

The State of American Entrepreneurship:

New Estimates of the Quantity and Quality of Entrepreneurship for 12 US States, 1988-2014[1]

Jorge Guzman, MIT

Scott Stern, MIT and NBER

MIT Innovation Initiative

Laboratory for Innovation Science and Policy

Preliminary and Incomplete!

Please do not Cite or Quote without Permission!

October 2015

[1] Acknowledgments: An earlier version of this paper, titled "Measuring Growth Entrepreneurship and the Performance of Entrepreneurial Ecosystems: Evidence from Five US States, 1988-2014" was prepared for the Kauffman Foundation New Entrepreneurial Growth Conference, June 2015. We are thankful to Erik Brynjolffson, Catherine Fazio, and Fiona Murray for comments and suggestions, to Open Corporates for providing data for New York and Michigan, and to RJ Andrews for his development of the visualization approach employed here. We also acknowledge and thank the Jean Hammond (1986) and Michael Krasner (1974) Entrepreneurship Fund and the Edward B. Roberts (1957) Entrepreneurship Fund at MIT and the Kauffman Foundation for financial support. All errors and omissions are of course our own.

© 2015 by Jorge Guzman and Scott Stern. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit is given to the source.

The State of American Entrepreneurship:
New Estimates of the Quantity and Quality of Entrepreneurship for 12 US States, 1988-2014

ABSTRACT

While official measures of business dynamism have seen a long-term decline, early-stage venture financing of new companies has reached levels not observed since the late 1990s, resulting in a sharp debate about the state of American entrepreneurship. Building on Guzman and Stern (2015a; 2015b), this paper offers new evidence to inform this debate by estimating measures of entrepreneurial quality based on predictive analytics and comprehensive business registries. Our estimates suggest that the probability of a significant growth outcome (either an IPO or high-value acquisition) is highly skewed and predicted by observables at or near the time of business registration: more than 80% of realized growth events are in the top 5% of our estimated growth distribution. This high level of skewness, consistent with a power law distribution for initial entrepreneurial potential, motivates the development of three new economic statistics that simultaneously account for both the quantity as well as the quality of entrepreneurship: the Entrepreneurial Quality Index (EQI, measuring the average quality level among a group of start-ups within a given cohort), the Regional Entrepreneurship Cohort Potential Index (RECPI, measuring the growth potential of firms founded within a given region and time period) and the Regional Entrepreneurship Acceleration Index (REAI, measuring the performance of a region over time in realizing the potential of firms founded there). We establish several new findings about the history and state of US entrepreneurship by examining these statistics for 12 US states (covering 47% of the overall economy) from 1988 through 2014. First, in contrast the secular decline in the aggregate quantity of entrepreneurship observed in series such as the Business Dynamic Statistics (BDS), the growth potential of start-up companies (RECPI relative to GDP) has followed a cyclical pattern that seems sensitive to the capital market environment and overall economic conditions. Second, while the peak value of RECPI is recorded in 2000, there is a sharp upward swing beginning in 2010 and, even after controlling for change in the overall size of the economy, the second highest level of entrepreneurial growth potential is registered in 2014. Third, the likelihood of start-up firms for a given quality level to realize their potential (REAI) declined sharply in the late 1990s, and did not recover through 2008. Importantly, state level RECPI strongly predicts future GDP growth, while the relationship between quantity-based measures is noisy. Our findings offer an empirical clarification of the state of American entrepreneurship: once one accounts for the quality of those firms that are founded, there is consistent evidence for a start-up boom over the last half-decade, though whether those firms grow and scale remains in doubt. Moreover, it is important to emphasize the striking variation across regions in entrepreneurial potential: Silicon Valley RECPI is now at an all-time high while Miami has yet to experience a meaningful start-up boom.

I. Introduction

> "The argument in favor of concern [over a potential tech bubble] is cyclical. The counterargument is that stuff works now. In 2000, you had fifty million people on the Internet, and the number of smartphones was zero. Today, you have three billion Internet users and two billion smartphones. It's Pong versus Nintendo."
>
> Marc Andreessen.
> "Tomorrow's Advance Man". *The New Yorker*. May, 2015.

While there is a long literature emphasizing the importance of entrepreneurship for economic growth, recent measures of entrepreneurship seem to accord more closely with economic and social progress. The pioneering studies of Haltiwanger and co-authors (Davis et al, 1996; Haltiwanger et al, 2013; Decker et al, 2014) moved attention away from simply counting the density of *small and medium sized firms* towards the measurement of the prevalence (and growth dynamics) of *young firms* (i.e., start-ups). Studying the growth patterns of firm age-groups within the same year-region-industry, these authors identified a systematic relationship between new firms and job creation: virtually all net job creation occurs in young firms. Other studies focusing on entrepreneurship institutions (e.g. venture capital) reinforce the relationship between entrepreneurship and economic performance: for example, Kortum and Lerner (2000) find that venture capital is associated with higher levels of innovation, and Samila and Sorenson (2011) find a robust positive effect of venture capital on aggregate income, employment, and rates of new establishments.

Despite these advances, a sharp divide has emerged between systematic population-level indices of entrepreneurial activity (such as the Business Dynamics Statistics database, hereafter BDS) and measures based on the financing and activities of start-up firms. On the one hand, Hathaway and Litan (2014a; 2014b; 2014c) use the BDS to document a secular decline in the rate of business dynamism and the "aging" of US private sector establishments, a theme echoed in the

job growth dynamics emphasized in Decker, et al (2014) along a muted stagnation in business founding rates reflected in the Kauffman Index of Entrepreneurial Activity through 2014. On the other hand, entrepreneurs, risk capital investors, and the media have highlighted what is often characterized as an "explosion" of start-up entrepreneurship, particularly by young people (Inc., 2015). Interest in entrepreneurial education at elite business and engineering schools seems to have risen dramatically in recent years, and there is even concern among practitioners and educators that the pendulum has swung too far and that we may be in the midst of an entrepreneurial "bubble." The rapid rise in early-stage financing, including venture capital as well as increased levels of angel financing and the growth of crowdfunding have only reinforced the impression that growth entrepreneurship is experiencing a very noticeable (and sustained) boom. As noted by renowned venture capitalist Tim Draper, "In the US, we are in the early days of a bubble. We have 3 to 4 good years to go." (UPSTART, 2014).

To put these differences in perspective, it is useful to consider the historical gap between these divergent views. In Figure 1, we compare, for the 12 US states that we will include in the main analysis the rate of firm births per year as measured by the Business Dynamics Statistics to the rate (relative to GDP) of successful growth firms founded in a particular year (i.e., the number of firms founded in a given year that achieved an IPO or significant acquisition within six years of initial business registration).[2] While the BDS shows a slow and steady decline of more than 50% (consistent with Hathaway and Litan), the realization of growth experienced a much sharper up-and-down cycle, with 1996 representing the most successful start-up cohort in US history,

[2] We focus on the experience of twelve US states (Alaska, California, Florida, Georgia, Massachusetts, Michigan, New York, Oregon, Texas, Vermont, Washington, and Wyoming, representing ~ 50% of the US economy) since these states will form the basis for the empirical analysis in the remainder of this paper. The broad aggregate patterns in Figures 1 and2 are qualitatively similar for the entire United States.

followed by a relatively stable level from 2001 to 2008. The difference is even more remarkable if we compare BDS firm births and economic growth. In Figure 2 we compare the rate of BDS firm births per year to GDP growth in the five years following the observation year. GDP growth also exhibits a sharp up-and-down pattern, with the period starting in 1995 (i.e. 1995 to 2000) representing the highest GDP growth in any five-year period in our sample.

Importantly, there seems to be a divergence between the steady decline in entrepreneurship and the more cyclical nature of high-value entrepreneurial exits and aggregate economic performance. Though analyses using Census data conclude that, within the same year-region-industry, young firms produce most employment growth (Haltiwanger, Jarmin and Miranda, 2013), yearly fluctuations in aggregate counts of firms births (from the same Census data) appear to hold little relationship to measures of economic performance.

How can we resolve this puzzle? Building on Guzman and Stern (2015a; 2015b), this paper breaks through this impasse by focusing not only on the quantity of entrepreneurship nor on highly selective measures of the rate of successful entrepreneurs but instead focus on the role of entrepreneurial "quality." While it has long been known that the growth consequences of start-up activity are concentrated in the outcomes associated with a very small fraction of the most successful firms (Cochrane, 2005; Kerr, Nanda, and Rhodes-Kropf, 2014), prior attempts to use population-level data to characterize the rate of entrepreneurship have largely abstracted away from initial differences across firms in the ambitions of their founders or their inherent growth potential. As emphasized by Hathaway and Litan, the challenge in directly incorporating heterogeneity is a fundamental measurement problem: "The problem is that it is very difficult, if

not impossible, to know at the time of founding whether or not firms are likely to survive and/or grow." (Hathaway and Litan, 2014b).

In our earlier work (Guzman and Stern, 2015a; 2015b), we responded to this challenge by developing a new method for the estimation of entrepreneurial quality that allows us to characterize regional clusters of entrepreneurship at an arbitrary level of granularity (placecasting), and examine the dynamics of entrepreneurial quality over time on a near real-time basis (nowcasting). Our approach combines three interrelated insights. First, a practical requirement for any growth-oriented entrepreneur is business registration (as a corporation, partnership, or limited liability company). These public documents allow us to observe a "population" sample of entrepreneurs observed at a similar (and foundational) stage of the entrepreneurial process (in this paper, from twelve US states comprising ~ 48% of total US economic activity over a 25-year period). Second, moving beyond simple counts of business registrants (Klapper, Amit, and Guillen, 2010), we are able to measure characteristics related to entrepreneurial quality *at or close to the time of registration.* These characteristics include how the firm is organized (e.g., as a corporation, partnership, or LLC, and whether the company is registered in Delaware), how it is named (e.g., whether the owners name the firm eponymously after themselves), and other early-stage milestones that might be a "digital signature" of a high-potential firm (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. This mapping allows us to form an 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).

We use this predictive analytics approach to propose three new statistics for the measurement of growth entrepreneurship: the Entrepreneurship Quality Index (EQI), the Regional Entrepreneurship Cohort Potential Index (RECPI), and the Regional Entrepreneurial Acceleration Index (REAI). EQI is a measure of *average quality* within any given group of firms, and allows for the calculation of the probability of a growth outcome for a firm within a specified population of start-ups. RECPI multiples EQI and the number of start-ups within a given geographical region (e.g., from a zip code or town to the entire five-state coverage of our sample). 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. As such, we will use RECPI (or RECPI / GDP) as our primary measure of the potential for growth entrepreneurship for a given start-up cohort. REAI, on the other hand, measures the ratio between the realized number of growth events for a given start-up cohort and the expected number of growth events for that cohort (i.e.., RECPI). REAI offers a measure of whether the "ecosystem" in which a start-up grows is conducive to growth (or not), and allows variation in ecosystem performance across time and at an arbitrary level of geographic granularity.

We calculate these measures on an annual basis for the twelve states included in our sample for the period from 1988-2014, documenting several key findings.[3] First, in contrast to the secular and steady decline observed in the BDS, RECPI has followed a cyclical pattern that seems sensitive to the capital market environment and overall economic conditions. Second, while the peak value of RECPI is recorded in 2000, there is a sharp upward swing beginning in 2010 and, even after

[3] We use a "nowcasting" index for the most recent cohorts which only use start-up characteristics available within the business registration data, and compare that index to an "enriched" index which captures events that might occur early within the life of a start-up such as the initial receipt of intellectual property

controlling for change in the overall size of the economy, the third highest level of entrepreneurial growth potential is registered in 2014. It is useful to note that the peak value of RECPI during the late 1990s is several years after the most "successful" cohort (1996), suggesting that the "financial guillotine" unleashed after the dot-com crash may have had the consequence of significantly reducing the ability of high-quality firms to realize their growth potential. Indeed, there is striking variation over time in the likelihood of start-up firms for a given quality level to realize their potential (REAI): REAI declined sharply in the late 1990s, and did not recover through 2008. Whether the most recent cohorts are able to realize their potential at rates similar to those achieved during the mid-1990s is yet to be seen. Relative to quantity-based measures from the BDS, regional variation in RECPI appears to hold a much stronger relationship to economic growth. In a yearly panel of twelve states, we find RECPI to successfully predict economic growth five and ten years later, while both BDS firm births and dynamism do not. Finally, there is striking variation across regions (and over time) in entrepreneurial potential. Consistent with Guzman and Stern (2015a), we document an extremely high and persistent level of entrepreneurial quality in the region traditionally associated with Silicon Valley (Santa Clara and San Mateo County) and Route 128, while we are able to observe a striking rise in the two urban cores closest to these suburban hotspots (San Francisco and Cambridge/Boston). At the same time, we also observe regions where the "start-up nation" has yet to take off: despite one of the highest levels of self-employment per capita in the nation, the Miami MSA has a low and declining level of RECPI.

Before turning to more general interpretations, we emphasize that our approach, though promising, does come with important limitations and caveats. First, and most importantly, we strongly caution against a causal interpretation of the regressors we employ for our predictive analytics – while factors such as eponymy and business registration form are a "digital signature"

that allows us to differentiate among firms in the aggregate, these are not meant to be interpreted as causal factors that lead to growth per se (i.e., simply registering your firm in Delaware is not going to directly enhance an individual firm's underlying growth potential). And, while we are encouraged by the robustness of our core approach across multiple states and time periods, we can easily imagine (and are actively working on identifying) additional firm-level measures (such as founder characteristics) which might allow for even more differentiation in quality, or accounting directly for changing patterns over time and space in the "drivers" of growth. Finally, while we focus here on equity growth outcomes, we do not provide any direct measure of the potential of firms in terms of employment growth (while these are likely highly correlated, it may be the case that a much more diverse range of start-ups contributes to employment growth relative to the highly skewed nature of equity growth outcomes).

Keeping in mind these caveats, our findings nonetheless do offer a new perspective on the state of American entrepreneurship. Most importantly, our results highlight that the recent shift in attention towards young firms (pioneered by Haltiwanger and co-authors) is enriched by directly accounting for initial heterogeneity among new firms. Even within the same industry, there is significant heterogeneity among new firms in their ambition and inherent potential for growth. Policies that implicitly treat all firms as equally likely candidates for growth are likely to expect "too much" from the vast majority of firms with relatively low growth potential, and might be focusing on a lever that is only weakly related to the economic growth they often seek. Second, the striking decline in REAI after the boom period of the 1990s is the first independent evidence for an often-cited concern of practitioners – even as the number of new ideas and potential for innovation is increasing, there seems to be a reduction in the ability of companies to scale in a meaningful and systematic way. Whether this is primarily a challenge for capital market, or

reflects systematic reductions in various aspects of ecosystem efficiency remains an important challenge for future research. Finally, our results highlight that the regional variation in start-up performance reflects very significant regional differences in both the underlying quality of ventures started in different locations (Silicon Valley has by far the highest EQI in the nation) but also that regions differ significantly in the ability of the regional entrepreneurial ecosystem to nurture high-potential companies in order to realize growth. Systematic and real-time measurement of both of these dimensions – entrepreneurial quality and ecosystem performance – can serve as powerful tools for policymakers and stakeholders seeking to accelerate and reinforce the impact of entrepreneurship on economic and social progress.

II. A Systematic Framework to Measure Growth Entrepreneurship and Ecosystem Performance

Assessing the state of entrepreneurship is not simply a data challenge but also requires conceptual clarity as to what one is trying to measure. While the vast majority of individuals who are classified as either self-employed or small business owners have neither the ambition nor inherent potential (in terms of their underlying business idea) for significant growth (Hurst and Pugsley, 2010), a very small fraction of start-ups have historically been associated with large economy-wide benefits from entrepreneurship (Kaplan and Lerner, 2010; Kerr, Nanda, and Rhodes-Kropf, 2014). In addition to accurate measures of the quantity of "average" entrepreneurs (as measured by, for example, the Kauffman Index of Entrepreneurial Activity), assessing the state of entrepreneurship requires measures that allow one to track and compare the *quality* of entrepreneurship (i.e., the potential for growth). As emphasized by Schoar, even if entrepreneurship has a net positive effect, policy efforts that aim to increase the supply of

entrepreneurship without regard to quality could have a negative economic effect: "I argue that unless we understand the differences between those two types of entrepreneurs more clearly, many policy interventions may have unintended consequences and may even have an adverse impact on the economy." (Schoar, 2010).

In Guzman and Stern (2015a; 2015b), we develop a novel methodology for estimating entrepreneurial quality and potential which has the potential to facilitate the calculation of entrepreneurial statistics that address the challenge laid out by Schoar, Hathaway and Litan, and others. Our approach combines three interrelated insights. First, because the challenges to reach a growth outcome as a sole proprietorship are formidable, a practical requirement for any growth-oriented entrepreneur is business registration (as a corporation, partnership, or limited liability company). This allows us to form a population sample of entrepreneurs "at risk" of growth at a similar (and foundational) stage of the entrepreneurial process. Second, we are able to potentially distinguish among business registrants through the measurement of 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 founders name the firm after themselves (eponymy), whether the firm is organized in order to facilitate equity financing (e.g., registering as a corporation or in Delaware), 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. More precisely, for a firm *i* born in region *r* at time *t*, with start-up characteristics $X_{i,r,t}$, we observe growth outcome $g_{i,r,t+s}$ s years after founding and estimate:

$$\theta_{i,r,t} = 1000 \times P\left(g_{i,r,t+s} \middle| X_{i,r,t}\right) = 1000 \times f\left(\alpha + \beta X_{i,r,t}\right) \quad (1)$$

Using this predictive model, we are able to *predict* quality as the probability of achieving a growth outcome given start-up characteristics at founding, and so estimate entrepreneurial quality as $\hat{\theta}_{i,r,t}$. 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 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).[4]

We build on this approach to propose three new entrepreneurship statistics capturing the level of entrepreneurial quality within 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 Entrepreneurial Quality Index. 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

[4] The practical requirement for estimating entrepreneurial quality in recent cohorts is the timeliness of observing the start-up characteristics, *X*. As in Guzman and Stern (2010b), we consider two different indices – a real-time "nowcasting" index that only includes information directly observable from the business registration form (and so can be calculated for firms as they register), and an informationally richer index that includes early-stage start-up milestones such as the acquisition or grant of a patent within the first year after founding, the granting of a trademark in the first year after founding, or mention in local media or news in the first year after founding. When one aggregates individual firm results in to aggregate indices, there is a very high level of concordance between indices based on these two approaches, a result that we use in interpreting the most recent findings using the nowcasting index to assess the current state of entrepreneurship on an essentially real-time basis.

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:

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

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 $X_{r,t}$ from the vector of observable start-up characteristics.

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:

$$RECPI_{r,t} = EQI_{r,t} \times N_{r,t} \quad (3)$$

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 potential of a given cohort of firms to generate growth events, the true number of realized growth events from that cohort will likely be different. An ecosystem that over performs relative to its expected number of growth firms will be contributing to the entrepreneurial performance of that region independent of the quality of its firms. This effect could be driven by both transitory elements (e.g. boom and bust business cycles) or fundamental elements of a region that deliver superior economic performance. We define REAI as the ratio of realized growth events to expected growth events:

$$REAI_{r,t} = \frac{\sum g_{i,r,t}}{RECPI_{r,t}} \quad (4)$$

A value of REAI above 1 would indicate that a region is providing extra value to its start-ups, while a value below 1 would indicate that, after controlling for their intrinsic quality, start-ups are less likely to achieve growth in this region than they would otherwise be in an average region-cohort of our sample. The index has two different interpretations. First, it indicates a region's ability to realize growth firms from a cohort's initial growth potential. Understanding these regional premiums would allow us, for example, to evaluate the ecosystem impact of interventions aimed at improving the performance of firms of similar quality. Second, it is also a measure of the value of a region for a new firm. A region with a higher value of REAI will be more effective in helping a new firm achieve growth conditional on the firm's entrepreneurial potential (i.e. the probability of a firm achieving growth equals $\hat{\theta}_{i,r,t} \times REAI_{r,t}$).

