Understanding Entrepreneurship:
Facilitating Academic Research with a Shared Data Repository*

Gregory W. Brown
UNC Chapel Hill, Kenan-Flagler Business School
gregwbrown@unc.edu

Paige P. Ouimet
UNC Chapel Hill, Kenan-Flagler Business School
Paige_Ouimet@unc.edu

David T. Robinson
Duke University, Fuqua School of Business
davidr@duke.edu

Ted Zoller
UNC Chapel Hill, Kenan-Flagler Business School
zoller@unc.edu

February 18, 2017

Keywords: Entrepreneurship, Start-ups, Venture Capital, Social Science Data

JEL Classification: G24, L26, O3,

* The authors thank Ashley Brown, David Borland, Huan Lian, and Jack Walker for the tremendous database and programming support they have provided for this project.

Executive Summary

UNC's Kenan Institute of Private Enterprise and the Duke University Innovation and Entrepreneurship (I&E) initiative have embarked on a joint initiative to build a data repository to facilitate empirical research in entrepreneurship. This paper outlines our motivation for this project as well as our long-term goals of promoting empirical research by lowering the costs to researchers of data collection and aggregation. We also outline the data that we have currently available and demonstrate the value of a resource like by documenting recent trends in startup activity in the US.

To facilitate easier access to the data, we have developed a web portal with data related to entrepreneurial research to facilitate further inquiry into entrepreneurial activity. Our long-term goal is to generate a data resource, updated regularly, that is as comprehensive as possible. Allowing researchers to easily contribute new data to the repository should also facilitate a shared understanding of how various factors and findings relate to one another.

An Overview of the Data

Background and Motivation

Entrepreneurship as a field of study has come of age. Not only is new business creation widely recognized as a critical element of economic growth by policy makers, financial institutions and business leaders, the phenomenon has attracted the interest of highly respected academics doing quality research. Increasingly, top academic journals in fields as varied as economics, finance, strategy, sociology, and public policy feature research in entrepreneurship and innovation.

At the same time, entrepreneurship as a research area also faces challenges related to its broadening importance. The study of entrepreneurship is in practice the study of how various factors related to traditional fields (finance, strategy, policy, etc.) and institutions (government, financial, educational, social, etc.) interact to generate a particular type of economic activity. Despite the inherent interdisciplinary nature of the field, academic researchers remain stuck in academic silos. There are important institutional reasons for this that no single team of researchers can change: specialized research skills and knowledge, tenure requirements that reward field-focused research and publication, as well as university and business school organizational structures all help to reinforce the siloed nature of academic research in the field of entrepreneurship.

Nevertheless, there are other impediments that we can change. Our effort is to address the lack of a centralized source for commonly used research data. This lack of a common source means researchers typically must re-invent the "data wheel" when wanting to extend previous research. It also leads to ambiguity when interpreting results from empirical studies that use similar measures but different data sources. When differences arise, it can be unclear if inconsistencies stem from changes in methodology or the underlying data.

To fill this gap, we have created a web-based, central data repository for critical data in entrepreneurship—a "one-stop shop" for entrepreneurship researchers. The data currently available can be accessed at:

http://www.kenaninstitute.unc.edu/EntrepreneurshipData/

The web-based interface allows users to select among variables, geographies, and dates. Once the desired data are selected, a custom dataset is generated and a web link is emailed to the user.

Our website features the ability to download time-series data panels for different measures of entrepreneurial activity, allowing the end user to vary the desired date-range and geographic focus. This facilitates easy comparison across different measures as well as the ability to detect trends in the time-series or cross-section.

In the current version of our repository, we have limited ourselves to the United States and started with two primary segments: state and metropolitan statistical area (MSA). State-level measures have the benefit of consistent definitions over time. However, the coarseness associated with using such large geographic bins may hamper the ability to investigate questions more specific to microclimates. Thus, we also include MSA-level data. We use the term MSA broadly. We have included micropolitan statistical areas, when available, and are working on expanding this to even more geographical definitions, such as counties and commuting zones.

In addition to access to raw data files, the database will provide data visualization tools to assist researchers in better understanding individual variables as well as relations between variables. To date, we have partnered with the Renaissance Computing Institute (RENCI.org) to provide prototypes of two tools (for a subset of our data). These are available at:

http://kipe.renci.org/streamGraph/
http://kipe.renci.org/connectedScatterplot/

The stream-graph can be utilized to observe the relative time variation in a large cross-section of states or MSAs. For example, clicking on the link and selecting "Firm age: 1" and "Metric: Job_Creation_Birth" reveals the dramatic time-series variation in job creation over time that can be attributed to firms 1-year old. The overall height of the graph is the total number and each band is the contribution from a specific state. For example, scrolling over the graph around 2007 reveals the significant contribution from Florida well after job creation in the rest of the country peaked. The recent low number suggests less job creation since the financial crisis from new firms.

Following the link to the connected scatterplot and completing these steps reveals interesting patterns in job creation by new firms by state:

1. Select "None" for states
2. Select Firm age = 2
3. X-axis = Estabs_Entry
4. Y-axis = Job_Creation_Births
5. Sequentially choose states: NC, MA, NY, TX, FL, CA, All

There is a clear (and expected) positive relation between new firms and job creation. However, this relation varies by year and state—especially to the upside where in most states there will be certain years where the number of jobs per firm jumps significantly. Perhaps most dramatically, the graph reinforces the importance of a small number of states for both the number of new firms as well as their impact on employment.

Defining Entrepreneurship

Given there is no standard definition of a start-up company or overall start-up activity, we strive to include a broad set of measures from many sources. From the Census Business Dynamics Statistics, we have collected data on firm counts by age by geography and year. Since all start-ups are not equal, we have also collected counts of new firms by initial employment for each geography-year pair. We have also collected data on the number of firms in a given industry and by geography-year. Another way to think about start-up activity is to look at the number of individuals employed at start-ups. We collect these data using the quarterly workforce indicators (QWI) created by the LEHD program at the US Census. We collected both the total count of employees at firms by age and geography-year as well as the fraction of total employment by firm age for each of these geography-year pairs. We have also included valuable data from past Kauffman surveys including the rate of new entrepreneurs and the opportunity share of new entrepreneurs.

Factors related to Startup Activity

Along with the data on entrepreneurship activity, we have also collected a broad set of variables that previous research has suggested may be correlated with start-up intensity. We group these variables into seven categories:

1. Business Climate and Government;
2. Demographics;
3. Economic;
4. Entrepreneurship;
5. Finance;
6. Geographic;
7. R&D and Technology.

