January 2021

Review of **JOEG-2020-449**: A New Method for Identifying and Delineating Spatial Agglomerations with Application to Clusters of Venture-Backed Startups

**Overview:**
This is an empirical study assessing the clustering of venture capital-backed startup firms in the US. The analysis uses a spatially precise data set consisting of very granular location data on early stage firms. This data is used with a spatial clustering algorithm to define custom geographies that more accurately reflect the concentration of venture backed entrepreneurial activities - particularly in the high-technology sector. A new methodology for identifying these clusters is proposed, which exploits the spatial precision of their underlying geographical data. The paper reveals several policy implications. While the authors do a decent job of situating this work in the literature, I do have some specific questions.

**Specific Observations and Questions:**
Page 6. The authors indicate that their HCA metric of choice is Euclidean distance, but there is no indication in the text of the coordinate system being used for the geocoded locations of their startups. Technically, appropriate distance measures using geographic coordinates (latitudes/longitudes) can only be derived via the 'great circle distance' formula. However, if the geocoded addresses are provided in projected coordinates (which are measured in linear units) then Euclidean metrics are appropriate. My recommendation is that the authors edit the text to indicate the specific projected coordinate system used for their startup locations - if the startup location coordinates were appropriately projected (preferably to some equidistant projection). If the authors inappropriately used geographic coordinates (lats/lons) in their HCA algorithms then the suggestion is that they redo their analysis with coordinates appropriately projected into linear units.

Page 7. The authors highlight prior research (on page 4) that attempts to assess economic clustering/-concentration based on 'exogenous' geographies such as MSAs or counties – but then indicate that census places (CPs) form the highest level clusters in their analysis. I'm not convinced that their work is free from exogenous determination, since the CP boundaries are the limiting factor determining which firms can be grouped together. Specifically, I would argue that groups of firms in neighboring cities (CPs) cannot be considered a cluster under their framework. It may be true that there is little to no spillover from one CP to another (as the authors suggest), but that doesn't mean that the approach proposed here is immune to exogeneity concerns. The authors should either make a stronger case for why this analysis doesn't suffer from exogeneity OR they should remove that assertion from the text.

Page 8-9. In general, the authors are able to identify very granular, sub-city level clusters that appear to be quite intuitive. I'm not overly concerned that their requirement that clusters form a 'convex hull' is too limiting for this analysis. However, I do think that we shouldn't treat these cluster boundaries as the 'hard and fast' truth of VC-backed startup concentrations. This is particularly important, since we know that policy is typically a blunt tool. I would prefer to think of these clusters as the kernels or seeds of VC-backed startup hotspots - which would allow policy makers to identify the general areas that they should apply their development incentives to.

Page 12. Going from one to two layers always produces the biggest second difference in $R^2$. By ruling out the first second difference does that mean that there are no one or two layer cities? Are there potential cases where that is too limiting? Can an entire city be a cluster?

Page 12. The last sentence of the final paragraph should say that Menlo Park has four clusters and Waltham has three.

Page 13. The elbow method approach is intuitive for the purpose at hand, and I like the (overlay) heat maps. I think they too are intuitive and informative.

Page 15. Cluster density may be an additional regression control to consider, particularly since you discuss is in section 7.2.

Page 17. The regression approach results are also intuitive. But what does the regression tell us if the dependent variable is the aggregate VC investment for the entire city? Is that refined enough? Isn't VC investment available for each firm, which means that you could aggregate VC investment per cluster? Or limit the dependent variable to VC investment going to firms in identified clusters.

Page 17. Directly comparing Figure 7 with Figure 5 is problematic since they are based on a different number of clusters (5 vs 3-4). We should expect them to be different by definition, and shouldn't assume that the differences stem from the different methodologies employed.

Page 18. There is no proper discussion of the 'layer index' results reported in Table 1. What do they indicate? Had a difficult time interpreting those results.

