Early-Stage Business Formation: An Analysis of Applications for Employer Identification Numbers*

Kimberly Bayard† Emin Dinlersoz‡ Timothy Dunne§ John Haltiwanger¶
Javier Miranda‖ John Stevens**

February 2018

Abstract

This paper reports on the development and analysis of a newly constructed dataset on the early stages of business formation. The data are based on applications for Employer Identification Numbers (EINs) submitted in the United States, known as IRS Form SS-4 filings. The goal of the research is to develop high-frequency indicators of business formation at the national, state, and local levels. The analysis indicates that EIN applications provide forward-looking and very timely information on business formation. The signal of business formation provided by counts of applications is improved by using the characteristics of the applications to model the likelihood that applicants become employer businesses. The results also suggest that EIN applications are related to economic activity at the local level. For example, application activity is higher in counties that experienced higher employment growth since the end of the Great Recession, and application counts grew more rapidly in counties engaged in shale oil and gas extraction. Finally, the paper provides a description of new public-use dataset, the "Business Formation Statistics," that contains new data series on business applications and formation.

---

*The views and opinions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau, the Federal Reserve Board, or the Federal Reserve Bank of Atlanta. All results have been reviewed to ensure no confidential information is disclosed. John Haltiwanger is also a Schedule A part time employee of the U.S. Census Bureau at the time of the writing of this paper. Part of this research was conducted when Timothy Dunne was with the Federal Reserve Bank of Atlanta. Veronika Penciakova provided expert research assistance. We thank conference and seminar participants at the 2017 NBER Summer Institute Meetings, the 2017 Federal Reserve Policy Summit, the 2016 Society for Economic Measurement Conference, the 2014, 2015, and 2016 Federal Reserve System Committees on Regional Analysis, Atlanta Fed RDC Research Workshop, U.S. Census Bureau, George Mason University's Schar School of Policy and Government, and Oberlin College for comments and suggestions.

†Federal Reserve Board
‡U.S. Census Bureau
§University of Notre Dame
¶University of Maryland
‖U.S. Census Bureau
**Federal Reserve Board

1 Introduction

Over the last two decades, there has been substantial expansion in the direct use of administrative records to better document business dynamics. Administrative record data on firms and establishments have traditionally formed the backbone of the business registers that provide the sampling frames for the vast majority of the surveys conducted by the U.S. Census Bureau and the Bureau of Labor Statistics. Improvements in computing and record linkage technology have enabled the use of these large micro data sets on establishments and firms for the measurement of business dynamics along a number of dimensions. The Bureau of Labor Statistics and the U.S. Census Bureau (especially the Center for Economic Studies) have developed new data sets to measure employment flows, provided new statistics on establishment and firm dynamics, and created matched employer-employee databases.[1] All of these data creation efforts have exploited the fact that administrative record data are generally comprehensive in nature and do not impose additional response burden on establishments and firms.

In a similar vein, the analysis described here also examines a comprehensive administrative data source with the potential to provide new and timely information about early-stage business dynamics and the life-cycle of entrepreneurs – applications for new Employer Identification Numbers (EINs) filed through IRS Form SS-4.[2] EINs are unique tax filing identification numbers that many types of business entities are required to use when submitting tax information to the Internal Revenue Service (IRS). EIN applications are often associated with the start of a new business, but they can also be related to other business needs such as banking requirements and changes in ownership and organizational type. Moreover, application responses include information on the intent of a business to engage in future economic activity. Applicants can indicate when they plan to begin paying wages and the expected number of employees they plan to employ over the next year. Responses to these inquiries may provide forward-looking information on business conditions.

The applications data are also very current. The Census Bureau receives information

[1] Decker, Haltiwanger, Jarmin, and Miranda (2014) provide a recent discussion of using Census Bureau data to measure young firm business dynamics. Dunne, Jensen and Roberts (2009) contain a set of papers that discuss the measurement of firm dynamics using the Longitudinal Business Database (LBD), Longitudinal Employer Household Dynamics (LEHD), and Business Employment Dynamics (BED).

[2] In related research, Guzman and Stern (2015, 2016) use information from state business registration records to document business formation for 34 states. They construct new measures of regional entrepreneurial activity to identify the quantity and quality of entrepreneurial activity, focusing on the likelihood of two rare events: a new business registrant transitions to an IPO or is involved in a high-value acquisition.

on applications from the IRS on a weekly basis. The timely nature of these data offers the potential to learn about the pace of business formation earlier than is available in other administrative datasets. Early information on business applications and startups could prove useful in augmenting existing survey data at the national and regional levels where the survey frames often pick up the entry of new businesses with a considerable lag.[3] The high frequency of the data on applications should also be a useful resource to researchers focused on business cycles and business formation. Recent research indicates that young firms may be particularly sensitive to business cycles (Fort, Haltiwanger, Jarmin and Miranda (2013)). For similar reasons, state and local economic analysts may be interested in such high-frequency data to help characterize business formation and economic conditions at the local level.[4] Adelino, Schoar, and Severino (2015) and Fort, Haltiwanger, Jarmin and Miranda (2013) show that self employment and young firm formation are affected by regional housing cycles. It is important to emphasize that application activity captures more than just business formation and may provide a proxy for general economic conditions at the local level as well.

That said, there are several challenges in using EIN applications to develop useful metrics of business formation and regional economic activity. In particular, many EIN applications have little to do with starting a new business, and simply aggregating the raw EIN applications would provide a noisy signal of startup activity. Even for those applications that indicate the reason for applying is to start a business, most will not become employer businesses.[5] To extract a more useful signal on business formation from the applications, response information from the EIN applications is used to model the likelihood that an application becomes an employer business. The output from this modeling exercise becomes the basis for a forward-looking measure of business formation. Another challenge is that raw applications sometimes include large spikes in application activity in specific locations related to certain tax filings that have nothing to do with business starts. Such volatility in filings can greatly affect quarter-to-quarter movements in application counts.

[3] Administrative data use by U.S. statistical agencies typically involves a substantial lag between the collection of the information and the processing of the records. For example, the Census Bureau's Business Dynamics Statistics program has a lag of 2 years, the Quarterly Workforce Indicators program provides information on job destruction and creation by firm age with a lag of one year, while BLS's Business Employment Dynamics program is more current with a lag of 7 months.

[4] Research by Glaeser and Kerr (2009) and Glaeser, Rosenthal and Strange (2009) discuss entrepreneurship within the context of urban and regional economic issues.

[5] Many businesses are "non-employers" that do not hire employees. Of particular interest in this paper is the transition of a non-employer business to an employer business – a job creator.