Ultimately these data allow us to measure important determinants of start-up activity such as measures of GDP, government spending and government revenues.[1] We include information on the age distribution of workers, as well as rates of immigration, number of scientists, and information on mean wages for different categories of workers.[2] The dataset includes information on exports, headquarters of large public companies, manufacturing intensity and strategic industry clusters.[3] We have added data on investment rates, grants and income.[4,5] We have collected

[1] Moretti and Wilson (2014) show a link between state subsidies and growth in the biotech sector.

[2] Bonte, Falck and Heblich (2009), Glaeser and Kerr (2009) and Ouimet and Zarutskie (2014) find that the age distribution of the local population is correlated with startup activity.

[3] Delgado, Porter and Stern (2010) documents the important role of industry clusters on entrepreneurship.

[4] Samilia and Sorenson (2011) find that VC investment in previous periods predicts new startup activity.

[5] Zucker, Darby and Brewer (1998) finds a correlation between federal grants and the count of biotech startups.

measures of housing supply, livability indices and statistics on crime and religion.[6] We also have estimates of R&D spending and counts of patents.[7] All of these data are collected at the geography-year level and with additional time and resources we hope be able to expand on this already considerable collection.

With a common set of variables available to all researchers, our work is intended to facilitate additive research efforts. Our goal is to assist researchers as they will no longer be required to collect and clean data from disparate sources when wishing to replicate earlier results. For example, a researcher exploring the relationship between tax incentives and start-up activity could, with the help of our website, also easily include controls for demographic factors. The full set of data variables are too numerous to list here but we have appended a technical document that provides further information and individual "read me" files that are specific to each data variable.

An Application of the Data

To illustrate the utility of a data repository such as this, we use our data to examine a fundamental question of current interest to policy makers, academics and practitioners alike: What is the current state of entrepreneurship in the United States? This question has generated considerable debate due to the well-documented importance of start-up firms driving innovation and growth in the real economy. As we discuss in detail below, several important indicators of start-up activity have exhibited troubling declines since the financial crisis of 2008-2009. However, other key measures of entrepreneurial activity, such as funding activity of young technology firms, have grown substantially over the same period. Understanding the basic trends, let alone the determinants, of entrepreneurial activity remains difficult because of a paucity of data and the lack of a common database utilized by empirical researchers examining this issue.

Using data collected to date, we have begun our own investigation exploring the status of entrepreneurial activity in the U.S. We document a 30% decline nationally in the number of new firms between 2006 to 2010, as measured by the BDS. Figure 1 plots both the national count of new firms (bars) as well as counts for key individual states (lines) across time. The national pattern is mirrored in California, North Carolina, Washington and, to a lesser extent, New York and Massachusetts. Since 2010, we observe a rebound in these rates in California, however in all states and nationally, the rate of new firm formation is well below the pre-crisis peak.

[6] Regional patterns in religion have been shown to influence the quantity and quality of startups in Doms, Lewis, and Robb (2010)

[7] Aharonson, Baum and Feldman (2007) show local R&D spending is correlated with startup activity.

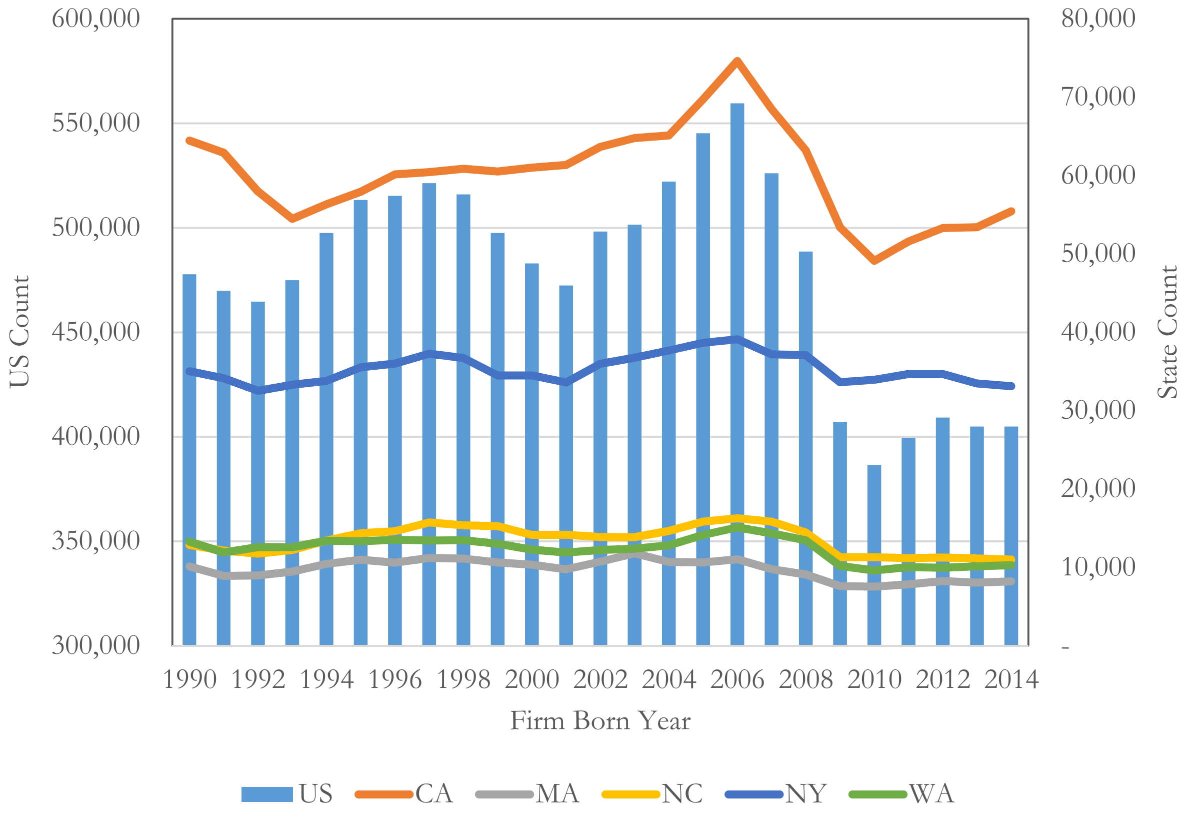

Figure 1. BDS Count of New Firms. This figure plots the count of startups at the national level (bars) or for selected states (lines). The left-hand scale captures national (US) counts; the right-hand scale captures state-level counts. States included in the plot are California (CA), Massachusetts (MA), North Carolina (NC), New York (NY), and Washington (WA). A startup is defined as a firm less than one year of age. These data are from the U.S. Census Business Dynamics Statistics (BDS).