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

Estimating the Relationship of RECPI to Economic Growth. These statistics are only useful if the underlying predictive analytics are informative, and provide insight into variation over time and space as to entrepreneurial quality beyond previous measures. We begin to develop these connections to broader economic performance outcomes by linking these measures of entrepreneurship to measures of economic growth. Specifically, for any region *r* in year *t*, we would like to estimate the equation:

$$\Delta Ln\left(GDP_{r,t}\right) = Ln\left(GDP_{r,t+s}\right) - Ln(GDP_r) = \alpha + \rho \times Ln(GDP_r) + \beta \times Ln(RECPI_{r,t}) + \gamma_r + \epsilon_{r,t}$$

Where $GDP_{r,t}$ represents economic output, γ_r are region fixed-effects, $RECPI_{r,t}$ is our measure of regional entrepreneurship potential and $\epsilon_{r,t}$ is an error term. This regression is a dynamic panel that controls for regression to the mean (Pritchett and Summers, 2014) and convergence effects (Barro, 1991) by including the current level of GDP, and were the coefficient β measures the regional elasticity in GDP growth from changes in entrepreneurship production. However, estimating this regression through OLS leads to biased estimates due to endogeneity in the relationship of economic growth and entrepreneurship (more and higher potential firms will register in regions that are likely to grow) and due to econometric issues of autocorrelation in the error term (Anderson and Hsiao, 1981; Holtz–Eakin, Newey, and Rosen,1988). To address these concerns in a preliminary way, we draw on a broad line of research that uses the Arrelano-Bond GMM lagged first-differences estimator to account for fixed differences across locations as well as the autocorrelation pervasive to growth data. Under the assumptions of Arrelano-Bond

(exogeneity of the lagged levels to the first-differences of the growth equation), the estimator offers a consistent estimate of the impact of particular measures of entrepreneurship on economic growth.

III. Data and Entrepreneurial Quality Estimation

As mentioned earlier, our analysis leverages publicly available business registration records, a potentially rich and systematic data for entrepreneurship and business dynamics. Business registration records are public records created when individuals register a business. This analysis focuses on the states of Alaska, California, Florida, Georgia, Massachusetts, Michigan, New York, Oregon, Texas, Vermont, Washington, and Wyoming, from 1988-2014. While it is possible to found a new business without business registration (e.g., a sole proprietorship), the benefits of registration are substantial, including limited liability, , various tax benefits, the ability to issue and trade ownership shares, credibility with potential customers, and the ability to deduct expenses. Furthermore, all corporations, partnerships, and limited liability must register with the state 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 state. In other words, our analysis excludes non-profit organizations as well as companies whose primary location is external to the state. Applied over the years 1988-2014, the resulting dataset is composed of 11,277,643

[5] This section draws on Guzman and Stern (2015a), where we introduce the use of business registration records in the context of entrepreneurial quality estimation.

observations.[6] For each observation we construct variables related to: (a) the 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.

Growth. The growth outcome utilized in this paper, 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.[7]

Start-Up Characteristics. The core of the empirical approach is to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures: (a) measures based on business registration observables, and (b) measures based on external indicators of start-up quality that are observable at or near the time of business registration. We review each of these in turn.

Measures Based on Business Registration Observables. We construct ten measures based on information observable in the 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 five 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

[6] The number of firms founded in our sample is substantially higher than the US Census Longitudinal Business Database (LBD), done from tax records. 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 a 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] Our results are robust to changes in the time allowed for a firm to achieve growth. See Guzman and Stern (2015, Supplementary Materials) for a subset of those robustness tests.

[8] Belenzon, Chatterji, and Daley (2014) perform a more detailed analysis of the interaction between eponymy and firm performance.

hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Since some of our states will not include the name of the firm manager, we also include *First Name* and *Last Name*. *First Name* is equal to 1 if the firm name includes the name of an individual in its name (e.g. "Joe's Pizza")[9], *Last Name* is equal to 1 if the firm name has a last name, and zero otherwise[10]. We expect each of these measures to be negatively correlated with both *Growth* and *Eponymy*.

Our last two 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 of less words, and zero otherwise.[11] We also define *Unique Name* equal to one if the firm name includes word that are very rare (exist less than five times) within the full list of all our firms.

We then create four 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 US Cluster Mapping Project ("US CMP") (Delgado, Porter, and Stern, 2015). We use a text analysis algorithm, trained with a large sample of over 10M firms, to identify firms that are likely to belong to certain groups of industries as defined in the US CMP. We develop nine such measures. The first four are associated with broad industry sectors and include whether a firm can be identified

[9] As a list of first names, we aggregate the Social Security Administration Baby Names lists from the years 1950 to 2000. To avoid issues of rare names with words that are more often seen as a common noun (e.g. Rainbow), we restrict our sample to names registered at least 1000 times throughout those 50 years.

[10] We use the list of top 50,000 last names provided by mongobay.com (http://names.mongabay.com/data/1000.html), and restrict our sample to those last names with at least 5000 occurences.

[11] 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.").

as local (*Local*), or traded (*Traded*), traded within high technology clusters (*Traded High Technology*) or traded within resource intensive industries (*Traded Resource Intensive*). The other five industry groups are narrowly defined to be 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.

Table 2 reports the summary statistics. A detailed description of all variables as well as the specific set of US CMP clusters that we use to develop each industry classification are provided in the Data Appendix.

Estimation of Entrepreneurial Quality. Before turning to the direct calculation of the indices outlined in Section II, we must first calculate our estimate for entrepreneurial quality for each firm.

To do so, we regress *Growth* on a set of start-up characteristics observable directly through the business registration records or related early-stage activities of growth-oriented start-ups. We begin in Table 3 with a series of univariate logit regressions of *Growth* on each of our measured

start-up characteristics. As mentioned earlier, these regressions (and all subsequent regressions) are conducted on a random 70% training sample of the complete 19995-2008 dataset, reserving 30% of the 1995-2008 data as a test sample. To facilitate the interpretation of our results, we present the results in terms of the odds-ratio coefficient and the pseudo-R^2.

These univariate results are suggestive. Various simple measures directly captured from the registration record (such as whether the firm is a Corporation or registered in Delaware, or is named after the founder or using less than two words) are each highly significant and associated with a large increase in the probability that a given firm achieves a growth outcome. For example, corporations are associated with almost a 4X increase in the probability of growth, and those that register in Delaware are associated with almost a 50X increase in the probability of growth. Conversely, firms named after their founders have only a 17% chance of a growth outcome relative to those with a non-eponymous name. Equally intriguing results are associated with measures of the degree of innovativeness and novelty of the start-up: *Patent* is associated with nearly a 150X increase in the probability of growth, and *Trademark* are each associated with more than a 75X increase in the probability of growth. Our name-based industry sectors are also relevant: for example, firms associated with IT and biotechnology are 6X and 17X (respectively) more likely to achieve growth, while firms associated with local industries are only 21% as likely to grow than non-local firms. In our univariate setting, we find all our measures to correlate with *Growth*, with the magnitudes of the effects being often economically large.

It is of course important to emphasize that each of these coefficients must be interpreted with care. While we are capturing start-up characteristics that are associated with growth, we are not claiming a causal relationship between the two: if a firm with low growth potential changes its

legal jurisdiction to Delaware that is unlikely to have any impact on its overall growth prospects.[12] Instead, Delaware registration is an informative signal, based on the fact that external investors often prefer to invest in firms governed under Delaware law, of the ambition and potential of the start-up, as observed at the time of business registration. Reliance on a univariate measure makes inference particularly tricky: in isolation, one cannot evaluate whether any particular start-up characteristic is more or less important than others.

We investigate the relationship between our measures and *Growth* in more detail in Table 4. Performing a separate analysis serves two goals: it allows to study the power of the relationship between our measures and growth by allowing us to include controls that might not be possible in our later prediction model (e.g. year fixed effects and state time trends), and it allows us to analyze measures that cannot be include in our predictive model because we do not have complete data (e.g. eponymy), but which are nonetheless interesting. Model (1) contains all the covariates that can be identified from firm name.

Finally, we move to develop our predictive models. The results for our two key specifications are in Table 5.[13] We report the incidence-rate ratio for each coefficient, which can

[12] One important concern in policy applications of this methodology, is that our measures might change incentives of firms, such that they try to "game" the result by select into high-quality measures they previously did not care about (e.g. changing its name form long to short). We note that this possibility, though real, is bounded by the incentives of the founders. For example, it is unlikely that a founder with no intention to grow would incur the significant yearly expense require to keep a registration in Delaware (which we estimate around $1000). Similarly, firms that signal in their name being a local business (e.g. "Taqueria") are unlikely to change their names in ways that affect their ability to attract customers. Finally, we also note that any effects from "gaming" would be short-lived since, as low quality firms select into a specific measure the correlation between such measure and growth – and therefore the weight our prediction model would assign to it – weakens.

[13] Guzman and Stern (2015a; 2015b) provide a much more detailed discussion of alternative specifications for this first-stage regression, using data for California and Massachusetts, Guzman and Stern (2015b), in particular, provides a detailed comparison of the predictive contribution from various data elements (e.g., the value of analyzing firm names versus focusing exclusively on corporate form). While worthy of further research on its own terms, the broad pattern of findings that we report are robust to the inclusion or exclusion of individual indicators, as highlighted by our comparison between the nowcasting index and specifications which draw on early-stage company milestones.

be interpreted as the predicted change in the likelihood of a growth event when a particular start-up characteristic is observed.[14] Our first specification (Model 1) uses only business registration observables. Three of our industry agnostic name-based measures appear to be informative, firms with a short name are 3.2 times more likely to grow, firms with a unique name are 13% more likely to grow, and firms whose name includes a proper name are 33% *less* likely to grow. While including a last name was informative in univariate results, the relationship goes away in the joint model. Our corporate structure measures are particularly informative. Corporations are 4 times more likely to grow and firms registered under Delaware jurisdiction are 46 times more likely to grow. Finally, our dummies indicating association to particular US CMP industry clusters are also informative. Firms whose names indicate inclusion in a local industry (such as "restaurant", "realtor", etc) are 42% less likely to grow, firms associated with traded industries are 12% more likely to grow, and firms specifically associated with resource intensive traded industries are 30% more likely to grow. Names associated with specific high-technology sectors are also associated with growth: firms related to biotechnology are 3.3 times more likely to grow, firms associated with IT 2.5 times, firms associated with semiconductors 2.3 times, and firms in ecommerce 1.5 times more likely to grow. The relationship with firms names related to medical devices, however, is insignificant.

We extend this specification in Model 2 to include observables associated with early-stage milestones related to intellectual property. The coefficients on the business registration observables are quite similar (though slightly reduced in magnitude), while each of the intellectual

[14] It is important to emphasize that we are not claiming a *causal* impact from these start-up characteristics (i.e., simply registering in Delaware does not necessarily causally change a firm's growth probability); instead, these factors are the "digital signatures" associated with start-ups and we are simply trying to develop a predictive algorithm that allows us to have a consistent (if noisy) estimate of their underlying entrepreneurial quality.

property observables is highly predictive. In particular, receiving a patent is associated with a 61 times increase in the likelihood of growth for non-Delaware firms, and the combination of Delaware registration and patenting is associated with a 269 times increase in the likelihood of growth (simply registering in Delaware without a patent is associated with only a 39X increase in the growth probability). Finally, firms successfully applying for a trademark in their first year after business registration are associated with a five times increase in the probability of growth.[15]

These two models offer a tradeoff. On the one hand, the "richer" specification (Model 2) involves an inherent lag in observability, since we are only able to observe early-stage milestones in the period after business registration (in the case of the patent applications, there is an additional 18-month lag due to the disclosure policies of the USPTO). While including a more informative set of regressors, Model 2 is not as timely as Model 1. Indeed, specifications that rely exclusively on information encoded within the business registration record can be calculated on a near real-time basis, and so provide the most timely index for policymakers and other analysts.[16] We will calculate indices based on both specifications; while our main historical analyses will be based off the results from Model 2, the Nowcasting model can be used to provide our best estimate of changes in the last few years.

V. The State of American Entrepreneurship

[15] It is worth noting that the coefficients in these two regressions are remarkably similar to what we found in previous research in California (Guzman and Stern, 2015a) and Massachusetts (Guzman and Stern, 2015b).

[16] It is also worthwhile to note that we can compare the historical performance of indices based on each approach – as emphasized in Figure 2 and 4, aggregate indices have a high level of concordance during the period in which a comparison is feasible, giving us some confidence in the trends predicted by the nowcasting index in the last few years.

We now proceed to evaluate trends in entrepreneurial potential (RECPI) in the United States. Our measures of regional entrepreneurship are calculated by aggregating estimates of entrepreneurial quality at the individual firm level to a measure that includes all start-ups founded in a particular cohort year.

Figure 3 shows our estimate of entrepreneurship in 2012 (the last year for which we have full data) for the United States. The data is presented by ZIP Code, with the size of each point representing to the number of firms in that ZIP Code and the color represents its average quality. Starting from the southwest region of the contiguous 48 states, entrepreneurship potential is clearly high in California, and is particularly high around the Bay Area. Potential drops quickly once we move into Oregon, which then improves somewhat on Washington (we do note that the level of quality of our ZIP Code data is not complete in the states of Washington, Ohio, and New York, but is complete in all other states). Wyoming shows much less density and generally low entrepreneurship, while Texas shows important clusters of high mass of entrepreneurship potential around Dallas and Houston, followed by Austin (a much smaller city, though still of high quality). The area around San Antonio shows a high number of firms but mostly low quality and the areas of El Paso and the Southern Plains (which houses important oil investments) have a smaller but visible entrepreneurship potential mass. In the Midwest, Michigan has small clusters of high quality around Detroit and Ann Arbor, while we skip Ohio given that our ZIP Code information in this state is not complete. In the Southeast, there is substantial entrepreneurship in both Florida and Georgia, thought the quality appears to be low, except, perhaps, for a slightly higher quality area around Atlanta, GA. In the Northeast, we skip over New York due to similar reasons. It is possible to appreciate the important mass of entrepreneurship potential around the Boston area, with a smaller but still visible mass around Central Massachusetts. While Vermont is known as a

highly innovative city, there is little indication of high entrepreneurial quality across the state. Finally, Alaska shows virtually no entrepreneurship except for a very small pocked of high quality around Juneau and another south of Anchorage (with Anchorage itself being rather low quality).

Figure 4 shows estimated Full RECPI and Nowcasted RECPI adjusted by GDP for an aggregate of twelve states (Alaska, California, Florida, Georgia, Massachusetts, Michigan, New York, Oregon, Texas, Vermont, Washington, and Wyoming), between 1988 and 2014. The confidence interval is estimated through a Monte Carlo process repeating our procedure for 100 bootstrapped random samples. Before analyzing trends in the indexes, we note that both indexes move very close to each other. Both indexes indicate a rise of entrepreneurial potential in the 1990s through the year 2000, with a substantial drop from the year 2000 to 2002 (albeit to a level 75% higher than that of 1990). This was followed by an increase through 2007, a sharp decline in 2008 and 2009 (likely linked to the onset of the Great Recession), and a steady and then increasingly robust upward climb starting in 2010.[17] Relative to quantity-based measures of entrepreneurship such as the BDS, RECPI is able to reflect broad and well-known changes in the environment for growth entrepreneurship, such as the dot-com boom and bust of the late 1990s and early 2000s. At the same time, RECPI offers an important comparison relative to simple measures of "success" (e.g., counting the number of high-value exits that occur in a given year or are associated with a given cohort) – though informative, success-oriented measures necessarily conflate the founding of growth-oriented firms and the process by which that growth potential is realized within particular ecosystems or for particular start-up cohorts. In Figure 4, we show a different measure of economic growth realized from a firm cohort: the amount of market value created from firms in that cohort. Specifically, we find the top 50 US firms by current (2015)

[17] These broad patterns closely accord with the patterns we found for Massachusetts in Guzman and Stern (2015b).

market value that were founded between 1988 and 2014, and plot the aggregate incidence and market value by founding year. Though noisier, this figure also shows a similar pattern of sharp rise in the later years of the 1990s, followed by a sharp drop in 2000-2001, a smaller rise up to 2004 and a weak economy afterwards. While there is obviously truncation bias in this analysis (firms in the last cohorts have had less time to grow) the sharp growth and decline in market value created in the 1990s and early 2000s is too much to be driven by bias. Instead, we offer a more economic interpretation, fluctuations in the market value created by different firm cohorts is driven by fluctuation in the underlying potential (RECPI) of those cohorts at birth.

Our index of entrepreneurial potential does show gaps relative to realized entrepreneurial performance, though the statistics of aggregate market value in Figure 4 as well as the number of growth firms in Figure 1 peak in the years 1996 and 1997, RECPI instead peaks in the year 2000. This highlights the sensitivity of entrepreneurial potential to credit market cycles where even though the cohorts of 1996 and 1997 had lower potential they where able to take advantage of the full dot-com bubble of the late 1990s to reach growth, while the peak RECPI start-up cohorts of 1999 and 2000 were never able to realize their potential due to the "financial guillotine" that followed the NASDAQ crash of March, 2000 (Nanda and Rhodes-Kropf, 2013, 2014).

Importantly, our RECPI estimates offer a new perspective on the "state" of entrepreneurship (at least for these twelve states). Specifically, our nowcasting index suggests that there has been a steep rise in entrepreneurial potential over the last several years, and 2014 is the first year to begin to reach the peaks of the dot-com boom. Indeed, it is useful to recall that our measure is *relative to GDP*: on an absolute scale, RECPI 2014 is at the highest level ever registered (327 in 20014 versus the previous peak of 312 in 2000). Finally, we emphasize that,

though there are small deviations, the Full RECPI and Nowcast RECPI have a very high concordance, and it is likely that a similar pattern for Full RECPI will emerge as we observe early-stage milestones of the most recent cohorts.

Overall, this evidence supports three interrelated conclusions. First, relative to a perspective emphasizing a worrisome secular decline in "shots on goal" (Hathaway and Litan, 2014b), our approach and evidence suggest that there has been a more variable pattern of entrepreneurship over the last 25 years, and that the last five years has been associated with an accumulation of entrepreneurial potential similar to that which marked the late 1990s. Second, this variation in potential has a clear relationship with later entrepreneurship performance of such cohorts using both measures of number of realized growth firms as well as market value created by firms in those cohorts. Finally, given the more gently sloped level of the entrepreneurial boom of recent years, it may be the case that this accumulation of entrepreneurial potential is more sustainable than earlier periods.

VII. Trends in the Effect of the US Entrepreneurial Ecosystem (REAI)

The relationship between entrepreneurship and economic progress, of course, depends on more than simply founding new enterprises, but scaling those enterprises in a way that is economically meaningful. This insight motivates our second set of findings where we examine "ecosystem" performance across the United States, as measured by the Regional Ecosystem Acceleration Index (REAI). 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 (i.e., REAI = Growth Events / RECPI). A value of 1 in the index indicates no ecosystem effect. A value above 1 indicates a positive ecosystem effect, and a value under 1 indicates a negative effect. In contrast to RECPI, this index reflects the impact of the economic and entrepreneurial environment in which a start-up cohort participates (i.e., the "ecosystem" in which it participates). This ecosystem will include the location in which the firm is founded (e.g., Silicon Valley versus Miami) as well as the environment for funding and growth at the time of founding. In Figure 5, to examine the changing environment for entrepreneurship in the United States (i.e., change in the US Ecosystem, as reflected in the twelve states for which we have data), we plot REAI over time from 1988-2008, and developed a projected measure of REAI for years 2009-2012. [18]

Three distinct periods stand out. The early portion of our sample saw a significant increase in REAI from a slight negative level to a peak of 1.6 for the 1996 cohort. This is consistent with our evidence from Figure 1, in which the 1996 start-up cohort was indeed the most "successful." This peak was followed by a steady decline through 2002, in which, conditional on the estimated quality of a given start-up, the probability of growth was declining as the result of the environment (i.e., time) in which that start-up was trying to grow. From 2003-2008, there is a period of stagnation, with REAI at or around 0.6. These differences are economically meaningful: a start-up for a given quality level is estimated to be 4 times more likely to experience a growth event in the six years after founding if they were founded in 1996 rather than in 2005. Finally, though still a preliminary estimate, we observe a weak resurgence the first increase in REAI for cohorts in 2009 to 2011, highlighting a potential improvement in the entrepreneurial ecosystem in recent

[18] 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 2008. For years 2009 to 2012, we estimate our model with a varying lag of $n = 2014 - year$ and calculate RECPI using such lag.

years in parallel with the boom in the availability of entrepreneurial finance. Still, while this rise is economically important, the level is still much lower than the decade of the 1990s.