To address these issues and develop more informative metrics, the EIN applications are first filtered and then modeled. The filtering removes application types that are either volatile, such as tax liens, or that change significantly over time, such as estate and trust tax filings.

The main contributions of the study are fourfold. First, we develop a set of new statistics based on EIN application activity. The statistics report the counts of business applications by state at a quarterly frequency for the period 2004:Q3-2016:Q4.[6] Second, we show how EIN applications can be used to model and project business formations at the state level. The modeling process refines the signal coming from the applications data and allows us to provide estimates of business formations at the quarterly frequency that are timely and forward looking. Third, we illustrate that application filings are linked to local economic conditions through a set of empirical exercises. These exercises show that application activity is correlated with standard measures of local economic conditions such as county employment growth and metropolitan house prices, but also correlated with local idiosyncratic effects such as the presence of shale oil and gas activity. Fourth, we construct microdata files for internal research at the Census Bureau and public use aggregate statistics that will augment existing information on early-stage entrepreneurial activity. The public use files include a set of variables that provide business application and formation counts across a number of different definitions, and a model-based projection series of business formation. The public use data, available at the Census Bureau's website under the name "Business Formation Statistics", currently provide statistics at the national and state levels at a quarterly frequency.[7] Future versions of the Business Formation Statistics may provide statistics at finer levels of geography (e.g. county) and possibly at higher frequencies.

The remainder of the paper proceeds as follows. The next section provides an overview of the business applications data, discusses linking the application data to the Census Bureau's Business Register, and reports on some of the basic patterns in the business applications data. The third section describes our empirical approach in using EIN applications to model business formation, and the fourth section reports on the results of this approach. The fifth provides an initial look at the relationship between application responses and employer size in the first year of business, while the sixth section links application activity

[6]While this paper reports on applications submitted through the end of 2016, the public use datasets (available at https://www.census.gov/programs-surveys/bfs.html) are updated on a regular basis. The initial posting of the public use data contains application filings through 2017:Q3.

[7]Visit the BFS website https://www.census.gov/programs-surveys/bfs.html.

to local economic conditions. The seventh section provides a brief description of the public-use data "Business Formation Statistics", and the last section concludes with a discussion of next steps. The paper also includes an appendix that documents the creation of the microdata files and the construction of the public use files.

2 SS-4 Applications Data

This research utilizes administrative data on Employer Identification Number (EIN) applications. The data contains the vast majority of EIN filings in the United States, including all of those associated with filings for a new employer business. Individuals request EINs through IRS Form SS-4. EINs are tax filing identification numbers used by business entities. All employer businesses in the United States must have an EIN to file payroll taxes. EIN applications are filed on a continuous basis in the United States, with most applications currently submitted on-line.[8] An EIN application form includes name and address of the applicant, business name and address (if available), reason for application, type of business entity, information on the principal activity of the business, plans to hire employees and planned date of initial wage payments, information on a prior EIN, and business start date. Appendix A provides a copy of the application.[9] The IRS collects the EIN applications – along with the assigned EIN – into a dataset of application filings and transmits the data to the Census Bureau on a weekly basis. The Census Bureau uses the application filings to support its Business Register (BR) program. The BR serves as the enumeration list for the Economic Census and is the sampling frame for other business surveys. More generally, it serves as the central storage for administrative business data at the Census Bureau and is the source of statistical products including the County Business Patterns and Business Dynamics Statistics. EIN applications provide information on new businesses and are used to keep the BR and associated sampling frames current.

This study uses the entire set of EIN applications transmitted to the Census Bureau in the period from 2004:Q3 through 2016:Q4. Earlier years of data are unfortunately not available at the Census Bureau, limiting the time-series aspects of our analysis.[10] Over this time period, the Census Bureau received 45.8 million application filings, averaging

[8] More than 85 percent of applications are currently submitted on-line. Other forms of application include phone, fax and mail.

[9] Table A.1 reports the specific response variables on the SS-4 form that are transmitted to the Census Bureau.

[10] In particular, the sample period available includes only one major recession.

about 69,000 applications per weekly file. The weekly files are referred to as cycles and each year contains 52 or 53 cycles. The analysis described here examines the applications at the quarterly frequency, aggregating cycles 1-13, 14-26, 27-39, and 40-53 into quarters one through four, respectively. There is strong seasonality in EIN applications, with higher application activity during the peaks of tax filing in the first and second quarters of the year.

It is important to note that EIN applications may miss some businesses that are formed as sole proprietorships with no employees. These businesses do not necessarily need an EIN, and can use instead the Social Security Number (SSN) of the business owner for tax purposes. Such businesses represent certain types of entrepreneurship, particularly independent contractors. Nevertheless, an EIN still has its advantages over an SSN for these type of businesses. For instance, the use of an EIN can preclude identity theft and help the business owner establish an independent contractor status and build a business credit history. In fact, there is a very large number of sole proprietorships in the EIN applications data used here, and many of them do not transition to an employer business during the sample period, indicating that many sole proprietorships with no plans to hire employees nevertheless obtain EINs for other purposes.

Application Pool Restrictions

At the outset, a number of restrictions are placed on the set of applications that are used to derive tabulations and to model business formation. We omit four broad types of applications from the analysis based on type of entity, industry, geography, and the observed concentration of applications from a specific source. With regard to type of entity, three groups are removed from the data – applications associated with tax-liens, trusts and estates. We omit these applications because they are generally unassociated with business formation and their presence in our data files varies over time. We see a large increase in tax lien filings in 2009 through 2011 and in estate and trust filings in 2008 into 2010. We also omit applications associated with a set of detailed industries within the agricultural, financial services and private household sectors. Applications from these specific industries have very low transition rates to employer businesses and are often quite volatile in terms of application volumes. Applications were not included if they were submitted by public entities (a very small number). Applications were also omitted that had missing state information (a small number) or came from outside the 50 states or the District of Columbia, such as Puerto Rico or the Virgin Islands. Finally, applications were

removed that came from concentrated filing spikes. A concentrated filing spike is defined as a group of EIN applications that appears in the same weekly cycle batch, comes from the same zip code, and has the same industry code. These filings are almost always related to some type of financial filing. The appendix provides more details on the individual restrictions imposed and the impact of each restriction on the number of applications in our data set. In total, 12 million applications are removed from the analysis, with roughly half the omitted cases associated with tax liens, trusts, and estates restrictions.[11] The resulting dataset contains 32.3 million applications, and is referred to as business applications (BA).

