Research Policy 51 (2022) 104437

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

The Startup Cartography Project: Measuring and mapping entrepreneurial ecosystems

RJ Andrews [a], Catherine Fazio [b], Jorge Guzman [c,*], Yupeng Liu [d], Scott Stern [e]

[a] Info We Trust, United States
[b] Boston University, Questrom School of Business, United States
[c] Columbia University, Columbia Business School, United States
[d] Rice University, Jones Graduate School of Business, United States
[e] MIT, Sloan School of Management and NBER, United States

A B S T R A C T

This paper presents the Startup Cartography Project (SCP), which offers a new set of entrepreneurial ecosystem statistics for the United States from 1988 to 2016. The SCP combines state-level business registration records with a predictive analytics approach to estimate the probability of "extreme" growth (IPO or high-value acquisition) at or near the time of founding for the population of newly-registered firms. The results highlight the ability of predictive analytics to identify high-potential start-ups at founding (using a variety of different approaches and measures). The SCP then leverages estimates of entrepreneurial quality to develop four entrepreneurial ecosystem statistics, including the rate of start-up formation, average entrepreneurial quality, the quality-adjusted quantity of entrepreneurship, and the entrepreneurial ecosystem performance associated with a given start-up "cohort." These statistics offer sharp insight into patterns of regional entrepreneurship, the correlation of quality (but not quantity) with subsequent regional economic growth and the evolution of entrepreneurial ecosystems over time. The SCP includes both a public-access dataset at the state, MSA, county, and zip code level, as well as an interactive map, the U.S. Startup Map, that allows academic and policy users to assess entrepreneurial ecosystems at an arbitrary level of granularity (from the level of states down to individual street addresses). The SCP and accompanying datasets may be found at: https://www.startupcartography.com/.

1. Introduction

Over the past two decades, there has been a dramatic increase in interest by both academic researchers and policymakers in the role of startup companies in regional economic performance (Saxenian, 1994; Feldman, 2001; Lerner, 2009). This interest reflects both increasing appreciation for the empirical linkage between the two (Feldman et al., 2005; Glaeser et al., 2015), and also the outsized success of particular regions such as Silicon Valley that have hosted waves of start-up firms and experienced a high and sustained level of innovation-driven entrepreneurial dynamism as a result. Relative to traditional economic development incentives (e.g., such as tax breaks for large employers), the promotion of regional entrepreneurship aims to nurture the establishment and expansion of new firms at a relatively low cost in order to benefit from the growth of (some of) those firms over time. For example, in the United States, a host of programs have been initiated over the past decade to foster entrepreneurial ecosystems, ranging from the US Economic Development Association Regional Innovation Strategies (i6) (EDA, 2010), to the Kauffman Foundation ESHIP Communities initiative (Kauffman Foundation, 2019a), to private sector efforts such as "Rise of the Rest." (Revolution, 2019). Perhaps not surprisingly, active debate exists around the design and structure of policies intended to promote regional "entrepreneurial ecosystems" (Feldman and Francis, 2004; Lerner, 2009; Audrestch and Lehmann, 2005; Stam, 2015).

Beyond important conceptual challenges in defining the nature of entrepreneurial ecosystems (Kauffman Foundation, 2019b; Feld, 2012;

We would like to thank Kevin Bryan, Maryann Feldman, Lee Fleming, Rem Koning, Astrid Marinoni, Javier Miranda, and Valentina Tartari for useful comments and feedback. We thank as well participants in the Kauffman Uncommon Methods and Metrics meetings, the Kentucky RISE Regions meetings, the MIT Regional Ecosystem Acceleration Program meetings, the MIT iEcosystems Conference, and the UpJohn Institute Conference on State and Local Financial Incentives. This work was developed thanks to the generous support of the Kauffman Foundation through its Uncommon Methods and Metrics grant. All errors and omissions are our own.

* Corresponding author.
E-mail addresses: rj@infowetrust.com (R. Andrews), cfazio@bu.edu (C. Fazio), jag2367@gsb.columbia.edu (J. Guzman), yupeng.liu@rice.edu (Y. Liu), sstern@mit.edu (S. Stern).

https://doi.org/10.1016/j.respol.2021.104437
Received 4 October 2021; Accepted 14 November 2021
Available online 6 December 2021
0048-7333/

Murray and Stern, 2015), the evaluation of entrepreneurial ecosystems involves an important empirical challenge: how can one measure the state of an entrepreneurial ecosystem at a point in time, track changes in that system over time, or make comparisons across regions (or within regions at different levels of geographic granularity)?

Confronting this challenge requires addressing three interrelated issues: skewness, lagged performance and multiple levels of geographic analysis. First, while the bulk of regional economic growth is linked to the scaling of young firms, startup growth, in and of itself, is heavily skewed. A relatively small number of "successes" from any given cohort of startups has a disproportionate impact on the overall cohort's economic performance. Measurement of an entrepreneurial ecosystem needs to somehow link the measurement of entrepreneurship to these potential skewed outcomes. Put another way, in evaluating the potential of an entrepreneurial ecosystem, it is important to measure not only the *quantity* of new startups being formed there, but also their potential for growth (i.e., their "entrepreneurial *quality*"). Second, the impact of an entrepreneurial ecosystem on a regional economy occurs only after a considerable lag in time between the founding of new firms there and the emergence of skewed outcomes (at least five but maybe as many as 10–20 years!). The combination of skewed outcomes and long time lags makes the assessment of an entrepreneurial ecosystem, and the timely evaluation of policies and programs intended to promote it, challenging even when the ecosystem is strong and programs are working as intended. Third, ecosystems occur at multiple levels of geographic analysis, ranging from small clusters of firms in a few individual locations (e.g., the ecosystem surrounding a university campus) to broader regions at the level of cities, counties, states, or even countries. Any empirical assessment of entrepreneurial ecosystems thus also needs to account for the fact that ecosystems can be evaluated (and will need to be measured) at multiple levels of analysis.

By and large, prior efforts to overcome these measurement challenges have relied on one of two broad approaches. One central approach focuses on measuring the quantity of new firms founded at a given point in time (within a fixed geographic domain). Most notably, the Longitudinal Business Database (LBD) (Jarmin and Miranda, 2002) provides enormously valuable insight into the number of new firms (with at least one employee) founded by year and state. However, despite its considerable strengths, the LBD (and related datasets such as the Business Dynamics Statistics) tend to abstract away from differences among firms at the time of founding, and the statistics that are made available may only be produced (at least for public use) at a relatively high rate of aggregation (such as a state). Alternatively, it is possible to simply condition the study of an entrepreneurial ecosystem on those firms resident there that satisfy a pre-determined performance criteria, such as the receipt of venture capital (less than 0.1% of all newly founded firms receive venture capital in any given year), or achievement of a certain level of employment growth within a limited time-frame (see Stangler and Bell-Materson, 2015). However, for many purposes, conditioning the analysis of an ecosystem on such milestones conflates the measurement of the rate of entrepreneurship found there with the assessment of overall entrepreneurial performance. If the rate of venture financing in a given region is low or declining, for example, does that imply that there is "too little" venture capital or too few firms found there with the potential to attract it?

The purpose of this paper is to introduce a new database and mapping platform, the Startup Cartography Project ("SCP"), that aims to address these challenges in an integrated manner for both academic and policy users.

Specifically, the core objective of the SCP is to provide a consistent, transparent and accessible data resource that allows for granular (as well as aggregated) and timely (as well as retrospective) measurements of entrepreneurial ecosystems. The SCP incorporates three broad elements. First, building on Guzman and Stern (2015, 2017, 2020), the SCP uses a predictive analytics approach to estimate, for any given startup, the probability of growth of that firm at or near the time of founding (a measure of its quality). Second, leveraging this measure of entrepreneurial "quality" for all firms, the SCP builds a set of novel entrepreneurship statistics that capture the quantity, quality and performance of any given set of firms, allowing for consistent measures of entrepreneurship across time and place. Finally, we then translate the core SCP statistics into an interactive mapping tool, the U.S. Startup Map, that allows for dynamic and interactive visualization of entrepreneurship at an arbitrary level of aggregation (from an individual street address up to the level of the United States).

Our predictive analytics approach builds upon and extends our own prior work (including Guzman and Stern (2015, 2017, 2020)) leveraging three core insights. First, because the challenges to growth as a sole proprietorship are formidable, a practical requirement for any growth-oriented entrepreneur that would contribute to an entrepreneurial ecosystem is business registration (as a corporation, partnership, or limited liability company). We take advantage of the public nature of business registration records to define a population sample of entrepreneurs observed at a similar (and foundational) stage of the entrepreneurial process. Second, moving beyond simple counts of business registrants (Klapper et al., 2010), we are able to measure characteristics related to entrepreneurial quality *at or close to the time of registration.* For example, we can measure start-up characteristics such as whether the firm is organized in order to facilitate equity financing (e.g., registering as a corporation or in Delaware), how the firm is named (e.g., whether it signals a high-tech sector versus a local focus) or whether the firm acquires or develops measurable innovations (e.g., a patent or trademark). Third, we leverage the fact that, though rare, we observe meaningful growth outcomes for some firms (e.g., those that achieve an IPO or high-value acquisition within six years of founding), and are therefore able to estimate the relationship between these growth outcomes and start-up characteristics. In other words, our approach implements a predictive analytics approach to entrepreneurship which allows us to estimate, for any given firm, its underlying level of quality (as linked to particular observables) at or near the time of founding.

The SCP applies this predictive analytics approach in the context of 49 U.S. states and Washington D.C. from 1988 to 2014, and 46 U.S. states within the year 2014–2016 (a significant extension beyond our earlier work). Consistent with Guzman and Stern (2015, 2017), we find that a small number of characteristics allow us to develop a robust predictive model that distinguishes firm quality. In an out-of-sample test, we find that 54% of realized growth outcomes occur in the top 5% of our estimated quality distribution (and nearly 37% in the top 1% of the estimated quality distribution). Moreover, we find that a small number of governance and intellectual property characteristics – Delaware registration, registering for a trademark or patent application – are the single largest factors predicting subsequent start-up performance. However, our work with policymakers and other analysts suggested that the ability to actually utilize a predictive analytics approach was more persuasive if we focused on a modified version of the model where we center only on those start-up characteristics that are closely linked with the legal and intellectual property environment surrounding the firm (i.e., Delaware registration, and the registration of patents or trademarks). For concision, we refer to the richer model incorporating such features as eponymy as the "academic" model, and we refer to the model that relies exclusively on institutional features as the "policy" model.

We then use these estimates to generate four aggregate economic statistics for the measurement of entrepreneurship: the Startup Formation Rate (SFR), the Entrepreneurship Quality Index (EQI), the Regional Entrepreneurship Cohort Potential Index (RECPI) and the Regional Ecosystem Acceleration Index (REAI). SFR is simply a measure of new firm formation (within a cohort of firms defined by a given time period and geographic scope). EQI is a measure of *average quality* within any cohort, allowing for the calculation of the probability of a growth outcome within a specified population of start-ups. RECPI multiplies SFR and EQI within a given geographic domain (e.g., a zip code such as 02,139 (in Cambridge, MA) or the entire state of Massachusetts),

yielding a measure of the quality-adjusted quantity of entrepreneurship within that ecosystem. Whereas EQI compares entrepreneurial quality across different groups (and so facilitates apples-to-apples comparisons across groups of different sizes), RECPI allows the direct calculation of the expected number of growth outcomes from a given start-up cohort within a given regional boundary. Finally, REAI, measured as the ratio of realized to expected growth events, is a measure of entrepreneurial ecosystem performance in accelerating startups after founding. While RECPI estimates the expected number of growth events for a given group of firms, over time we can observe the realized number of growth events from that cohort. This difference (reflected in REAI) can be interpreted as the relative ability of firms within a given region to grow, conditional on their initial entrepreneurial quality. Variation in ecosystem performance could result from differences across regional ecosystems in their ability to nurture the growth of start-up firms, or changes over time or location in financing availability, economic conditions, or economic policies or programs.

We construct these statistics at the state, MSA, county and zip code level, and illustrate the potential of these data for regional analysis by undertaking a descriptive examination of the 100 largest MSAs in the United States. A few key findings stand out. On the one hand, both the average level of EQI and RECPI/Population are much higher for key regions that have been traditionally associated with growth entrepreneurship, such as the Bay Area (San Francisco and San Jose, CA), as well as Boston, MA and Austin, TX. At the same time, the SCP captures (in a timely way) the recent growth in entrepreneurial ecosystems such as Provo, UT, and Denver, CO. In addition, our descriptive analysis offers a novel lens through which to view the linkage between entrepreneurship and regional economic growth. Whereas the quantity of entrepreneurship is essentially uncorrelated with subsequent regional economic growth, the quality of entrepreneurship in a given ecosystem is strongly correlated with subsequent regional economic performance.