This pattern is both striking and worrisome. Over the past years, there has been increasing understanding of the role that successful entrepreneurship plays as an engine for economic progress, and increased public involvement in supporting start-up activity and nurturing regional entrepreneurial ecosystems. Yet, despite that attention, the emergence from the Great Recession seems to have not been driven by (nor helped) the start-up cohorts founded in the late 2000s. Preliminary evidence shows that more recent cohorts experience a more favorable set of outcomes, but how favorable still remains an open question, and understanding the factors that facilitate more favorable outcomes for a given level of RECPI are an important agenda for future research.

IV. Understanding Regional Entrepreneurship

We now shift our level of focus from the entrepreneurship production at the national level to the regional level, and its relationship to regional economic performance. Figure 6 shows RECPI for each of our twelve states, both adjusted by GDP and in levels. In levels of entrepreneurial potential, large populous states like California, Florida and Texas have very high values of RECPI, with California having a value twice as high as the next state. It is possible to also see sharp rises in RECPI in the dot-com boom in California and Massachusetts as well as the more recent natural resource boom in Alaska and Wyoming. However, adjusted for GDP, the level of RECPI/GDP for California is not the highest. Instead, Vermont has a much higher RECPI than any other state with values over twice as large as the next state. These results parallel patenting statistics that find

California has the highest number of total yearly patents US P.T.O. patents while Vermont has the highest rate of patenting per capita in the nation.

Assessing the importance of regions in RECPI. Next, we move to evaluate the relevance of regions in our estimates of entrepreneurial quality. To do so, we re-run our full prediction model but include state fixed-effects and year fixed-effects and assess by how much do the coefficients vary. The results are presented in Table 6 and, without requiring us to comment on each coefficient in detail, they vary remarkably little.

Estimating the relationship between regional entrepreneurship and economic growth. We now move to understand the relationship between our regional measures and economic performance. As there are obvious omitted factors that affect both entrepreneurship and economic growth (e.g. the available investment opportunities), our goal is not claim causality controlling for these factors, but instead establish whether our index can successfully *predict* future economic growth. Even if these omitted variables exist, a positive correlation would still allow us to establish a macroeconomic (non-causal) correlation between entrepreneurial investment and economic growth, a mechanism often theorized but which existing measures of economic performance still have not connected. Further, predictive ability, independent of causality, is no small policy feature: an economic statistic that can estimate current economic potential and preempt future economic performance can be of substantial value for policymakers, practitioners, and empirical researchers.

We build a panel of states and measures of EQI, RECPI, number of new firms in the business registry (births), and real gross state product (GSP) – the state level equivalent of GDP. We also include the number of firm births reported in the Business Dynamics Statistics (BDS) for each state and the rate of business dynamism, to allow comparisons between RECPI and quantity-

based measures. All measures are in their natural logarithm, except for EQI and business dynamism which are percentage measures.

Table 7 shows the summary statistics and correlation matrix for this dataset. We include a synthetic measure of output growth (*GSP Growth*) equal to the change in GSP in the following five years (calculated as a forward lag) to also assess the correlation between our measures and economic growth. Three measures, BDS births, registry births, and RECPI have a substantial positive correlation with GSP, all of them above 0.8, highlighting how as the state economy grows so does economic activity, while reallocation holds no relationship to GSP and, interestingly, quality (EQI) has a negative correlation. However, once we look at GSP growth instead of GSP levels, these effects change substantially: the only measure correlated with forward growth is quality, with a coefficient of 0.18, while all the others have insignificant correlations.

While these correlations are interesting, concerns of regional level variation might make the results spurious. To assess the relationship in more detail, we estimate a regression model with region fixed effects. To control for issues related to error-term correlation in dynamic panels, we use Arellano-Bond GMM (Holtz-Eakin, Newey, and Rosen, 1988; Arellano and Bond, 1991).

Table 8, shows regression estimates of the relationship between economic growth and each of our measures, using GSP five years later (GSP Growth) [19] as the dependent variable. It also includes a Hansen J-test for overidentifying restrictions and the Arellano-Bond m_2 measure of autocorrelation of first-differences to assess whether the Arellano-Bond identifying condition is met. To increase confidence in our results, we cluster standard errors at the state level to control for added correlation, use the small sample version of standard errors, and use two-step GMM

[19] Instead of estimating a regression that has growth in output as a dependent variable and a control for the current level, our regression just uses the level as a dependent variable. For the purposes of estimating β these two are equivalent as $GSP_{r,t+s} = \rho \times GSP_r + \beta \times RECPI_{r,t} + \gamma_i + \epsilon_{r,t}$ is equivalent to $GSP_{r,t+s} - GSP_{r,t} = (\rho + 1) \times GSP_r + \beta \times RECPI_{r,t} + \gamma_i + \epsilon_{r,t}$

which has better small sample properties. Though Arellano-Bond originally recommend using all prior lags as instruments, a short panel like ours would have issues of many instruments (see Roodman, 2009a) and hence we limit our instruments only to second and third lags (first lags are correlated by definition in first-differences)[20]. We identify second and third lags as having low correlation to later lags based on our analysis of the Akaike's information criterion (AIC) and log-likelihood relationships between the dependent variables and GSP growth.

Models 1 through 4 all fail the m_2 test with over 95% confidence. This test fails of the error terms of the lagged differences are correlated with the first differences, and thus making the Arellano-Bond estimator unidentified for these regressions. Model 5, estimating the relationship between RECPI and economic growth, is the only one that does not fail the m_2 test. The coefficient for RECPI is positive and significant and its magnitude – 0.2 – is also economically relevant. Given the log-log structure of our regressions, the coefficient can be interpreted as the elasticity: an increase of 10% in RECPI leads to an increase of 2% in GDP over the full five year period.

In conclusion, though important measurement error, simultaneity and endogeneity challenges exist in understanding the causal relationship between entrepreneurship and economic growth, a simpler question to ask is what type of entrepreneurship *predicts* economic growth. In our analysis, we are able to identify the relationship between RECPI and economic growth clearly, though not others except for business dynamism, whose coefficient we find on the opposite direction.

RECPI Silicon Valley: A Case Study. While our results so far have focused on the aggregate experience across twelve (relatively diverse) US states, many questions about the state of entrepreneurship are particularly concerned with specific regional ecosystems, perhaps none

[20] We also test including third and fourth lags simultaneously with second lags and find the results to be virtually unchanged.

more so than Silicon Valley. We therefore calculate RECPI over time solely for the combined MSA coverage of San Jose-Sunnyvale-Santa Clara as well as San Francisco-Oakland-Fremont regions, and plot the results (on an absolute scale) in Figure 7.[21] The overall pattern of results is quite similar to that of the aggregate RECPI in Figure 2, with a sharp increase in RECPI Silicon Valley during the dot-com boom, an equally sharp drop from 2000-2002, stagnation through 2010, followed by a sharp increase since 2010. While the overall directional shifts are the same, the levels are quite different. In particular, the boom in RECPI since the bottom of the Great Recession has been as steep (if not steeper) than during the late 1990s, and Nowcasted RECPI Silicon Valley is more than 50% higher than was ever realized during the dot-com boom (indeed, RECPI Silicon Valley has exceeded its dot-com peak every year since 2011). Of course, the very rapid increase in recent year may indeed be cause for concern (suggesting a bubble that, like the 1990s, cannot be sustained).

VI. Mapping Entrepreneurial Quality Across Metropolitan Areas

As a final piece of analysis, we look at the changing nature of the micro-spatial distribution of *average* entrepreneurial quality (EQI) for a few key geographic areas in our sample. Figures 8A-8D show maps of EQI at the ZIP Code level, for five areas across 4 different years: the Boston metropolitan area, the San Francisco Bay area, the City of San Francisco, and the Miami metropolitan area. Each map represents a snapshot of entrepreneurial quality during the year in question. Looking across snapshots of quality for a particular city gives a sense of the evolution

[21] While a full analysis of economic impact would properly "deflate" RECPI by the overall size of the economy (as we did in Figure 2), it is useful to consider the absolute numbers to capture the perspective of individual observers of a regional ecosystem who may be benchmarking their experience against an earlier time period.

of the ecosystem. While one might expect each region to follow a similar pattern, we see important heterogeneity in changes in entrepreneurial quality across regions and time periods.

Figure 8A shows the Boston metropolitan area. In 1988, we find entrepreneurial quality concentrated around the Route 128 corridor, a pattern documented in the detailed analyses of Massachusetts growth entrepreneurship by Saxenian (1992) and Roberts (1991). As the Boston area moves into the dot-com boom, the amount of entrepreneurial quality increases in both the central and neighboring districts while continuing to be centered around Route 128. However, over the past decade, the center of high-quality entrepreneurship has shifted. There is still high quality entrepreneurship around Route 128, but Cambridge (particularly Kendall Square) and areas of Boston (such as the Innovation District) have emerged as the leading areas in terms of intensive entrepreneurial quality in the Boston region.[22]

Figures 8B and 8C take a deep-dive on the San Francisco Bay Area (with Figure 8A capturing the entire Bay Area and Figure 8B focusing in on the City of San Francisco in particular). First, the initial state of entrepreneurial quality in 1988 is relatively modest, with a narrow set of areas near San Jose and Sunnyvale accounting for the entirety of a "Silicon Valley" effect. The 1990s saw both an upgrade of entrepreneurial quality in the South Bay, with a boom particularly around Stanford and Berkeley. Consistent with Figure 3, the drop-off in entrepreneurial quality was much more muted after the dot-com crash than in many other places, with a particular striking rise in overall quality by 2012. More importantly, we see a shift over the past decade in the rise of entrepreneurial quality in San Francisco, extending beyond a few districts (as in 2000); by 2012,

[22] In Guzman and Stern (2015b) we have also documented this pattern of migration from Route 128 to Cambridge by estimating yearly average quality for both regions, We also document micro-geographical patterns at the level of individual addresses, highlighting the heterogeneity that exists around the "MIT Ecosystem" (e.g., comparing buildings around Kendall Square from the more retail entrepreneurship around Central Square and Cambridgeport.

more than half of the zip codes in San Francisco registered a level of entrepreneurial quality that places them in the top 5% of the distribution of all zip codes throughout the 25-year sample period.

Beyond these hotspots, Figures 8D and 8E document patterns in regions that have yet to experience the type of entrepreneurial ecosystem development as Boston or the Bay Area. In Los Angeles (Figure 8D), we see a rather unconnected set of entrepreneurial quality hotspots (most located adjacent to key universities or research centers such as UCLA or UC-Irvine) and, despite extensive policy discussion in Los Angeles aimed at boosting the environment for entrepreneurship, the level and intensity of entrepreneurial quality is lower in 2012 than it was in 1996.[23] Finally, Figure 8E documents the entrepreneurial ecosystem in Miami over time: even during the height of the dot-com boom, there was relatively little shift in the overall entrepreneurial quality of any region, and over time, there has been an erosion of relative quality in this region. By 2012, most of the Miami area has low entrepreneurial quality (outside the top quartile). This result stands in sharp contrast to previous results that have found this same area to have the highest level of self-employment (e.g. Glaeser, 2007),[24] thus highlighting the importance of focusing on quality rather than intensity of new firm formation in analyses of entrepreneurial ecosystems.

In short, while our sample as a whole experienced a pattern of growth in the 1990s, decline in the early 2000s, and then growth from 2009-2014, the experience within metropolitan areas varied widely. Interestingly, areas that have historically been associated with start-up led growth, such as Silicon Valley and the Boston area, have both continued to thrive, though there has been a

[23] See also Guzman and Stern (2015a) for a specific break-out of Los Angeles from 2008-2012.

[24] Specifically, Gleaser (2007) finds that the top three MSAs (using the 2000 Census definitions) in the United States by rates of self-employment are West Palm Beach-Boca Raton-Delray Beach, FL, Miami-Hialeah, FL, and Fort Lauderdale-Hollywood-Pompano Beach, FL. Here we use the updated 2012 MSA definitions and present the Miami-Fort Lauderdale-West Palm Beach, FL MSA, which is (basically) the same area.

shift in both regions towards their urban cores; other regions such as Miami have yet to experience a start-up boom.

VIII. Conclusion

Regional entrepreneurship can be measured in many different ways for different purposes. Measuring entrepreneurship by its ability to reflect investment and serve as a leading indicator to future economic growth is perhaps the most important one, but population-level measures of regional entrepreneurship appear to hold little relationship to economic growth. Existing quantity based measures show a secular decline in entrepreneurship and do not appear to track the economic cycles of the last three decades, including the often commented boom in growth entrepreneurship during the 1990s and after the Great Recession, as well as the lower potential period of the early 2000s.

Using a quality-based approach pioneered in Guzman and Stern (2015a; 2015b) with business registration records for twelve states, we focus on the systematic measurement of entrepreneurial quality to create synthetic entrepreneurship indexes at the regional and national level. Not simply a matter of data, a focus on entrepreneurial quality allows us to focus on a more rigorous examination of variation over time and across places in the potential from a given start-up cohort (RECPI) and the ability of an entrepreneurial ecosystem to realize that potential over time (REAI).

This approach yields several novel findings:

- Rather than a pattern of secular long-term decline, start-up potential has followed a cyclical pattern that seems sensitive to the capital market environment and overall economic conditions.
- The cyclical pattern in RECPI tracks well variation in overall economic performance – measured as regional GDP growth in the following years – as well as tracking measures of the economic contribution of entrepreneurial cohorts (e.g. the market value created by each cohort).
- While the peak value of RECPI is recorded in 2000, there is a sharp upward swing beginning in 2010 and, even after controlling for change in the overall size of the economy, the third highest level of entrepreneurial growth potential is registered in 2014.
- However, there exists gaps between expected and realized entrepreneurial production, the likelihood of start-up firms for a given quality level to realize their potential (REAI) was highest for firms founded around 1995, declined sharply in the late 1990s, and did not recover through 2008. Though preliminary evidence shows a recent improvement in REAI, whether the most recent cohorts are able to realize their potential at rates similar to those achieved during the mid-1990s is yet to be seen.
- AT the regional level, Arellano-Bond regressions establish a strong correlation between our measure of entrepreneurship production (RECPI) and regional GDP growth, while other measures are not consistently positive. Business dynamism turns out to be negatively correlated to economic growth though whether that result is spurious is an important separate question of research.
- Finally, there is striking variation across regions (and over time) in entrepreneurial potential: Silicon Valley RECPI is now at an all-time high (and has been at a level beyond

its dot-com peak since 2011) while Miami has yet to experience a meaningful start-up boom. The state of American entrepreneurship appears heterogeneous in both the cross section and time trends, with implications for the future development of regions.

Our analysis thus indicates that *both* changes in entrepreneurial potential and ecosystem effects are economically important in US entrepreneurial performance. Relative to the 1900s (without the dot-com boom and bust of 1998-2002), we observe a three to four-fold drop in the US ecosystem performance while observing very little drop in overall entrepreneurial potential. This suggests that, in the current US environment, policies oriented towards improving the performance of existing entrepreneurs might be better guided than policies focused on increasing the supply of quality entrepreneurship, which already appears robust.

MIT President Rafael Reif recently emphasized the importance of scaling in advocating for a new US innovation model: "To tackle our biggest societal challenges, we need an innovation pipeline that delivers every drop." (Reif, 2015). We believe that the first step towards realizing this ambition is developing a systemic understanding of entrepreneurial ecosystems that allow for meaningful linkages between regional investments policies and institutions, the founding of high-quality entrepreneurial firms with the ambition to take ideas to impact, and the ability to realize that potential on a systematic basis. The economic statistics outlined here are a step in this direction.

REFERENCES

Arellano, Manuel and Stephen Bond. (1991) "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations" *Review of Economic Studies*. 58 (2) p. 277-97.

Axtell, Robert. (2001) "Zipf Distribution of US Firm Sizes". *Science* Vol. 293 no. 5536 pp. 1818-1820

Belenzon, Sharon, Chatterji, Aaron and Brendan Daley. (2014) "Eponymous Entrepreneurs" *Working Paper*

Cochrane, John H. (2005) "The risk and return of venture capital" *Journal of Financial Economics*. 75. 3-52

Davis, Steven, and John Haltiwanger. (1992) "Gross Job Creation, Gross Job Destruction, and Employment Reallocation". *The Quarterly Journal of Economics.* 107 (3): 819-862

Decker, Ryan, Haltiwanger, John, Jarmin, Ron and Javier Miranda. (2014) "The Role of Entrepreneurship in US Job Creation and Economic Dynamism". *Journal of Economic Perspectives.* 28(3): 3-24

Gabaix, Xavier. (1999) "Zipf's Law for Cities: An Explanation" *Quarterly Journal of Economics.* August. P739-767

Gabaix, Xavier (2009). "Power Laws in Economics and Finance," Annual Review of Economics, Annual Reviews, vol. 1(1), pages 255-294, 05.

Gibrat, Robert. (1931) *Les Inégalités Économiques*, Paris, Librairie du Recueil Sirey

Glaeser, Edward. (2007) "Entrepreneurship and the City". *NBER Working Paper #13551.*

Glaeser, Edward, Sari Pekkala Kerr, and William R. Kerr. (2012) "Entrepreneurship and Urban Growth: An Empirical Assessment with Historical Mines". *NBER Working Paper #18333*

Guzman, Jorge, and Scott Stern. (2015a) "Where is Silicon Valley?" *Science*. Vol. 347. Issue #6222.

Guzman, Jorge, and Scott Stern. (2015b) "Nowcasting and Placecasting Entrepreneurial Quality and Performance" *NBER Working Paper #20954*

Haltiwanger, John, Ron Jarmin, and Javier Miranda. (2013) "Who Creates Jobs? Small versus Large versus Young" *The Review of Economics and Statistics*. 95 (2). 347-361.

Hathaway, Ian, and Robert E. Litan. (2014a) "Declining Business Dynamism in the United States: A Look at States and Metros". *Economic Studies at Brookings Series.* May, 2014.

Hathaway, Ian, and Robert E. Litan. (2014b) "Declining Business Dynamism: It's For Real," *Economic Studies at Brookings Series.* May, 2014.

Hathaway, Ian, and Robert E. Litan. (2014c) "The Other Aging of America: The Increasing Dominance of Older Firms," *Economic Studies at Brookings Series.* July, 2014

Holtz-Eakin, Douglas, Newey, Whitney and Harvey S. Rosen (1988) "Estimating Vector

Autoregressions with Panel Data" *Econometrica*. Vol. 56, No. 6 pp. 1371-1395

Kaplan, Steven, and Josh Lerner. (2010) "It Ain't Broke: the Past, Present, and Future of Venture Capital". *Journal of Applied Corporate Finance.* 22 (2). 36-47

Kerr, William, Nanda, Ramana, and Matthew Rhodes-Kropf. (2014) "Entrepreneurship as Experimentation". *Journal of Economic Perspectives.* 28 (3): 25-48.

Kortum, Samuel, and Josh Lerner. (2000) "Assessing the contribution of venture capital to innovation" *RAND Journal of Economics.* 31 (4). 674-692

Krugman, Paul (1996) "Confronting the Mystery in Urban Hierarchy" *Journal Of The Japanese And International Economies* 10,399–418

McFadden, Daniel. (1974) "Conditional logit analysis of qualitative choice behavior". *Frontiers in Econometrics. Chapter 4*. Academic Press. p. 105-142.

Nanda, Ramana, and Matthew Rhodes-Kropf. (2014) "Financing Risk and Innovation". HBS Working Paper 11-013.