With these restrictions in place, Figure 1 presents the number of quarterly business applications between 2004:Q3 and 2016:Q4, on a non-seasonally adjusted basis and seasonally adjusted basis. There are 650,000 business applications, on average, per quarter. This is a large number compared to the annual number of employer business startups in the U.S. (about 450,000 per year) in the Business Dynamics Statistics (BDS) over the same period of time. In terms of the general patterns in the series, there is a rise in the number of applications between the end of 2004 and 2007. Application counts drop off as the economy fell into recession during 2008 and 2009, but have since recovered. In 2016, the number of applications was about 12 percent higher than in 2007. In addition, the seasonal nature of applications, discussed earlier, is clearly present in the figure.

Business Register Match

We match the application data to employer records in the Business Register (BR) for the period 2003 through 2014. The BR contains the complete set of businesses operating in the United States that have employees or payroll. We utilize the employer-business universe and match the new EINs to the set of firms identified as new employer businesses (firm age equals zero) based on first payroll observation, by the Longitudinal Business Database program.[12] The match process is straightforward as both sets of data contain EINs. The match to the BR allows us to identify which applications become new employer businesses and the quarter in which they begin to pay employees, denoted as the startup quarter.[13]

[11] The transition rate to employer status of deleted cases is 1.0 percent, whereas the transition rate of the remaining applications is 18.4 percent. Concentrated filing spikes have a transition rate of only 0.5 percent.

[12] Haltiwanger, Jarmin and Miranda (2013) discuss the measurement of firm age using Census data. Age zero firms include EINs that first begin paying workers in a year but exclude new EINs that are associated with an older firm. Older employer firms are often associated with new EINs by mergers and acquisitions, through corporate spinoffs, or by changes in organization type. In short, firm age is constructed based on the age of the oldest establishment linked to the firm, identified through the LBD file.

[13] The BR data are only available with a lag and this paper uses BR records through the fourth quarter

At this point, the match is to the employer business universe only, though it is likely that many applications for new businesses end up as non-employer businesses. Davis et. al. (2009) discuss the transition of non-employers to employer firms using the BR. We plan to examine this aspect of the applications data more fully in the future. In addition, we aim to study business formations by existing firms (firm age greater than zero) in future work. In principle, one can model all three modes of transition for an EIN application using application characteristics: becoming a non-employer business, an entirely new employer business, or a new employer business from an existing non-employer business.

The matched data face two types of censoring. Applications that occur late in our sample are increasingly right-censored with respect to identifying those that become employer business births. For example, for the sample used for the analysis in this paper, an application received in the first quarter of 2014 only has a 4-quarter window within which to identify a transition to an employer, whereas an application received in the first quarter of 2010 can have as many as 20 quarters. Our approach to dealing with right censoring is to examine the likelihood an application becomes an employer business birth over a specific window of time (discussed in more detail below). The applications data also face left censoring. Currently, the applications data begin in 2004:Q3. This left-censoring means that early in the sample period only a subset of business births can be linked to EIN applications. Over time the impact of left-censoring diminishes, since the likelihood an application becomes an employer business birth declines as an application ages.

An additional feature of the BR-application match is that a small number of matched EIN applications have payroll data that pre-dates the application submission. This type of timing typically happens because the business register data begins showing payroll in the first quarter, whereas the application submission appears in a later quarter. However, there is also a small number of cases where payroll and employment activity occurs in the year before the application is submitted.[14] In the analysis presented here, the application date is typically used as the quarter of record.

Figure 2 shows a histogram of the distribution of application transitions to employer business births by age of application. The sample of applications includes all matched filings

of 2014. Thus, while our applications run through the fourth quarter of 2016, the BR data allow us only to match records through the end of 2014. Note, however, that the newest BR data is incorporated annually into the public use files as it arrives.

[14] Some smaller firms can choose to file their payroll taxes on an annual basis. In this case, payroll is likely spread across all four quarters, even if the business started up after the end of the first quarter. In addition, there are a number of businesses that file retroactively for an EIN.

through the end of 2012, allowing eight additional quarters of transition for the 2012:Q4 applications.[15] The horizontal axis reports the age of the application in number of quarters since submission. The bars over the negative values capture the business employer births that appear in the business register prior to the application quarter. The spike at zero indicates that the modal employer-application match occurs in the same quarter as when the application is submitted. About 75 percent of matched application-employer births occur within the first four quarters including the negative tail. The fact that a majority of applications that transition to employer status do so relatively quickly after the initial submission suggests that recent EIN application activity will be a good proxy for near-term business formation activity.

Using this information on the time to transition to an employer business, we construct two indicator variables for each application that identifies whether an application becomes an employer business within four quarters of submission or within eight quarters of submission. Throughout the analysis, the 4-quarter window is the main focus because it captures the majority of employer business births as seen in Figure 2 and allows us to measure application transitions through 2013. The 4-quarter window uses the last year of matched BR-applications data (2014) to measure the transition to employer business status of applications received in 2013. The longer 8-quarter window uses the 2013 and 2014 matched data to identify transitions for applications received in 2012. Given the fact that the matched file is already quite dated compared to the very current applications data, we focus primarily on the 4-quarter window to keep the lag between the analysis sample and the incoming data as short as possible.[16]

To give a sense of how cohorts of applications match to employer births, Figure 3 shows the age distribution of applications for all employer births that occur in 2013. For this set of employer births, we can identify all applications that become employer births going back to 2004 and for the small number that arrive in 2014. There were 415,000 employer births

[15] The bars at -4 and 16 include all observations where the quarter of transition and application quarter is less than or equal to -4 and greater than or equal to 16, respectively.

[16] Since applications arrive at different weeks within a quarter, the length of the transition window identified at a quarterly frequency will vary. To control for this, we randomly assign each application the opportunity to look ahead for an additional quarter depending on the week the application arrives in the quarter. An application that arrives in a late week in the quarter has a greater chance of looking ahead an additional quarter to identify transition to employer business status than an application that arrives early in the quarter. In this way, we preserve the length of the 4-quarter window look ahead period across the entire application sample. Appendix A provides more details on the randomization process and the construction of the 4-quarter window.

in 2013 matched to applications, with only a small number of employer births in 2013 not matching to an EIN application. About three-quarters of employer births in 2013 are from applications received from 2012:Q1-onward. An implication is that recent applications account for the bulk of employer births within a year.

2.1 Application Response Information

The analysis uses information contained on applications to both model the likelihood an application becomes an employer business and to create application data series. The data reported on an application are limited in detail. An application has inquiries about the type of entity (legal form of organization), the reason for applying, plans to hire workers, principal business activity, business start date, previous EIN, along with name and address information. The inquires are largely check box responses except for the name, address and date fields that provide very useful information in assessing whether an application is more or less likely to become an employer business. Tables 1-5 provide the distribution of responses for a key set of variables from the business applications (BA) received over the period 2004:Q3-2016:Q4.