Figure 1 illustrates an important feature of our database. In particular, the data are structured in a manner that facilitates state-level comparisons over time, as well as comparisons between state-level trends and national trends.

We find a similar trend when looking at the share of employment in startups. Figure 2 plots the fraction of workers at firms less than 2 years old (age=0-1) for the U.S. as a whole (bars) and for select states (lines), as estimated by the QWI. Looking at employment through 2015, we see no evidence of a rebound following the crisis. This decline in employment share is all the more surprising given the historical tendency for small firms to generate significant employment during the early stages of prior economic recoveries.

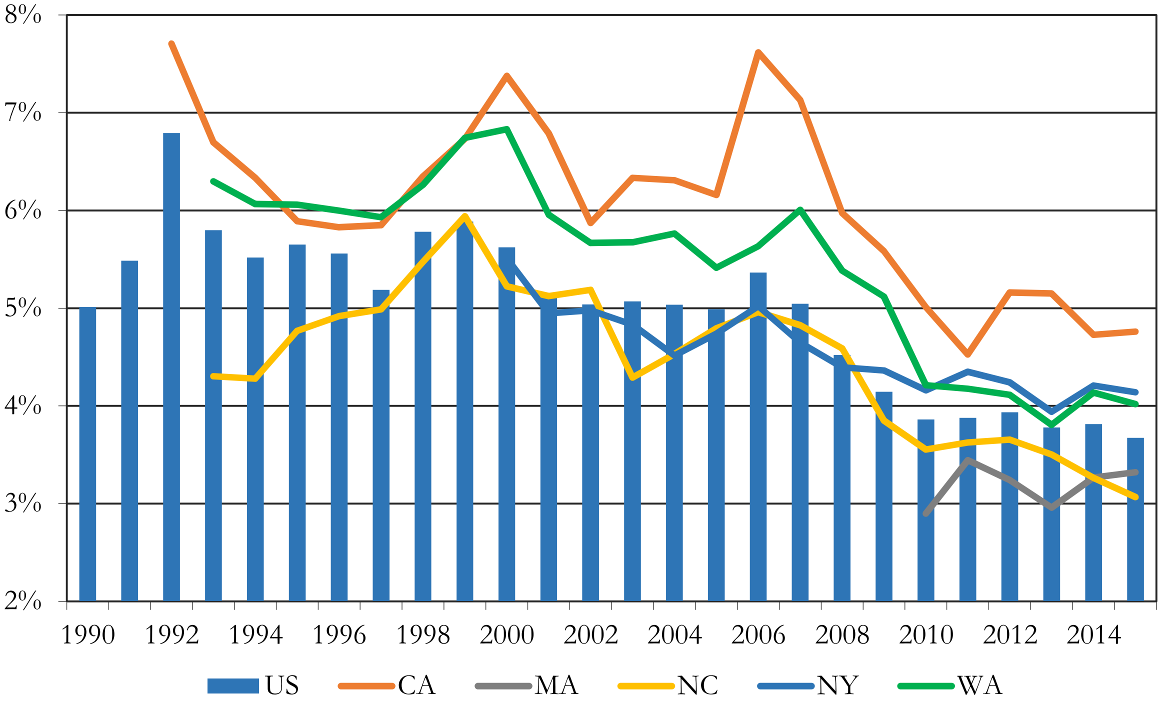

Figure 2. Share of Employment at Startups. This figure plots the percentage of total employment at start-up firms at the national level (bars) and for selected states (lines). Employment is measured as of the first quarter for each year. A startup is defined as a firm less than two years of age. Coverage of states increases over time and the national data are calculated using the summary of all available data for a given year. These data are obtained from the U.S. Census Quarterly Workforce Indicators (QWI).

The declines in activity documented in Figures 1 and 2 do not tell the whole story though. In fact, the trends look completely different when we examine measures of investment rates going into the entrepreneurial sector. Figure 3 plots investment by venture capital (VC) firms for the U.S. as a whole (bars) and select states (lines), as measured by Thomson Reuters and prepared by SSTI. VC investment dollars show a dip around the financial crisis but then a strong recovery with 2015 posting the third highest national investment level in the last two decades. Levels in Massachusetts and New York—two regions that have historically been known as VC hubs—also rebounded sharply in 2014-2015.