Quah, J. K.-H. and Strulovici, B. (2012), "Aggregating the Single Crossing Property". *Econometrica*, 80: 2333–2348. doi: 10.3982/ECTA9869

Roberts, Edward. (1991) "Entrepreneurs in High Technology: Lessons from MIT and Beyond". Oxford University Press. 1st Edition.

Samila, Sampsa, and Olav Sorenson. (2011) "Venture Capital, Entrepreneurship, and Economic Growth". *The Review of Economics and Statistics.* 93(1). 338-349.

Saxenian, Anna Lee. (1992) "Regional Advantage: Culture and Competition in Silicon Valley and Route 128". Harvard University Press.

Sutton John. (1997), "Gibrat's Legacy", *Journal of Economic Literature*, 35(1), 40-59

UPSTART Business Journal. (2014) "The great VC bubble fight of 2014: Tim Draper says yes, Marc Andreessen says no". January 13, 2014. *Accessed on June 5, 2015*. http://upstart.bizjournals.com/money/loot/2014/01/13/tim-draper-says-bubble-marc.html

Table 1 - US States in Sample. 1988-2015

Rank	*State*	*GDP*	*Share of GDP*	*Cumulative GDP*	*Cumulative Share*
1	*California*	$2,287,021	13.1%	$2,287,021	13.1%
2	*Texas*	$1,602,584	9.2%	$3,889,605	22.3%
3	*New York*	$1,350,286	7.8%	$5,239,891	30.1%
4	*Florida*	$833,511	4.8%	$6,073,402	34.9%
10	*Georgia*	$472,423	2.7%	$6,545,825	37.6%
12	*Massachusetts*	$462,748	2.7%	$7,008,573	40.3%
13	*Michigan*	$449,218	2.6%	$7,457,791	42.8%
14	*Washington*	$425,017	2.4%	$7,882,808	45.3%
25	*Oregon*	$229,241	1.3%	$8,112,049	46.6%
46	*Alaska*	$60,542	0.3%	$8,172,591	46.9%
49	*Wyoming*	$48,538	0.3%	$8,221,129	47.2%
52	*Vermont*	$30,723	0.2%	$8,251,852	47.4%

Aggregate Statistics

Total States	12
Total GDP	$8,251,852
US GDP	*$17,411,875.00*
Share GDP	47%

All states are included from year 1988 to 2015 except for New York where we only have data until 2013. Future versions will likely also include Missouri, North Carolina and Ohio, leading to a total GDP coverage of 54%.

Table 2: Summary Statistics

	N	Mean	Min	Max	Sum
Growth	17,428,280	0.032%	0	1	5,521
Year	17,428,280	2004.6	1988	2015	3.49E+10
Corporation	17,428,280	53.81%	0	1	9,378,927
Short Name	17,428,280	45.97%	0	1	8,011,422
Firm Name Has Last Name	17,428,280	15.56%	0	1	2,711,118
Firm Name Has First Name	17,428,280	12.89%	0	1	2,246,443
Firm Name is Unique	17,428,280	13.66%	0	1	2,381,465
Eponymous	17,428,280	11.78%	0	1	1,597,890
Delaware	17,428,280	2.92%	0	1	509,620
Patent	17,428,280	0.18%	0	1	30,534
Trademark	17,428,280	0.12%	0	1	21,226
Local	17,428,280	18.63%	0	1	3,246,754
Traded Resource Intensive	17,428,280	13.70%	0	1	2,387,752
Traded	17,428,280	54.60%	0	1	9,516,617
Biotech Sector	17,428,280	0.20%	0	1	35,027
Ecommerce Sector	17,428,280	4.80%	0	1	836,372
IT Sector	17,428,280	2.21%	0	1	385,367
Medical Dev. Sector	17,428,280	2.89%	0	1	503,691
Semiconductor Sector	17,428,280	0.05%	0	1	8,285
Observations	17,428,280				

Summary statistics for all business registration records in data. Growth is defined as achieving IPO or Acquisition in 6 years (see Guzman and Stern, 2015a for more information on the definition).

Table 3: Logit Univariate Regressions

Logit univariate regressions of *Growth* (IPO or Acquisition within 6 years) with each of the observables we develop for our dataset. Incidence rate ratios reported; Standard errors in parentheses * p<0.05 ** p<0.01 *** p<0.001

Firm Name Measures:

Variable	*Univariate Coefficient*	*Pseudo R2*
Firm Name has Proper Name	0.385***	0.004
	(0.0277)	
Firm Name has Last Name	0.487***	0.003
	(0.0289)	
Firm Name is Unique	1.560***	0.002
	(0.0652)	
Short Name	3.867***	0.022
	(0.151)	
Eponymous	0.174***	0.006
	(0.0209)	

Corporate Form Measures:

Variable	*Univariate Coefficient*	*Pseudo R2*
Corporation	3.874***	0.018
	(0.175)	
Delaware	48.49***	0.163
	(1.633)	

IP Measures:

Variable	*Univariate Coefficient*	*Pseudo R2*
Patent	150.4***	0.101
	(5.982)	
Trademark	75.07***	0.03
	(4.586)	
Observations	9249832	

Industry Measures:

Variable	*Univariate Coefficient*	*Pseudo R2*
Local	0.221***	0.01
	(0.0169)	
Traded Resource Intensive	1.329***	0.001
	(0.0579)	
Traded	1.349***	0.001
	(0.0457)	
Biotech Sector	16.94***	0.006
	(1.624)	
Ecommerce Sector	1.997***	0.002
	(0.115)	
IT Sector	5.968***	0.013
	(0.300)	
Medical Dev. Sector	3.012***	0.004
	(0.180)	
Semiconductor Sector	17.69***	0.002
	(3.151)	
Observations	9249832	

Table 4: Joint Determinants of Growth. Logit Regression

This table presents the role of different measures in determining growth when presented together.

	(1)	(2)	(3)	(4)
Short Name	3.373***	3.240***		
	(0.174)	(0.152)		
Firm Name has Last Name	1.263**	1.002		
	(0.108)	(0.0752)		
Firm Name has Proper Name	0.705***	0.675***		
	(0.0676)	(0.0616)		
Firm Name is Unique	1.102	1.184**		
	(0.0605)	(0.0610)		
Eponymous	0.299***			
	(0.0421)			
Corporation	4.056***	3.288***		
	(0.237)	(0.183)		
Delaware	46.13***	44.60***		
	(2.108)	(1.921)		
Patent			107.4***	103.0***
			(6.645)	(6.397)
Trademark			11.06***	12.15***
			(1.212)	(1.389)
Local	0.627***	0.584***		
	(0.0618)	(0.0545)		
Traded Resource Intensive	1.345***	1.263***		
	(0.0813)	(0.0718)		
Traded	1.119*	1.098*		
	(0.0555)	(0.0510)		
Biotech Sector	3.021***	3.677***		
	(0.395)	(0.468)		
Ecommerce Sector	1.481***	1.448***		
	(0.113)	(0.106)		
IT Sector	2.507***	2.495***		
	(0.194)	(0.185)		
Medical Dev. Sector	1.135	1.113		
	(0.102)	(0.0970)		
Semiconductor Sector	2.101**	2.374**		
	(0.570)	(0.637)		
Year Fixed-Effects	No	Yes	No	Yes
Observations	5079588	6474876	6474876	6474876
Pseudo R-squared	0.228	0.230	0.113	0.136

Table 5: Prediction Models. Nowcasting and Full Information Model

The models used for prediction of entrepreneurial quality are presented below. The nowcasting model uses only information present in the business registration records and thus can be done in almost real-time, the full information model includes information that we are only able to observe with a lag. In this case, we incorporate only intellectual property but in previous worked we have also included media mentions (Guzman and Stern, 2015b).

	(1)	(2)
Name Based Measures:		
Short Name	3.192***	2.715***
	(0.150)	(0.132)
Last Name	1.042	1.094
	(0.0782)	(0.0837)
First Name	0.672***	0.703***
	(0.0613)	(0.0651)
Unique Name	1.130*	1.086
	(0.0579)	(0.0571)
Corporate Structure Measures:		
Corporation	4.228***	3.698***
	(0.228)	(0.201)
Delaware	45.71***	
	(1.965)	
Intellectual Property Measures:		
Trademark		5.080***
		(0.461)
Interactions:		
Patent Only		63.28***
		(6.190)
Delaware Only		38.70***
		(1.854)
Patent and Delaware		268.7***
		(19.18)
US CMP Cluster Dummies:		
Local	0.581***	0.628***
	(0.0543)	(0.0593)
Traded Resource Intensive	1.298***	1.295***
	(0.0732)	(0.0744)
Traded	1.117*	1.247***
	(0.0518)	(0.0590)
Biotech Sector	3.265***	2.196***
	(0.411)	(0.301)
Ecommerce Sector	1.492***	1.310***
	(0.109)	(0.0978)
IT Sector	2.542***	2.025***
	(0.190)	(0.155)
Medical Dev. Sector	1.114	0.866
	(0.0969)	(0.0775)
Semiconductor Sector	2.291**	1.212
	(0.604)	(0.334)
Observations	6474876	6474876
Pseudo R-squared	0.217	0.263

Table 6: Including State and Year Fixed-Effects in Prediction Model

	(1)	(2)	(3)	(4)
Name Based Measures:				
Short Name	2.942***	2.536***	3.240***	2.740***
	(0.139)	(0.124)	(0.152)	(0.133)
Firm Name has Last Name	1.058	1.082	1.002	1.042
	(0.0800)	(0.0835)	(0.0752)	(0.0800)
Firm Name has Proper Name	0.646***	0.690***	0.675***	0.703***
	(0.0590)	(0.0640)	(0.0616)	(0.0654)
Firm Name is Unique	1.061	1.035	1.184**	1.131*
	(0.0548)	(0.0546)	(0.0610)	(0.0599)
Corporate Structure Measures:				
Corporation	4.386***	3.820***	3.288***	2.809***
	(0.236)	(0.208)	(0.183)	(0.157)
Delaware	40.87***		44.60***	
	(1.867)		(1.921)	
Trademark		4.874***		5.290***
		(0.439)		(0.496)
Interactions:				
Patent Only		52.29***		59.37***
		(5.180)		(5.848)
Delaware Only		35.40***		37.30***
		(1.796)		(1.794)
Patent and Delaware		208.1***		305.5***
		(15.36)		(21.93)
US CMP Cluster Dummies	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	No	No
Year Fixed Effects	No	No	Yes	Yes
Observations	6474876	6474876	6474876	6474876
Pseudo R-squared	0.235	0.276	0.23	0.278

Table 7: Summary Stats and Correlation of State Level Panel

Panel A: Summary Statistics

Variable	*Obs*	*Mean*	*SD*	*Min*	*Max*
GSP Growth (5 years)	252	0.13	0.09	-0.15	0.35
GSP	312	12.36	1.39	9.73	14.54
Reallocation Rate	300	29.44	4.30	20.20	45.60
BDS Firm Births	300	9.22	1.34	6.76	11.22
Bus. Registration Births	312	9.78	1.99	3.81	12.58
EQI (quality)	312	0.0007	0.0009	0.0001	0.0053
RECPI	312	2.07	1.79	-3.52	4.80

Panel B: Correlation Matrix

	GSP Growth	*GSP*	*Reallocati Rate*	*BDS Firm Births*	*Births*	*EQI*	*RECPI*
GSP Growth	1						
GSP	-0.0391	1					
Reallocation Rate	0.1166	0.1092	1				
BDS Firm Births	0.05	0.9686*	0.2070*	1			
Bus. Registration Births	-0.0736	0.8942*	-0.0616	0.8871*	1		
EQI (quality)	0.1832*	-0.4620*	-0.0094	-0.4202*	-0.6206*	1	
RECPI	0.0065	0.8606*	-0.0945	0.8798*	0.9108*	-0.2972*	1

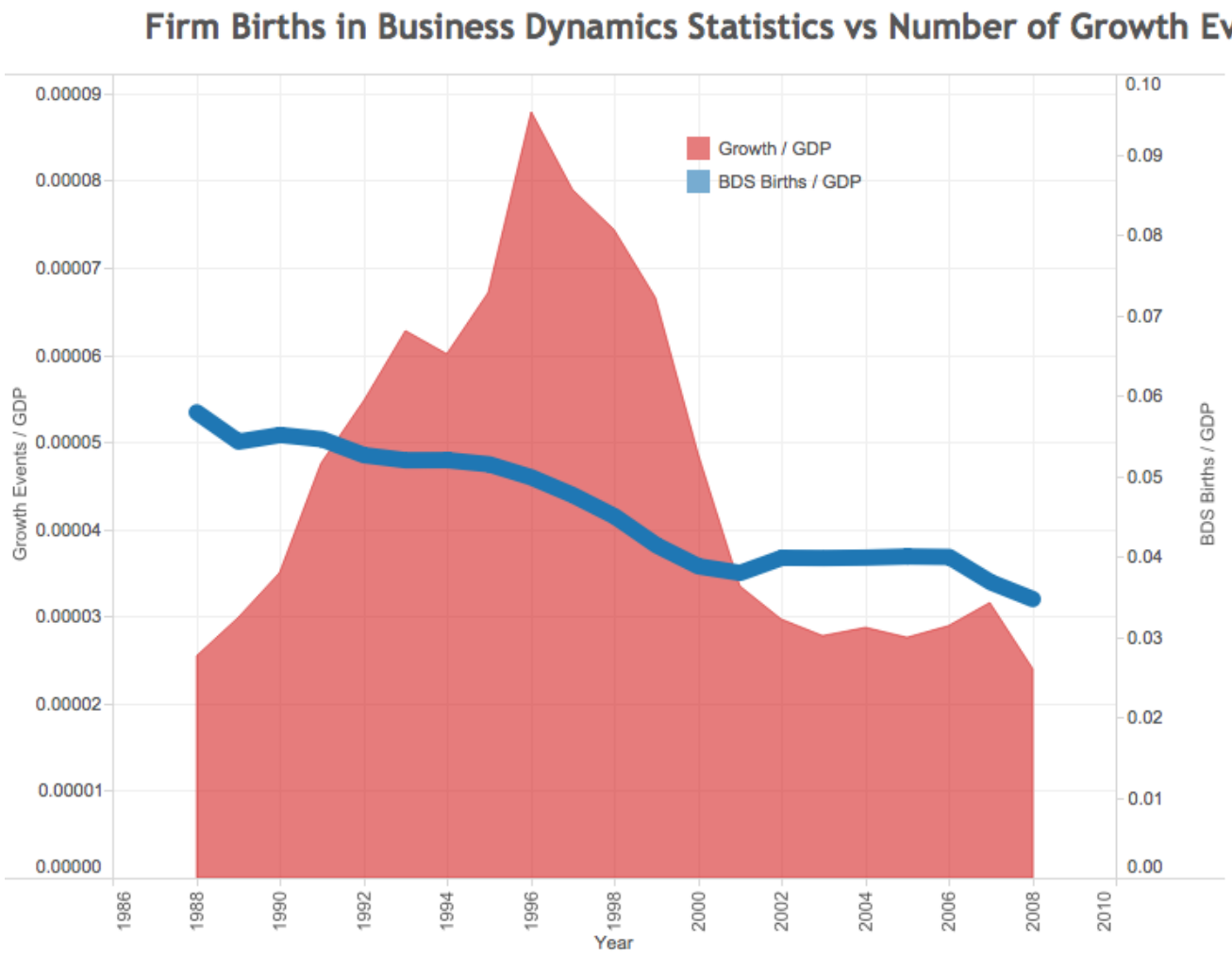
* p < .05

Table 8: Regression of Entrepreneurship Potential on Growth

Arellano-Bond estimator used to control for autocorrelation in the error term in dynamic panels, using the Stata command xtabond2 (Roodman, 2009b). Dependent variable is GSP 5 years later. Model estimated with two-step GMM, standard errors reported in parenthesis with small sample correction, and clustered at the state level. We limit lags to second lags to avoid issues of many instruments as suggested in Roodman (2009a), lags are selected looking at the AIC and log-likelihood statistics of VAR models of our measures on GSP growth. All regressors except for EQI are in their natural logarithms as is the dependent variable. + p<.1 * p<.05 ** p<.01

	(1)	(2)	(3)	(4)	(5)
`	0.760*	0.773*	0.563*	0.775*	0.420*
	(0.0947)	(0.0917)	(0.148)	(0.0961)	(0.186)
Firm Births (BDS)	-0.0269				
	(0.146)				
Reallocation Rate (BDS)		-0.000615			
		(0.00173)			
Firm Births Bus. Registration			0.0750		
			(0.0542)		
EQI				15.54	
				(37.15)	
RECPI					0.200*
					(0.0551)
N	240	240	240	240	240
Hansen J-Test	10.92	10.84	10.99	10.72	10.88
# of Instruments	38	38	38	38	38
AR(2) Autocorrelation Test	2.637	2.475	2.471	2.534	1.302
AR(2) P-Value	0.00836	0.0133	0.0135	0.0113	0.193

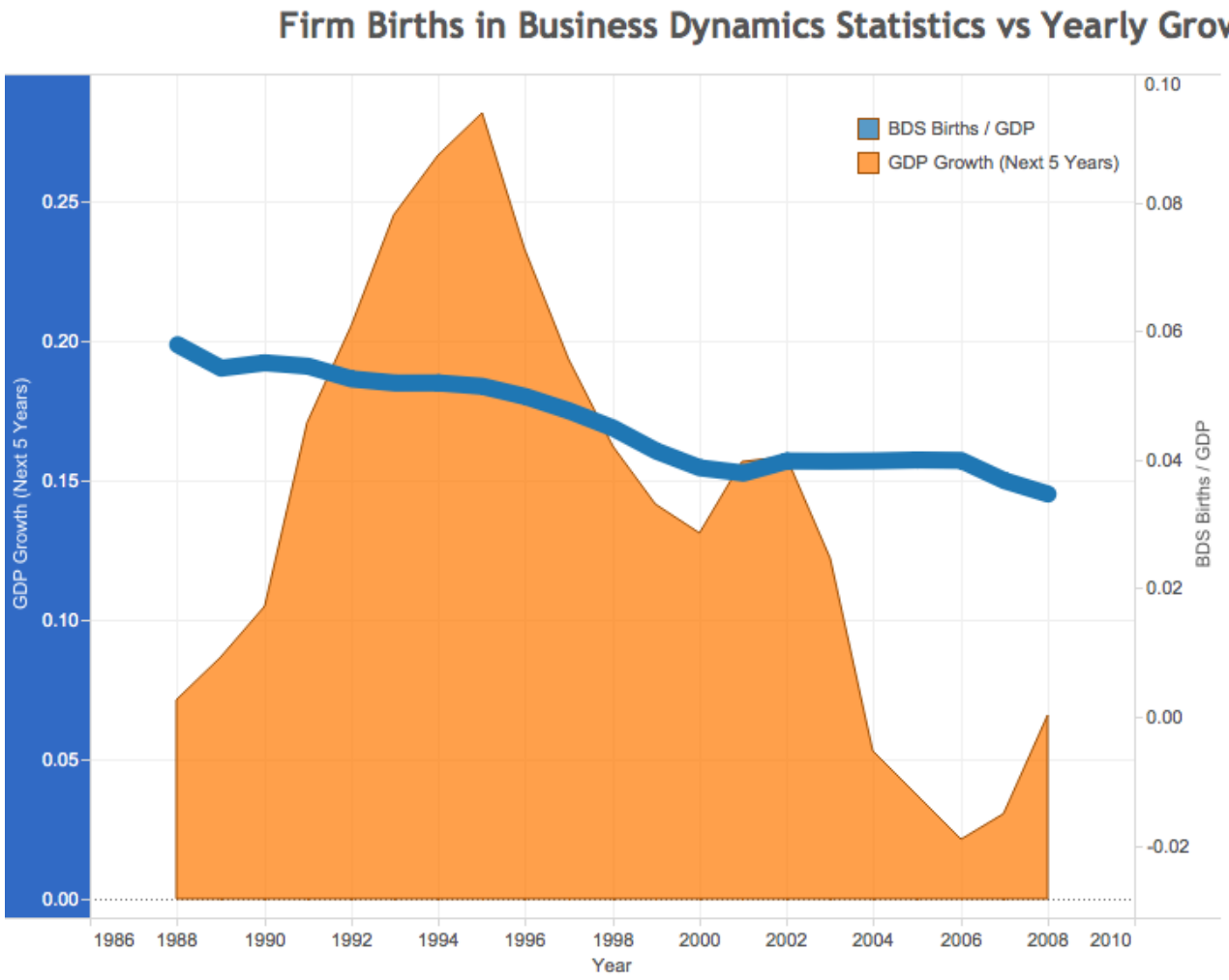
FIGURE 1



Firm Births vs Growth Events per Cohort. The rate of firm births over GDP shows a long-term secular decline in (GDP adjusted) entrepreneurship from 1988 to 2012, as already evidenced by Hathaway and Litan (2014), and has almost no relationship to shorter cycles of US economic performance like the boom of the late 1990s or the recession of the early 2000s. On the other hand, an entrepreneurship measure based on the number of growth firms from a cohort (adjusted by GDP) matches those cycles better, showing an increase from 1988-1996, a drop from 1996-2001 (though still at very high levels) and a relatively flat performance from 2002-2008., while not showing any descernible pattern of secular decline in entrepreneurship.