Table 1 presents the frequency distribution of applications by type of entity (SS-4 form: Line 9) for applications received, the percent of applications received by the end of 2013 that become employer births within the 4-quarter window, and the percent of applications received through the end of 2012 that become employer births within an 8-quarter window. Nearly 57 percent of all applications come from traditional business organizations – sole proprietorships, partnerships and corporations. There is also an "Other" category. In all, 13.6 percent of business applications become employer businesses within the 4-quarter window, while 15.9 percent transition within the 8-quarter window. Corporate filings are two and one-half to three times as likely to transition to an employer business within the specified windows than sole proprietorships, partnerships or the"Other" category. A large fraction of applications fall into the "Other" category, including many that are associated with LLCs.[17] Table 2 presents the same information for the inquiry on "reason for applying" for an EIN (SS-4 form: Line 10). The responses to this question are less promising in terms of capturing variation across applications, as over 87 percent of cases are recorded as "start a new business". However, there are some smaller response categories that translate

[17]There does exist some write-in information for the "Other" category that has been coded and may allow us to distinguish amongst cases in the "Other" category.

into relatively high employer business startup rates – hiring employees, changing organizational type and the purchase of a business – all have relatively high rates of conversion to an employer business (two to four times the average). In our projection model, we include all of these response categories for completeness. Our inclusion of changing organizational type and the purchase of a business warrants further comment. We include these categories since they reflect in part the transition from an existing non-employer to an employer business. However, existing employer businesses that undergo an EIN change for one of these reasons will by construction not transit to be an early stage employer business. We seek to capture the heterogeneous outcomes with respect to these response categories with the rich set of interactions that we include in our model specification below.[18]

Table 3 provides a breakout by industry. Each application is assigned to a NAICS industry based on the response to inquiries on principal activity of the business and the principal line of merchandise sold, services provided, etc. The actual industry coding of applications is somewhat uneven in the data. For some sectors such as construction, the NAICS coding is quite detailed – often coding applications down to the 6-digit industry code. Alternatively, applications in manufacturing often contain less detailed industry coding (1-digit manufacturing level). In the table, we provide a coarse breakout across 1-digit NAICS sectors. Among the applications received by 2013, 8.3 percent have missing industry codes. The most common applications are associated with wholesale/retail and financial and business services. With regards to business formation rates, manufacturing, education and health, and leisure and hospitality industries have relatively high conversion rates to employer status. Within these broad sectors, there can be considerable differences across more detailed industries in application transition rates to employer business status. For instance, applications associated with the offices of health practitioners, retail stores, and restaurants have particularly high transition rates.

Table 4 provides information on the date of the first wage payment (SS-4 form: Line 15) reported on an EIN application. The information Census receives is a date field – month and year. We record whether the date is filled in or is missing (an indicator variable). A wage date is present for about 24 percent of applications. Of the applications received that report a wage date, 40.2 percent become an employer business within 8-quarters of application submission. Looking at applications without a wage date given, only 7.0 percent become employers within 8-quarters. As discussed later, the wage date variable turns out

[18] For users of the data, the inclusion of these categories implies that application series include cases with transitions by existing employers.

to be the best single predictor of transition to a new employer business for an application.

Table 5 presents the same statistics for a group of applications that we refer to as a high propensity applications. We define a high-propensity application to be one that satisfies one or more of the following criteria: (1) is a corporate entity; (2) indicates on the application that they are hiring employees, purchasing a business or changing organizational type; (3) provides a wage date; or (4) has a NAICS industry code in manufacturing (31-33), retail stores (44), health care (62), or restaurants/food service (72). High propensity applications make up about half of all applications and have an 8-quarter transition rate of 27.0 percent. Applications that are not part of this group have a 8-quarter transition rate of 3.8 percent. The set of high propensity of applications are provided as a distinct data series in our national and state data release files.

The patterns in the tables suggest that response information on applications may prove useful in screening applications for their likelihood to become an employer business. The conversion rate from an application to an employer business varies systematically with responses on industry, the wage date, reason for applying, and type of entity. This information, along with other information submitted in applications, will be used to assign propensity scores to applications in order to develop a closer link between application volumes and new business formation and to construct data series that may be more closely related to economic activity.

2.2 Patterns of Applications and Transitions over Time

We present the time series patterns of application volumes for five application groups: all business applications (BA), high-propensity business applications (HBA), business applications with planned wages (WBA), business applications from corporations (CBA), and business applications from industries with high rates of transition.[19] The last three groups of applications (HBA, CBA, and selected industries) are proper subsets of the high-propensity business applications (HBA).

For all application series, Figure 4 shows annual applications data from 2005 to 2016 with each series indexed to 1 in 2005. Looking at the overall business applications series, applications were rising prior to the Great Recession and during the recession. Since the end of the recession, they have expanded by about 20 percent. The rise in the overall

[19] The first four series on this list (BA, HBA, WBA, CBA) are part of the main publicly available data tables are planned to be released on a regular basis. See Appendix B.

series stands in sharp contrast to patterns observed in the wage date series. Recall, the wage date series (bottom line in the chart) includes only those applications that indicate a planned date to begin paying employees. The number of wage date applications fell sharply starting in 2007 and has not recovered. Corporate filers, a subset of applications with a higher propensity to become employer businesses, also decline but not as steeply as the wage date cases. The one group of high propensity application that did not experience a sharp decline is based on industry grouping. In particular, applications in the health service and food service industries have increased steadily during the recovery. The index of high-propensity applications comes in the middle. The number of these applications remains below pre-recession levels, leveling off during the recovery. Finally, though not directly comparable, we show an index of firm births that comes from the Business Dynamics Statistics (BDS) program. This index is based on the number of firm births observed in a year and declines by almost 30 percent from 2006 to 2010. From 2010 and 2014 (the last published data used for this paper's analysis), the index shows a slight increase but the rate of firm births remains well below the rate observed prior to the Great Recession.[20] The pattern exhibited by the index of high-propensity applications is similar to the pattern of the index of BDS firm births.

Clearly, changes in the overall number of applications are not going to be able to explain the firm birth patterns observed in the BDS series. Changes in the composition of applications and/or application-specific transition rates will be necessary to explain these patterns. In particular, the link between EIN applications and business formation will depend not only the number and type of business applications but also on the transition of business applications to employer businesses. Figure 5 presents cumulative transition rates from application to employer business for six different annual cohorts of applications using applications in the business application sample. For the more recent cohorts, the data are increasingly right censored. The cumulative transition rate has clearly shifted down over time, with the 2012 application cohort having the lowest transition rates across the cohorts. At four quarters out, the 2012 cumulative transition rate is, on average, .05 lower than that of the 2006 cohort, representing a substantial drop. A key issue is whether the shift in transitions is due primarily to application characteristics (for example, the decline in the number of wage date applications observed in Figure 4) or due to a general shift down in transition rates across all application types.