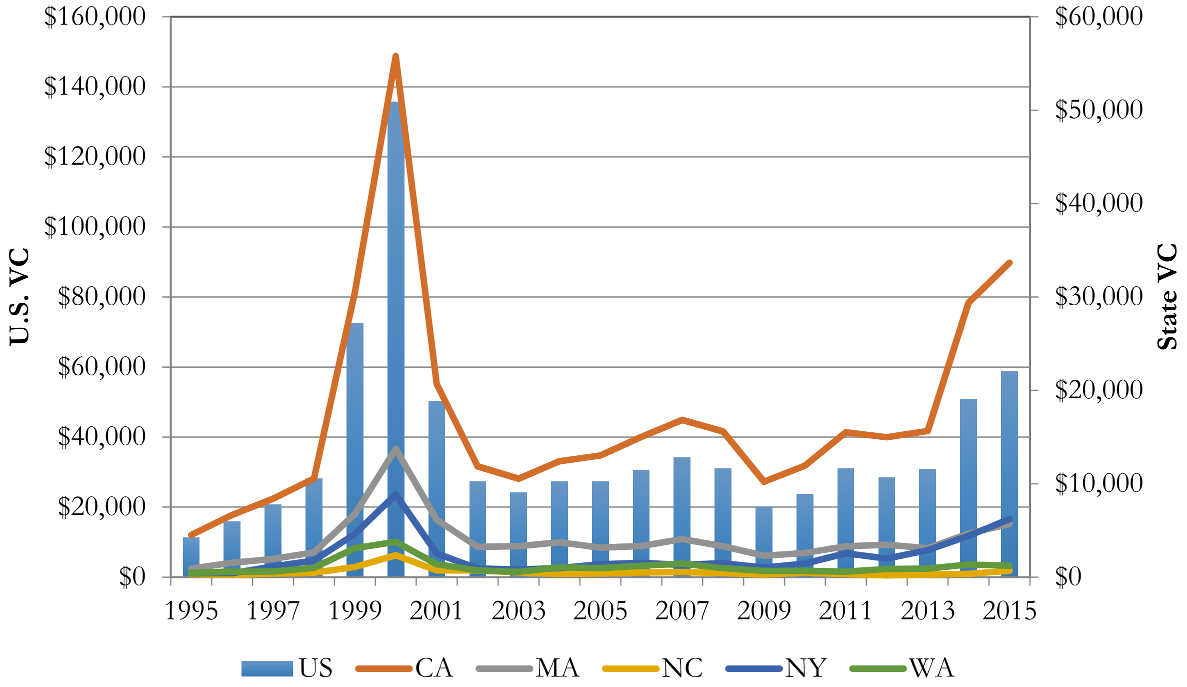

Figure 3. VC Dollars Invested. This figure plots dollar investments by venture capital funds at the national level (bars) and for select states (lines) in millions of USD. The left-hand scale is for national rates of investment; the right-hand scale is for state level investment. All investments are inflation adjusted and reported in 2015 dollars using the CPI-All Items index. Data are from Thomson Reuters and prepared by SSTI.

While only a small percentage of startups will ever receive VC funding, it is a valuable barometer of entrepreneurial activity as VC dollars are targeted towards those startups with the highest growth potential. Another measure of high growth startups is the count of initial public offerings of equity (IPOs).[8] IPO rates can vary substantially year-by-year, as shown in Figure 4. There is a dramatic decline in IPO rates in 2001 after the technology boom turned into a bust. Rates increased again in the 2000s before falling in 2008-2009 during the crisis. Since 2009, IPO activity improved so that 2014 experienced the most activity since 2000 (though 2015 experienced a decline).

[8] IPO rates are used as a measure of successful startups in Guzman and Stern (2015).

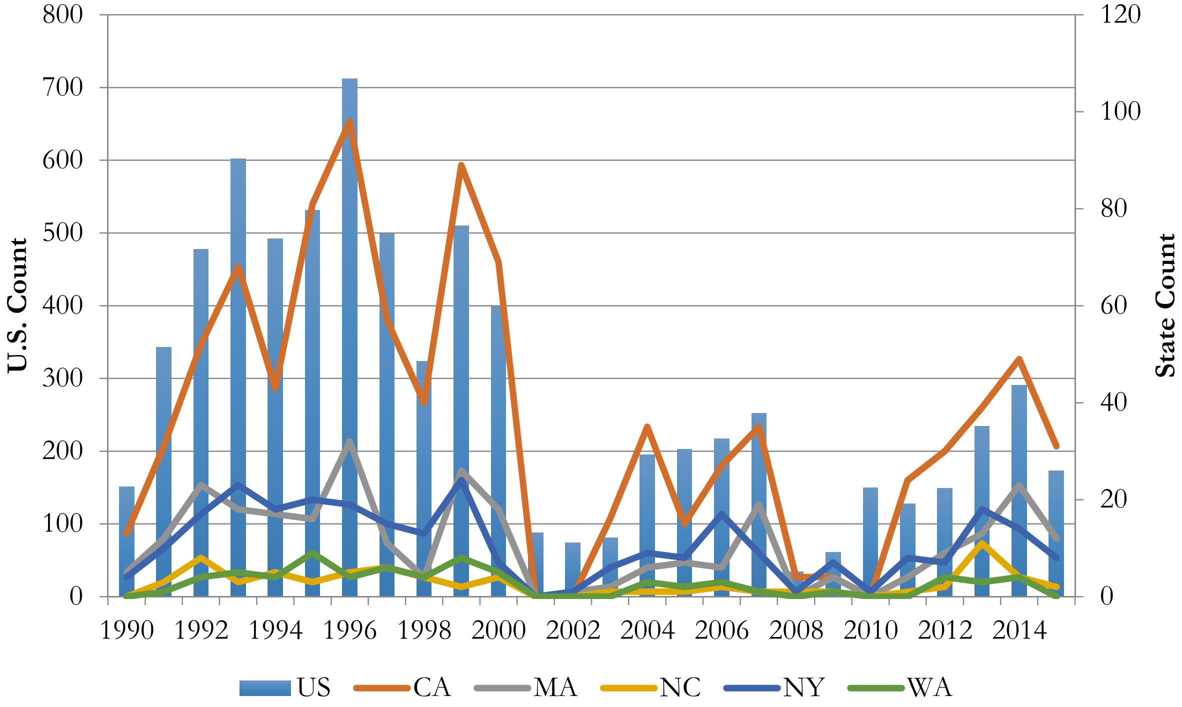

Figure 4. Count of IPOs. This figure plots the number of initial public offerings (IPOs) at the national level (bars) and for select states (lines). The left-hand scale is for national counts; The right-hand scale is for state counts. Data are from Loughran and Ritter (2004). IPOs are matched to states using the location of the firm's headquarters as identified by Compustat.

The conflicting measures of entrepreneurial activity suggest that one overall measure may hide important variation within types of startups. VC dollars and IPO activity tend to be clustered in high-tech and biotech industries, suggesting different patterns in startup activity across industries. Alternatively, the data suggest that a more efficient resource allocation has led to more successful startups even if there are fewer total startups.

To explore whether the mix of startup activity has changed following the crisis, we look at survival rates. We measure the two-, three- and five-year survival rates for new firms born in a given year. We find a striking pattern that while fewer firms were established after the Great Recession, the new firms which were established have much higher survival rates. In fact, as shown in Figure 5, the three-year survival rate of new establishments started in 2011 is the highest over the past two decades.