**Analysis done only for 12 states in our sample. Firm Birth statistics from the US Census Business Dynamics Statistics.*

FIGURE 2



Firm Births vs GDP Growth. The rate of firm births over GDP shows a long-term secular decline in (GDP adjusted) entrepreneurship from 1988 to 2012, as already evidenced by Hathaway and Litan (2014), and has almost no relationship to variation in GDP. GDP growth, estimated as the growth realized over the following five years, starts low but then raises rapidly up to the year 1995, then drops towards the year 2000, and drops again further during the Great Recession.

**Analysis done only for 12 states in our sample. Firm Birth statistics from the US Census Business Dynamics Statistics.*

FIGURE 3

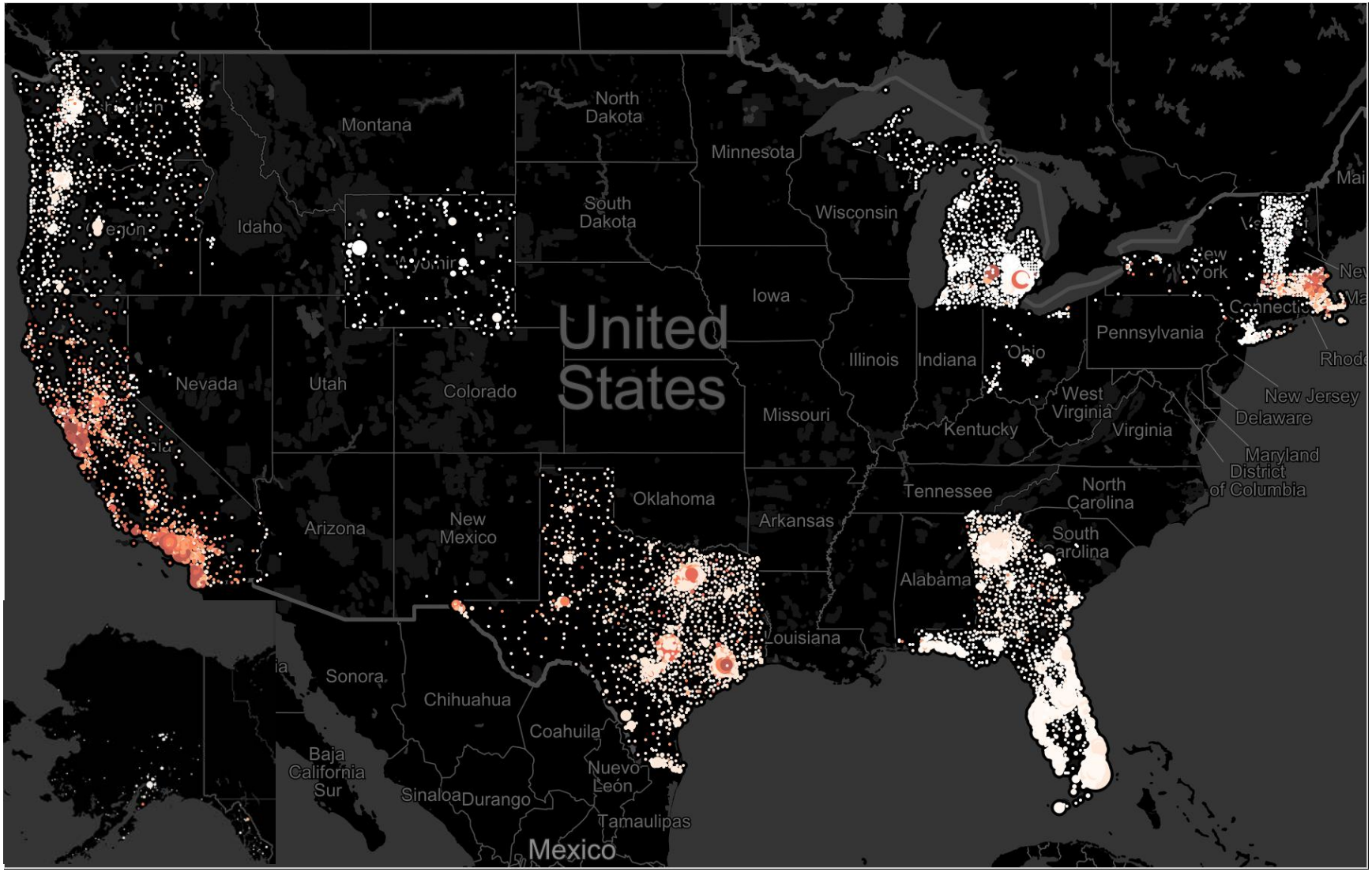
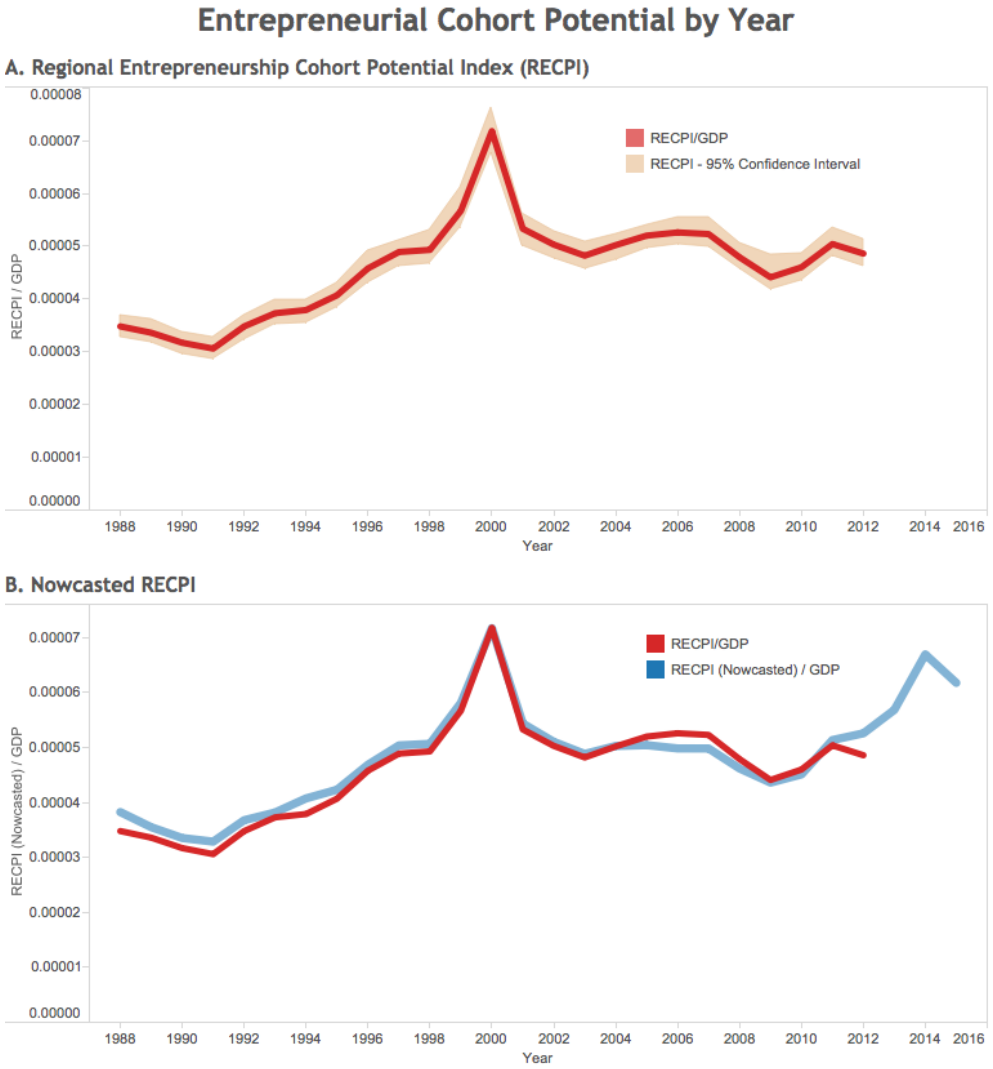

Montana
North Dakota
Minnesota
South Dakota
Idaho
Wisconsin
Iowa
United States
Nevada
Utah
Colorado
Illinois
Indiana
Ohio
Pennsylvania
West Virginia
Kentucky
Virginia
New Jersey
Delaware
Maryland
District of Columbia
Missouri
Oklahoma
Arkansas
Tennessee
North Carolina
South Carolina
Alabama
Louisiana
Arizona
New Mexico
Sonora
Chihuahua
Coahuila
Nuevo León
Tamaulipas
Baja California Sur
Sinaloa
Durango
Mexico

FIGURE 4



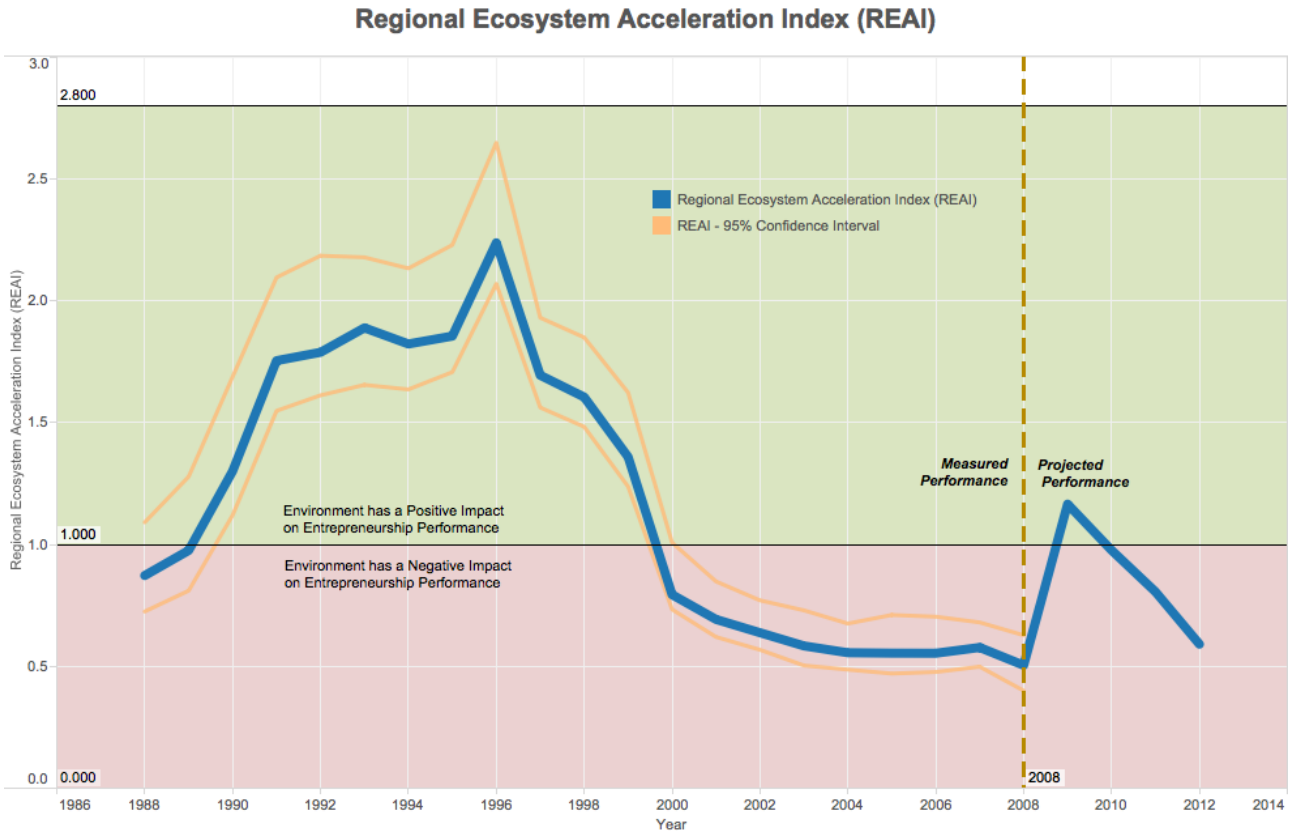
Entrepreneuership Potential by Year. The Regional Entrepreneuship Cohort Potential Index (RECPI) measures the estimated number of growth events for a cohort of firms given their startup characteristics. Mechanically, it is estimated by multiplying the average quality of firms in the cohort (EQI) times the number of firms. RECPI uses an entrepreneurial quality estimate build using firm observables from business registration records and intellectual property, which is observed with a lag. Panel (A) shows the yearly estimate of RECPI using the full model in Table 3 (Model 7), adjusted by GDP to account for differences in the size of the economy. This model includes intellectual property measures. Since IP measures are only observed with a lag, our index is only estimated up to 2012. Our index shows a marked increase in potential in the years leading up to the year 2000, followed by a sharp drop, then a slow rise up to 2007, a drop during the Great Recession, and a recovery. Panel (B) shows our Nowcasting Index, estimated without intellectual property measures. The index keeps a close pattern to the full RECPI. After 2012, the index shows a shapr increase in RECPI in 2013 and 2014 with a subsequent drop in 2015 (though albeit a noisier estimate as it uses projected numbers).

GDP is estimated as the sum of Gross State Product for each of the states in our sample. GSP estimates provided by the Bureau of Economic Analysis. Confidence intervals estimated through a bootstrap procedure, replicating our procedure with 100 random samples (with replacement) of the same size as our analysis sample.

*2015 estimates use average quality of observed firms in 2015 times a predicted number of firms based on an autoregressive model specified as Obs(t) = b0 + b1*Year+b2*Obs(t-1)+e0 for years 2002 to 2014. Since EQI is not homogeneous throughout the calendar year (normally increasing in the fall) the RECPI 2015 projection likely biased downward.*

***Notes:** Index excludes Texas and New York due to need of data refresh. Small dent at RECPI 2012 needs to be corrected by adding 2013 patents.*

FIGURE 5



Regional Ecosystem Acceeleration Index (REAI). REAI measures the regional abilty to convert entrepreneurial potential into realized growth events. It is formally defined as the ratio of realized to expected growth events: this difference is adjudicated to ecosystem effects in achieving this outcome. A value above 1 indicates the ecosystem has a positive effect on entrepreneurial firm performance, and a value below one indicates it has a negative effect.

The United States was the most favorable ecosystem for firms started between 1995 and 2000. For the cohort of 1996, REAI reaches the value of 2.26. It then starts to drop precipitously, crossing into the negative zone for cohorts after the year 1999. We project performance for years after 2008 by comparing running estiamtes using smaller windows of growth (for example, for 2009, we look at expected and realized growth events 5 years after registration). Our results suggest that REAI incrased for cohorts in 2009 and then drops again for later cohorts, though not as low as previous levels. Given the recent startup boom highlighted in Figure 3, this result appears to be lower than expectations might suggest.

..

FIGURE 6

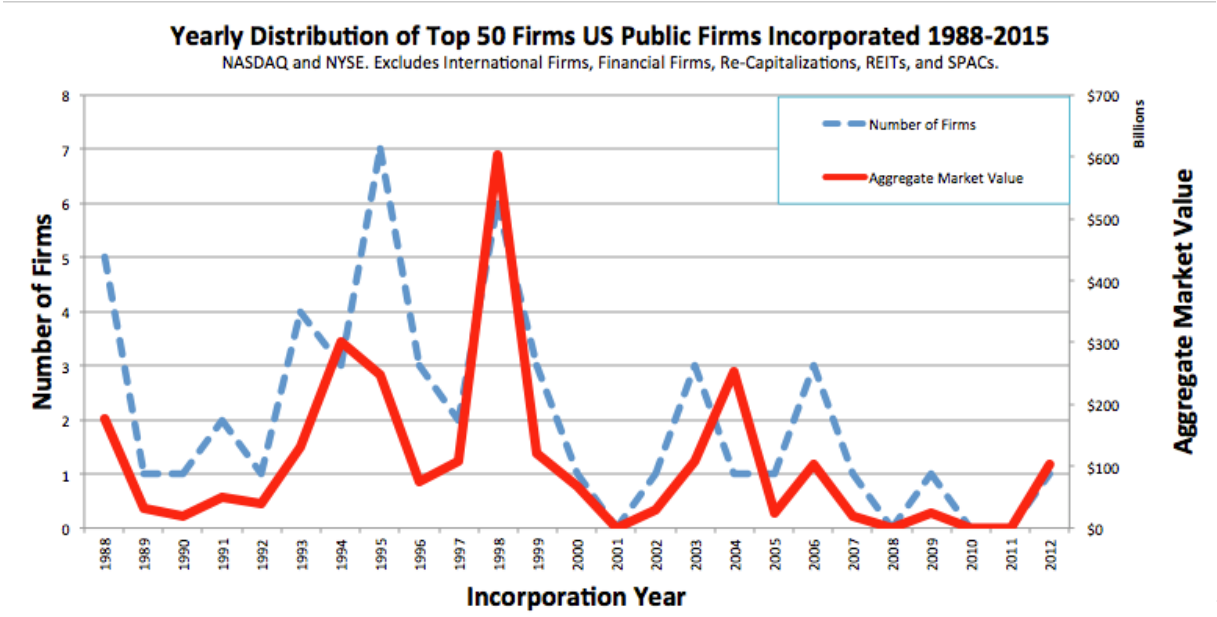
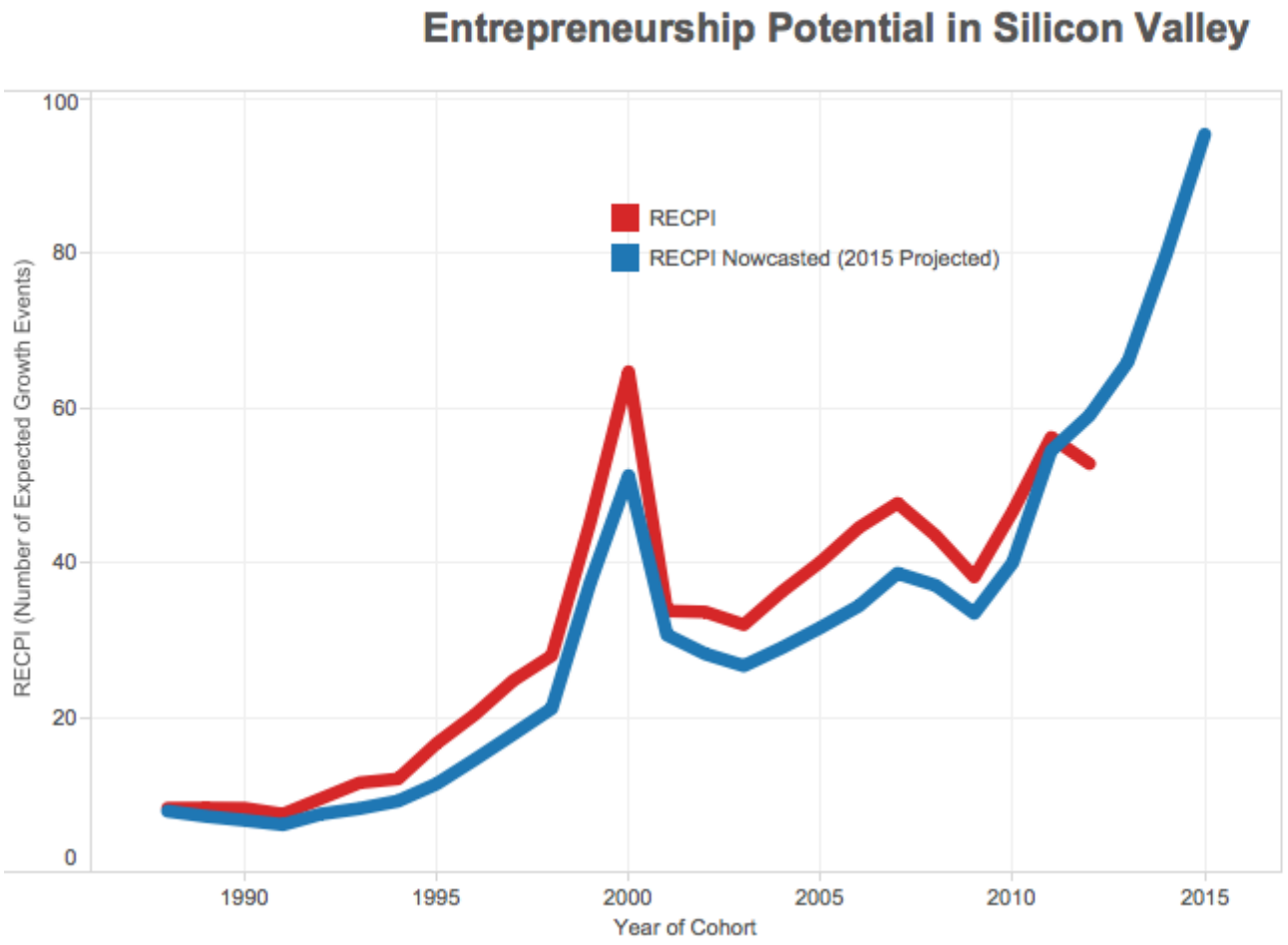


