, which contained five startups, indicated with red crosses. As of 2020, five distinct agglomerations overlapped the original 2000 area and there were 39 startups, indicated with green crosses, in these five agglomerations. A sixth agglomeration with five additional startups also emerged just to the East.

7.4. Industry effects

Another interesting application is to consider the role of industry. MSAs differ dramatically in the industrial composition of the startup population. Dividing the population into information and communications technology (ICT), life sciences including biotechnology and 'other', we observe that in both the San Francisco and San Jose MSAs about 83% of the ventures are in ICT with only 10% and 13%, respectively, in life sciences. Large MSAs with a relatively large number of life science startups include Durham-Chapel Hill (53%), New Haven (51%) and San Diego (45%). The only fairly large MSA with >40% 'other' is Detroit.

Using the HCA-regression method, it is possible to investigate the extent to which ventures in the same industry cluster together. Figure 9 shows 2020 agglomerations for Kendall Square in Cambridge, downtown Boston and Boston's North End, identifying startups by industry.

A visual inspection of clusters maps suggests that ventures in the same industry tend to cluster together and many agglomerations can be identified as ICT agglomerations (as in Boston's North End) or life sciences agglomerations (as in Kendall Sq.). Downtown areas appear to support the most mixed-industry agglomerations.

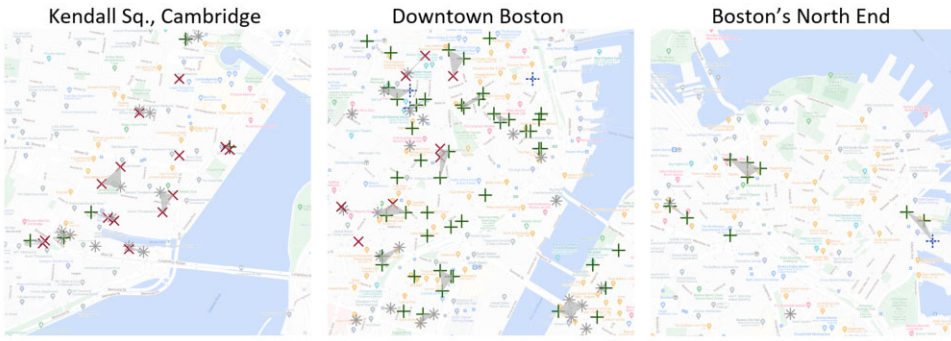

Figure 9. Industry effects for Kendall Square in Cambridge, downtown Boston, and Boston's North End, 2020.
Notes: solid plus (+) signs (green) = ICT, crosses (X) (red) = life sciences, dashed plus signs (blue) = other. Agglomerations are shaded (gray). Locations with multiple startups are shown as gray asterisks, without industry information.

8. Concluding remarks and further applications

The main objective of this article is to introduce a new approach, based on HCA, for identifying and delineating agglomerations 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 agglomerations of venture-backed startups at potentially policy-relevant scales. The estimated areas are relatively small, with a median size on the order of a small neighborhood, although with substantial variance and skewness. Jurisdictions sometimes apply support policies to areas of comparable size although typically target areas which are larger.

Our other variant is the elbow method, which answers a different question. Specifically, it asks how best to identify and delineate a small number of areas within some larger exogenously given area (such as an MSA) based on locational information. It estimates areas that typically cover a significant fraction of the starting geography.

The elbow method confirms that startups are not randomly or uniformly distributed throughout a geographic region such as an MSA. There are many reasons for such non-uniformity of startup locations (or other types of economic activity) in addition to Marshallian agglomeration economies, such as local topography (river confluences, etc.) and transportation networks. The elbow method cannot distinguish between these reasons and identifies groupings of startups based on purely locational information. Starting with MSAs, the agglomerations estimated by the regression method are almost always inside elbow method areas and often coincide with high density cores of the elbow areas.

We create heat maps using both methods, showing how the estimated agglomerations evolve over time. Observing agglomerations over time allows us (and policy makers) to identify stable agglomeration cores and also allows identification of important trends in startup activity that would not be apparent from a static 'snapshot'.

In the introduction, we mention areas that have received agglomeration support. Based on our analysis, supporting broad areas, as with the UAMMI, is not likely to capture agglomeration economies efficiently. Even smaller, neighborhood-level areas such as the CIC in St Louis and the (recently inaugurated) OCID may be too large. In our regression

method analysis, roughly half of the CIC corresponded to an agglomeration when it was founded in 2002. However, by 2020, all of the startups in the CIC were concentrated in a small fraction of its area. Likewise, the OCID is many times bigger than the smallest area needed to contain its startups. Both the CIC and OCID are supported by tax increment financing (i.e., tax concessions to developers and landholders), so future analysis could assess whether these policies cover an excessively large area.

These examples also point to the possible role of office parks or industrial parks in our analysis. The CIC is primarily an industrial park, whereas the OCID is not. We do not have systematic information regarding whether startups in our data are in industrial parks, although we do have information about many specific examples. While industrial parks are fairly common, our information suggests that the majority of startups are not in industrial parks and the majority of our estimated agglomerations using either method are not coincident with or based on industrial parks. Also, the development of industrial parks is endogenous and is partly a response to Marshallian agglomeration economies. Industrial parks supported by the public sector, the private sector or by universities might be viewed as attempts to internalize positive externalities between firms.