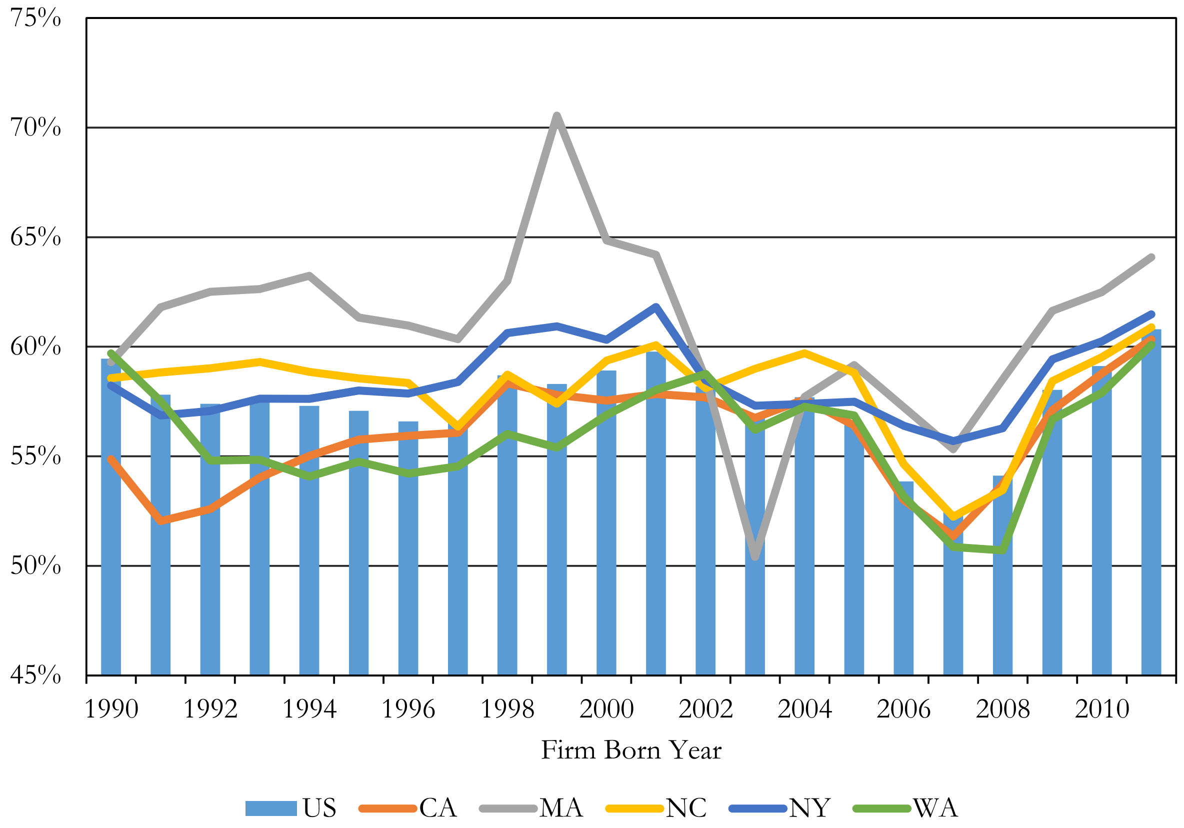

Figure 5. Start-ups 3-Year Survival Rate. This figure plots the percent of startups established in a given year that survived for at least three years, as measured at the national level (bars) or for selected states (lines). These data are from the U.S. Census Business Dynamics Statistics (BDS).

Together these results offer some interesting insights on the state of entrepreneurial activity in the U.S. that is important not just to academic researchers, but to policy makers and investors. They are consistent with several competing views of the state of entrepreneurship in the US.

One view is that entrepreneurial activity is becoming more efficient. As discussed by Chatterji and Robinson (2016), entrepreneurship may be getting more selective, potentially even more efficient. Consider the following sports analogy. It might be intuitive—especially to a novice—to think that the best way to make more three-pointers in a basketball game would be for every player on the court to take as many shots from three-point range as possible. The more shots on basket, the greater the number of points. This logic might seem sensible, but, in fact, the opposite is probably true: players with different skills specialize, some in long-range shooting, others in play under the basket. Perhaps we are seeing a similar specialization in entrepreneurship, in which individuals with high-tech skills are disproportionately selecting into entrepreneurship, while those with blue-collar skills are disproportionately selecting into wage employment. This is certainly one interpretation suggested by the data at hand.

This is a potentially profound result because it implies that financial and labor inputs are more efficiently allocated in a way that should lead to better overall economic outcomes from lower deadweight costs (e.g., investing in businesses that will fail). Under this interpretation, policy makers should celebrate the decline in entrepreneurship because more selective entrepreneurship means more efficient resource allocation.

Of course, this is not the only interpretation that is supported by these data. Another interpretation would be that innovation is getting much more difficult than in the past. Indeed, recent work by Nicholas Bloom, Chad Jones, John van Reenen, and Michael Webb points out that it now takes 75 times the amount of R&D research man-hour equivalents to keep up with Moore's Law—that computer chip speed doubles every 18 months—than it did when Moore's Law was first conjectured in the early 1970s. Across a wide range of industrial R&D settings, they show that the cost of maintaining constant growth rates has increased dramatically. The data we have shown here also lend to this interpretation, because they suggest that while the amount of resources devoted to innovation is higher than ever before, increasing fewer individuals find it attractive to undertake the task of pursuing innovation.

While these conclusions demand much more careful analysis than we can provide here, they illustrate the central point of our project: understanding entrepreneurial activity requires utilizing data from a variety of primary sources that are not currently available in any one data repository. Without well organized, readily accessible data, we cannot even begin to debate properly.

Going Forward

Our pilot database is a modest first step in what we hope is a larger sustained effort. We seek to create a widely used repository for data utilized in examining a range of important issues related to entrepreneurship. We close with two additional research topics (each deserving of many separate projects) that could each benefit immediately from our efforts.

The first topic is **Economic Spillovers.** Entrepreneurial activity has positive and negative spillovers to the broader economy. Successful innovations will generally increase productivity and economic output per capita. This may be done by devising new goods and services, more efficient business methods, and/or new markets. However, entrepreneurial activity can have (especially short-run) negative spillovers. For example, a new firm may reduce industry employment as it introduces more efficient processes that require fewer employees and thus dislocates workers at existing firms. This might happen in the domestic economy through improvements in labor productivity (e.g., through process improvement or capital investment) or by outsourcing certain functions. In either case, overall economic value may increase, and thus overall wealth would increase, but it is possible that income inequality increases as well if economic rents are concentrated among a small set of entrepreneurs. Data on labor markets, income, new and old firm growth, and productivity are required to carefully examine this issue.

The second is **Optimal Ecosystems.** Substantial recent research has examined the importance of a network of individuals, organizations or institutions for robust entrepreneurial activity. Research has documented the importance of many factors including technology infrastructure, access to skilled and experienced labor, accommodative regulation and economic policy, and a business culture accepting of experimentation and failure, among many others (see Hwang and Horowitt, 2012). It is widely accepted that there is not a single optimal mix of factors for any one geography, however, the trade-offs across factors are not well understood. For example, in a state with a relatively restraining economic (e.g., tax) policy, what other factors best compensate for this short coming? Likewise how does the optimal mix of ecosystem factors depend on largely fixed factors (e.g., geography, certain types of infrastructure, existing labor pool, etc.). Understanding more precisely how factors interact to create optimal ecosystems will aid policymakers in devising the most impactful public investments.