FIGURE 6

FIGURE 7

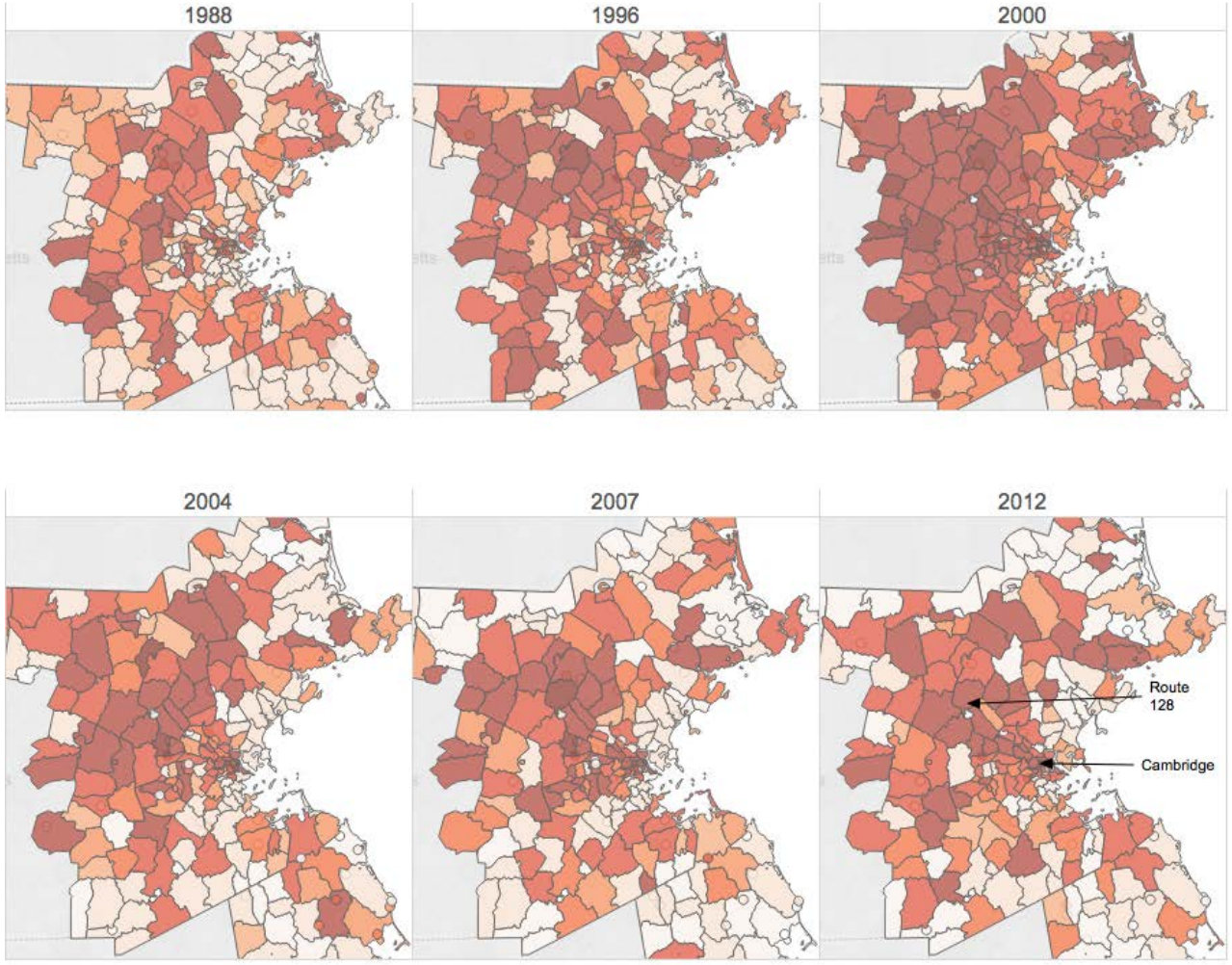


RECPI for Silicon Valley. While the overall US economy has not experienced a boom in entrepreneurial performance similar to that of the year 2000, the specific region of Silicon Valley has seen substantial growth in the number of expected growth firms created for recent cohorts. The level of growth firms attained in 2000 is close in the full index by the year 2012, and surpassed in the nowcasted index by 2013. In 2015, the projected number of growth firms is almost 50% higher than the 2000 level.

Silicon Valley is defined as all firms created in the San Francisco-Oakland-Freemont, CA MSA and the San Jose-Sunnyvale-Santa Clara, CA MSA.
We receive our data up to May of 2015 and project 2015 levels of entrepreneurship by multiplying the 2015 RECPI by 12/5.

FIGURE 8A

Entreprenurial Quality in Boston. 1988-2012



Entrepreneurial Quality around the Boston Area. Map of entrepreneurial quality for all ZIP codes around the Boston area. In 1988, we find entrepreneurial quality concentrated around the Route 128 corridor, a pattern already documented in the detailed analyses of Massachusetts growth entrepreneurship by Saxenian (1992) and Roberts (1991). As the Boston area moves into the dot-com boom, the amount of entrepreneurial quality increases in both the central and neighboring districts while continuing to be centered around Route 128. However, as the region moves to 2012, the amount of high quality ZIP Codes appears to be about the same as the early 1990s, but the spatial location of such quality has changed. While there is still high quality in Route 128 the central cities of Cambridge and areas of Boston have emerged as hotspots of entrepreneurship.

ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).

FIGURE 8B

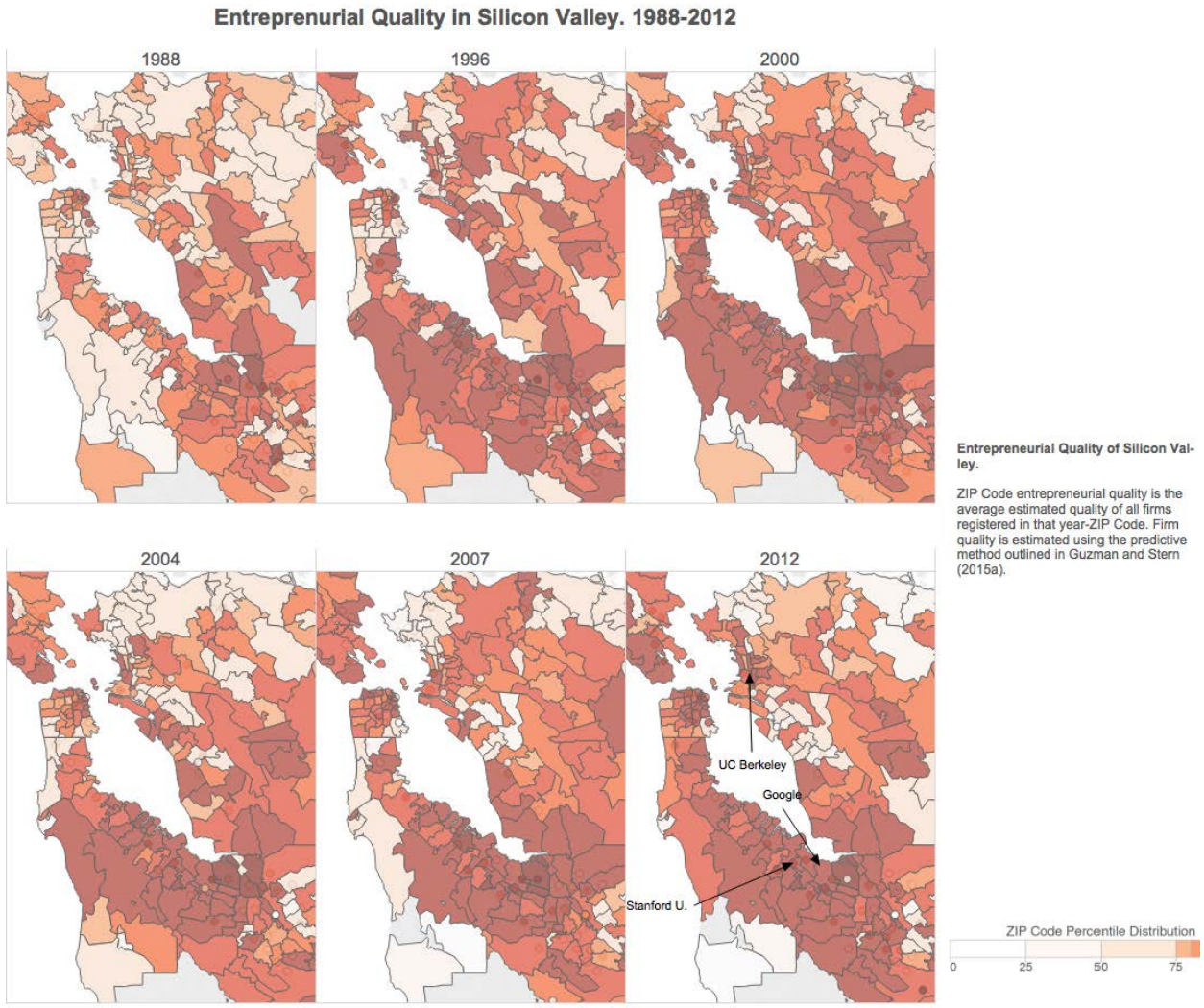
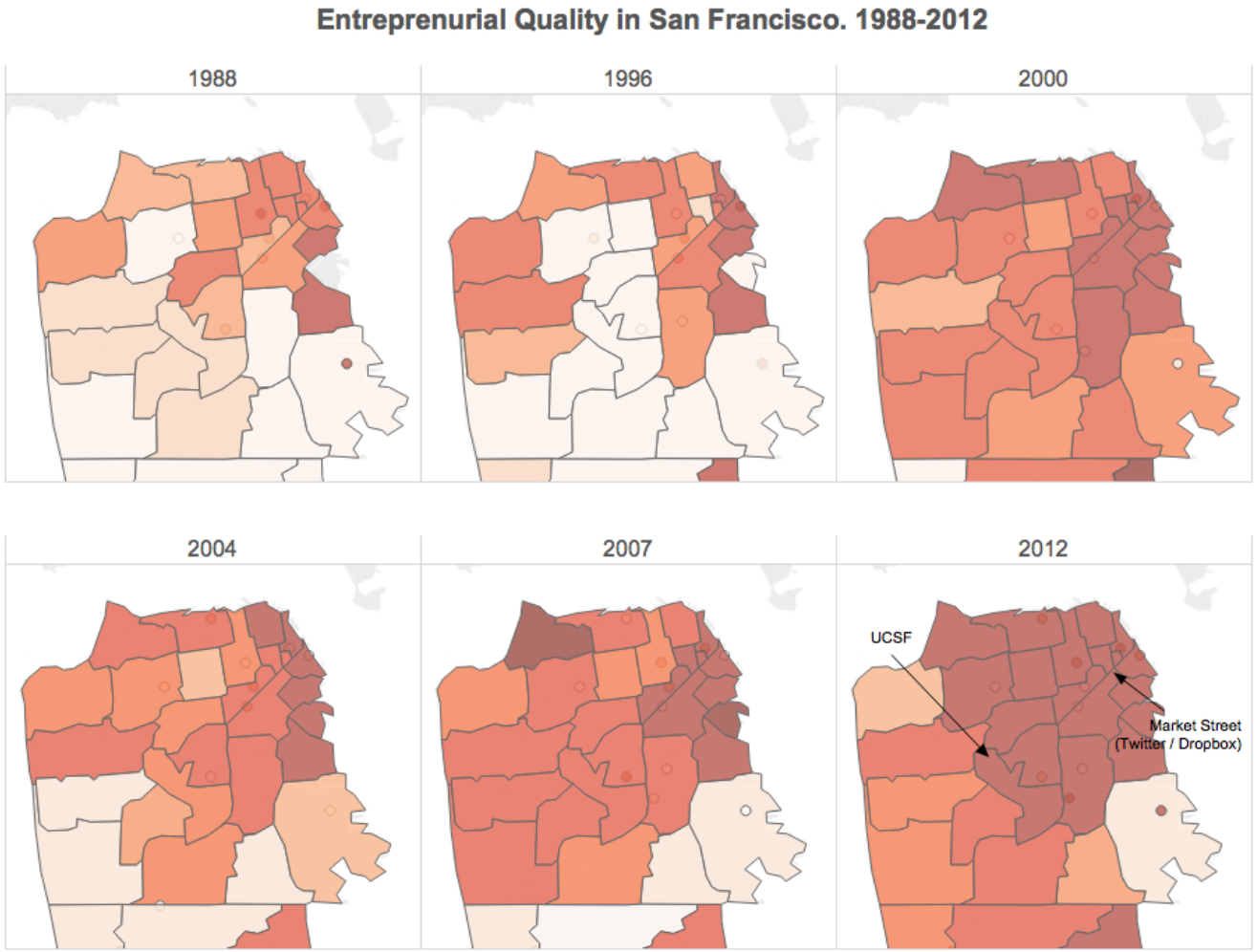

Entreprenurial Quality in Silicon Valley. 1988-2012
1988
1996
2000
2004
2007
2012
UC Berkeley
Google
Stanford U.
Entrepreneurial Quality of Silicon Valley.
ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).
ZIP Code Percentile Distribution
0
25
50
75

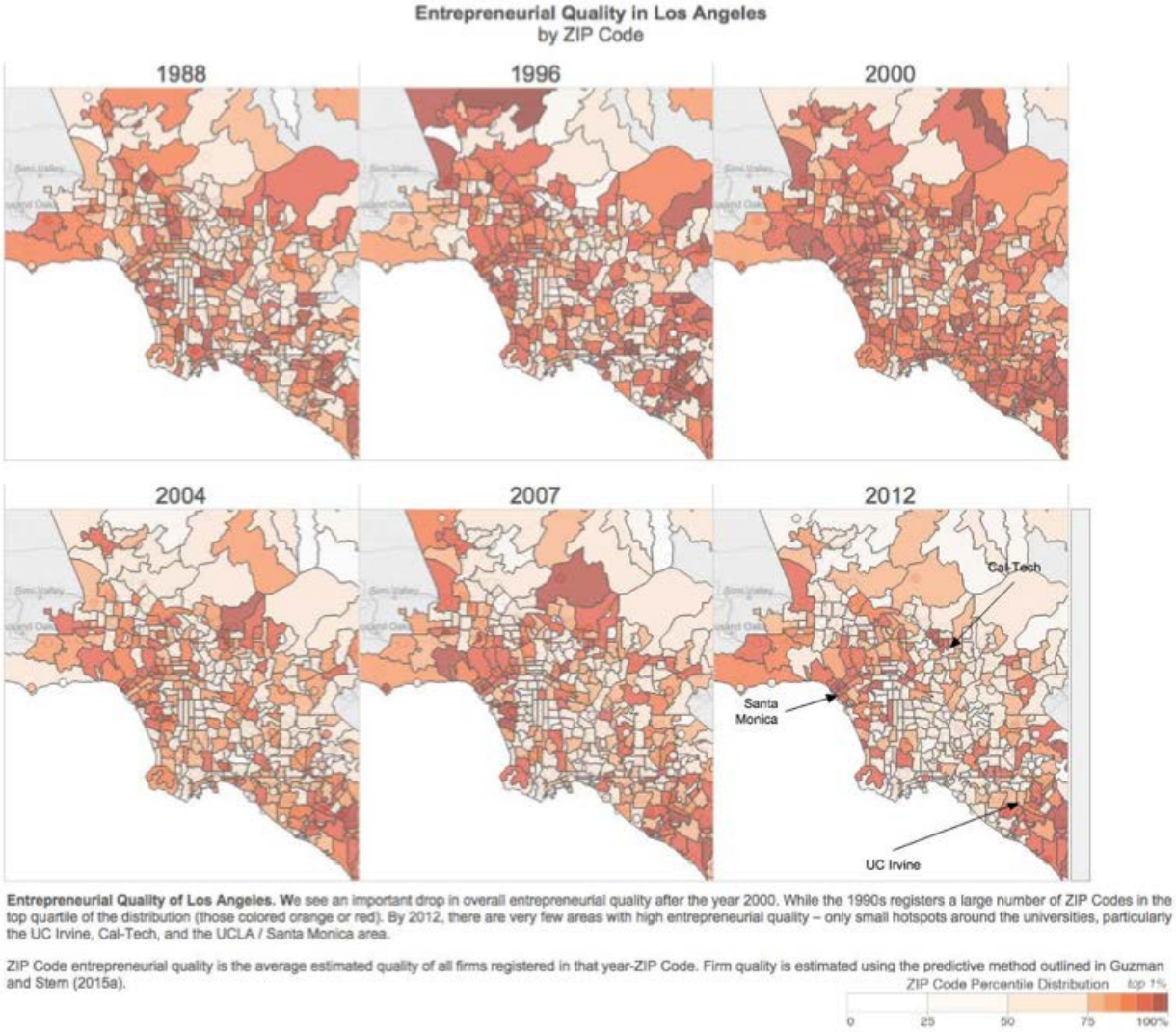
FIGURE 8C



Entrepreneurial Quality of San Francisco. While San Francisco has not historically had much of the high quality entrepreneurship with which the Bay Area has been associated, many media accounts and data on VC investments suggest it has recently seen a boom in startup activity (e.g. Florida, 2013). We find entrepreneurial quality increasing monotonically in San Francisco from 1988-2012. And, while the entrepreneurial quality that existed in 1988 was concentrated in only a small set of ZIP Codes, by 2012, there are several areas of very high quality around Market Street (where new heavyweights like Twitter and Dropbox are located) and almost half of the city itself has increased its quality to relatively high levels (within the top 5% of the year-ZIP Code observations in our sample).

ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).