Finally, we use these statistics to build an interactive map, the U.S. Startup Map, that visualizes entrepreneurial ecosystems across time and place. Specifically, the U.S. Startup Map allows individual users to choose both the timeframe for analysis (i.e., a given year) as well as the level of geographic granularity (ranging from the United States down to the level of individual street addresses), and provides a visualization of both the SFR and EQI for that chosen geography. As suggested earlier, feedback from policy users suggested that we adopt the "policy" model for our visualization, since users of the interactive map are more likely to be interested in identifying start-up populations linked to institutional factors such as Delaware registration, or trademark or patent applications. By helping to stakeholders to see the results of quantitative academic empirical research, our work with the U.S. Startup Map also holds broader implications for policy and practice.

This paper builds upon and extends our prior work measuring the quantity and quality of entrepreneurship (Guzman and Stern, 2015, 2020; Fazio et al., 2017). Specifically, we lay the foundation for broader use, and significantly extend the term and coverage, of the released SCP data. SCP datasets now encompass 49 states (and Washington, D.C.) through 2014 and 46 states through 2016 (up from the business registration records for 32 states from 1988 to 2014 previously analyzed in Guzman and Stern, 2020) and account for 99.6% of US GDP (vs. 81% in Guzman and Stern, 2020). The scope of SCP datasets has likewise been increased to include equity growth events and patent and trademark applications corresponding to business registrations from additional states and time periods. Building on Guzman and Stern (2020), we also develop a more comprehensive set of entrepreneurial ecosystem metrics, including RECPI and SFR (through 2016) and REAI (which measures the performance of entrepreneurial ecosystems over time) up through 2012. In addition, we introduce the "policy model" as a complement to our "academic model" based on stakeholder feedback. Finally, we establish the basis for the U.S. Startup Map, visualizing a consistent and understandable set of digital markers of startup quality. We detail how color palette and assignment based on the policy model assist stakeholders in grasping entrepreneurial quality separate from startup formation and make comparisons possible across towns, regions and time.

The rest of this paper proceeds as follows. Section II presents our approach for constructing entrepreneurial ecosystem statistics. Section III reviews data and estimation. Section IV calculates the key entrepreneurial quality statistics and overviews some broad descriptive patterns we observe from the SCP metrics. Section V describes the application of the SCP for policy, and the development of an interactive and dynamic visualization tool, the U.S. Startup Map, that allows users to assess entrepreneurial ecosystems in any time period of their choosing and at an arbitrary level of granularity. Section VI concludes.

1.1. Measuring entrepreneurial ecosystem quantity, quality and performance[1]

A central challenge in assessing entrepreneurial ecosystems (for a wide range of both academic and policy questions) is the development of measures of entrepreneurial ecosystems that are (at least potentially) comparable across different ecosystems and over time. A central contribution of the Startup Cartography Project is to introduce a set of consistent measures that account not only for the quantity of entrepreneurs but also for the quality of those entrepreneurs at or near the time of founding. Our approach leverages that fact that while there are a very large number of new businesses established at any point in time (and so attempting to categorize them through an external assessment would be burdensome), entrepreneurs themselves have information about both their underlying idea and ambition, and make choices at the time of founding consistent with their objectives and potential for growth. Specifically, by starting with the entire population of business registrants (a relatively low-stakes requirement for any business in the United States that has ambition to move beyond self-employment), it is possible to use a predictive analytics approach that relates the ultimate performance of start-up firms to initial early-stage choices by the entrepreneur that are also observable at or around the time of founding as a "digital signature" for each firm. We measure entrepreneurial quality by estimating the relationship between observed growth outcomes and start-up characteristics using the population of at-risk firms. For a firm i born in region r at time t, with at-birth start-up characteristics $H_{i,r,t}$, we observe growth outcome $g_{i,r,t+s}$ s years after founding and estimate :[2]

$$\theta_{i,r,t} = P\left(g_{i,r,t+s}\middle|H_{i,r,t}\right) = f\left(\alpha+\beta H_{i,r,t}\right) \quad (1)$$

This model allows us to *predict* quality as the probability of achieving a growth outcome given start-up characteristics at founding, and so estimate entrepreneurial quality as $\widehat{\theta}_{i,r,t}$. As long as the process by which start-up characteristics map to growth remain stable over time (an assumption which is itself testable), this mapping allows us to form an

[1] This section draws upon but also extends the discussion of the estimation of entrepreneurial quality, as well as the development of statistics for a "cohort" of start-ups from Guzman and Stern (2015, 2017, 2020). Relative to those earlier works, this paper both extends the range of statistics reported and the coverage of data that are now incorporated into the SCP, and also links the statistics to the U.S. Startup Map, a novel interactive mapping tool.

[2] The key assumption we make is that the relationship between characteristics of startups and growth outcomes is stable during our time period, and stable across U.S. states. While there is no need for this to be necessarily true, we demonstrate in Appendix C and prior work that this appears to be the case in our data. For example, the correlation between a quality measure estimated by running a regression for each state independently, and our quality measure estimated from a single national regression is 0.82, and no state has a lower correlation than 0.5. In Guzman and Stern (2020), we also report, in our appendix, a similar exercise across years finding significant stability. However, care should be taken when going beyond our specific geography and time periods: the process through which the relationship of observables to outcomes changes over longer periods of time and outside the U.S. is not well understood and a central question for any future work that expands on our approach.

estimate of entrepreneurial quality for any business registrant within our sample (even those in recent cohorts where a growth outcome (or not) has not yet had time to be observed).

The predictive analytics approach implemented in (1) (the primary focus of our prior work) allows us to recover an estimate for the entrepreneurial quality of any given firm at (or near) the time of founding. However, we then need to undertake a second step in which we form consistent and meaningful entrepreneurial ecosystem metrics that allow for comparisons across different ecosystems and across time. Specifically, the Startup Cartography Project provides users with four key entrepreneurship statistics capturing the rate of formation of registered firms, the level of entrepreneurial quality for a given population of start-ups, the potential for growth entrepreneurship within a given region and start-up cohort, and the performance over time of a regional entrepreneurial ecosystem in realizing the potential performance of firms founded within a given location and time period.

The Startup Formation Rate (SFR) represents the quantity of for-profit, new business registrants within a given population. It mimics other quantity based measures available such as the Business Dynamics Statistics (BDS) or the Global Entrepreneurship Monitor (GEM).

The Entrepreneurial Quality Index (EQI). To create an index of entrepreneurial quality for any group of firms (e.g., all the firms within a particular cohort or a group of firms satisfying a particular condition), we simply take the *average* quality within that group. Specifically, in our regional analysis, we define the *Entrepreneurial Quality Index* (EQI) as an aggregate of quality at the region-year level by simply estimating the average of $\theta_{i,r,t}$ over that region:

$$EQI_{r,t} = \frac{1}{N_{r,t}} \sum_{i \in \{I_{r,t}\}} \theta_{i,r,t} \tag{2}$$

where $\{I_{r,t}\}$ represents the set of all firms in region r and year t, and $N_{r,t}$ represents the number of firms in that region-year. To ensure that our estimate of entrepreneurial quality for region r reflects the quality of start-ups in that location rather than simply assuming that start-ups from a given location are associated with a given level of quality, we exclude any location-specific measures $H_{r,t}$ from the vector of observable start-up characteristics.[3]

The Regional Entrepreneurship Cohort Potential Index (RECPI). From the perspective of a given region, the overall inherent potential for a cohort of start-ups combines both the quality of entrepreneurship in a region and the number of firms in such region (a measure of quantity). To do so, we define *RECPI* as simply $EQI_{r,t}$ multiplied by the number of firms in that region-year:

$$RECPI_{r,t} = EQI_{r,t} \times N_{r,t} \tag{3}$$

Since our index multiplies the *average* probability of a firm in a region-year to achieve growth (quality) by the number of firms, it is, by definition, the expected number of growth events from a region-year given the start-up characteristics of a cohort at birth. This measure of course abstracts away from the ability of a region to realize the performance of start-ups founded within a given cohort (i.e., its ecosystem performance), and instead can be interpreted as a measure of the "potential" of a region given the "intrinsic" quality of firms at birth, which can then be affected by the impact of the entrepreneurial ecosystem, or shocks to the economy and the cohort between the time of founding and a growth outcome.

The Regional Ecosystem Acceleration Index (REAI). While RECPI estimates the *expected* number of growth events for a given group of firms, over time we can observe the *realized* number of growth events from that cohort. This difference can be interpreted as the relative ability of firms within a given region to grow, conditional on their initial entrepreneurial quality. Variation in ecosystem performance could result from differences across regional ecosystems in their ability to nurture the growth of start-up firms, or changes over time due to financing cycles or economic conditions. We define REAI as the ratio of realized growth events to expected growth events:

$$REAI_{r,t} = \frac{\sum g_{i,r,t}}{RECPI_{r,t}} \tag{4}$$

A value of REAI above one indicates a region-cohort that realizes a greater than expected number of growth events (and a value below one indicates under-performance relative to expectations). REAI is a measure of a regional performance premium: the rate at which the regional business ecosystem supports high potential firms in the process of becoming growth firms.

Together, SFR, EQI, RECPI, and REAI offer researchers and regional stakeholders the ability to undertake detailed evaluations (over time, and at different levels of geographic and sectorial granularity) of entrepreneurial ecosystem performance.

2. Data and estimation

The foundational data source for the SCP are state-level business registration records, a potentially rich and systematic data set for the study of entrepreneurship. Business registration records are public records created endogenously when an individual registers a new business as a corporation, LLC or partnership. Our data covers 49 states and Washington, D.C. from 1988 to 2014, and 46 states and Washington, D. C. from 2014 to 2016.[4] While it is possible to found a new business without business registration (e.g., a sole proprietorship), the benefits of registration are substantial, and include limited liability, various tax benefits, the ability to issue and trade ownership shares, and credibility with potential customers. Furthermore, all corporations, partnerships, and limited liability companies must register with a Secretary of State (or Secretary of the Commonwealth) in order to take advantage of these benefits: the act of *registering* the firm triggers the legal creation of the company. As such, these records reflect the population of businesses that take a form that is a practical prerequisite for growth.[5] Concretely, our analysis draws on the complete population of firms satisfying one of the following conditions: (a) a for-profit firm in the local jurisdiction or (b) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the local state. In other words, our analysis excludes non-profit organizations as well as companies whose primary location is not in the state. The resulting dataset contains 39,460,805

[3] Three particular features of EQI are notable. First, while the general form of [(EQI)]_(r,t) is a panel format, it is possible to construct a cross-sectional distribution of quality at a moment in time (i.e., [(EQI)]_(r,t_0)) to facilitate analyses such as spatial mapping. Second, the level of geographical aggregation is arbitrary: while the discussion of a "region" may connote a large geographic area, it is possible to calculate EQI at the level of a city, ZIP code, or even individual addresses. Finally, we can extend EQI in order to study an arbitrary grouping of firms (i.e., we do not need to select exclusively on geographic boundaries). For example, we can examine start-ups whose founder share a common demographic characteristic (e.g., sex), or firms that undertake a specific strategic action (e.g., engage in crowdfunding).

[4] These are all U.S. states except for Delaware from 1988-2014 and all U.S. States except for Delaware, Illinois, South Carolina and Michigan from 2014-2016 (these three states significantly increased the fees and/or administrative burden with using state-level registration data for the most recent years)

[5] This section draws on Guzman and Stern (2015, 2017, 2020), where we introduce the use of business registration records in the context of entrepreneurial quality estimation. Please also see data appendices in those earlier papers.

observations.[6] For each observation we construct variables related to: (a) a growth outcome for each start-up; (b) start-up characteristics based on business registration observables; and (c) start-up characteristics based on external observables that can be linked directly to the start-up. We briefly review each one in turn. We provide a more detailed summary relating to each observable in our data appendix. *Growth.* The growth outcome utilized in the SCP, *Growth*, is a dummy variable equal to 1 if the start-up achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration, as reported in Thomson Reuters SDC database.[7] During the period of 1988 to 2010, we identify 15,362 firms that achieve growth, representing 0.06% of the total sample of firms in that period. *Start-Up Characteristics*. At the center of our analysis is an empirical approach to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures of start-up characteristics: (a) measures based on business registration data observable in the registration record itself, and (b) measures based on external indicators of start-up quality that are observable at or near the time of business registration. *Measures Based on Business Registration Observables.* We construct twelve measures based on information observable in business registration records. We first create two binary measures that relate to how the firm is registered, *Corporation*, whether the firm is a corporation rather than an LLC or partnership, and *Delaware Jurisdiction*, whether the firm is registered in Delaware. We then create two additional measures based directly on the name of the firm. *Eponymy* is equal to 1 if the first, middle, or last name of the top managers is part of the name of the firm itself.[8] We hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Our second measure relates to the structure of the firm name. Based on our review of naming patterns of growth-oriented start-ups versus the full business registration database, a striking feature of growth-oriented firms is that the vast majority of their names are at most two words (plus perhaps one additional word to capture organizational form (e.g., "Inc.")). We define *Short Name* to be equal to one if the entire firm name has three or less words, and zero otherwise.[9]

We then create several measures based on how the firm name reflects the industry or sector within which the firm is operating, taking advantage of the industry categorization of the U.S. Cluster Mapping Project ("US CMP") (Delgado et al., 2016) and a text analysis approach. We develop eight such measures. The first three are associated with broad industry sectors and include whether a firm can be identified as local (*Local*), or traded (*Traded*), or traded within resource intensive industries (*Traded Resource Intensive*). The other five industry groups are narrowly defined high technology industries that could be expected to have high growth, including whether the firm is associated with biotechnology (*Biotech Sector*), e-commerce (*E-Commerce*), other information technology (*IT Sector*), medical devices (*Medical Dev. Sector*) or semiconductors (*Semiconductor Sector*).