[20] This decline in new firm formation has been documented in a set of papers by Decker, Haltiwanger, Jarmin and Miranda (2014, 2015).

3 Projecting Business Startups

In this section, we describe our approach to modeling business startup activity from the EIN applications data. The goal of the exercise is to assess whether application information can be used to provide early information on startup activity for the U.S. economy, as a whole, and for states, individually.

Let N_{gt} be the number of business applications (BA) in any region g at time t.[21] The total number of business formations that materialize during the period t to $t+k$ from the applications at time t is given by

$$S_{gt+k} = \sum_{i=1}^{N_{gt}} I_{igt+k} \tag{1}$$

where I_{igt+k} is a realization of a Bernoulli random variable that governs whether application i turns into an employer business between t and $t+k$. The probability distribution function for I_{igt+k} is given by

$$I_{igt+k} = \begin{cases} 0, & \text{with prob. } 1 - P_{igt+k} \\ 1, & \text{with prob. } P_{igt+k} \end{cases} \tag{2}$$

where P_{igt+k} is the probability that application igt turns into an employer business between t and $t+k$. Then,

$$E[S_{gt+k}] = \sum_{i=1}^{N_{gt}} E[I_{igt+k}] = \sum_{i=1}^{N_{gt}} P_{igt+k}. \tag{3}$$

To estimate $E[S_{gt+k}]$, we need an estimate of $P_{igt+k} = E[I_{igt+k}]$. To do so, we can model P_{igt+k} as a function of application-level variables, Z_{igt}, and a set of unknown parameters, β_{gt}. The probability that an application transitions to an employer business can then be estimated as

$$\hat{P}_{igt+k} = F(Z_{igt}; \hat{\beta}_{gt}), \tag{4}$$

[21] As discussed above, applications are received on a weekly basis. In the current exercise, the data are aggregated to a quarterly frequency. However, the model uses the weekly cohort information in the estimation. The discussion that follows treats the cohort as quarterly.

where F is either a linear function (a linear probability model (LPM)) or the $c.d.f.$ of the normal distribution (a probit specification). $\hat{\beta}_{gt}$ is an estimate of the unknown parameters, β_{gt}, based on the LPM or probit model. The predicted application-level probabilities, $\hat{P}_{igt+k}$, can be used to construct an estimate of the expected number of business formations by time $t+k$ as

$$\hat{S}_{gt+k} = \sum_{i=1}^{N_{gt}} \hat{P}_{igt+k}. \tag{5}$$

This approach amounts to reweighting each application by the predicted probability (propensity score) that the application becomes an employer business between t and $t+k$.

In the current analysis, the time frequency is quarterly and k is set to four quarters. As discussed above, a window of four quarters allows a long enough time period for an application to become a business but does not result in a significant loss of information on application transitions due to right censoring.[22] The small number of applications that become employer businesses before the application date t are also considered as having started up within the four quarter window.

The set of predictors, Z_{igt}, contains sets of indicator variables based on the application response information. These variables include indicators for the type of entity, reason for applying, industry (6-digit NAICS) and the wage date variables discussed above. The empirical models also include variables that code the week of application submission within the year, the business start date, the limited liability status of the business entity (SS-4 form: Line 8), the presence of a prior EIN (SS-4 form: Line18), and whether the application indicates a trade name (SS-4 form: Line 2), an executor's name (SS-4 form: Line 3), or a distinct business address on the application (SS-4 form: Line 6). Appendix A provides a detailed discussion of variable construction. Z_{igt} also contains a rich set of interactions between industry, the wage date, type of entity, and reason for applying response variables. The interactions include two-digit industry interacted with the wage date variable, 1-digit industry interacted with type of entity controls, and interactions between the wage date variable and type of entity and reason for applying controls.

The geographical unit of analysis is at the state level and the model is estimated individually by state and pooled across all states. The pooled specifications contain state fixed

[22] An 8-quarter window is also examined and the results are quite similar. The main difference is that the 8-quarter window captures a greater share of new businesses but cuts off our estimation sample one year earlier.

effects. For the pooled model, the number of estimated parameters is close to 600. For the models estimated individually by state, the total number of parameters estimated exceeds 25,000.[23] Our approach is flexible with respect to the geographic unit of aggregation, g. For example, we could estimate the models and construct projections at the CBSA level.[24]

The proposed measure of projected business startups, $\hat{S}_{gt+k}$, is forward looking, providing an estimate of the number of new business startups that will appear from a given cohort of applications in a specific geographic area over a particular horizon. What it does not provide is an estimate of the total number of business startups that will appear within a specific time window, for example, a quarter. This type of measure requires a different approach – aggregating the propensity to become an employer business across all at-risk applications for a specific time period (i.e., a given quarter). Our data allow for the construction of this type of measure for later years, when almost all applications that generate employer firms in a given quarter fall completely in the time window 2004q3-2016q4.[25]

Finally, a key point in our empirical strategy is that we only utilize information submitted on an application. This approach is important because in order to publish "real time" business formation statistics based on incoming applications, there is little external data available (especially at the regional level) to incorporate into the estimation approach. In addition, from a model-fit perspective including additional information may result in a loss of predictive accuracy. For instance, we estimated versions of the empirical models by including the Philadelphia Federal Reserve Bank's state-level coincident and leading economic indices to control for regional economic activity. The inclusion of such aggregate series, however, reduced the projection accuracy of the models. Furthermore, the inclusion of various time trends also resulted in a worse model fit in terms of both in and out-of-sample predictions.

[23] The number of parameters per state can vary as some detailed industries and specific interactions do not appear in every state.

[24] While re-estimating the model for different levels of geography is straightforward, aggregating across different frequencies (weekly or monthly) is somewhat more complicated as the Business Register data on employer businesses are quarterly.

[25] We are currently developing models to provide direct estimates of the number of startups that appear in a given quarter from the application data. For the 2012-onward birth cohorts, almost all firm births can now be observed in the application data. That is not true of the earlier firm birth cohorts as applications are left censored, as discussed above.