A common thread uniting these questions is that the marriage of time-series and geospatial data from a range of publicly available and proprietary sources can address key issues, but data access is a bottleneck that slows down the scientific process. Our project aims to break this bottleneck, stimulating the creation of knowledge valuable to both policy makers and researchers alike.

Our vision is to build a repository that can be augmented through user contributions. Thus, as the community of scholars using the repository grows, so can the repository itself grow. Indeed, many empirical researchers may wish to release key variables from recent research to our repository as a way of increasing the impact of their own research.

References

Adelino, Ma and Robinson. 2016. Firm Age, Investment Opportunities, and Job Creation. *Journal of Finance*, forthcoming.

Aharonson, Baum and Feldman. 2007. Desperately seeking spillovers? Increasing returns, industrial organization and the location of new entrants in geographic and technological space. Industrial and Corporate Change.

Bonte, Falck and Heblich. 2009. The impact of regional age structure on entrepreneurship. *Economic Geography*. 269-287.

Bloom, Nicholas, Chad Jones, John van Reenen and Michael Webb, 2016. Are ideas getting harder to find? Working Paper, Stanford University.

Chatterji, Aaron and David T. Robinson, 2016. The golden age of entrepreneurship. *TechCrunch*.

Delgado, Porter and Stern. 2010. Clusters and entrepreneurship. *Journal of Economic Geography*. 1-24.

Doms, Mark, Ethan Lewis and Alicia Robb. 2010. Local labor force education, new business characteristics, and firm performance. *Journal of Urban Economics*. 61-77.

Glaeser, Ed and William Kerr. 2009. Local industrial conditions and entrepreneurship: How much of the spatial distribution can we explain? *Journal of Economics and Management Strategy*. 623-663.

Guzman, Jorge and Scott Stern. 2015. Where is Silicon Valley? *Science*. 606-609.

Hwang, Victor and Greg Horowitt. 2012. The Rainforest: The Secret to Building the Next Silicon Valley. Los Altos Hills: Regenwald. ISBN 978-0615586724.

Loughran, Timothy and Jay Ritter. 2004. Why has IPO underpricing changed over time? *Financial Management*, 5-37.

Moretti and Wilson. 2014. State incentives for innovation, star scientists and jobs: Evidence from biotech. *Journal of Urban Economics*. 20-38.

Ouimet, Paige and Rebecca Zarutskie. 2014. Who Works for Startups? The Relation between Firm Age, Employee Age and Growth. *Journal of Financial Economics*. 386-407.

Samila and Sorenson. 2011. Venture capital, entrepreneurship and economic growth. *Review of Economics and Statistics*. 338-349.

Zucker, Darby and Brewer. 1998. Intellectual human capital and the birth of US biotechnology enterprises. *American Economic Review*. 290-306.

Technical Appendix

This note accompanies "Understanding Entrepreneurship: Facilitating Academic Research with a Shared Data Repository" by Brown, Ouimet, Robinson and Zoller and provides further detail on key data sources and methodological approaches. We start by reviewing the main data sources. We describe the data with an emphasis on the pros and cons of the given source, methods, and, where appropriate, mention other available data for future collection efforts. We conclude with a section discussing how we handle changes in MSA definitions over time.

Business Dynamics Statistics (BDS) from the US Census Bureau

The Business Dynamics Statistics (BDS) is sourced from the Longitudinal Business Database (LBD), compiled by the Center for Economic Studies (CES) division of the US Census. LBD tracks most establishments with at least one paid employee in the US, starting in 1976. Paid employees working in railroad or agricultural production industries, as well as government employees and individuals working at private households are excluded from the sample. Employees are measured as of March 12 and include both full- and part-time workers employed as of March 12. Entrepreneurship in the form of self-employment or proprietors and partners of unincorporated businesses are not included in this sample.

This data has a number of advantages, among them the long time series and broad geographic coverage. However, there is significant left-censoring in the early years. For example, in 1977, the count of new (year 0) firms is available, however, all other firms are categorized in one age group "left censored." In 1978, the count of new (year 1) firms which have survived for one year is available and all other firms are categorized in one age group as "left censored." By 1987, the count of all firms 10 years or younger are separately identified. By 1997, the count of all firms 20 years or younger are separately identified.

We provide counts of firms, counts of establishments, and job creation by MSA-year and by state-year. We also measure the two-, three- and five-year survival rate for new firms born in a given year using firm counts data. The three-year survival rate (S) for firms born in year t is calculated as: $S_t = N_{t+3} / N_t$ where N_t is the number of 0-year-old firms at t year and N_{t+3} is the number of 3-year-old firms at t+3 year. For example, there are 51,597 age 0 firms in California in 2011 and 31,128 of those firms make it to 3 years in 2014, so the percent of start-up three-year survival rate for firms born in California in 2011 is 0.60 (31,128/51,597).

Additional data is available from: http://www.census.gov/ces/dataproducts/bds/data.html including breakdowns by industry. Moreover, disaggregated data is available to approved researchers through the CES.

Quarterly Workforce Indicators (QWI) from the US Census Bureau

The Quarterly Workforce Indicators (QWI) derives from the Longitudinal Employer-Household Dynamics (LEHD) database, a Census database which covers over 95% of U.S. private sector jobs. The QWI provides labor market statistics including employment, earnings and other measures of employment flows based on firm characteristics (industry, age, size) and worker demographics information (sex, age, race, ethnicity, education). These estimates are available by state, metropolitan/micropolitan areas and for the US as a whole.

Advantages of the data are the long time series and broad geographic coverage. However, the extent of the time series varies by state. The availability of historical data depends on the year and quarter when a given state joined the Census partnership. The earliest state time series begin in 1990. There is also some variation in the latest available data by state, as each state reports updated data at different intervals.

We provide the number of jobs, as measured on the last day of each quarter, by MSA-year and state-year by using the following files: "qwi_(state)_rh_fa_gs_ns_op_u"; and "qwi_(state)_rh_fa_gm_ns_op_u".

Further data is available from: http://lehd.ces.census.gov/data/#qwi including the total number of jobs on the first day of each quarter (instead of the last day) and the count of people employed in a firm at any time during the quarter (instead of the count of jobs). Data is also available by county and Workforce Investment Board (WIB) areas.