Measures based on External Observables. We construct two measures related to start-up quality based on intellectual property data sources from the U.S. Patent and Trademark Office. *Patent* is equal to 1 if a firm holds a patent application within the first year and 0 otherwise. We include patents that are filed by the firm within the first year of registration and patents that are assigned to the firm within the first year from another entity (e.g., an inventor or another firm). Our second measure, *Trademark*, is equal to 1 if a firm applies for a trademark within the first year of registration.[10]

Table 1 groups these measures in five categories: outcome variables, name-based observables, intellectual property observables and industry measures (US CMP Clusters and US CMP High-Tech Clusters), and reports the summary statistics and sources for these measures. Appendix A includes a detailed discussion of the specific set of US CMP clusters used to develop each industry classification and the relative difference in the means of our variables between firms that grow and firms that do not grow. Appendix Table C4 reports the coefficients of univariate logit regressions on growth for each of these variables.

3. Entrepreneurial quality models

We use this data to estimate two alternative logit regression models that allow one to examine how the presence or absence of a startup characteristic correlates with the probability of growth: the 'academic model', which includes all measures, and the policy model which exclusively utilizes jurisdiction (i.e., *Delaware*), legal form (*Corporation*), and intellectual property measures (*Patent* and *Trademark*) only.[11]

Table 2 reports our results for the academic model for all registered firms in the dataset between 1988 and 2010. The results are striking. We find an extremely strong (and robust) correlation between startup characteristics and the probability of growth. Substantial changes in the predicted likelihood of a growth outcome are associated with characteristics observable at founding from business registration records as well as characteristics observable with a lag (e.g., patent and trademark applications). On the one hand, startups founded as corporations are 190% *more* likely to grow. Similarly, firms with short names are close to 120% *more* likely to grow. On the other, eponymous firms are roughly 70% *less* likely to achieve an equity growth outcome. Startups that apply for a patent or trademark in their first year after founding are 2300% and over 273%, respectively, more likely to achieve an equity growth

[6] The number of firms founded in our sample is substantially higher than the U.S. Census Longitudinal Business Database (LBD), done from tax records. For example, for Massachusetts in the period 2003-2012, the LBD records an average of 9,450 new firms per year and we record an average of 24,066 firm registrations. We have yet to explore the reasons for this difference. However, we expect that it may be explained, in part by: (i) partnerships and LLCs that do not have income during the year do not file tax returns and are thus not included in the LBD, and (ii) firms that have zero employees and thus are not included in the LBD.

[7] Although the coverage of IPOs is likely to be nearly comprehensive, the SDC data set excludes some acquisitions. SDC captures their list of acquisitions by using over 200 news sources, SEC filings, trade publications, wires, and proprietary sources of investment banks, law firms, and other advisors (Churchwell, 2016). Barnes, Harp, and Oler (2014) compare the quality of the SDC data to acquisitions by public firms and find a 95% accuracy; Netter, Stegemoller, and Wintoki (2011), perform a similar review. While we know this data not to be perfect, we believe it to have relatively good coverage of 'high value' acquisitions. Further, none of the cited studies found significant false positives, suggesting that the only effect of the acquisitions we do not track will be simply an attenuation of our estimated coefficients.

[8] Belenzon et al (2017; 2019), perform a more detailed analysis of the interaction between eponymy and firm performance, highlighting name as a signal chosen by entrepreneurs given differences in growth intention.

[9] Companies such as Akamai or Biogen have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., "New England Commercial Realty Advisors, Inc.").

[10] We aggregate patent filings by and assignments to new startups in the construction of our patent measure given the relatively rare nature of patents in the data. As stated in Gans et al. (2003) and Ziedonis and Hsu (2008), the disambiguation of startup patent filings and assignments, and exploration of their correlation with predicted growth outcomes is an important subject for future research.

[11] Given that growth outcomes for startups are extremely rare, we considered whether use of a rare event logit estimation would be more appropriate (King and Zeng, 2001). As our data includes over 15,000 growth events, it presents a sufficiently large sample of these rare events such that the maximum likelihood estimation of logistic regression analysis would not suffer from small sample bias. (Allison, 2012).

Table 1
Summary Statistics.

Measure	Source	Description	Mean	Std. Dev.
Outcome Variables				
Equity Growth (IPO or Acquisition)	SDC Platinum IPO and M&A.	1 if firm has an equity growth event in the first 6 years.	0.0006	0.024
Corporate Form Observables				
Corporation	Business Reg.	1 if a firm is a corporation (not an LLC or partnership)	0.454	0.248
Delaware	Business Reg.	1 if the firm's jurisdiction is Delaware	0.021	0.020
Name-Based Observables				
Short Name	Business Reg.	1 if the firm's name length is 3 words or less (including firm type (e.g. "inc."))	0.461	0.248
Eponymous	Business Reg.	1 if the firm's name includes the president or CEO first or last name.	0.079	0.073
Intellectual Property Observables				
Patent	USPTO	1 if the firm acquires for a patent application within 1 year of founding.	0.0018	0.0018
Trademark	USPTO	1 if the firm acquires for a trademark within 1 year of founding.	0.0015	0.015
Industry Measures (US CMP Clusters)				
Local Industry	Estimated from name	If firm name is associated to a local industry.	0.194	0.156
Traded	Estimated from name	If firm name is associated to a traded industry.	0.538	0.249
Resource Intensive Industry	Estimated from name	If firm name is associated to a resource intensive industry.	0.128	0.111
Industry Measures (US CMP High-Tech Clusters)				
Biotechnology	Estimated from name	If firm name is associated to the Biotechnology industry cluster.	0.002	0.002
E-Commerce	Estimated from name	If firm name is associated to the E-Commerce industry cluster.	0.049	0.046
IT	Estimated from name	If firm name is associated to the IT industry cluster.	0.021	0.143
Medical Devices	Estimated from name	If firm name is associated to the Medical Devices industry cluster.	0.027	0.026
Semiconductor	Estimated from name	If firm name is associated to the Semiconductor industry cluster.	0.0004	0.0004
Observations			39,460,805	

This table represents our full dataset, comprised of all registered firms registered within the years 1988 and 2014 in Washington D.C. and 49 US states (excluding Delaware), and 46 states (excluding Delaware, Illinois, Michigan, South Carolina) within the year 2014 and 2016. These states account for 99.6% of US GDP in 2014. All measures defined in detail in Section III of this paper. Business registration records are public records created endogenously when a firm registers as a corporation, LLC, or partnership. Equity Growth is a binary indicator equal to 1 if a firm achieves IPO or acquisition within 6 years and 0 otherwise. Growth is only defined for firms born in the cohorts of 1988 to 2010. Growth IPOs include only 'true' startup IPOs; we exclude all financial IPOs, REITs, SPACs, reverse LBOs, re-listings, and blank check corporations. IP observables include both patents and trademarks filed by the firm within a year of founding, as well as previously filed patents assigned to the firm close to founding. All business registration observables, IP observables, are estimated at or close to the time of firm founding. Further information on all measures and our approach more generally, can be found in Guzman and Stern (2018).

outcome within 6 years of founding.[12] Moreover, these changes in predicted probabilities are multiplicative in nature: a startup that registers in Delaware *and* applies for a patent in its first year is over 93 times more likely to grow than a firm that only registers in its home state and does not apply for intellectual property protection.[13,14]

Not surprisingly, findings from the policy model are comparable. As reported in Table 3, forming as a corporation, registering in Delaware, and filing for a patent or trademark within the first year are correlated with increases in the likelihood of a growth outcome of 100%, 2780%, 1783% and 309%, respectively. And, like the academic model, the predictive power of these startup characteristics is multiplicative in nature. A startup that is registered in Delaware and files for both a patent and trademark within its first year is 856 times more likely to achieve a growth outcome than one that does not.

Robustness and Predictive Quality. In Fig. 1, we evaluate the predictive quality of the academic model estimates by undertaking a tenfold cross-validation test (Witten and Frank, 2005),[15] and report the out-of-sample share of realized growth outcomes at different portions of the entrepreneurial quality distribution. The results are striking. On average, 63% of all growth firms are included within the top 10% of our estimated growth probability distribution. 54% and 37% of all growth outcomes are included within the top 5% and 1% of the estimated entrepreneurial quality distribution, respectively. Growth, however, is still a relatively rare event even among those startups with the highest estimated entrepreneurial quality: the average firm within the top 1% of estimated entrepreneurial quality distribution has only a 2.1% chance of realizing a growth outcome. Figure A1 demonstrates that estimates generated by the Policy model are similarly robust in terms of predictive quality. The top 10% of the distribution of estimated entrepreneurial

[12] Preliminary models which center on one or two startup characteristics find similar, albeit slightly higher correlations. Corporations and firms registered in Delaware are 267% and 2,554% more likely to achieve a growth outcome, respectively. Firms with short names are 173% more likely to grow, while eponymous firms are close to 80% less likely to achieve a growth outcome. Those startups that apply for a patent or trademark in their first year are 5,091% and 624% more likely to achieve an equity growth outcome, respectively.

[13] It is very important to emphasize that these startup characteristics are not the *causal* drivers of growth, but instead are "digital signatures" that allow us to distinguish firms in terms of their entrepreneurial quality. Registering in Delaware or filing for a patent will not guarantee a growth outcome for a new business, but the firms that historically have engaged in those activities have been associated with skewed growth outcomes.

[14] The precision allowed by our definition of quality comes nonetheless at a cost. Our definition does not allow us to include all the richness of social outcomes through which companies help communities or individuals. In principle, however, a richer version of our approach that includes multiple outcomes and a larger number of observables might be able to achieve this result.

[15] Specifically, we divide our sample into 10 random subsamples, using the first subsample as a testing sample and the other 9 to train the model. For the retained test sample, we compare realized performance with entrepreneurial quality estimates from the model resulting from the 9 training samples. We then repeat this process 9 additional times, using each subsample as the test sample exactly once. This approach allows us to estimate average out of sample performance, as well as the distribution of out of sample test statistics for our model specification.

Table 2
Academic Model Predictive Analytics Model of Equity Growth Dependent Variable: Equity Growth Logit model. Incidence Rate Ratios Reported.

	Preliminary Models			Full Model
	(1)	(2)	(3)	(4)
Corporate Governance Measures				
Corporation	3.671***			2.867***
	(0.0776)			(0.0612)
Delaware	26.54***			
	(0.479)			
Name-Based Measures				
Short Name		2.729***		2.228***
		(0.0485)		(0.0415)
Eponymous		0.200***		0.288***
		(0.0112)		(0.0162)
Intellectual Property Measures				
Patent			51.91***	
			(1.556)	
Trademark			7.235***	3.731***
			(0.412)	(0.191)
Patent - Delaware Interaction				
Patent Only				23.58***
				(1.049)
Delaware Only				17.38***
				(0.358)
Patent and Delaware				94.31***
				(3.468)
US CMP Clusters				Yes
US CMP High-Tech Clusters				Yes
N	26,969,231	26,969,231	26,969,231	26,969,231
R-squared	0.135	0.050	0.087	0.184

We estimate a logit model with Growth as the dependent variable. Growth is a binary indicator equal to 1 if a firm achieves IPO or acquisition within 6 years and 0 otherwise. Growth is only defined for firms born in the cohorts of 1988 to 2010. This model forms the basis of our entrepreneurial quality estimates, which are the predicted values of the model. Incidence ratios reported; Robust standard errors in parenthesis. * $p<0.05$ ** $p<0.01$ *** $p<0.001$.

quality includes 49% of all growth firms.[16] Appendix C, 'Evaluating the Robustness of Predictive Models', includes further assessment on the validity of our estimates, including a geographic validation, differences across model ROC scores, the univariate coefficients for each of our main

Table 3
Policy Model Predictive Analytics Model of Equity Growth Dependent Variable: Equity Growth Logit model. Incidence Rate Ratios Reported.