Page 21. What are the indications for Figure 8? Specifically, are there prescribed policy implications for cities that fall below the (blue dotted) regression line?

**Concluding Remarks and Questions:**

The paper is positioned to make relevant and material policy suggestions. Both approaches, the HCA-Regression and the elbow method, produce clusters that are data driven. The indications for these defined clusters are quite clear: the implication is that persistent cluster areas should be targeted for infrastructure investment and other economic development policies (subsidies, tax concessions).

I think the paper would be augmented by identifying and investigating clusters that have grown in either the number of firms or in the increase in VC investment over time. These would be, arguably, the most productive clusters in terms of generating locational qualities conducive to VC investment - and they may reveal specific factors that could better inform policy makers.

I would also like to know how the resultant clusters presented here compare with the 'geographies of innovation' mentioned in the introduction. For example, do these clusters relate in any way to the concentrations of firms associated with the U.S. Regional Innovation Cluster program, the Great Plains Tech and Manufacturing Cluster, or the Oklahoma City Innovation District?

This is a very interesting paper with a clear idea: how to use cluster analysis to define the micro-agglomerations of startups. While the idea that agglomeration tends to happen at very small scales has been around for a while (the Allen Curve; Arzhagi and Henderson, 2008; Saxenian, 1994; Duranton and Overman, 2005; Davis and Dingel, 2019)—and is very intuitive to anybody that recognizes the role of neighborhoods—an effective way of defining startup micro-geographies at different levels has remained elusive.

The author uses hierarchical cluster analysis (HCA), and presents what I think is one of the first serious approaches to solve this problem. I think the ideas he has developed here are interesting, and may be some of the first steps towards a more serious advance in the measurement of the micro-geography of entrepreneurship and innovation.

I do, however, have some major comments.

1. My most important comment is that there are two central issues in the approach as presented.
    a. The first issue is that the author uses ‘Census Places’, which are municipalities such as Waltham, MA, Cambridge, MA, and so forth, as the highest level of agglomeration from which the algorithm runs. But this level of aggregation is wrong because it necessarily misses so many micro-clusters that occur at the border of municipalities.
        i. For example, consider the definitions of Waltham, MA included in the paper. There are startups around the Newton side of the Waltham-Newton border, and the Burlington side of the Burlington-Waltham border that are likely part of the same startup neighborhoods found in these two corners, but are excluded. Similarly, the middle cluster in the Menlo Park, CA, maps in the paper, is very close (and related) to University Ave. in Palo Alto, and may be best thought of as belonging to the same cluster.
            1. From research background, the Menlo Park issue is also apparent in Kerr and Kominers (2015), who show knowledge flow patterns flow across Silicon Valley, and are not constrained to remain within any of those municipalities.
        ii. For this method to be successful, I think it needs to be able to allow these inter-municipality clusters to also emerge, and possibly needs to start all the way at the MSA level.
        iii. I do recognize that this imposes significant computational costs (which the author mentions increases exponentially), but it is precisely being able to show this can be done that makes the contribution of the author in this paper.
            1. I also believe that, because the author has already been able to do his clustering approach in ‘mega municipalities’ like Dallas, Houston, New York, or Miami, this should be computationally do-able for almost any MSA.

    b. The second issue is with one of the methods for choosing the right level of aggregation (i.e. the optimal layer). The author’s ‘elbow’ method makes a lot of sense to me: using changes in the variance explained to find the layer when information gains reduce (sort of the second derivative equal to zero). But the regression method---which was finding the number of layers where the cluster characteristics best predict total VC fundraised—

did not make too much sense personally because it was unclear why layer characteristics such as number of nodes and edges should correlated at all to total VC fundraised.

i. I would recommend dropping this method or adding a lot more literature on why choosing agglomerations that maximize the correlation to clustering features would make sense conceptually.