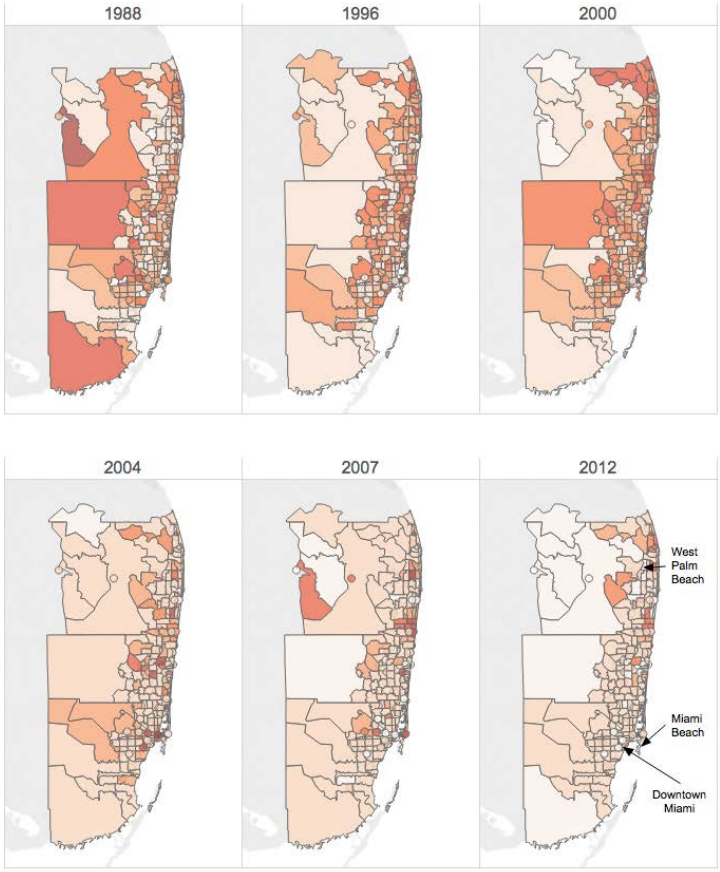
FIGURE 8D



Entrepreneurial Quality of Los Angeles. We see an important drop in overall entrepreneurial quality after the year 2000. While the 1990s registers a large number of ZIP Codes in the top quartile of the distribution (those colored orange or red). By 2012, there are very few areas with high entrepreneurial quality – only small hotspots around the universities, particularly the UC Irvine, Cal-Tech, and the UCLA / Santa Monica area.

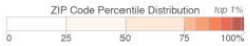
ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).

Entrepreneurial Quality in Miami. 1988-2012



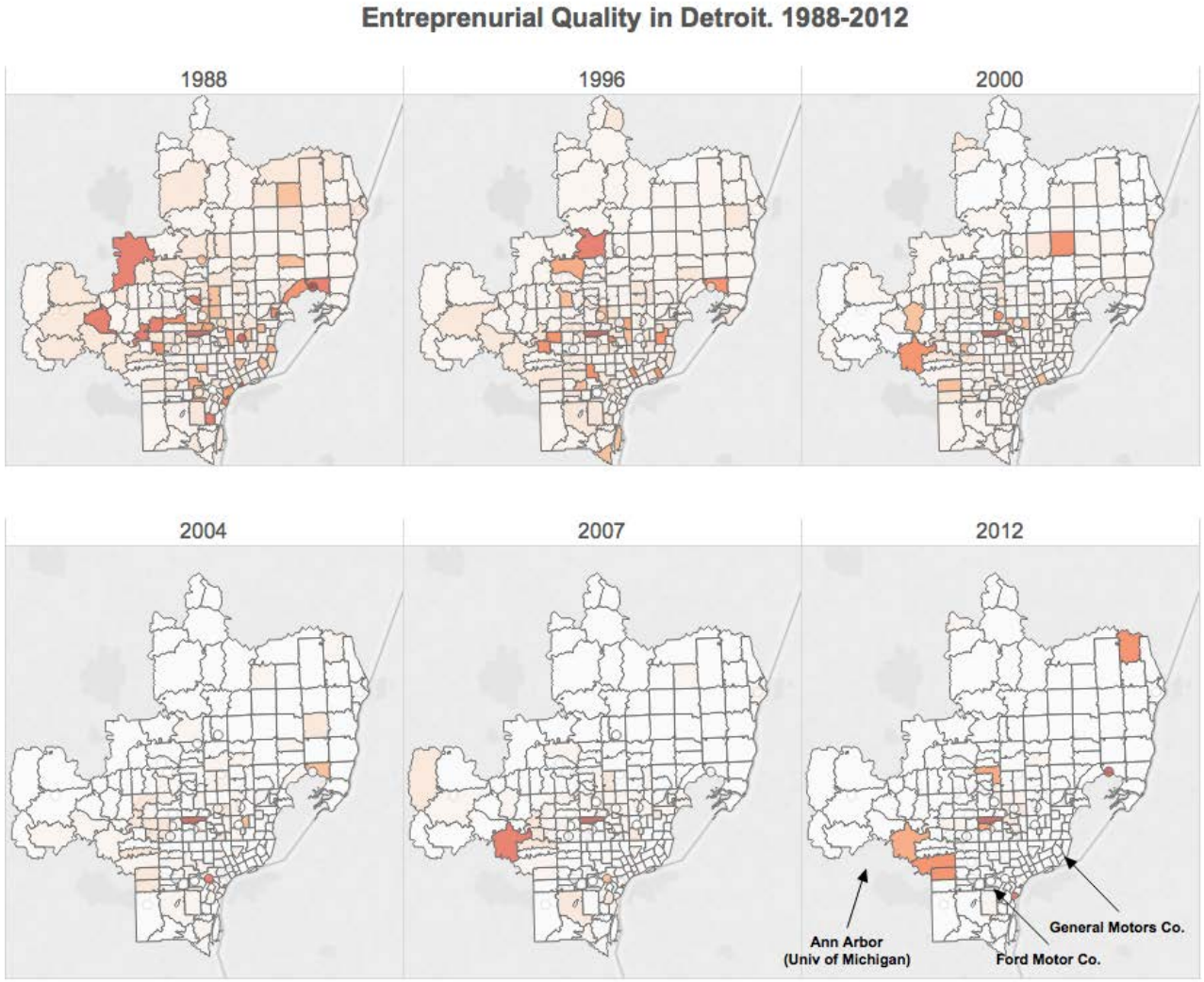
Entrepreneurial Quality of Miami.

ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).



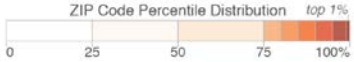
SUPPLEMENTARY MATERIALS AND APPENDIX

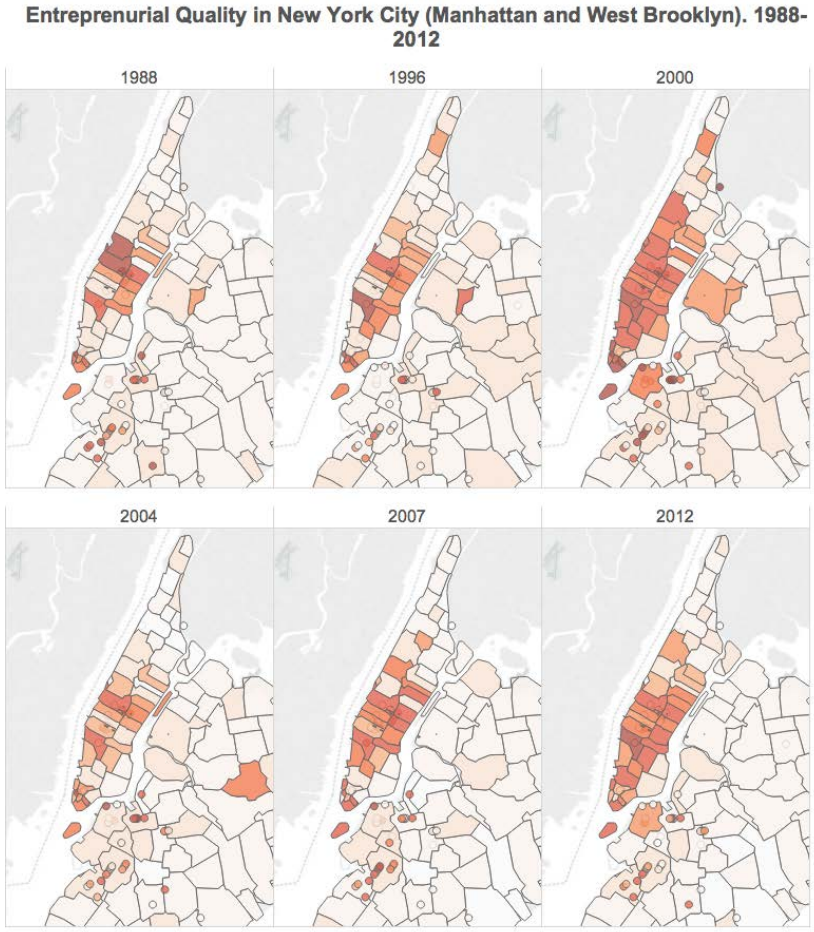
APPENDIX I. CITY GRAPHS



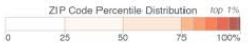
Entrepreneurial Quality of Detroit. Though entreprenuerial quality in Detriot is low throughout the entire time period, it is not stable, having visible peaks and troughs. In 1988 Detroit exhibited small pockets of high quality areas around the center of the city, to the west, and on the north short of Lake St. Clair. This quality had a persistent drop over the next decade and, by 2004, Detriot had dismal levels of quality with (virtually) no entrepreneurial quality anywhere. By 2007, and further in 2012, a small region to the west of Detriot (and close to Ann Arbor and University of Michigan) has emerged with mild levels of quality.

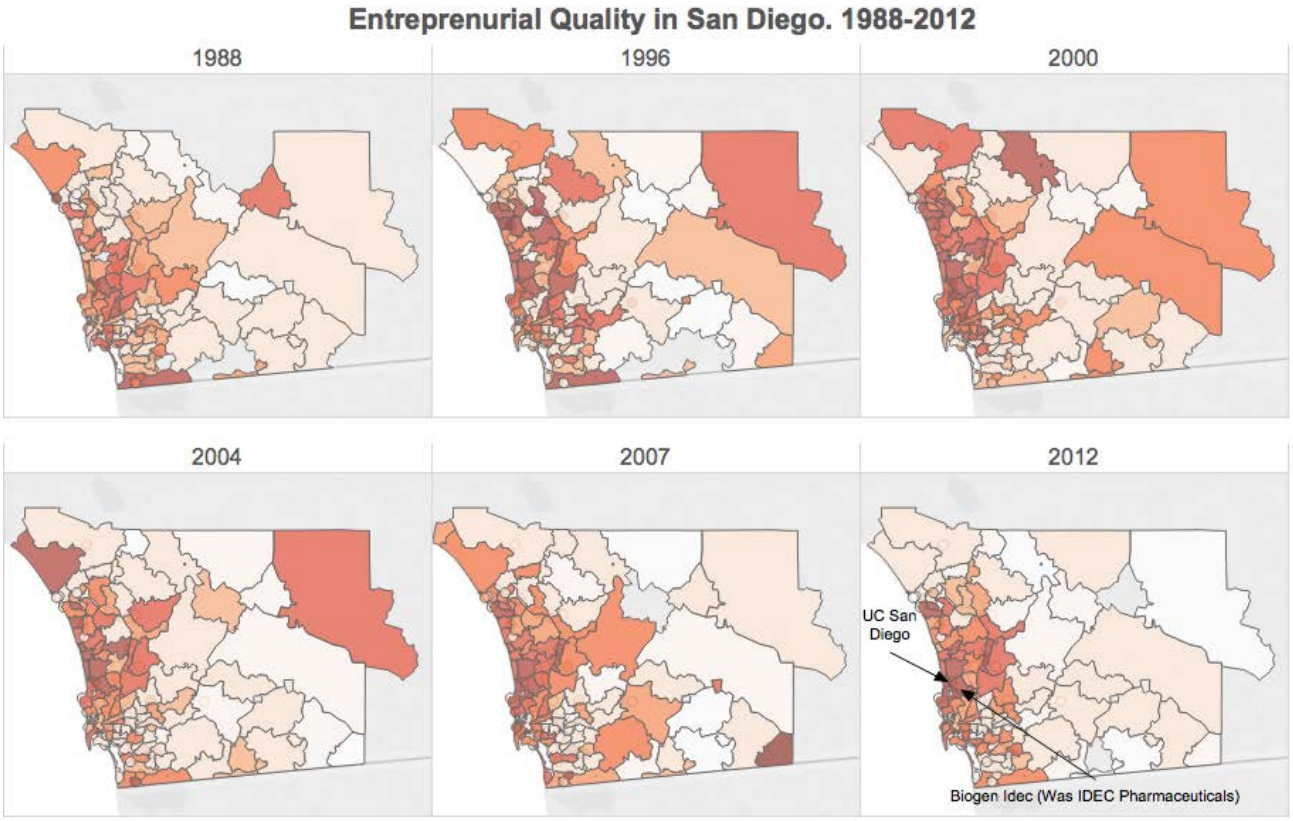
ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).





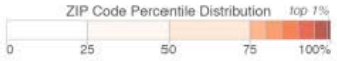
Entreprenurial Quality for Manhattan and Brooklyn. Entrepreneurial quality by ZIP Code for a select number of years between 1988 and 2012. The color represents the average quality of that ZIP Code relative to the full distribution of ZIP Code - Year observations in the aggregated dataset.

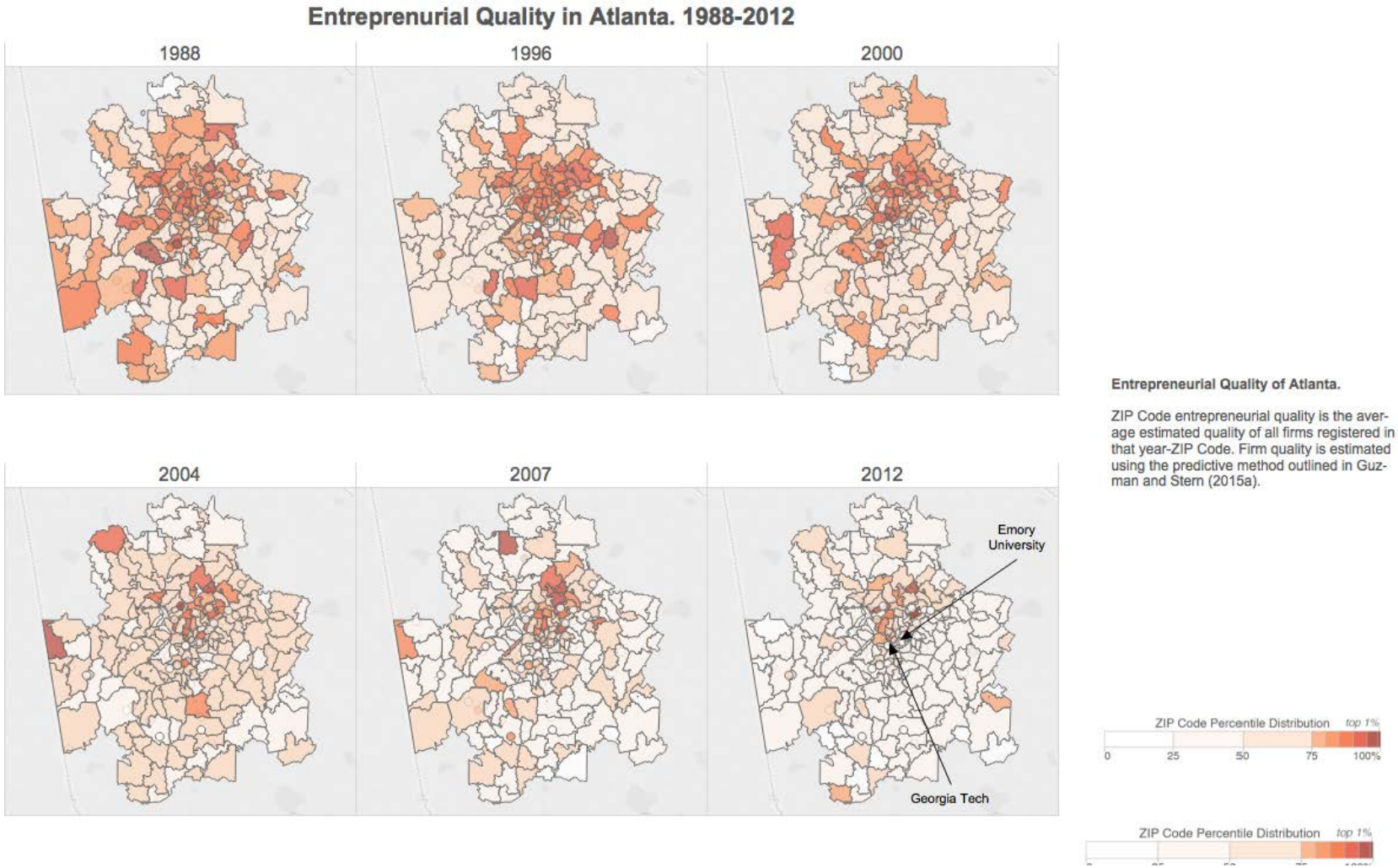




Entrepreneurial Quality of San Diego.

ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).




Entreprenurial Quality in Atlanta. 1988-2012
1988
1996
2000
2004
2007
2012
Emory University
Georgia Tech
Entrepreneurial Quality of Atlanta.
ZIP Code entrepreneurial quality is the average estimated quality of all firms registered in that year-ZIP Code. Firm quality is estimated using the predictive method outlined in Guzman and Stern (2015a).
ZIP Code Percentile Distribution top 1%
0
25
50
75
100%
ZIP Code Percentile Distribution top 1%

APPENDIX II

Evaluating Entrepreneurial Quality Estimates

The strength of the inference we can draw from entrepreneurial quality estimates relies critically on how well predicted quality measures map true firm potential. In this section, we proceed to evaluate such estimates, both at the national level and across states and time periods. A common approach in economics could be to use the estimated R^2 from our model, or compare the distribution of quality vs realized growth outcomes in the training sample. However, a well-known problem in predictive algorithms is over fitting – the case where a model has good predictive capacity within the training sample but has poor predictive power outside such sample. As such, we instead follow from machine learning approaches and test our prediction out of sample, using the 30% random sample of observations which have not been used in the estimation but for which we observe both the growth outcome and start-up characteristics. In particular, using only data from the test sample (but relying on the estimates from Table 4 to estimate entrepreneurial quality), Figure A1 in our appendix presents the relationship between the distribution of realized growth events versus the distribution of firm-level entrepreneurial quality. The results are striking. 70% of all growth firms are in the top 5% of our estimated growth probability distribution, and 52% are within the top 1% (interestingly, these results are extremely similar to the findings for California from Guzman and Stern (2015)). To be clear, growth is still a relatively rare event even among the elite: the average firm within the top 1% of estimated entrepreneurial quality has only a X% chance of realizing a growth outcome.

Even if our model has strong predictive capacity, another potential source of concern could be heterogeneity within subsamples. Specifically, if one state (California) holds a disproportionate

number of growth outcomes, or if growth outcomes occur disproportionately on a small number of years (the late 1990s), it is possible that our model is mostly fitting that region or time-period but does not have the external validity to work outside of the training years and states. If so, our prediction of quality in future years would be poor even if such predictions are good in the sample years.

We begin testing the accuracy across states in Table A1. We perform three different tests. In Column 3, we estimate the share of state growth firms in the top 5% of the state quality distribution using our 30% training sample. All states appear to separate growth firms in an within a small percentage at the top of the distribution[25]. The share of firms in the top 5% is highest in New York (85%) and Oregon (83%), and lowest in Florida (46%) and Washington (62%); California (68%) is only around the median, and there does not appear to be a discernible relationship between this statistic and the distribution of venture capital or high technology clusters. Our second test evaluates to what extent do our observables characterize the growth process in a region. To do so, we re-run our full information model (Model 1 of Table 4) separately for each state and calculate the pseudo-R^2 of each model. Once again, variation in this measure appears to be stable, with our measures having important relationship to growth outcomes in all states. The highest R^2 value is for Massachusetts (32%) and the lowest for Florida (18%), with all other states being between 23% and 31%. Finally, we measure the relationship between entrepreneurial quality estimated from these states specific models to our global quality measure. In column 5 we report the correlation between the two. [26] All correlation measures are high, with

[25] We are unable to estimate this measure for Alaska, Vermont and Wyoming due to the low number of growth firms that the states have.