	(1)	(2)	(3)
Independent Effects			
Delaware	28.80***		
	(0.742)		
Patent	18.83***		
	(0.634)		
Trademark	4.087***	4.312***	
	(0.240)	(0.230)	
Delaware, Patent Interactions			
Delaware = 1, Patent = 0		38.66***	
		(1.001)	
Delaware = 0, Patent = 1		81.20***	
		(3.770)	
Delaware = 1, Patent = 1		470.8***	
		(18.35)	
Delaware, Patent, Trademark Interactions			
Delaware = 1, Patent = 0, Trademark = 0			40.20***
			(1.075)
Delaware = 0, Patent = 1, Trademark = 0			96.26***
			(4.448)
Delaware = 0, Patent = 0, Trademark = 1			620.8***
			(22.30)
Delaware = 1, Patent = 1, Trademark = 0			41.53***
			(3.241)
Delaware = 1, Patent = 0, Trademark = 1			326.8***
			(21.00)
Delaware = 0, Patent = 1, Trademark = 1			137.7***
			(23.76)
Delaware = 1, Patent = 1, Trademark = 1			856.9***
			(66.53)
Corporation	1.902***	2.278***	2.445***
	(0.0415)	(0.0547)	(0.0606)
N	26,969,231	26,969,231	26,969,231
pseudo R-sq	0.134	0.138	0.141

Robust standard errors reported in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

measures, and the robustness of our regression to high end employment outcomes.[17]

Entrepreneurial Ecosystem Statistics. We then use our measures of estimated quality to develop economic indices that simultaneously account for both the quantity and the quality of entrepreneurship (and which are outlined in the empirical framework section):

- SFR—the Startup Formation Rate—the quantity of new business registrants within a given population.
- EQI—the Entrepreneurial Quality Index—the average growth potential (or "quality") of any given group of new firms.
- RECPI—the Regional Entrepreneurship Cohort Potential Index—the number of startups within a particular location or region expected to later achieve a significant growth outcome.
- REAI—the Regional Entrepreneurship Acceleration Index—the ability of a region to convert entrepreneurial potential into realized growth.

[16] Of course, other measures of growth can also be used as outcome variables. Guzman and Stern (2020) finds that "these broad patterns of results also hold if one focuses on … alternative growth measures such as the realization of more than 500 employees within the first six years after founding." There, the authors "take advantage of a dataset of employment levels for more than 10 million firms available from Infogroup USA between 1997 and 2014 [and] construct two new outcome variables, Employment Growth 500 and Employment Growth 1000, each equal to 1 for all firms recorded as having greater than 500 or 1000 or more employees, respectively, within 6 years, and 0 otherwise." As robustness checks, the authors then: (1) compare their baseline entrepreneurial quality model using Growth and Employment Growth measures, finding their estimates similar in sign and relative magnitude; (2) use the model with the lower level of concordance (Employment Growth 500) as an alternative baseline for their predictive approach, finding the correlation between predictive analytics to be .84; and (3) examine how the incidence of Employment Growth 500 is predicted by their estimates of entrepreneurial quality, again finding striking similarity. In their appendix, the authors also provide detailed disaggregation for different types of equity outcomes including only IPOs, dropping all acquisitions lower than $100M, and increasing the outcome window to include also late IPOs such as those occurring after 9 years. As well, Catalini et al., 2019 study further the predictive relationship between our observables and venture capital.

[17] Further information on the univariate relationship and predictive accuracy of our estimates can be found in Guzman and Stern (2017) and Guzman and Stern (2020).

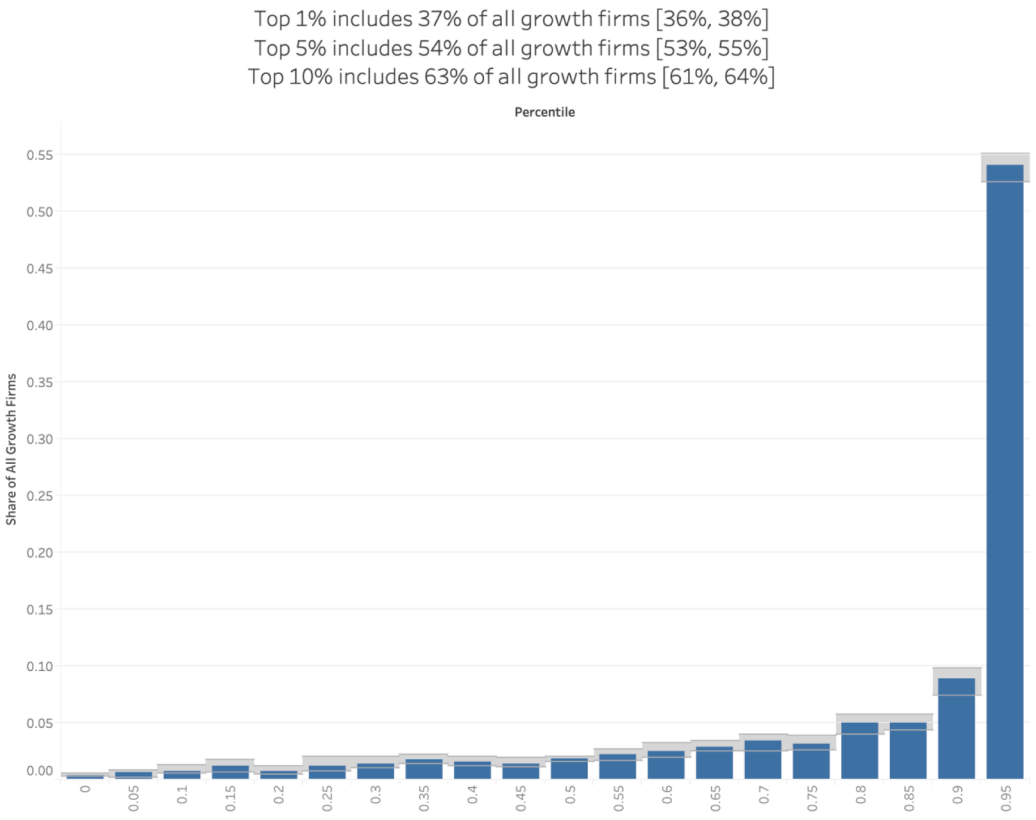


Fig. 1. Validating Entrepreneurial Quality Predictions 10-Fold Out of Sample Test (Academic Model).

Each index calculates a different quantitative measure of entrepreneurship that can be aggregated and systematically compared across entrepreneurial ecosystems. The EQI, RECPI, and REAI indexes offer quantitative measures that incorporate the *quality* of entrepreneurship. Each gives a better indication than possible under traditional methods about how skewed the distributions of growth potential and likely growth outcomes are (and whether and to what extent a greater number of small to medium-sized businesses could be expected to catalyze the same growth outcomes as a high-potential growth firm).[18] Additionally, REAI systematically quantifies the ratio of *realized* to *expected* growth events for a given cohort of new firms, providing an indication of whether the ecosystem in which a cohort of new firms is located is conducive to growth (or not). As such, these indexes offer policymakers and stakeholders a better view of whether and to what extent their regions are generating startups with high-growth potential and to what extent they are helping or hampering these firms' efforts to realize their potential after founding.

Aggregating Across Locations. Finally, we aggregate our estimates for four levels of locations—national, state, MSA, county and ZIP Code. For national and state level indexes, we aggregate all firms in our sample in each year from 1988 to 2016, while for county, MSA, and ZIP Code level indexes, we use all firms that have valid ZIP Code information to form the aggregation.[19] Rather than changing the MSA definitions through time, we stick to the 2013 MSA definitions for consistency in our time-series.

Publicly Available Datasets. We make publicly available datasets aggregated at the national, state, MSA, county and ZIP Code level, which can be downloaded in tab-delimited text files, Stata compatible files, and R compatible files. These can be found on the Startup Cartography Project Harvard Dataverse and through the Startup Cartography Project website (http://www.startupcartography.com). In each, we provide SFR, EQI, RECPI and REAI estimates by year, state-year, MSA-year, county-year, and ZIP Code – year for 49 states (all except Delaware and Washington D.C.) from 1988 to 2014, and 46 states (all except Delaware, Michigan, Illinois, South Carolina and Washington D.C.) through 2016. The Entrepreneurship_National.tab includes 6 variables and 29 observations (one per year). The Entrepreneurship_by_State.tab includes 7 variables and 1444 observations. The Entrepreneurship_by_MSA.tab includes 8 variables and 10,428 observations. The Entrepreneurship_by_County.tab includes 12 variables and 88,049 observations. The Entrepreneurship_by_ZIP_Code.tab includes 8 variables and 824,770 observations. While the U.S. Startup Map (discussed in Section VI) provides visualization down to the level of individual street addresses,

[18] The level of skewness of entrepreneurial quality is highly informative. It indicates how much more likely a startup at the high end of the entrepreneurial quality distribution is to grow than an average firm. If skewness were low, then adding several average firms could have as much regional impact as one high-growth-potential firm. But, if skewness is high (as the findings indicate), then a much larger number of firms with average growth potential is needed to generate the expected impact of one high-potential firm. Given the level of skewness observed, almost 4,000 local limited liability companies (average firm) are needed to generate the same potential as only one new Delaware corporation with an early patent and trademark. Put another way, initial ambition/potential for growth is a key dimension of heterogeneity across new firms. The subset of high-potential-growth startups is very small and fundamentally different than the vast majority of new firms

[19] Specifically for the county level index, we use the registered ZIP Code of each company to identify each county using the HUD USPS ZIP Code crosswalk files. We then aggregate across each county and year to estimate EQI, RECPI and REAI based on the observed outcomes in each one.

Table 4
Summary Statistics by Regions.

Measure	State		MSA		County		Zip Code	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Firm Quantity (SFR)	27,327.43	38,161.03	3103.40	9987.76	349.06	1865.98	42.04	100.52
Firm Quality (EQI)	0.00045	0.00040	0.00049	0.00049	0.00036	0.00050	0.00047	0.0016
Quality-adjusted Quantity (RECPI)	14.79	36.14	1.74	7.20	0.20	1.79	0.023	0.10
Equity Growth (IPO or Acquisition)	12.84	28.57	1.52	5.93	0.17	1.54	0.020	0.19
Number of Regions	50	N.A.	362	N.A.	3138	N.A.	38,264	N.A.
Observations	1444	N.A.	10,428	N.A.	88,049	N.A.	824,770	N.A.

This table represents our full dataset, of all registered firms registered within the years 1988 and 2014 in Washington D.C. and 49 US states (excluding Delaware), and 47 states (excluding Delaware, Illinois, Michigan, South Carolina) within the year 1988 and 2016. These states account for 99.6% of US GDP in 2014. MSA is Metropolitan Statistical Area. All measures defined in detail in Section III of this paper.

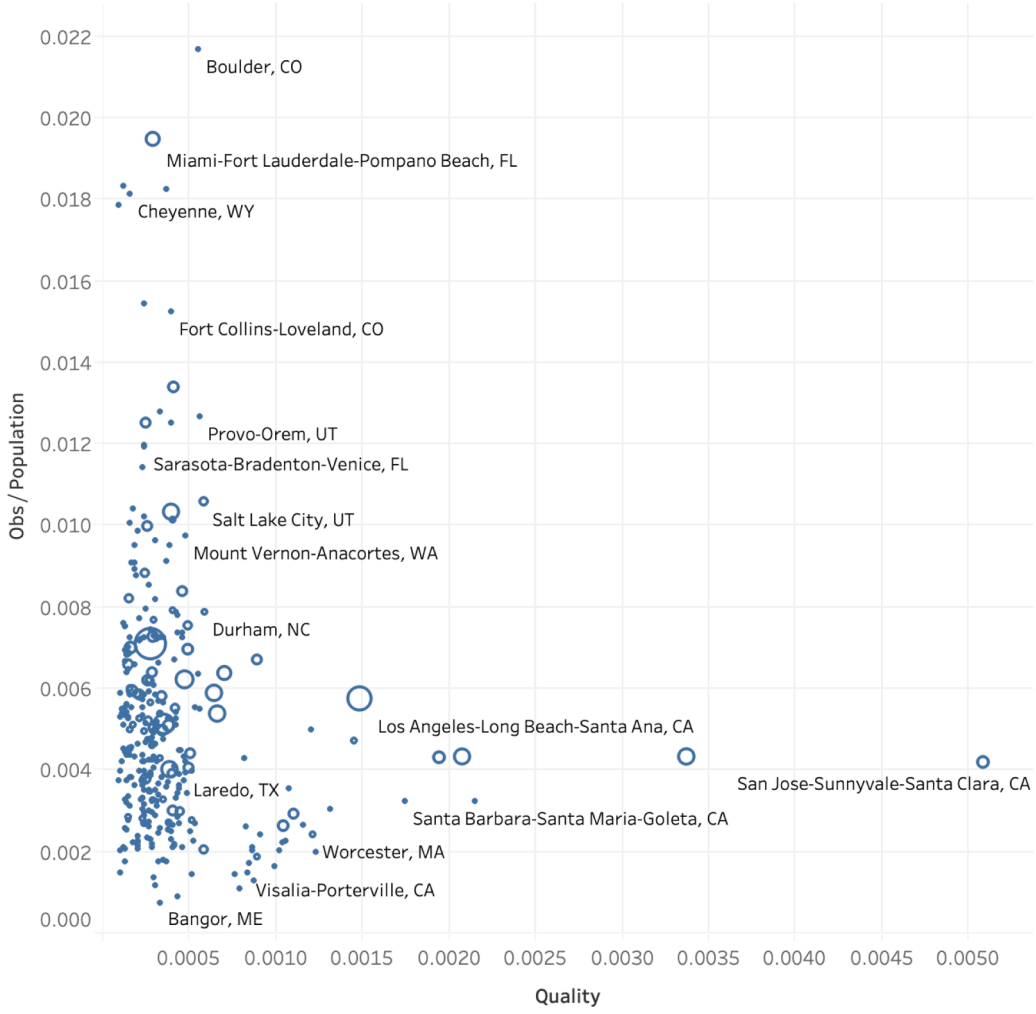


Fig. 2. Population-adjusted Quantity and Estimated Quality by MSA.