Data from Clustermapping

This data is from the U.S. Cluster Mapping Project led by Harvard Business School's Institute for Strategy and Competitiveness in partnership with the U.S. Department of Commerce and U.S. Economic Development Administration. The project provides over 50 million open data records on industry clusters and regional business environments in the U.S. Data available through the cluster mapping project is primarily sourced externally with inputs from the U.S. Census Bureau, the U.S. Bureau of Labor Statistics, the Bureau of Economic Analysis, the U.S. Patent and Trademark Office, the National Science Foundation, Moody's economy.com, WISERTrade, VentureDeal, and Fortune.

Advantages of the data are the wide variety of economic indicators and broad geographic coverage. However, the website is limited as it only allows downloading data for one selected year and for one variable and region type.

We provide variables in the category of performance, business environment, and demographics & geography by state-year and by MSA-year. Further data is available from: http://clustermapping.us including data by clusters at national level.

Initial Public Offering (IPO) data

We collect data on Initial Public Offerings (IPOs) from a website maintained by Jay Ritter of the University of Florida (https://site.warrington.ufl.edu/ritter/ipo-data/) Data on the headquarters of the IPO firm is available by reading the SEC filings which accompany the issuance. These filings can be found using the "search Edgar" tool provided by the SEC (https://www.sec.gov/edgar/searchedgar/webusers.htm).

Venture Capital Data from the State Science & Technology Institute (SSTI)

We collect Venture Capital (VC) investment data from the State Science & Technology Institute (SSTI), a national nonprofit organization that supports prosperity through science, technology, innovation and entrepreneurship. The SSTI provides a measure of total deals and total venture capital dollars by state using input from the PricewaterhouseCoopers (PwC)/National Venture Capital Association (NVCA) Moneytree Report.

The SSTI VC dollars datasets consist of four datasets in 6-year intervals, starting in 2007 up to 2010. There is also a dataset of the period 1995-2010, published in 2011. However, this more historical data cannot be directly matched with the later datasets. Also, there is a category labeled "Unknown" with dollar figures at the bottom of each of the SSTI reports that we are not able to find documentation from either SSTI.

We provide VC dollars dataset that combines the 1995-2010 dataset with the 2010-2015 dataset to get VC dollars data for the 1995-2015 period. And we use annual CPI-U index numbers from the Bureau of Labor Statistics (BLS) website to calculate inflation-adjusted VC dollars in 2015 dollars. Alternative VC data is available from the PwC Moneytree website: https://www.pwcmoneytree.com/HistoricTrends/CustomQueryHistoricTrend including VC figures for 1995-2016 for each sate.

Data from DELTA database of IPEDS

This data is from the Delta Cost Project Database, a longitudinal database derived from the Integrated Postsecondary Education Data System (IPEDS) surveys on institutional characteristics, finance, enrollment, completions, graduation rates, and staffing for academic years 1986-87 through 2012-13.

Advantages of the data are the long time series and large number of observations. However, there are missing counts from some institutions in some years. To avoid creating a bias when we aggregate at the state-level, we backfill where missing. We backfill those missing data using the data of the previous year to count the total number of faculty and employees per state-year. For example, we replace the 2003 faculty with 2002 faculty if the institution still exists in 2003. For percentage variables, we take both mean and median to aggregate by state.

We provide counts of admissions, counts of employees, counts of faculty, fall cohort, graduation rate, counts of institutions, research costs, research and public service expenses and grants, and research share by state-year. Further data is available from: http://www.deltacostproject.org/delta-cost-project-database with more detail breakdowns.

Other Data Sources

We have also collected data from Annual Survey of Entrepreneurs (ASE), Kauffman Index of Entrepreneurship Series, Bureau of Economic Analysis (BEA), Tax Foundation, National Science Foundation (NSF), and Federal Bureau of Investigation (FBI). For more details about these data sources, please refer to the individual readme files.

Notes about MSA Data Uploading to the Database

Some of our MSA data files are based on 2013 OMB metropolitan area definitions, while others are based on 2009 definitions. Differences include adding new identifiers with new MSAs, removing identifiers, and keeping the same identifiers but the cities within that MSA are changed. In order to be consistent, for those datasets based on 2009 definitions we only keep the MSAs that are exactly the same with the MSAs based on 2013 definitions and upload them to the database. There are total 295 MSAs out of 381 MSAs that match between the two definitions.

Variable List and Definitions

This document provides definitions for the variables included in the UNC's Kenan Institute of Private Enterprise and the Duke University Innovation and Entrepreneurship (I&E) Entrepreneurship Database. Each variable is assigned an unique variable ID. For more information, please refer to the individual read me files. Currently, the MSA data only includes 2013 Metropolitan Statistical Areas.

8 **Poverty Rate**: Ratio of individuals that are below the designated national poverty line.

10 **Manufacturing Intensity Rate**: Ratio of manufacturing jobs to all jobs.

11 **Average Firm Size**: Average establishment size.

12 **Scientific Degrees**: Total science & engineering doctorates awarded.

13 **Patent Count**: A count of all patents granted.

14 **Rate of New Entrepreneurs**: Ratio of the local adult population that became entrepreneurs each month, measured as a 3-year moving average.

15 **Startup Density**: Number of startup firms per 1,000 firm population. Startup businesses are defined as firms less than one-year old and employing at least one person besides the owner.

16 **Opportunity Share of New Entrepreneurs**: Ratio of new entrepreneurs who were not unemployed before starting their business, measured as a 3-year (state) or a 5-year (metro area) moving average.

17 **Foreign Employment**: Jobs created through foreign direct investment (FDI).

18 **Taxes Per GDP**: State and local taxes as percent of GDP.

19 **Unemployment Rate**: Unemployment rate.

20 **Total Employee Headcount**: Persons identified by the post-secondary institution as employee

21 **Innovation Rate**: Utility patents per 10k employees.

22 **Count of Institutions**: Count of post-secondary institutions.

23 **Net International Migration**: Ratio of net international migration to total population.

24 **Population by Age - 0 to 4**: Population by Age - Ages 0 to 4 (Preschool).

25 **Population by Age - 5 to 17**: Population by Age - Ages 5 to 17 (School Age).

26 **Population by Age - 18 to 24**: Population by Age - Ages 18 to 24 (College Age).

27 **Population by Age - 25 to 44**: Population by Age - Ages 25 to 44 (Young Adult).

28 **Population by Age - 45 to 64**: Population by Age - Ages 45 to 64 (Older Adult).

29 **Population by Age - 65 and older**: Population by Age - Ages 65 and Older (Older Adult).

30 **Admissions Count**: The total number of first-time, degree/certificate-seeking undergraduate students who have been granted an official offer to enroll in a college or university.