2. My second important comment is that this paper is missing a significant amount of conceptualization and comment on the literature all the way from the top.

   a. The use of the word ‘cluster’ from the author is inconsistent with its use across the literature, but it is very central to the exercise. What the author identifies are very small geographies, something sort-of like a startup neighborhood or a micro-geography, but ‘cluster’ has always been taken to be very large geographic areas encompassing one or more MSAs such as Silicon Valley or Research Triangle (e.g. Porter, 1990; Delgado et al, 2014; Kerr and Robert-Nicoud, 2020; Chatterji et al). Sometimes even bigger.

   i. So I think even though the paper does formally use computational clustering, the author needs to use a term for his geographies that more directly describes what he measures without causing confusion, something like ‘startup neighborhoods’ or ‘startup microgeographies’, perhaps.

   b. This paper needs a significantly deeper connection to the literature on the nature of agglomeration in startups, and particularly emphasizing how localized high end entrepreneurship. The author is only focusing so far on overall measures of agglomeration (Glaeser and Ellison, 1997; Duranton and Overman, 2005), but I expected a much better literature review on how important the *micro* agglomeration of startups, and the person-to-person connections are. I think an obvious place to start is Arzaghi and Henderson (2008), but I encourage the author to at least consider closely Kerr and Kominers (REStat, 2015), Roche (REStat, 2019), some of the seminal work by Audrestch and Feldman, and Davis and Dingel (AER, 2018). There’s a lot more, but I think this is at least a good start.

   i. As well, the paper should lay out a clearer need for the importance of measuring and observing startup micro agglomerations over time, and the benefits of that above all the other measures we already have: cities, counties, ZIP Codes, MSAs, Census Tract, Census Block which are also intended to measure agglomerations at different levels.

   ii. Ideally, the paper would also spend time considering the advantages of these prior definitions compared to this new method.

These are the two major concerns. I do have a series of other comments that I think could also be useful.

3. While agglomeration benefits are important, there’s many other reasons why companies cluster together in space. Indeed, the north-west cluster identified in the paper in Waltham is all around a big office park (Hobbs Brook Office Park). Maybe the startups are just there because

that's the only place with real office space in Waltham? Similarly, Kendall Square may be a big startup neighborhood with the CIC and MIT, but it's also the only place that allows significant office construction in Cambridge – where else are startups supposed to show up in the data?

   a. Obviously, real estate use policy and overall economic benefits are interrelated and grow symbiotically in economic development plans. The cause-effect of these is not something for the present paper to solve.
   b. But, I do worry that a fair critique of the method here is that all it does is find where a cities' office parks are. I think it would be important for this paper if the author can provide thoughtful comments against this critique.

4. Can the author do some thinking around a different measure of distance than Euclidean? My assumption is that distance is a proxy for 'time to travel', maybe it would be possible to develop direct "time to travel" estimates in google maps for some of these cities and then compare the overlap between the clusters created by both types of distance?
   a. I think this would be a nice robustness to the paper that would also help strengthen any concerns that the author has failed to account for 'congestion' costs of agglomeration.

5. Kerr and Kominers show that the patent citations clusters are overlapping, but the HCA algorithm the author uses is divisive and does not allow overlap. I think the author should include a more thorough discussion of how to think about the fact that clusters could be overlapping in principle, but are not allowed by his approach.

6. I did notice that while the US Census provides 29,573 municipalities, the author only has enough startups to apply the method to 198 of them. This means the data / approach only works in 1% of the set of US cities. This is probably not a problem, agglomeration is truly very skewed, but it is something that warrants a more meaningful discussion as to what do to with the other 99% of the municipalities. For example, can they be included if one had better data? Or is this a natural result of the clustering of innovation and entrepreneurship in space?

7. This paper contribution should include an open dataset or open source code to re-create the startup neighborhoods created by the author. Since its value is mostly as a general purpose tool, I think its scientific contribution is critically dependent on how likely are other researchers to use it.

To summarize, I think the paper is very interesting but also has some meaningful issues that need to be addressed. Good luck with this interesting paper!