Fig. 3. . Does Quantity Predict Economic Growth? Population-adjusted Quantity and Average GDP Growth by MSA. B. Does Quality Predict Economic Growth? Entrepreneurial Quality and Average GDP Growth by MSA.

confidentiality and data release restrictions do not allow us to make the data at the level of individual firms publicly available.[20]

4. SCP ecosystem statistics: a first look

The calculation of SFR, EQI, RECPI and REAI at different levels of geographic agglomeration and across time enables researchers and policy makers to evaluate different entrepreneurial ecosystems and regional trends in ecosystem statistics. Table 4 reports summary statistics by region at the state, MSA, county and ZIP Code level. Across all region-years, there were on average 27,327 startups formed per state per year, 3103 per MSA per year, 349 per county per year and 42 startups per ZIP Code per year, respectively. On average, the growth potential of an average start-up (or EQI) was low, with the probability of a growth outcome ranging from 1 in 2128 at the ZIP Code level to 1 in 2040 at the MSA level. The average expected number of growth outcomes (RECPI) ranged from 0.20 at the county level to 14.8 at the state level. RECPI correlated closely with the actual number of growth outcomes later observed (which ranged from 0.17 at the county level to 12.84 at the state level).

In Figs. 2 and 3, we provide simple comparisons of the estimated average SFR (entrepreneurial quantity) per capita and EQI (entrepreneurial quality) by MSA, to average GDP growth rate with a 5- year-lag. In Fig. 2, we find that entrepreneurial *quantity* per capita and entrepreneurial *quality* are not highly correlated at the higher range of their respective distributions. We observe both a low SFR per capita and EQI for most MSAs (e.g., both a low number of new registered businesses and low estimated probability of achieving a growth outcome). The highest average EQI observed is in the Silicon Valley region, but the number of startups formed per capita there is in the lower range of observations. Boulder, Missoula and Miami boast among the highest range of entrepreneurial quantity (on a per capita basis), but lowest in quality. In Fig. 3A, we find a weak, but slightly positive correlation between SFR per capita and GDP growth by MSA, when fitted by Real GDP of MSAs. On the other hand, as shown in Fig. 3B, the correlation between EQI and

[20] A number of data vendors and individual states from which we sourced business registration data, including Bizapedia, Open Corporates, Arizona, Arkansas and Illinois, place restrictions on commercial use of the data and/or the redistribution of the data in its original form. Similarly, we are prohibited from redistributing Thompson Reuters SDC and Infogroup USA data. Academics interested in access to the micro-data for specific research projects may reach out to us to discuss the possibility of accessing the microdata pursuant to existing license agreements.

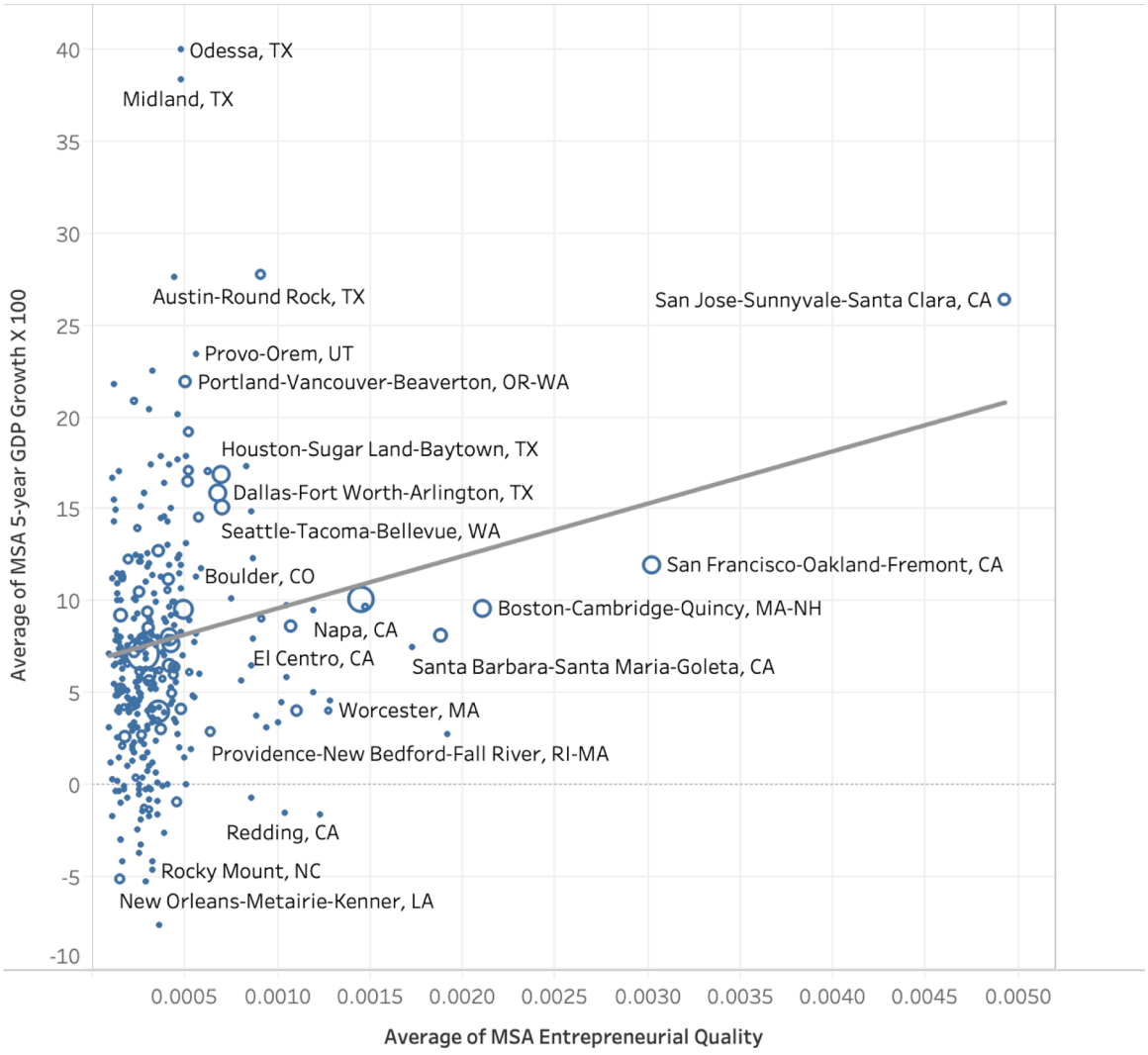


Fig. 3. (*continued*).

GDP growth is significantly positive. We observe MSAs with a higher EQI also have higher average GDP growth in the 5 year period following the firm founding.[21] While we note that these comparisons are not causal, and many unobservable factors could be contributing to these observed trends, relative to Fig. 3A, we can see that the GDP growth rate is more correlated with the quality, than the quantity, of regional entrepreneurship.

Next, in Table 5, we report the top 15 ranked MSAs across our measures for three distinct four-year periods, 1990 to 1994, 2000 to 2004, and 2010 to 2014. Table 5a is RECPI per capita. The ranking of cities we observe is consistent with common understanding of the geographic distribution of growth oriented entrepreneurial activity in the U.S. The San Jose-Sunnyvale-Santa Clara, CA MSA ranks first in all periods, and other common startup hubs such as Boston-Cambridge-Quincy, MA-NH, San Francisco-Oakland-Fremont, CA, or Austin-Round Rock, TX, are always found among the top 15 MSAs. We also observe changes across our time periods that are consistent with the evolution of the U.S. . For example, by 2010, the Boston area had moved from second to fourth place on our list, and small college towns like Madison, WI, or Bend, OR, which are present in 1990, fall off our list. On the other hand, San Francisco-Oakland-Fremont, CA, and San Diego-Carlsbad-San Marcos, CA, Austin-Round Rock, TX MSA, and the Seattle-Tacoma-Bellevue, WA, move up the rankings across our time periods. Overall, these RECPI per capita rankings show patterns consistent with our common understanding of the geography of U.S. entrepreneurship.

Table 5b considers the top 15 regions by average quality over our three time periods. While we continue to see the San Jose-Sunnyvale-Santa Clara, CA MSA rank at the top of the distribution, some smaller cities, such as Bangor, ME or Salinas, CA, are also included (perhaps because they have a small number of high-quality firms which meaningfully impacts the average estimated entrepreneurial quality of start-ups founded there). Observationally, the relationship between our measures and the common knowledge of the geography of entrepreneurship appears noisier than RECPI per capita, suggesting that there is merit to including quantity together with quality estimates in measures of regional entrepreneurship.

Finally, in Table 5c we focus on the distribution of REAI, the performance of the ecosystem conditional on the predicted number of exit events from RECPI. The types of locations we observe here appear differ substantially from the quality-based rankings, suggesting this type of ecosystem performance is distinct (and potentially uncorrelated) from

[21] Specifically, this figure considers a panel data where, for each observation, we include the average quality of firms founded in the region-year and the GDP growth over the five subsequent years. We only include years since 2001 because MSA GDP estimates are not available before 2001. We stop in 2013 to be able to observe GDP growth over the five subsequent years. The plotted point in the scatterplot is the average of each variable for all years in our data.

Table 5a
Rankings of Top Average RECPI / Resident Population by MSA.

	Top MSAs of Average RECPI / Resident Population between 1990 and 1994	Top MSAs of Average RECPI / Resident Population between 2000 and 2004	Top MSAs of Average RECPI / Resident Population between 2010 and 2014
1	San Jose-Sunnyvale-Santa Clara, CA	San Jose-Sunnyvale-Santa Clara, CA	San Jose-Sunnyvale-Santa Clara, CA
2	Boston-Cambridge-Quincy, MA-NH	San Francisco-Oakland-Fremont, CA	San Francisco-Oakland-Fremont, CA
3	Miami-Fort Lauderdale-Pompano Beach, FL	Boston-Cambridge-Quincy, MA-NH	San Diego-Carlsbad-San Marcos, CA
4	Salt Lake City, UT	Boulder, CO	Boston-Cambridge-Quincy, MA-NH
5	Bridgeport-Stamford-Norwalk, CT	San Diego-Carlsbad-San Marcos, CA	Los Angeles-Long Beach-Santa Ana, CA
6	Atlanta-Sandy Springs-Marietta, GA	Los Angeles-Long Beach-Santa Ana, CA	Boulder, CO
7	San Francisco-Oakland-Fremont, CA	Austin-Round Rock, TX	Provo-Orem, UT
8	Barnstable Town, MA	Provo-Orem, UT	Santa Cruz-Watsonville, CA
9	Los Angeles-Long Beach-Santa Ana, CA	Oxnard-Thousand Oaks-Ventura, CA	Austin-Round Rock, TX
10	Ocala, FL	Miami-Fort Lauderdale-Pompano Beach, FL	Salt Lake City, UT
11	St. George, UT	Santa Barbara-Santa Maria-Goleta, CA	Miami-Fort Lauderdale-Pompano Beach, FL
12	Provo-Orem, UT	Salt Lake City, UT	Denver-Aurora, CO
13	Madison, WI	Fort Collins-Loveland, CO	Santa Barbara-Santa Maria-Goleta, CA
14	Bend, OR	St. George, UT	Oxnard-Thousand Oaks-Ventura, CA
15	San Diego-Carlsbad-San Marcos, CA	Denver-Aurora, CO	Seattle-Tacoma-Bellevue, WA

This table represents the ranking of top 15 MSAs in average quality-adjusted quantity (RECPI) / Resident Population across three time periods. Resident population by MSA is aggregated from 1990, 2000 and 2010 U.S Census population by Zip Code (ZCTAs). We exclude regions in Arkansas and Michigan due to the lack of firm addresses at the Zip Code level in current Startup Cartography Project.

RECPI. For the period of 2010 to 2014, however, there does seem to be a pattern: the best performing regions include many fracking locations, which had a strong localized shock on economic activity.

In Fig. 4, we select five high achieving MSAs in terms of EQI and find striking differences in MSA entrepreneurial quality across MSAs and across time from 1988 to 2014. The movement of the line indicates the level of entrepreneurial quality for the region, while the line thickness indicates the average real MSA GDP within these years. Interestingly, these five MSAs all begin at similar levels in 1988 and follow a similar trajectory—starting to grow in the early 1990s and then peaking in 2000. The San Francisco Bay Area MSAs have consistent high EQI compared to other regions across years. The San Jose-Sunnyvale-Santa Clara, CA MSA has the highest entrepreneurial quality every year and nearly double its 1988 level in 2014, while the San Francisco-Oakland-Fremont, CA MSA starts from a second lowest EQI in 1988 and gradually grows to the second highest MSA of EQI. The Boston-Cambridge-Quincy, MA-NH MSA, however, starts at the second highest EQI level in 1988 but is superseded by San Francisco after 2003. It has a level of basically half that of San Francisco's EQI by 2014. On the other hand, the New York-Northern New Jersey-Long Island, NY-NJ-PA MSA—which has the largest average real MSA GDP of all, shares a similar level of EQI with the Austin-Round Rock, TX MSA both in 1988 and 2014. The EQI of Austin MSA reaches its highest level in 2007 but then drops closer to New York's level of EQI in 2008 and largely remains there.