The analysis presented here could easily be extended to address additional questions of interest. For example, as suggested by the editor, it would be possible to link this data to patent citation data to investigate whether patent citation connections are more significant within estimated agglomerations than between agglomerations (after adjusting for distance).

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

Also, as noted by a reviewer, it would useful to try using travel time as a distance metric to compare with the results obtained using geographic distance. Full implementation of that extension is not feasible due to cost and computational considerations. We did purchase travel time data from Google Maps for one moderately sized MSA in one year (Portland, OR in 2020) and found similar results to those obtained using geographic distance except for a couple of minor differences due to the effect of rivers on travel times.

The methods we propose here could also be applied much more broadly. The underlying population could be something like car accidents, criminal activity or instances of a particular disease. The methods could also be applied to non-geographic data. For example, we could estimate clusters of product varieties using these methods.

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

Supplementary material

Supplementary data for this article are available at *Journal of Economic Geography* online.

Acknowledgements

We are very grateful to two anonymous referees and the co-editor for exceptionally helpful comments. 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 the financial support from Social Sciences and Humanities Research Council of Canada (SSHRC) grant 435-2017-0627.

Conflict of interest

The authors have no conflicts of interest to support.

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: 758–799.

Al Kindhi, B., Sardjono, T. A., Purnomo, M. H., Verkerke, G. J. (2019) Hybrid *k*-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis. *Expert Systems with Applications*, 121: 373–381.

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

Arzaghi, M., Henderson, J. V. (2008) Networking off Madison avenue. *Review of Economic Studies*, 75: 1011–1038.

Audretsch, D. B., Feldman, M. P. (1996) R&D spillovers and the geography of innovation and production. *American Economic Review*, 86: 630–640.

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: 17–24.

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.

Buzard, K., Carlino, G. A., Hunt, R. M., Carr, J. K., Smith, T. E. (2020) Localized knowledge spillovers: evidence from the spatial clustering of R&D labs and patent citations. *Regional Science and Urban Economics*, 81: 1–20.

Carlino, G., Kerr, W. R. (2015) Agglomeration and innovation. In G. Duranton, J. V. Henderson and W.C. Strange (eds) *Handbook of Regional and Urban Economics.* Vol. 5, pp. 349–404. Amsterdam: North-Holland.

Cesario, E., Vinci, A., Zhu, X. (2020) Hierarchical clustering of spatial urban data. In Y. Sergeyev and D. Kvasov (eds) *Numerical Computations: Theory and Algorithms*. NUMTA 2019. Lecture Notes in Computer Science. Vol. 11973. Springer.

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

Contreras, P., Murtagh, F. (2015) Chapter 6: Hierarchical clustering. In C. Hennig, M. Meila, F. Murtagh and R. Rocci (eds) *Handbook of Cluster Analysis*, pp. 103–120. New York: Chapman and Hall.

Da Rin, M., Hellmann, T., Puri, M. (2013) A survey of venture capital research. In G. Constantinides, M. Harris and R. Stulz (eds) *Handbook of the Economics of Finance.* Vol. 2, pp. 573–648. Amsterdam: North-Holland.

Davis, D. R., Dingel, J. I. (2019) A spatial knowledge economy. *American Economic Review*, 109: 153–170.

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

Egan, E. J. (2021) A framework for assessing municipal high-growth high-technology entrepreneurship policy. *Research Policy*, 104292.

Ellison, G., Glaeser, E. L., Kerr W. R. (2010) What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *American Economic Review*, 100: 1195–1213.
Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011) *Cluster Analysis*. 5th edn. Wiley.
Ellison, G., Glaeser, E. L. (1997) Geographic concentration in US manufacturing industries: a dartboard approach. *Journal of Political Economy*, 105: 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: 1117–1143.
Guzman, J., Stern, S. (2015) Where is Silicon Valley? *Science*, 347: 606–609.
Kaplan, S. N., Lerner, J. (2016) Venture capital data: opportunities and challenges. In J. Haltiwanger, E. Hurst, J. Miranda and A. Schoar (eds) *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, pp. 413–431. Chicago: University of Chicago Press.
Kerr, W. R., Kominers, S. D. (2015) Agglomerative forces and cluster shapes. *Review of Economics and Statistics*, 97: 877–899.
Ketchen D. J., Shook C. L. (1996) The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17: 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: 42–48
Roche, M. P. (2020) Taking innovation to the streets: microgeography, physical structure, and innovation. *Review of Economics and Statistics*, 102: 912–928.
Rosenthal, S. S., Strange, W. C. (2003) Geography, industrial organization, and agglomeration. *Review of Economics and Statistics,* 85: 377–393.
Thorndike, R. L. (1953) Who belongs in the family? *Psychometrika*, 18: 267–276.
von Thünen, J. H. (1826) *Von Thünen's Isolated State*. Pergamon Press (translated into English and reprinted, 1966).
Verstraten, P., Verweij, G., Zwaneveld, P. J. (2019) Complexities in the spatial scope of agglomeration economies. *Journal of Regional Science*, 59: 29–55.
Wentzensen N., Wilson L. E., Wheeler C. M., Carreon J. D., Gravitt P. E., Schiffman M., Castle P. E. (2010) Hierarchical clustering of human papilloma virus genotype patterns in the ASCUS-LSIL triage study. *Cancer Research*, 70: 8578–8586.