[26] Another potential approach to test the difference in predictive measures between quality estimated with a state and national model would be to look at the distribution of the difference between these two measures ($d_i = \theta_{i,state} - \theta_i$ and test for $H_0 : d_i = 0$. However, because the state model implicitly includes a state fixed effect this would counfound quality and ecosystem effects.

the highest one being in California (.902) and the lowest in Michigan (.572), all other states are between .740 and .888. In conclusion, while there is variation in state performance each of these three test, we find our estimate of quality with a national index to hold good predictive capacity at the state level.

We repeat the same three tests for each year in table A2. The robustness of our model across years appears to be even higher than the robustness across states. The share of top 5% varies from 59% to 80%, the pseudo R^2 from 21% to 32% and the correlation of predicted quality from .822 to .953. Interestingly both the best predictive accuracy (share in top 5%) and the best fit between our observables and growth do not occur in the late 1990s but in the years 2005 to 2008. Both the stability across a long period of time and the fact that this accuracy appears to be improving gives us confidence in the quality of our predictions in the years following 2008, where growth is unobserved.

TABLE A1

Table A1: Quality of Predictive Algorithm By State (30% Test Sample)

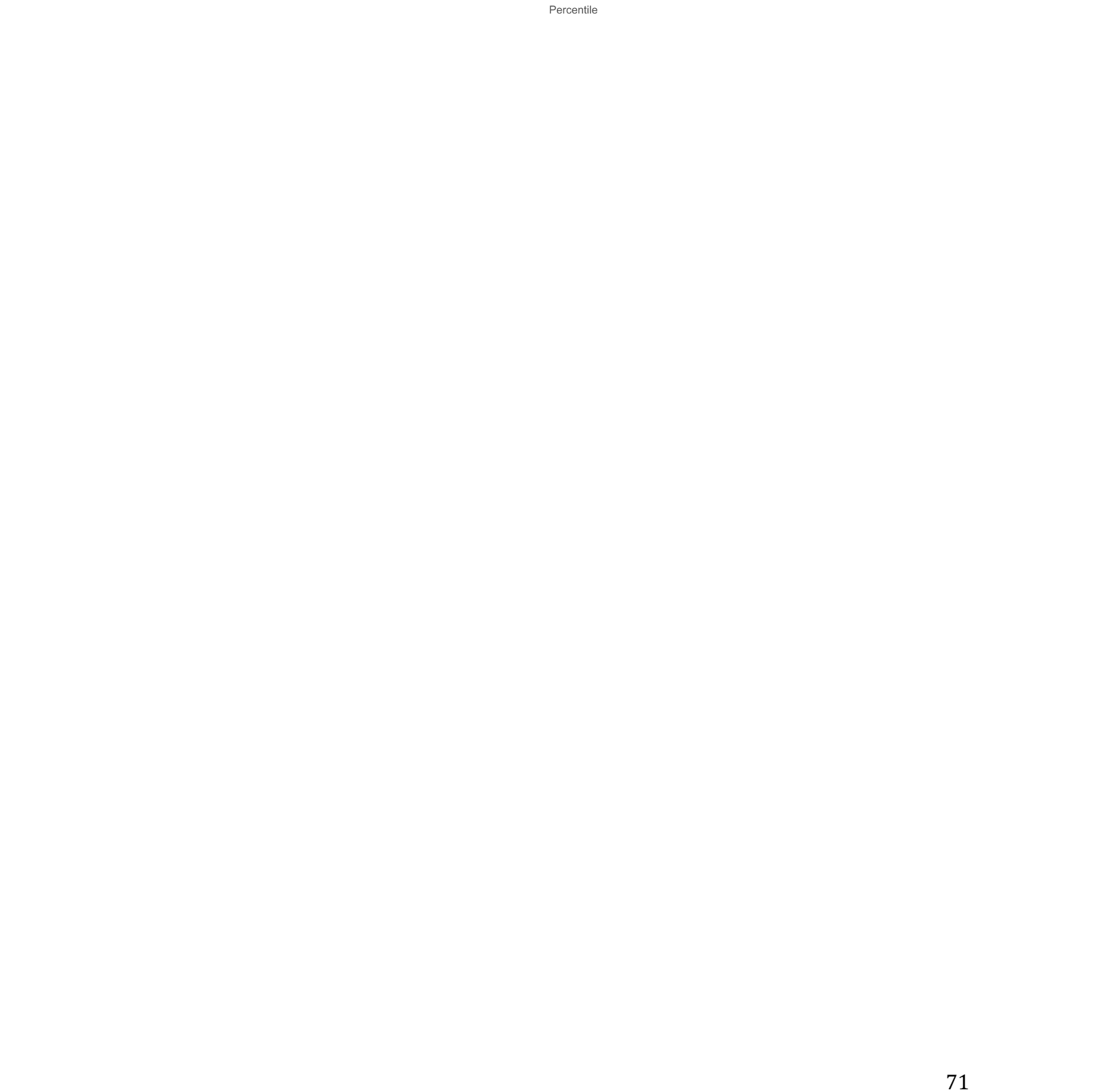
	1	***2***	***3***	***4***	***5***
State	*Total Growth Firms*	*Growth Firms in Training Sample*	*Share of Growth Firms Top 5% of Test Sample*	*Pseudo R^2 of Training Model*	*Correlation with Single State Quality*
Alaska	1	1	-		
California	1587	1126	68%	31%	0.902
Florida	296	226	46%	18%	0.740
Georgia	154	96	60%	26%	0.856
Massachusetts	320	237	76%	32%	0.807
Michigan	91	54	62%	24%	0.572
New York	295	209	85%	26%	0.888
Oregon	40	28	83%	-	-
Texas	536	383	63%	23%	0.818
Vermont	4	3	-	-	-
Washington	153	111	60%	26%	0.788
Wyoming	6	5	-	-	-

TABLE A2

Table A2: Quality of Predictive Algorithm By Cohort (30% Test Sample)

	1	***2***	***3***	***4***	***5***
Cohort Year	*Total Growth Firms*	*Growth Firms in Training Sample*	*Share of Growth Firms Top 5% of Test Sample*	*Pseudo R^2 of Training Model*	*Correlation with Single State Quality*
1995	287	210	60%	21%	0.907
1996	395	278	67%	27%	0.933
1997	367	253	65%	26%	0.920
1998	365	264	59%	25%	0.903
1999	343	240	75%	31%	0.935
2000	281	201	69%	29%	0.953
2001	185	142	72%	28%	0.884
2002	161	120	71%	28%	0.897
2003	161	106	75%	26%	0.895
2004	178	130	69%	25%	0.822
2005	185	124	79%	33%	0.894
2006	185	132	77%	32%	0.851
2007	221	155	80%	32%	0.880
2008	169	124	78%	27%	0.891

FIGURE A1



DATA APPENDIX

Our data set is drawn from the complete set of business registrants in California, Florida, Massachusetts, Texas, and Washington state, from 1988 to 2014. Our analysis draws on the complete population of firms satisfying one of the following conditions: (i) a for-profit firm whose jurisdiction is in the source state or (ii) a for-profit firm whose jurisdiction is in Delaware but whose principal office address is in the state. The resulting data set is composed of 11,277,643 observations. For each observation, we construct variables related to (i) the growth outcome for the startup, (ii) measures based on business registration observables and (iii) measures based on external observables that can be linked to the startup.

Growth outcome. The growth outcome utilized in this paper, *Growth*, is a dummy variable equal to 1 if the startup achieves an initial public offering (IPO) or is acquired at a meaningful positive valuation within 6 years of registration. Both outcomes, IPO and acquisitions, are drawn from Thomson Reuters SDC Platinum[27]. Although the coverage of IPOs is likely to be nearly comprehensive, the SDC data set excludes some acquisitions. However, although the coverage of significant acquisitions is not universal in the SDC data set, previous studies have "audited" the SDC data to estimate its reliability, finding a nearly 95% accuracy (Barnes, Harp, and Oler, 2014). We observe 2,427 positive growth outcomes for the 1995–2005 start-up cohorts), yielding a mean for *Growth* of 0.0006. In our main results, we assign acquisitions with an unrecorded acquisitions price as a positive growth outcome, because an evaluation of those deals suggests that most reported acquisitions were likely in excess of $5 million. In prior analyses, we drop firms where the acquisition price is not reported; neither the pattern of coefficient estimates nor our qualitative findings regarding the geography of growth entrepreneurship is affected by this choice. All of our results are also robust to including firms registered in Delaware into cohorts associated with their initial Delaware registration date (rather than their state registration date).

Start-up characteristics. The core of the empirical approach is to map growth outcomes to observable characteristics of start-ups at or near the time of business registration. We develop two types of measures: (i) measures based on business registration observables and (ii) measures based on external indicators of start-up quality that are observable at or near the time of business registration. We review each of these in turn.

Measures based on business registration observables. We construct six measures of start-up quality based on information directly observable from the business registration record. First, we create binary measures related to how the firm is registered, including *corporation,* whether the firm is a corporation (rather than partnership or LLC) and *Delaware jurisdiction*, whether the firm is incorporated in Delaware. C*orporation* is an indicator equal to 1 if the firm is registered as a corporation and 0 if it is registered either as an LLC or partnership.[28] In the period of 1995 to 2005, 0.08% of corporations achieve a growth outcome versus only 0.02% of noncorporations. *Delaware jurisdiction* is equal to 1 if the firm is registered under Delaware, but has its main office in the source state (all other foreign firms are dropped before analysis). Delaware jurisdiction is favorable for firms which, due to more complex operations, require more certainty in corporate law, but it is associated with extra costs and time to establish and maintain two registrations.

[27] Thomson Reuters's SDC Platinum is a commonly used database of financial information. More details are available at http://thomsonreuters.com/sdc-platinum/

[28] Previous research highlights performance differences between incorporated and unincorporated entrepreneurs (Levine and Rubinstein, 2013).

Between 2001 and 2006, 2.8% of the sample registers in Delaware; 57% of firms achieving a growth outcome do so.

Second, we create four measures that are based on the name of the firm, including a measure associated with whether the firm name is eponymous (named after the founder), is short or long, is associated with local industries (rather than traded), or is associated with a set of high-technology industry clusters.

Drawing on the recent work of Belenzon, Chatterji, and Daley (2014) (BCD), we use the firm and founder name to establish whether the firm name is eponymous (i.e., named after one or more of the founders). *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.[29] We require names be at least four characters to reduce the likelihood of making errors from short names. Our results are robust to variations of the precise calculation of eponymy (e.g., names with a higher or lower number of minimum letters). We have also undertaken numerous checks to assess the robustness of our name matching algorithm. 13.5% of the firms in our training sample are eponymous [an incidence rate similar to BCD], though less than 2% for whom *growth* equals one. It is useful to note that, while we draw on BCD to develop the role of eponymy as a useful start-up characteristic, our hypothesis is somewhat different than BCD: we hypothesize that eponymous firms are likely to be associated with lower entrepreneurial quality. Whereas BCD evaluates whether serial entrepreneurs are more likely to invest and grow companies which they name after themselves, we focus on the cross-sectional difference between firms with broad aspirations for growth (and so likely avoid naming the firm after the founders) versus less ambitious enterprises, such as family-owned "lifestyle" businesses.

Our second measure relates to the length 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."). Companies such as Google or Spotify have sharp and distinctive names, whereas more traditional businesses often have long and descriptive names (e.g., "Green Valley Home Health Care & Hospice, Inc."). We define *short name* to be equal to one if the entire firm name has three of less words, and zero otherwise. 46.65% of firms within the 1995-2005 period have a short name, but the incidence rate among growth firms is more than 80%. We have also investigated a number of other variants (allowing more or less words, evaluating whether the name is "distinctive" (in the sense of being both noneponymous and also not an English word). While these are promising areas for future research, we found that the three-word binary variable provides a useful measure for distinguishing entrepreneurial quality.

We then create four measures based on how the firm name reflects the industry or sector that the firm within which the firm is operating. To do so, we take advantage of two features of the US Cluster Mapping Project (Delgado, Porter, and Stern, 2015), which categorizes industries into (a) whether that industry is primarily local (demand is primarily within the region) versus traded (demand is across regions) and (b) among traded industries, a set of 51 traded clusters of industries that share complementarities and linkages. We augment the classification scheme from the US Cluster Mapping Project with the complete list of firm names and industry classifications contained in Reference USA, a business directory containing more than 10 million firm names and industry codes for companies across the United States. Using a random sample of 1.5 million

[29] For corporations, we consider top managers only the current president, for partnerships and LLCs, we allow for any of the two listed managers. The corporation president and two top partnership managers are listed in the business registration records themselves.

Reference USA records, we create two indices for every word ever used in a firm name. The first of these indices measures the degree of localness, and is defined as the relative incidence of that word in firm names that are in local versus non-local industries (i.e., $\rho_i = \frac{\sum_{j=\{\text{local firms}\}} 1[w_i \subseteq \text{name}_j]}{\sum_{j=\{\text{non-local firms}\}} 1[w_i \subseteq \text{name}_j]}$). We then define a list of Top Local Words, defined as those words that are (a) within the top quartile of □□ and (b) l
0.01% within the population of firms in local industries (see Guzman and Stern, (2015, Table S10) for the complete list). Finally, we define local to be equal to one for firms that have at least one of the Top Local Words in their name, and zero otherwise. We then undertake a similar exercise for the degree to which a firm name is associated with a traded name. It is important to note that there are firms which we cannot associate either with traded or local and thus leave out as a third category. Just more than 18% of firms have local names, though only 4.7% of firms for whom growth equals one, and while 55% of firms are associated with the traded sector, 60% of firms for whom growth equals one do.

We additionally examine the type of traded cluster a firm is associated with, focusing in particular on whether the firm is in a high-technology cluster or a cluster associated with resource intensive industries. For our high technology cluster group (Traded High Technology), we draw on firm names from industries include in ten USCMP clusters: Aerospace Vehicles, Analytical Instruments, Biopharmaceuticals, Downstream Chemical, Information Technology, Medical Devices, Metalworking Technology, Plastics, Production Technology and Heavy Machinery, and Upstream Chemical. From 1995 to 2005, while only 5% firms are associated with high technology, this rate increases to 16% within firms that achieve our growth outcome. For our resource intensive cluster group, we draw on firms names from fourteen USCMP clusters: Agricultural Inputs and Services, Coal Mining, Downstream Metal Products, Electric Power Generation and Transmission, Fishing and Fishing Products, Food Processing and Manufacturing, Jewelry and Precious Metals, Lighting and Electrical Equipment, Livestock Processing, Metal Mining, Nonmetal Mining, Oil and Gas Production and Transportation, Tobacco, Upstream Metal Manufacturing. While 14% of firms are associated with resource intensive industries, and 13% amongst growth firms.

Finally, we also repeat the same procedure to find firms associated with more narrow sets of clusters that have a closer linkage to growth entrepreneurship in the United States. We specifically focus on firms associated to Biotechnology, E-Commerce, Information Technology, Medical Devices and Semiconductors. It is important to note that these definitions are not exclusive and our algorithm could associate firms with more than one industry group. For Biotechnology (Biotechnology Sector), we use firm names associated with the US CMP Biopharmaceuticals cluster. While only 0.21% of firms are associated with Biotechnology, this number increases to 2.9% amongst growth firms. For E-commerce (E-Commerce Sector) we focus on firms associated with the Electronic and Catalog Shopping sub-cluster within the Distribution and Electronic Commerce cluster. And while 5% of all firms are associated with e-commerce, the rate is 9.5% for growth firms. For Information Technology (IT Sector), we focus on firms related to the USCMP cluster Information Technology and Analytical Instruments. 2.6% of all firms in our sample are associated with IT, and 15% of all growth firms are identified as IT-related. For Medical Devices (Medical Dev. Sector), we focus on firms associated with the Medical Devices cluster. We find that while 3.3% of all firms are in medical devices, this number increases to 9% within growth firms. Finally, for Semiconductors (Semiconductor Sector), we focus on the sub-

cluster of Semiconductors within the Information Technology and Analytical Instruments cluster. Though only 0.06% of all firms are associated with semiconductors, 1.15% of growth firms are.

Measures based on external observables. We construct two measures related to start-up quality based on information in intellectual property data sources. Although this paper only measures external observables related to intellectual property, our approach can be utilized to measure other externally observable characteristics that may be related to entrepreneurial quality (e.g., measures related to the quality of the founding team listed in the business registration, or measures of early investments in scale (e.g., a Web presence).

Building on prior research matching business names to intellectual property (Balasubramanian and Sivadasan, 2009; Kerr and Fu, 2008), we rely on a name-matching algorithm connecting the firms in the business registration data to external data sources. Importantly, because we match only on firms located in California, and because firms names legally must be "unique" within each state's company registrar, we are able to have a reasonable level of confidence that any "exact match" by a matching procedure has indeed matched the same firm across two databases. In addition, our main results use "exact name matching" rather than "fuzzy matching"; in small-scale tests using a fuzzy matching approach [the Levenshtein edit distance (Levenshtein, 1965)], we found that fuzzy matching yielded a high rate of false positives due to the prevalence of similarly named but distinct firms (e.g., Capital Bank v. Capitol Bank, Pacificorp Inc v. Pacificare Inc.).

Our matching algorithm works in three steps.

First, we clean the firm name by:

- expanding eight common abbreviations ("Ctr.", "Svc.", "Co.", "Inc.", "Corp.", "Univ.", "Dept.", "LLC.") in a consistent way (e.g., "Corp." to "Corporation")
- removing the word "the" from all names
- replacing "associates" for "associate"
- deleting the following special characters from the name: . | ' " - @ _

Second, we create measures of the firm name with and without the organization type, and with and without spaces. We then match each external data source to each of these measures of the firm name. The online appendix contains all of the data and annotated code for this procedure.

This procedure yields two variables. Our first measure of intellectual property captures whether the firm is in the process of acquiring patent protection during its first year of activity. *Patent* is equal to 1 if the firm holds a patent application in the first year. All patent applications and patent application assignments are drawn from the Google U.S. Patent and Trademark Office (USPTO) Bulk Download archive. We use patent applications, rather than granted patents, because patents are granted with a lag and only applications are observable close to the data of founding. Note that we include both patent applications that were initially filed by another entity (e.g., an inventor or another firm), as well as patent applications filed by the newly founded firm. While only 0.25% of the firms in 1995–2005 have a first-year patent, 25% of growth firms do.

Our second intellectual property measure captures whether a firm registers a trademark during its first year of business activity. *Trademark* is equal to 1 if a firm applied for a trademark within the first year, and 0 otherwise. We build this measure from the Stata-ready trademark DTA file developed by the USPTO Office of Chief Economist (Graham et al, 2013). Between 1995 and 2005, 0.11% of all firms register a trademark, while 8% of growth firms do.

REFERENCES

N. Balasubramanian, J. Sivadasan, (2009) “NBER Patent Data-BR Bridge: User guide and technical documentation” *SSRN Working paper #1695013*

B. Barnes, N. Harp, D. Oler, (2014) “Evaluating the SDC mergers and acquisitions database” *SSRN Working paper #2201743*

S. Belenzon, A. Chatterji, B. Daley. (2014) “Eponymous entrepreneurs” Working paper,

M. Delgado, M. Porter, S. Stern, (2014) “Defining clusters in related industries” *NBER Working Paper #20375*

S. Graham, G. Hancock, A. Marco, A. F. Myers, (2013) “The USPTO case files data set: Descriptions, lessons and insights” *SSRN Working Paper #2188621*

W. R. Kerr, Shihe Fu, (2008) "The Survey of Industrial R&D--Patent Database Link Project." *J. Technol. Transf.* **33**, no. 2

V. I. Levenshtein, (1965) "Binary codes capable of correcting deletions, insertions, and reversals." *Doklady Akad. Nauk SSSR* **163**(4): 845–848

R. Levine, Y. Rubinstein, (2013) “Smart and illicit: Who becomes an entrepreneur and does it pay?” *NBER Working Paper #19276*