Trends in the Effect of the US Entrepreneurial Ecosystem (REAI). Regional performance depends not simply on the founding of new high potential enterprises, but also the scaling of those enterprises to generate

Table 5b
Rankings of top average quality by MSA.

	Top MSAs of Average Quality between 1990 and 1994	Top MSAs of Average Quality between 2000 and 2004	Top MSAs of Average Quality between 2010 and 2014
1	San Jose-Sunnyvale-Santa Clara, CA	San Jose-Sunnyvale-Santa Clara, CA	San Jose-Sunnyvale-Santa Clara, CA
2	Santa Cruz-Watsonville, CA	San Francisco-Oakland-Fremont, CA	San Francisco-Oakland-Fremont, CA
3	Bangor, ME	Boston-Cambridge-Quincy, MA-NH	Santa Cruz-Watsonville, CA
4	San Francisco-Oakland-Fremont, CA	Huntsville, AL	San Diego-Carlsbad-San Marcos, CA
5	Santa Barbara-Santa Maria-Goleta, CA	San Diego-Carlsbad-San Marcos, CA	Boston-Cambridge-Quincy, MA-NH
6	Boston-Cambridge-Quincy, MA-NH	Worcester, MA	Santa Barbara-Santa Maria-Goleta, CA
7	San Diego-Carlsbad-San Marcos, CA	Santa Barbara-Santa Maria-Goleta, CA	Los Angeles-Long Beach-Santa Ana, CA
8	Oxnard-Thousand Oaks-Ventura, CA	Santa Cruz-Watsonville, CA	Santa Rosa-Petaluma, CA
9	Worcester, MA	Oxnard-Thousand Oaks-Ventura, CA	Oxnard-Thousand Oaks-Ventura, CA
10	Portland-South Portland-Biddeford, ME	Los Angeles-Long Beach-Santa Ana, CA	Salinas, CA
11	Modesto, CA	Pittsfield, MA	Napa, CA
12	Salinas, CA	Santa Rosa-Petaluma, CA	Sacramento–Arden-Arcade–Roseville, CA
13	Santa Rosa-Petaluma, CA	Napa, CA	Pittsfield, MA
14	Lewiston-Auburn, ME	Modesto, CA	Worcester, MA
15	Los Angeles-Long Beach-Santa Ana, CA	San Luis Obispo-Paso Robles, CA	Vallejo-Fairfield CA

This table represents the ranking of top 15 MSAs in average quality across three time periods. Resident population by MSA is aggregated from 1990, 2000 and 2010 U.S Census population by Zip Code (ZCTAs). We exclude regions in Arkansas and Michigan due to the lack of firm addresses at the Zip Code level in current Startup Cartography Project.

Table 5c
Rankings of Top Average REAI by MSA.

	Top MSAs of Average REAI between 1990 and 1994	Top MSAs of Average REAI between 2000 and 2004	Top MSAs of Average REAI between 2010 and 2014
1	Sioux Falls, SD	Owensboro, KY	Morgantown, WV
2	State College, PA	Bangor, ME	Hattiesburg, MS
3	Reno-Sparks, NV	Cedar Rapids, IA	Charleston-North Charleston, SC
4	Las Vegas-Paradise, NV	Alexandria, LA	Midland, TX
5	Anderson, SC	Anchorage, AK	Springfield, IL
6	Sheboygan, WI	Dothan, AL	Bismarck, ND
7	Ocala, FL	Omaha-Council Bluffs, NE-IA	Glens Falls, NY
8	St. Joseph, MO-KS	Idaho Falls, ID	Omaha-Council Bluffs, NE-IA
9	Oklahoma City, OK	Bloomington-Normal, IL	St. Joseph, MO-KS
10	Corvallis, OR	Morgantown, WV	Las Cruces, NM
11	Owensboro, KY	Columbia, MO	Alexandria, LA
12	Tuscaloosa, AL	Lafayette, LA	Odessa, TX
13	Lake Charles, LA	Rochester, NY	Pascagoula, MS
14	Dubuque, IA	Carson City, NV	Greenville-Mauldin-Easley, SC
15	Decatur, AL	St. Cloud, MN	Lincoln, NE

This table represents the ranking of top 15 MSAs in average REAI across three time periods. Resident population by MSA is aggregated from 1990, 2000 and 2010 U.S Census population by Zip Code (ZCTAs). We exclude regions in Arkansas and Michigan due to the lack of firm addresses at the Zip Code level in current Startup Cartography Project.

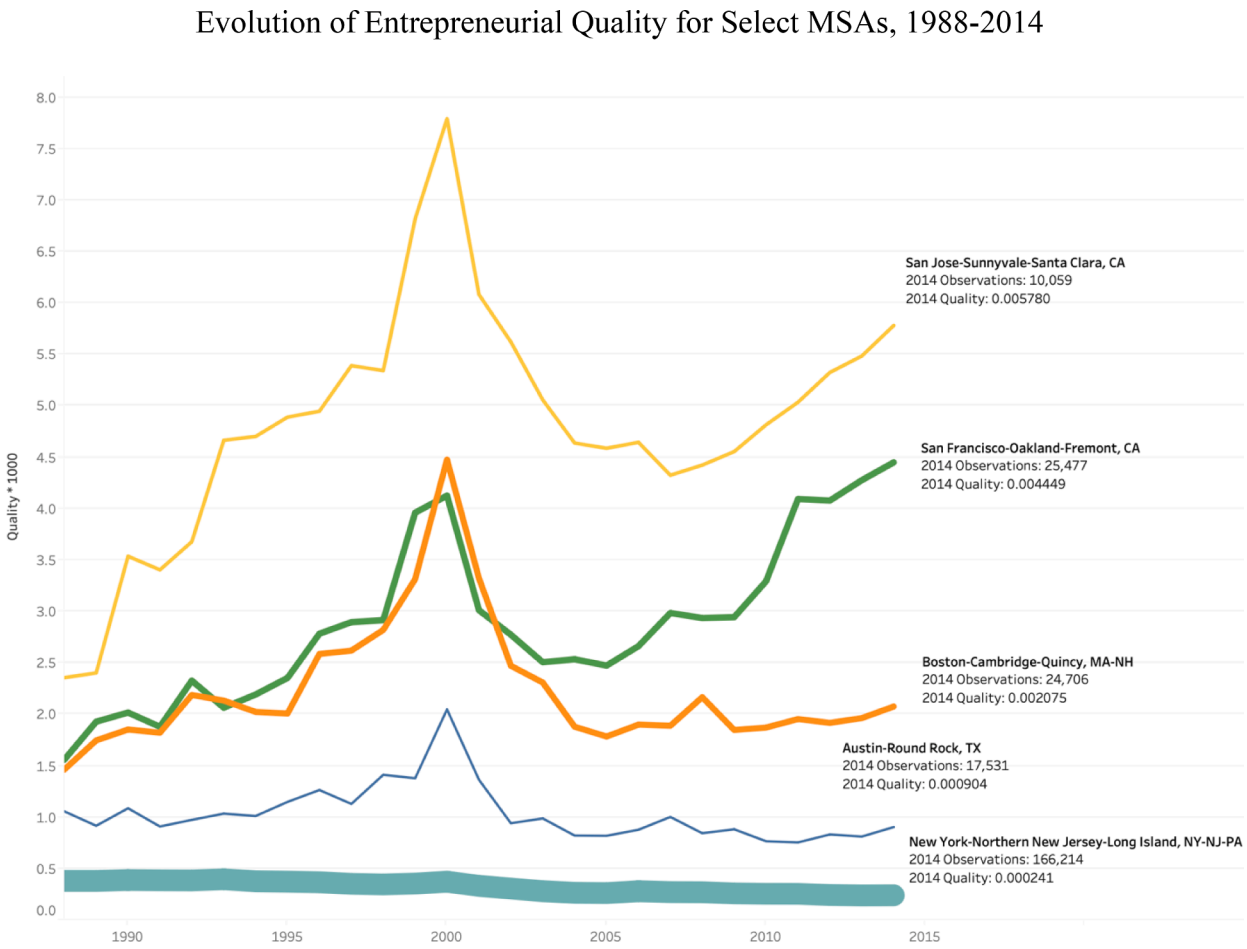


Fig. 4. Evolution of Entrepreneurial Quality for Select MSAs, 1988–2014.

employment and economic activity. In Fig. 5, we assess the performance of the US ecosystem by plotting REAI to examine "ecosystem" performance of the United States during our time period, using predicted values for the years 2011 to 2014.[22] We estimate confidence intervals by repeating our procedure on thirty bootstrap samples, and also including the maximum and minimum of each value in the graph. REAI captures the relative ability of a given start-up cohort to realize its potential, relative to the expectation for growth events as measured by RECPI. A value above 1 indicates a positive ecosystem effect, and a value under 1 indicates a negative effect. In contrast to RECPI, REAI reflects the impact of the economic and entrepreneurial environment in which a start-up cohort participates (i.e., the "ecosystem" in which it participates).

Three distinct periods stand out in this graph. The early portion of our sample saw a significant increase in REAI from a slight negative level to a peak of 1.60 for the 1995 cohort. Relative to the rest of the years that we observe, startups born in 1995 were 60% more likely to achieve an equity growth outcome conditional on their estimated quality. This peak was followed by a steady decline of REAI over the subsequent decade, and the index turns negative in the year 2001. It continues this negative decline from 2001 to 2007, with REAI moving from 0.93 down to 0.66. Putting these values together creates meaningful differences for startup cohorts: a start-up at a given estimated quality level was 2.4 times more likely to experience a growth event if it was founded in 1995 rather than in 2007. Our index then begins to recover after 2007, and turns once again positive in 2012. Though these last few years are only preliminary estimates due to the natural time-lags inherent in observing startup growth, it would appear that there is a significant increase in the performance of the US entrepreneurial ecosystem, reaching a level higher than all prior estimates by 2014.

5. The U.S. startup map

The U.S. Startup Map is an interactive visualization and map of SFR and EQI from 1988 to 2016 (in states and years where data is available). The map enables users to geographically explore SCP data and analysis in a web browser with familiar mouse click and touch gestures. It broadens the impact of the Startup Cartography measures by allowing users to better see the growth potential of entrepreneurship in their ecosystem. As mentioned earlier, based on feedback from policy users, the U.S. Startup Map uses the policy model as its basis for assessment of any given location, in order to focus on a consistent and understandable set of digital markers of start-up quality.

Assigning colors by these measures does not take advantage of all available startup characteristics leveraged in our academic model. For example, founder-firm eponymy does not impact color assignment. This is a deliberate choice to make the color palette, and the map, more accessible to all stakeholders. Earlier map iterations assigned colors according to a more complex algorithm. This inflicted a burden on users to understand the algorithm before using the map. We found this burden to be counter-productive to the map's goals.

Entrepreneurial *quantity*—SFR—is a necessary foundation for the map. But it is not sufficient. Alone, entrepreneurial *quantity* produces an

[22] Because our approach requires that we observe the *realized* growth firms we can only measure our index with a 6-year lag, thus, up to 2010. For years 2011 to 2014, we estimate our model with a varying lag of $n = 2016 - year$ and calculate RECPI using such lag.