3.1 Estimation Details

To select an empirical model to project business startups, we explore differences in empirical specifications and the length of the estimation sample window. With regard to model specifications, we examine differences in the functional form of the empirical model (probit or linear probability model), the benefits of including a more saturated model in terms of the interactions among variables, and the gains from estimating models for individual states. With respect to the length of the estimation sample, we begin by using a large estimation sample from 2006 through 2012 to generate projections for the period 2013-2016 and then shrink the estimation sample by omitting the earlier years. In particular, we want to investigate whether a smaller sample that uses only the most recent data provides a more accurate projection model of business startups.

To assess model performance, we construct a modified root mean-squared percent deviation (RMSD) statistic on a quarterly basis at the state level for in-sample and out-of-sample periods. The $RMSD_t$ statistic is based on the percent deviation between the actual and projected number of startups using the 4-quarter window definition at the state-level in a quarter. The percentage deviation is used to adjust for the fact that geographic areas vary markedly in the volume of business formations. The statistic is constructed as

$$RMSD_t = \sqrt{\frac{1}{G_t}\sum_{g=1}^{G_t}[\frac{\hat{S}_{gt} - S_{gt}}{S_{gt}} \cdot 100]^2} \tag{6}$$

where G_t is the number of geographic areas (51) in time period t under study. A lower $RMSD_t$ means a more accurate projection. The current analysis presents both an unweighted measure of the $RMSD_t$ and a state size-weighted $RMSD_t$, where the weight is based on the total number of applications received in a state in a quarter. Model selection is determined by comparing the $RMSD_t$ statistic both in-sample and out-of-sample. For the out-of-sample exercises, we limit the estimation sample up through 2012 and then use projections and data from 2013 to construct the out-of-sample $RMSD_t$.

4 Results

Because of the large number of estimated parameters, we cannot report individual parameter values and standard errors for our control variables. However, a small set of variables and application responses play the largest role in explaining the propensity of an applica-

tion to become a business. These include the wage date variable, industry controls, and several key response categories from the type of entity and reason for applying inquiries. The wage date variable provides the strongest signal of transition to an employer business. Holding other factors constant, the difference in the probability of becoming an employer within one year of filing for applications reporting a wage date versus applications not reporting a wage date averages about 23 probability points across specifications. As discussed above, applications from corporations and applications from individuals purchasing a business, or changing organization type had generally higher propensities to become an employer business. In addition, we found significantly higher propensities associated with applications from multi-member limited liability companies (LLCs). There are also clear patterns across industries. Applications from industries such as health care practitioners, restaurants, and manufacturers have higher transition rates.

We investigated a range of preliminary specifications before settling on the core specifications. Models with weekly time effects (52 or 53 week specific dummies) perform better than models with quarter effects. Model performance improves with the inclusion of more detailed industry controls. We found that the models perform very poorly with the inclusion of time-trend terms - linear, quadratic or cubic. Trend terms generate significant differences across estimation samples and generate large projection errors as one moves away from the end of the in-sample data. The weak performance of the trend variables may be related to the relatively short data series that are currently available to us and that this short time series includes the Great Recession. The trend variables pick up the sharp drop in applications in 2008 and 2009 and then have a tendency to extend this drop in the out-of-sample exercises.

4.1 Model Evaluation

Table 6 presents the RMSDs across a set of alternative empirical specifications. A lower RMSD means that our state-level prediction errors are smaller. The column labelled "2013(U)" presents the unweighted mean of the RMSD for all quarters in 2013 (out-of-sample). The second column labelled "2013(W)" presents a weighted version of the same statistic, where the weighting is based on the state's application volume. The base model is the LPM, with a full set of interactions, estimated individually at the state level. The LPM is used to search across a broad set of empirical models, as it runs more quickly compared to a probit model. As the model specification is narrowed, we augment our estimation and

examine the performance of probit specifications.

We first examine the sample window length. The sample estimation window length is allowed to shrink from seven years, a sample that includes applications from 2006 through 2012, to two years (2011-2012). Out-of-sample RMSDs decline as the window shrinks until the window length is three years, though there is not much difference between the RMSDs at the three- or four-year window length. The 2010-2012 model has a RMSD of 5.29 percent. A comparison of the weighted versus unweighted results shows lower weighted RMSDs, indicating more accurate projections in the larger states, but the same general pattern holds. The final line in the upper block of Table 6 uses an average of predictions from three yearly samples from 2010-2012. The RMSDs are slightly higher than the model that pools across the three years. The next block takes the three year estimation window and steps back in time using 2008-2010 and 2009-2011. This allows us to assess how the model performs in projections that are further away in time from the estimation sample window. The RMSD shows a modest rise. The model that uses the 2008-2010 estimation sample has a 0.50 percentage point larger unweighted RMSD compared with the 2010-2012 estimation sample. The last block in the upper panel steps back even further and uses estimation samples from 2005-2007 and 2006-2008. Here, we see a significant decline in model performance, as RMSDs rise sharply. Models based on early data overestimate the number of business formations in 2013 and generate relatively large RMSDs.[26]

Using the three-year estimation window (2010-2012), the middle panel of Table 6 provides the RMSD statistics for a set of alternative specifications. The first row reports the results from a model that pools across states and contains no interactions. The RMSD statistic is somewhat higher than that of the base model presented in the top panel, but only by 0.3 percentage points. The next row includes full interactions but estimates the model pooled across states, while the third row drops the interaction terms but estimates the model individually by states. The RMSD statistics decline slightly in comparison to the non-interacted, pooled model (row 1). Finally, the last row in the table shows the results from estimating the within-state, full interactions model using a probit specification. The RMSD is slightly smaller than the LPM version of this model.[27] This probit specification is the model that we use in forming our projections. Nevertheless, empirically there are

[26] This gap is also consistent with the findings reported in Table 7 below, which indicates a substantial drop in transition rates between 2006 and 2007, controlling for application types.

[27] Probit models are estimated for a number of different specifications. In general, they slightly outperform the LPM models based on the RMSD statistic but the differences tend to be small.

only small differences in model performance across the various specifications and window lengths. To see this, Figure 6 presents projection series for a three-year estimation window altering the sample years and functional form. The underlying models are estimated individually by state with a full set of interactions. For each model, the projected number of employer business startups both in- and out-of sample is presented, in addition to the actual business formations (denoted by circles). The projections include forecasts and backcasts depending on the particular estimation sample employed. In general, the models track very closely, except in the earlier years where the models backcasting over a longer period (2010-2012 and 2011-2013 samples) project a lower rate of business formation than models employing an early sample. The projections based on the 2006-2008 sample miss to the high side. Still, it is encouraging that across the various estimation samples, the models yield very similar results in the out-of-sample projection period from 2014 to 2016.