31 **Faculty Count**: Persons identified by the post-secondary institution as faculty.

32 **Fall Cohort**: The group of students entering in the fall term.

33 **Population Density**: Population density.

34 **Unionization Rate**: Ratio of workers represented by unions.

35 **Advanced Scientific Workers**: Ratio of employed science, engineering and health doctoral holders to total population.

36 **Employment Cluster Strength**: Percent of traded employment in strong clusters.

37 **Kauffman Index of Entrepreneurial Activity**: Ratio of individuals (ages 20-64) who start a business where they worked more than 15 hours and who didn't own a business in the previous month.

38 **HQ of Fortune 1000 Firms**: Count of Fortune 1000 firms headquartered in the area.

39 **Corporate Taxes as Percentage of GDP**: Ratio of state and local net income tax to GDP.

40 **GDP per capita**: Gross domestic product (GDP) per capita.

41 **Military Payroll per capita**: Military personnel wages and expenditures.

42 **Annual Wage**: Average payroll divided by total employment in a particular year.

43 **Employed S.E.H. Doctorates**: The count of employed doctorate holders.

44 **Labor Force Productivity**: Gross domestic product (GDP) per civilian labor force participant.

45 **Labor Mobilization**: Proportion of the working age population in the workforce, calculated as the civilian labor force divided by the civilian non-institutional population.

46 **Net Domestic Migration**: Ratio of net domestic migration to total population.

47 **Research Costs**: Total expenditures on research

48 **Research and Public Service Expenses and Grants**: Total expenditures on research and public services related expenses, scholarships and fellowships.

49 **Exports as Percent of GDP**: Ratio of exports to GDP.

50 **Count of IPOs**: Count of IPOs.

51 **Venture Capital Dollars per $10,000 GDP**: Venture capital per $10,000 GDP.

52 **Federal R&D Funding per capita**: Federal funding for R&D per capita.

53 **Total R&D Expenditures per capita**: Total R&D expenditures per capita.

54 **Count of New Firms**: A count of the number of age 0 firms.

55 **GDP (millions of current dollars)**: The measure of the market value of all final goods and services produced within a state/metropolitan area in a particular period of time.

56 **State Subsidies (millions of current dollars)**: The monetary grants paid by government agencies to private business or to government enterprises at another level of government.

57 **State Government Employment**: All individuals gainfully employed by and performing services for a government.

58 **State Government Payroll**: The gross payroll includes all salaries, wages, fees, commissions, bonuses, or awards paid to employees during the one month pay period that includes the date of March 12.

59 **Wages and Salaries (thousands of current dollars)**: The remuneration receivable by employees (including corporate officers) from employers for the provision of labor services.

60 **Venture Capital Deals**: Count of US venture capital deals by state.

61 **Venture Capital Dollars (current dollars)**: Total US venture capital dollars invested by state.

62 **Graduation Rate**: Ratio of full-time, first-time, degree/certificate-seeking undergraduate students graduating within 150 percent of normal time.

63 **Research Share**: The ratio of spending on research to all other spending at higher education institutions.

64 **Startups 2-Year Survival Rate**: Ratio of startups established in a given year that survived for at least two years.

65 **Startups 3-Year Survival Rate**: Ratio of startups established in a given year that survived for at least three years.

66 **Startups 5-Year Survival Rate**: Ratio of startups established in a given year that survived for at least five years.

67 **Violent Crime per 100,000 Inhabitants**: Violent crime is defined as any of the following offenses: murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault.

68 **Property Crime per 100,000 Inhabitants**: In the Uniform Crime Reporting Program, property crime includes the offenses of burglary, larceny-theft, and motor vehicle theft. The object of the theft-type offenses is the taking of money or property, but there is no force or threat of force against the victims.

69 **State Business Tax Climate Index**: The Tax Foundation's *State Business Tax Climate Index* is a hierarchical structure built from five components: Individual Income Taxes, Sales Taxes, Corporate Taxes, Property Taxes, and Unemployment Insurance Taxes.

70 **Startups End-of-Q1 Employment Counts**: Estimate of the number of jobs at firms aged 0-1, as measured on the final day of quarter one of the given year.

71 **Count of New Establishments**: A count of the number of age 0-1 establishments. An establishment is a single physical location where business is conducted or where services or industrial operations are performed.

72 **Job Creation**: Count of all jobs created over the last 12 months.

73 **Number of Young Firms**: Number of firms with paid employees and less than 2 years in business.

74 **Number of Employees for Young Firms**: Number of employees at firms with less than 2 years in business.

75 **Annual Payroll for Young Firms**: Annual payroll for firms with less than 2 years in business.

76 **Number of American Indian and Alaska Native-owned Young Firms**: Number of American Indian and Alaska Native-owned firms with paid employees and less than 2 years in business.

77 **Number of Employees for American Indian and Alaska Native-owned Young Firms**: Number of employees for American Indian and Alaska Native-owned firms with less than 2 years in business.

78 **Annual Payroll for American Indian and Alaska Native-owned Young Firms**: Annual payroll for American Indian and Alaska Native-owned firms with less than 2 years in business.

79 **Number of Asian-owned Young Firms**: Number of Asian-owned firms with paid employees and less than 2 years in business.

80 **Number of Employees for Asian-owned Young Firms**: Number of employees for Asian-owned firms with less than 2 years in business.

81 **Annual Payroll for Asian-owned Young Firms**: Annual payroll for Asian-owned firms with less than 2 years in business.

82 **Number of Black or African American-owned Young Firms**: Number of Black or African American-owned firms with paid employees and less than 2 years in business.

83 **Number of Employees for Black or African American-owned Young Firms**: Number of employees for Black or African American-owned firms with less than 2 years in business.

84 **Annual Payroll for Black or African American-owned Young Firms**: Annual payroll for Black or African American-owned firms with less than 2 years in business.

85 **Number of Female-owned Young Firms**: Number of female-owned firms with paid employees and less than 2 years in business.

86 **Number of Employees for Female-owned Young Firms**: Number of employees for female-owned firms with less than 2 years in business.

87 **Annual Payroll for Female-owned Young Firms**: Annual payroll for female-owned firms with less than 2 years in business.

88 **Number of Veteran-owned Young Firms**: Number of veteran-owned firms with paid employees and less than 2 years in business.

89 **Number of Employees for Veteran-owned Young Firms**: Number of employees for veteran-owned firms with less than 2 years in business.

90 **Annual Payroll for Veteran-owned Young Firms**: Annual payroll for veteran-owned firms with less than 2 years in business.