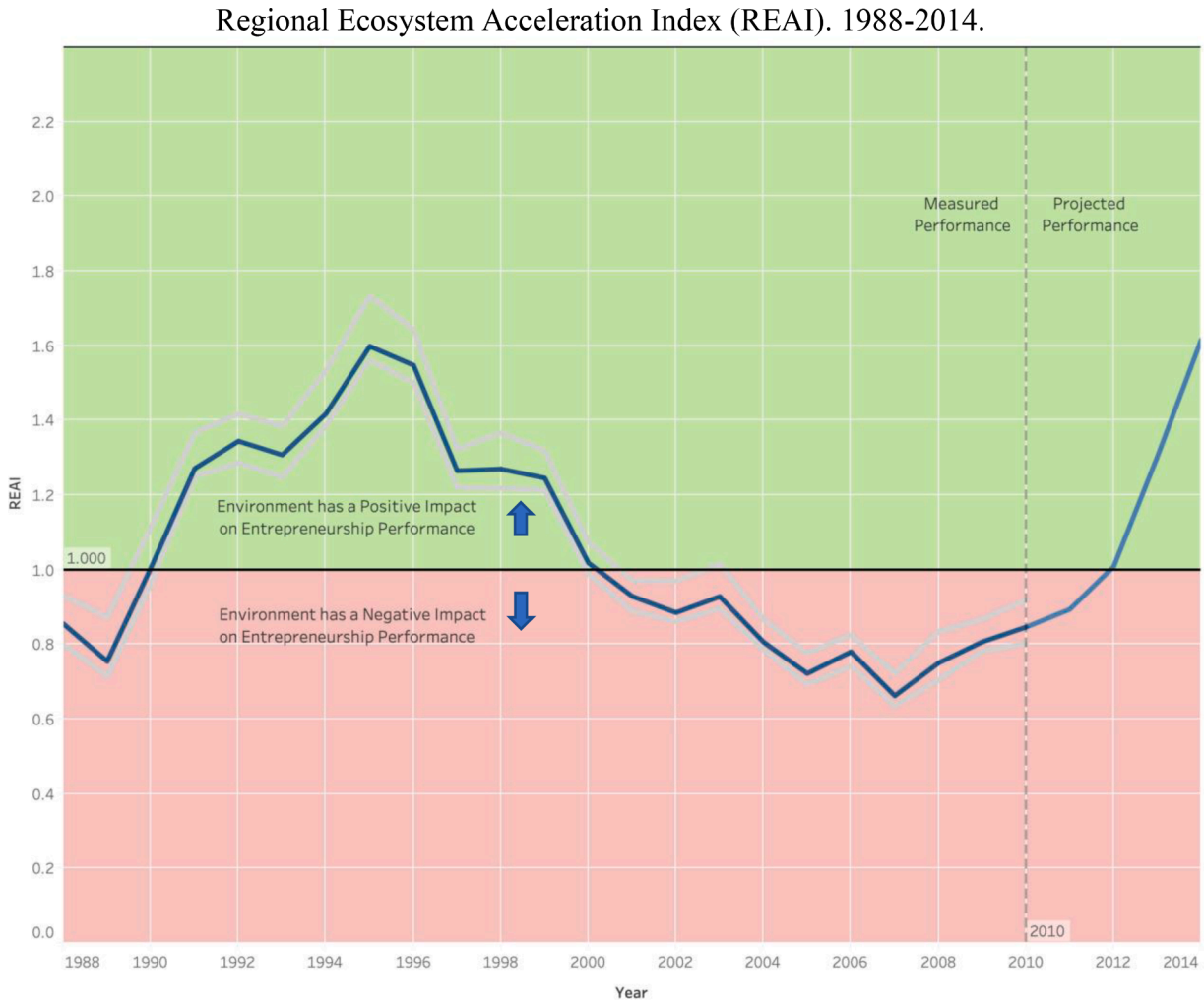


Fig. 5. Regional Ecosystem Acceleration Index (REAI). 1988–2014.

image that closely matches a general population map. For the map to bring the Startup Cartography Project data to life it must also show entrepreneurial *quality*. Panel A of Fig. 6 presents a national view of the U.S. Startup Map. The larger section on the right displays the map. Users can explore the map with zoom, pan, search, and select interactions. To the map's left is the legend section. The legend introduces the map and contains dynamic controls. These controls include a timeline scrollbar filter and an option to add various contextual data to the map. Only a single year is displayed in a single view.

The map places new business registrations at the location associated with their registration. New businesses are represented as circles. Quantity of registrations is visualized with circle (or *bubble*) size. Larger bubbles represent greater numbers of newly registered firms. The color of the circle corresponds to the quality percentile of new business registration(s) at that location. There is a direct correlation between the number of new businesses and the number of pixels in the displayed circle (i.e., circle area).

The U.S. Startup Map zoom level determines how registrations aggregate into bubbles. Panel B of Fig. 6 shows the four map zoom levels: State, Metro, City and Address. Zoomed all the way out, the contiguous United States are seen and registrations aggregate at the state level. Zoom in and businesses aggregate into metropolitan statistical areas. Zoom in further and businesses aggregate into cities. Zoom all the way in, to a neighborhood level, and businesses aggregate at individual addresses. At the neighborhood level each bubble represents an individual address. A larger address bubble often represents a large building where many businesses were registered during a given year.[23]

Selecting a jurisdiction circle reveals a pop-up *tooltip* that lists the jurisdiction (state, metro, or city), quality percentile, quantity of new business registrations, and year displayed on the map. Address-level tooltips are not displayed. Panel C of Fig. 6 provides an example of tooltip detail for Palo Alto California.

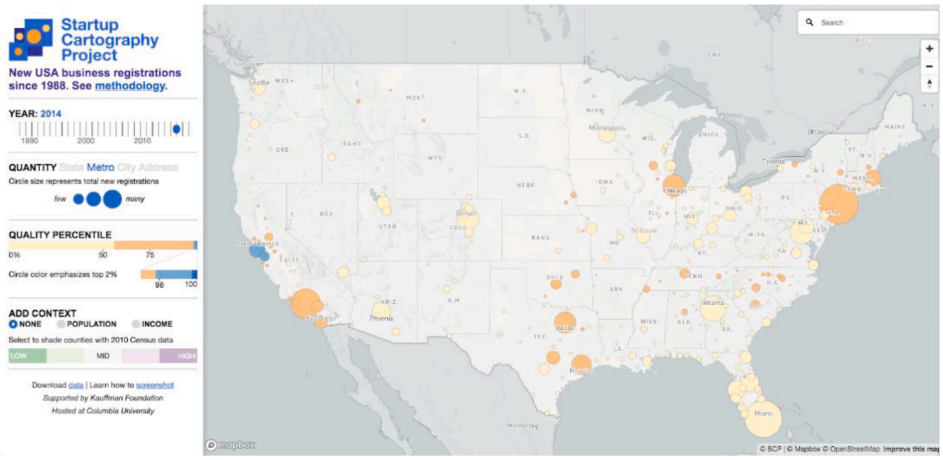
The color palette for the map is *grouped* into two quality classes. Blues are associated with high growth potential entrepreneurship. Oranges are associated with local entrepreneurship, businesses not expected to experience a growth outcome.

Each individual business registration's specific color is determined by the presence of specific or multiple impactful measures included in our policy model: LLC (pale orange) or corporation (orange), Delaware registration (pale blue), patent or trademark (blue), and a combination of at least two high quality measures (dark blue). Figure A2 shows the semantic color table for the map.

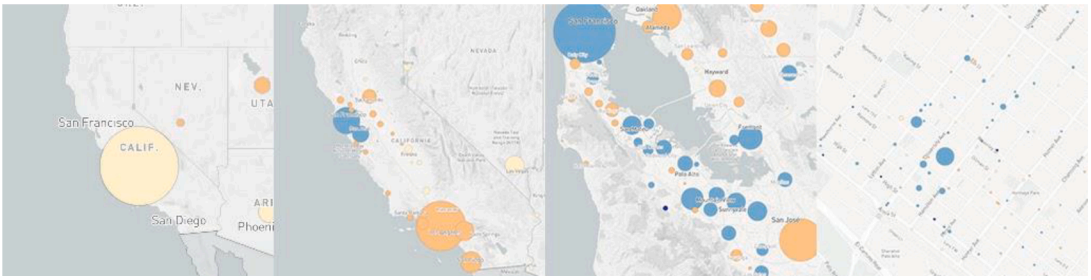
As shown in Figure A2, the map legend displays colors on a percentile scale (below), sizing color buckets according to the distribution of registrations. Color buckets are not regularly spaced across the percentile spectrum (e.g., with breakpoints at 25, 50, and 75%) because the actual portion of impactful measures is not regular. For example, the pale orange color that corresponds to the lowest 56% of the palette

[23] To learn more about our approach to encoding quantity with circle size and encoding quality with a semantic color palette see Info We Trust chapters six (Infuse Meaning) and seven (True Colors) (Andrews, 2019).

Panel A. Startup Cartography Project Map



Panel B. Map zoom levels (L-R): State, Metro, City, Address



Panel C. Tooltip Detail: Palo Alto, CA

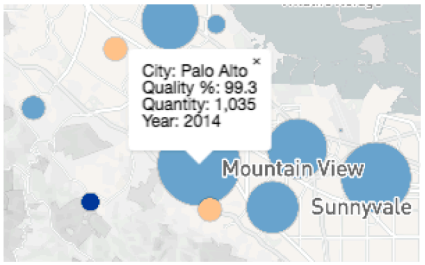


Fig. 6. *Panel A.* Startup Cartography Project Map.

directly represents the observed 56% of new business registrations that are LLCs (with no other impactful measure).

The color palette breakpoints were determined at the individual registration level. These same numerical breakpoints are extended across all aggregation levels (city, metro, state). The direct association with impactful measures (Patent, Trademark, etc.), however, does not similarly extend. Each aggregate bubble is colored by its quality percentile, relative to the rest of the nation. A pale orange city represents a city somewhere in the lower 56% of all cities, not a city only composed of LLCs.

Aggregating data, in this map's case into jurisdiction bubbles, is a necessary way of viewing a field as large as the United States. However, summary aggregation poses a risk of missing significant outliers. In this case we are concerned that an interesting cohort of high-quality businesses might go undetected, lost in a large city.

A new bubble design addresses this concern. It employs nested rings to reveal the entrepreneurial composition of individual cities. Each registration is now represented by its associated color. This ring design closes the semantic gap mentioned above between aggregate color assignments and impactful measures. Fig. 7 compares city ring views of Silicon Valley and Phoenix, both at the same zoom level.[24]

Conclusion

This paper presents the Startup Cartography Project (SCP), which offers a new set of entrepreneurial ecosystem statistics (SFR, EQI, RECPI and REAI) for the United States from 1988 to 2016. The SCP includes both a public-access dataset at the state, MSA, county, and zip code level, as well as an interactive map, the U.S. Startup Map, that permits academic and policy users to assess entrepreneurial ecosystems at an

[24] Real world testing was an important part of the map development process. We used a design thinking approach where we developed solutions for visualization through observation of users and product iteration. As cartographers familiar with the data, it is important not be blinded by our own knowledge. Real users showed us what was confusing and which aspects of the map did not work as expected. For example, the city rings now a focal point of our design were arrived at through observing stakeholders interacting with and questioning the map.

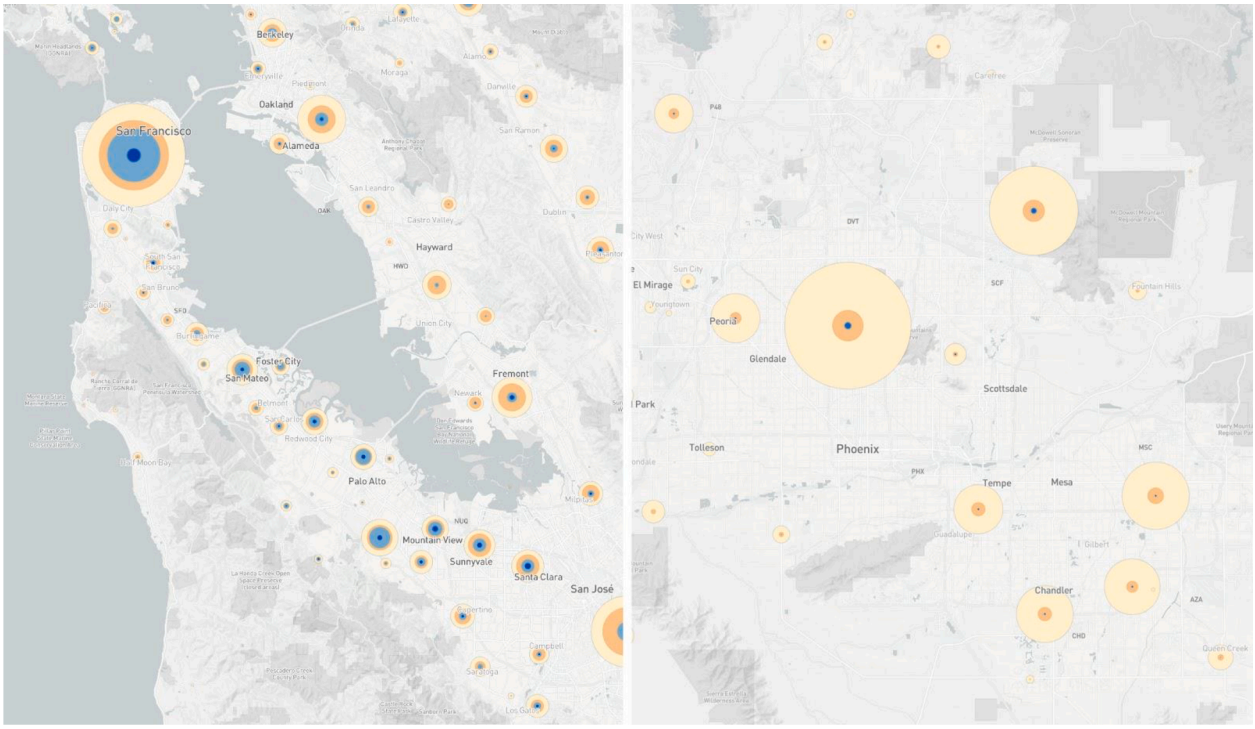


Fig. 7. Two City Ring Views at Same Zoom Level: Silicon Valley vs. Phoenix, AZ.

arbitrary level of granularity (from the level of states down to individual street addresses). The SCP's consistent, transparent and accessible data and visualization inform debate around the design and structure of policies intended to promote regional "entrepreneurial ecosystems" (Feldman and Francis, 2004; Lerner, 2009; Audrestch and Lehmann, 2005; Stam, 2015) by addressing the issues that make systematic measurement challenging and enabling evaluation on a granular (as well as aggregated) and timely (as well as retrospective) basis. By estimating the growth potential (or entrepreneurial quality) of startups at or near the time of founding, SCP indexes provide a view of the skew of entrepreneurship most correlated with later regional economic growth. SCP indexes enable the assessment of entrepreneurial potential prior to the emergence of outcomes through predictive analytics (and the study of impact without selecting on desired outcomes). They permit evaluation of entrepreneurial ecosystems at multiple levels of geographic analysis, empowering academics and policymakers to consider the power of place in novel ways. Taken together, our quality-oriented approach can yield significant insights in both research and policy, and has set the stage for a more nuanced and comprehensive understanding of entrepreneurial ecosystems and the role that entrepreneurship policy plays in economic development and regional resilience.