Figure 7a presents the business formation series from 2006-2013, spliced together with a set of the projection series for 2014:Q1-2016:Q4. The projections based on the 2011-2013 probit model use the most current data. The spliced series are seasonally adjusted using the X-12 seasonal adjustment procedure. The figure shows that the projections are relatively close to one another, even with non-overlapping estimation samples. Figure 7b shows a close-up view of the series from 2009-2016:Q4.

As mentioned above, the models were also estimated with an 8-quarter window. Figure 8 shows the seasonally-adjusted spliced series for the 4-quarter and 8-quarter windows. The 8-quarter window, estimated from the 2010-2012 data sample, tracks the path of the 4-quarter window closely but contains, as expected, a higher number of business formations.

4.2 Decomposition Analysis

Overall, the similarity across the projections suggests that the estimated parameters from the various empirical models are relatively stable over the alternative estimation samples. We now examine this issue more fully by carrying out a decomposition analysis.

The significant drop in transitions to employer status over time could be due to a change in average transition rate of an application to employer status, to changes in the composition of applications, or to a combination of both factors. To assess these factors, we first estimate a version of the model pooled across states that included year effects. The sample is the full sample of applications matched to the Business Register, 2004-2013. Table 7 presents the year effects from the LPM model where the omitted year is 2004.

The probability an application transitions to an employer business is higher in the early years but from 2007 onward there is little difference in the probabilities. The continued shift down in the cumulative transitions that occur after 2006, as seen in Figure 4, is then accounted for largely by changes in the composition of applications that occurs over the 2007-2013 period, an issue we explore next.

Formally, one can decompose the change in the number of business startups into a part due to the change in the number of applications and a part that reflects the change in the probability an application becomes an employer business. The change in the aggregate number of business startups between periods $t-m$ and t can be written as

$$\Delta(S_t) = P_t \cdot N_t - P_{t-m} \cdot N_{t-m} \tag{7}$$

where P_t is the average probability an application from cohort t becomes an employer business within a specific time period and N_t is the number of applications in a cohort t. The expression can be rewritten as

$$\Delta(S_t) = \Delta(P_t) \cdot N_t + \Delta(N_t) \cdot P_{t-m} \tag{8}$$

by adding and subtracting $P_{t-m} \cdot N_t$ and rearranging terms. The first term measures the contribution of a change in the average probability that an application becomes an employer business to the change in the aggregate number of the startups. The second term measures the contribution of the change in the overall number of applications on startup activity. The change in the average probability that an application becomes an employer, $\Delta(P_t)$, can be further decomposed. Following Fairlie (2005), the average probability can be written as a function of Z, the full set of application characteristics, and a set of parameters to be estimated β. The change in the average probability between $t-m$ and t, $P_t - P_{t-m}$, is written as

$$\left[\sum_{i=1}^{N_{it}} \frac{F(Z_{i,t};\hat{\beta}_{t-m})}{N_t} - \sum_{i=1}^{N_{i,t-m}} \frac{F(Z_{i,t-m};\hat{\beta}_{t-m})}{N_{t-m}}\right] + \left[\sum_{i=1}^{N_{it}} \frac{F(Z_{i,t};\hat{\beta}_t)}{N_t} - \sum_{i=1}^{N_{it}} \frac{F(Z_{i,t};\hat{\beta}_{t-m})}{N_t}\right]. \tag{9}$$

The first term in the expression is the contribution due to a change in the Z's between $t-m$ and t and the second term reflects the contribution due to a change in the estimated

parameters between the two time periods. This is a Blinder-Oaxaca decomposition modified for a limited dependent variable model. For our overall approach to work well, the change in business startups needs to be driven primarily by either changes in the number of applications received or changes in the probability an application becomes an employer business due to changes in the composition of applications. In the latter case, this would reflect stability in estimated parameters in the projection models and a small relative contribution of the second term in equation (9).

To assess the relative importance of the sources of change in business startups, we examine the change in startup activity from 2006 $(t-m)$ to 2013 (t).[28] The probit model framework described above is used as the basis of the estimation. The models are estimated separately for the 2006-2008 and 2011-2013 periods and the estimated coefficients from the two sets of probits are used in the decompositions.

The results are reported in Table 8. The top panel of Table 8 reports the first part of the decomposition that divides the change in the number of business startups into the fraction due to changes in the probability that an application becomes an employer and the fraction due to the change in the number of applications. We measure the overall net change in number of startups, under the 4-quarter window definition, as a decline of 146,000 between 2006 and 2013. Roughly 92 percent of the decline is due to a decline in the likelihood an application becomes a startup and the remaining 8 percent is due to a decline in the overall number of applications.

The lower panel of Table 8 reports the second part of the decomposition that divides the overall change in the probability an application becomes an employer business into the fraction due to the change in application characteristics, Z, and the fraction due to changes in the parameter estimates, $\hat{\beta}$. The mean probability an application becomes an employer business within the 4-quarter window declines by .034 probability points between 2006-2008 and 2011-2013, with 78.6 percent of the decline explained by changes in application characteristics and he remaining 21.4 percent is due to shifts in $\hat{\beta}$.

The above analysis uses the 2006-2008 parameters and the 2011-2013 characteristics to weight the first and second terms, respectively, in the Blinder-Oaxaca decomposition. However, the choice of the weights is arbitrary and the decomposition can be restated using the 2011-2013 parameter estimates as the weight in the first term and 2006-2008 characteristics as the weight in the second term of expression (9). These results are presented in the

[28] The analysis period is limited to 2013, as that is the last year that we can construct our forward-looking measure of business formation.

second row of the lower panel in Table 8.[29] The results are nearly identical, indicating the choice of weighting period does not impact the decomposition results in our application.

While not shown, when we produce the same decomposition for more recent time intervals (2007-2009 to 2011-2013, for example), a larger fraction of the decline in the probability an application transitions to an employer business is explained by application characteristics. The increase in the contribution of application characteristics reflects, in part, the patterns observed in Figure 5, where the application cohorts of 2005 and 2006 have significantly higher transition rates than later cohorts. Accompanying the decline in transition rates is an increase in the average duration until an application becomes an employer business. Figure 9 shows this increase in seasonally-adjusted average durations measured in quarters for both the 4-quarter window births and 8-quarter window births. For 4-quarter window births, the increase is roughly 2.5 weeks, and for 8-quarter window births, the increase is about 3.25 weeks. In general, what we have seen is that after the onset of the Great Recession the time it takes for an application to become a business has lengthened, reducing somewhat the fraction of employer births that will be captured within the fixed-length windows. Much of this shift is picked up by application characteristics, as shown above, but some also reflects changes in the underlying parameters, holding application characteristics constant. We plan to explore further the nature of the durations in future work.