Research insights. Relative to quantity and outcome-based measures, research leveraging the SCP's quality-oriented approach has already offered significant insight into both entrepreneurship and entrepreneurial ecosystems at the macro, regional and firm level. At a macro-level, Guzman and Stern (2020) document that, relative to the secular decline in business dynamism emphasized in Decker et al. (2014), quality-adjusted quantity (RECPI/GDP) has actually experienced a more cyclical pattern over the past thirty years, with the level of start-up potential in 2015 near the potential observed during the 1990s. Fazio et al. (2017) further emphasize the key importance and variation of the U.S. entrepreneurial ecosystem in allowing firms to scale. Consistent with other research on the size distribution of firms across economies (Hsieh and Klenow, 2014), these results indicate that it is not enough for the United States (or regional ecosystems for that matter) to produce high potential firms, the United States must also foster an environment that allows them to grow.

Moving to the regional level, our simple correlations of quality and GDP growth document the critical role that entrepreneurial quality (though not quantity) plays in predicting economic performance and presents a significant opportunity for follow-on work. Recent research has begun to use SCP measures to study the impact of R&D tax policy on new firm formation (Fazio et al., 2019), the role of universities and other knowledge intensive institutions in shaping local entrepreneurship (Tartari and Stern, 2019), the returns to entrepreneurial migration (Guzman, 2018), and the consequences of regional entrepreneurship for local economic inequality (Marinoni and Voorheis, 2019). This research appears to us only the beginning of a rich avenue of further inquiry, and we hope the release of this public dataset facilitates further research along these lines.

At a firm level, a quality-oriented approach lends itself to answering some of the main questions in entrepreneurship research by allowing progress in the separation of entrepreneurial quality from the process of selection. Existing work has focused on understanding how the process of selection shapes the gender gap in entrepreneurship (Guzman and Kacperczyk, 2019), and the benefits of investment by venture capitalists (Catalini et al., 2019a, 2019b).

Policy and practitioner applications. With our streamlined policy model and sharper focus on measures that more concretely differentiate between the startup formation of local and high-growth potential firms, the SCP provides stakeholders with a much clearer view of the potential and trajectory of startup formation in their respective ecosystems. Given the possibility that entrepreneurial quality is a leading indicator for other outcomes in regional performance, tracking EQI, for example, would allow government analysts to measure and support entrepreneurial quality, and to observe entrepreneurial dynamics in a more proactive and informed way. Not simply a tool for direct measurement, our methodology allows government organizations (e.g., the Small Business Administration) to design and evaluate interventions that focus on the quality of entrepreneurship rather than only increasing rates of firm formation, thus facilitating an approach that could potentially increase the impact of entrepreneurship interventions.

The U.S. Startup Map will assist policy makers in forming consensus with ecosystem stakeholders in evaluating and designing

entrepreneurship initiatives. By dynamically visualizing entrepreneurial quality, the map "sets the table" for policy makers and other stakeholders to reach a shared assessment of the as-is state of their entrepreneurial ecosystem (including latent opportunities and bottlenecks). The SCP reflects, in part, the outgrowth of work started in the context of the MIT Regional Entrepreneurship Acceleration Program, which has now worked with more than 50 high-level regional stakeholder teams around the world on identifying and implementing programs to enhance entrepreneurial ecosystems; the SCP has the potential to allow policymakers and practitioners come to a clearer shared understanding of their ecosystem, and therefore pursue policies and programs that are more likely to accelerate ecosystems and their impact on regional economic and social progress.

The U.S. Startup Map is valuable in many ways. It attracts attention as a salient introduction to the Project. Everyone arrives to the map with a wealth of geographic knowledge and is usually delighted to see how a new layer of information intersects with their ready understanding. It engages users at all levels with millions of data observations and sophisticated empirical analysis. There is simply something wonderful about playing with novel data on a map. On first encounter a new user is likely to investigate what the data look like around their hometown or current office. From there, user engagement often proceeds to visually testing little hypotheses, eager to see if reality matches expectations.

The U.S. Startup Map also serves as a crude (and intuitive) validator of data, displaying that the data exists and it is rich. Seeing tens of millions of business registrations dynamically snap into formation according to your interaction impresses the eye. Likewise, the map is a crude quality check on the data system. Unlike a lonely typo in a book, a single map bubble out of place (e.g., an address in the middle of a body of water) casts doubt upon the entire project.

At its best, the U.S. Startup Map acts as a central object of discussion between anyone engaged with the project. As a shared artifact, it provides a common ground for people with different models of understanding to come together, discuss, and imagine a better vision together. In this sense, the map is a coordination mechanism that fosters discussion about local and high growth firms. In addition to facilitating discussion, the map also creates opportunity for insight discovery. Comparisons are possible crosstown and across regions. Patterns can be detected, especially across time as one scrolls through the years and sees the entrepreneurial activity of a location change. These insights are most powerful at the intersection of the map's display and the local knowledge of an interested stakeholder. Their special context and vested interest bring the map's data to life.

We believe the opportunities for the SCP and the U.S. Startup Map to help advance research and policy understand and improve entrepreneurship ecosystems are just emerging. Our approach highlights the significant potential of business registration records, a data source that has been used sparingly and only in an aggregated form by economists. We encourage further efforts by states to collect somewhat more granular information about the objectives of an enterprise (e.g., industry codes or founder addresses) in connection with business registration and to make business registration records more easily accessible. The lack of standardization and the uneven level and scope of digitization of business registration records across states remains a significant barrier to scaling business registration analysis across the entire United States. We look forward to further use of SCP measures and the U.S. Startup Map by researchers, policymakers and other stakeholders and the insights that work will bring in seeding and scaling entrepreneurship ecosystems and fostering the growth of regional economies.

Declaration of Competing Interest

The author declares that he has no relevant or material financial interests that relate to the research described in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.respol.2021.104437.

References

Allison, Paul. (2012) "Logistic regression for rare events." *Statistical horizons*. Available at: https://statisticalhorizons.com/logistic-regression-for-rare-events (Accessed on October 1, 2021).

Andrews, R.J., 2019. Info We Trust: How to Inspire the World with Data. Wiley.

Audretsch, D.B., Lehmann, E.E, 2005. Does the knowledge spillover theory of entrepreneurship hold for regions? Res. Policy 34 (8), 1191–1202. https://doi.org/10.1016/j.respol.2005.03.012.

Barnes, Beau, Harp, Nancy, Oler, Derek. (2014) "Evaluating the SDC mergers and acquisitions database" SSRN Working Paper #2201743.

Belenzon, Sharon, Chatterji, Aaron, Daley, Brendan, 2017. Eponymous entrepreneurs. Am. Econ. Rev. 107 (6), 1638–1655.

Belenzon, Sharon, Chatterji, Aaron, Daley, Brendan, 2019. Choosing between growth and glory. Manag. Sci. Articles Adv. 1–25.

Catalini, Christian, Jorge Guzman, and Scott Stern. 2019. "Hidden in plain sight: venture growth with and without venture capital". Working Paper.

Catalini, Christian, Jorge Guzman, and Scott Stern. 2019. "Passive vs active growth: evidence from founder choices and venture capital investment". Working Paper.

Churchwell, C. 2016. *SDC: M&A database.* Baker library-fast answers.

Decker, Ryan, Haltiwanger, John, Jarmin, Ron, Miranda, Javier, 2014. The role of entrepreneurship in US job creation and economic dynamism. J. Econ. Perspect. 28 (3), 3–24.

Delgado, Mercedes, Porter, Michael, Stern, Scott, 2016. Defining clusters in related industries. J. Econ. Geogr. 16 (1), 1–38.

EDA. 2010. "Regional innovation strategies". Economic development agency. Available at: https://www.eda.gov/oie/ris/ (Accessed on January 21, 2020).

Fazio, Catherine, Guzman, Jorge, Murray, Fiona, Stern, Scott, 2017. MIT Innovation Initiative Policy Report.

Fazio, Catherine, Jorge Guzman and Scott Stern. 2019. "The impact of state-level R&D tax credits on the quantity and quality of entrepreneurship". NBER Working Paper #26099.

Feld, Brad., 2012. Startup Communities: Building an Entrepreneurial Ecosystem in Your City. Wiley, p. 224.

Feldman, Maryann., 2001. The entrepreneurial event revisited: firm formation in a regional context. Ind. Corp. Change 10 (4), 861–891.

Feldman, Maryann, Francis, Johanna, 2004. Homegrown solutions: fostering cluster formation. Econ. Dev. Q. 18 (2), 127–137.

Feldman, Maryann, Francis, Johanna, Bercovitz, Janet, 2005. Creating a cluster while building a firm: entrepreneurs and the formation of industrial clusters. Reg. Stud. 39 (1), 129–141.

Glaeser, Edward, Kerr, Sari Pekkala, Kerr, William, 2015. Entrepreneurship and urban growth: an empirical assessment with historical mines. In: The Review of Economics and Statistics, 2. MIT Press, pp. 498–520.

Guzman, Jorge. 2018. "Go west young firm: agglomeration and embeddedness in startup migrations to Silicon Valley". SSRN Working Paper #3175328.

Guzman, Jorge, Kacperczyk, Olenka, 2019. Gender gap in entrepreneurship. Res. Policy 48 (7), 1666–1680.

Guzman, Jorge, Stern, Scott, 2015. Where is Silicon Valley? Science 347. Issue #6222.

Guzman, Jorge, Stern, Scott, 2017. Nowcasting and placecasting entrepreneurial quality and performance. In: Haltiwanger, John, Hurst, Erik, Miranda, Javier, Schoar, Antoinette (Eds.), Measuring Entrepreneurial Businesses: Current Knowledge and Challenges, pp. 63–109.

Guzman, Jorge, Stern, Scott, 2020. "The state of American entrepreneurship: evidence of the quantity and quality of entrepreneurship for 32 US States, 1988-2014″. Am. Econ. J.: Econ. Policy. Forthcoming.

Hsieh, Chang-Tai, Klenow, Peter J, 2014. The life cycle of plants in India and Mexico. Q. J. Econ. 129 (3), 1035–1084.

Jarmin, Ron, and Javier Miranda. 2002. "The longitudinal business database." SSRN Working Paper #2128793.

Kauffman Foundation, 2019a. 2018 state report on early-stage entrepreneurship. Kauffman Foundation – Kauffman Indic. Entrep.

Kauffman Foundation. 2019. "ESHIP communities". Available at: https://www.kauffman.org/what-we-do/entrepreneurship/entrepreneurial-communities (Accessed on January 21, 2020).

King, Gary, Zeng, Langche., 2001. Logistic regression in rare events data. Polit. Anal. 9, 137–163.

Klapper, Leora, Amit, Raphael, Guillen, Mauro. (2010) "Entrepreneurship and Firm Formation Across Countries". NBER Volume: International Differences in Entrepreneurship. Edited by Josh Lerner and Antoinette Schoar.

Lerner, Josh., 2009. Boulevard of Broken Dreams: Why Public Efforts to Boost Entrepreneurship and Venture Capital Have Failed—And What to Do about It. Princeton University Press, p. 240.

Marinoni, Astrid, and John Voorheis. 2019. "Who gains from creative destruction? evidence from high-quality entrepreneurship in the United States". Working Paper.

Murray, Fiona, Stern, Scott, 2015. Linking and leveraging. Science 348 (6240), 1203.

Netter, Jeffry, Stegemoller, Mike, Wintoki, M Babajide, 2011. "Implications of data screens on mergers and acquisitions analysis: a large sample study of mergers and acquisitions from 1992 to 2009″. Rev. Financ. Stud. 24 (7), 2316–2357.

Saxenian, AnnaLee., 1994. Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University Press.
Stam, Erik., 2015. Entrepreneurial ecosystems and regional policy: a sympathetic critique. Euro. Plann. Stud. 23 (15), 1759–1769.
Stangler, Dane, and Bell-Materson. 2015. “Measuring entrepreneurial ecosystems”. Report.