4.3 State-level Analysis

The aggregate results are encouraging. The empirical models are relatively stable and produce reasonable out-of-sample projections. However, a key goal of the project is to produce regional statistics on business formation. To that end, we focus on the performance of the models at the individual state level. Figure 10 provides information on the unweighted distribution of percent differences between the projected number of startups and the actual startups at the state level. The state-level percent differences are presented for each quarter in the form of a box-whiskers plot. The quarters presented are 2009:Q1 to 2013:Q4, with the projection errors for 2013 coming from the out-of-sample exercise. The box shows the interquartile range, the white line segment within the box the median, and the length of the whisker is an indicator of the spread in the upper and lower tails of

[29] It is a well known problem that the results from the Oaxaca-Blinder decomposition can be sensitive to choice of base period.

the distributions. Each box is constructed based on 51 data points – the 50 states and the District of Columbia. There are some outliers on the tails of the distribution with percent differences close to +/- 15 percent. The variation in prediction errors also increases in the out-of-sample periods. Nevertheless, the interquartile range is bounded roughly by a prediction error of +/-5 percent.

As a second check, we examine the correlation between the growth in actual startups and the growth in projected startups at the state-level. The growth in startups is not explicitly modeled, only the levels are; examining the fit of the growth rates for the model's projections provides a separate evaluation of the model. The out-of-sample projections and the actual startups from 2013:Q1-2013:Q4 are used to construct annual growth rates at a quarterly frequency in startup activity at the state level. Figure 11 presents a scatter plot of the growth in the actual number of business startups versus the growth in the projected number of business startups. Each dot represents the annual growth rate at a quarterly frequency at the state level for the period 2013:Q1-2013:Q4 (four growth rates for each of the 50 states and DC). The figure includes a 45 degree line, along with a regression fit line. The regression line lies very close to the 45 degree line, indicating there is no systematic bias in projections. The correlation of the actual and projected transitions is 0.680, implying that the growth in modeled applications does yield good information about the growth in startup activity at the state-level.

We also constructed correlations between the growth in actual startups and the growth in three of our main application series: overall business applications, the wage date cases, and high propensity applications. Here, we want to assess whether the modeling exercise significantly improves the information on business formation in comparison to the information directly available in the various applications series. The correlations in the quarterly growth rates between startups and the three application series are smaller in the out-of-sample period (2013) compared to the correlation between actual startups and projected startups. The contemporaneous correlations between startup growth and the growth in business applications, wage date cases, and high propensity applications are 0.37, 0.65 and 0.55, respectively. While the correlation between the growth in wage date cases and start-up growth is nearly the same as the projections in the 2012-2013 data, this is not always the case. We typically find the out-of-sample projection correlations are .03 to .10 higher than the in-sample wage date correlations.

Overall, the results at the national and state level suggest that the modeling of the application data to produce new measures of business formation is a promising approach

to extracting a stronger signal than one gets from looking only at the relatively noisy business applications. While we have not focused on the specific trends in the business projections series, Figure 7 suggests that business formation remains muted through 2016 compared to the levels seen prior to the Great Recession. This result agrees with the general patterns of job creation rates reported in recent Business Employment Dynamics and Quarterly Workforce Indicators data that show continued low levels of the job creation by young firms through 2015 and the middle of 2016.

5 Employment of Business Startups

This section briefly reports on the link between application responses and the initial employment levels of business startups. We ask the question: Are some application types associated with larger businesses at startup? The matched application-business register data allows for the measurement of employment at the firm level at the quarterly frequency. For each application record that transitions to an employer business, the maximum employment that is observed in the first four quarters of the firms' life is used as the measure of startup employment size. The underlying data include all applications that transition to employer status in 2012 or 2013 within the eight-quarter window, encompassing over 700,000 births.[30]

Table 9 shows the mean of employment size of startups broken down by five different application categories: with planned wages, legal form of organization status, LLC applications, high propensity applications, and for selected industries. Average employment size in the first year is 2.2 employees larger for applications that provide a wage date than those that do not provide a wage date. Partnerships have higher average employment size than sole proprietorships or corporate entities. Multi-member LLC's have an average startup employment size of 10.5 employees, almost 4.5 employees greater than non-LLC members. The high-propensity applications are, on average, 2.2 employees larger than non-high propensity applications that become employer businesses. The final four rows report how employment size varies by selected industry sectors, focusing on sectors with the largest

[30] There are a small number of new EIN filings that are observed with an initially large level of employment. Many large employment cases are associated with employment leasing firms, spinoffs, administrative actions, or merger and acquisition activity. We exclude these initial large employer cases from the startup size calculations. Some of these cases may be identified as non-births as a result of future processings of the Longitudinal Business Database, which the public-use product, Business Formation Statistics, incorporates in annual updates.

and smallest average employment sizes. The manufacturing and hospitality-leisure sectors have relatively large startup sizes, while the wholesale trade, retail trade, and personal service sectors have relatively small startup sizes. The patterns suggest that one possible avenue of future research is to examine whether application characteristics could be used to generate projections of the employment activity of startups. For this exercise, the information on the expected maximum number of employees in the next 12 months likely will be useful (SS-4 Form: Line 13). Our preliminary analysis indicates that there is a high degree of positive correlation between the expected and the actual maximum number of employees in the first year of a business, conditional on transitioning to an employer.

6 Applications and Regional Economic Activity

The above analysis shows that the application activity can be linked directly to business formation. A second potential use of the application data is as direct measures of economic activity, especially at the regional level. A key shortcoming of many measures of regional economic activity is that they are not very timely or, if timely, face substantial revisions. Series based on EIN applications could be produced on a quarterly or monthly basis subject to little revision (except for the seasonal factors and updates of industry codes). A main question, however, is whether application activity is correlated with local economic conditions.

As a first step to addressing this question, the application data has been geocoded to the state, county, census tract, and block levels. Over 99 percent of applications were coded to the state and county level, and 85 percent of applications were coded down to the census tract and block levels. Using the geocoded data, we first provide a broad assessment of the variation in applications and transitions over time and across states, with an eye on how states fared before and after the Great Recession. We then explore in more detail the correlation between economic activity and business applications over time at the county and core-based statistical area (CBSA) levels.

6.1 State-Level Variation in Applications and Transitions

To show how application activity and transitions to employer businesses have changed over time and across states, we present two sets of heat maps. The set of maps in Figure 12 depict how high-propensity applications vary across states and over time. The quarterly

number of high-propensity applications is normalized by state population, and states are grouped into six categories based on the level of high-propensity business applications per capita (per 1,000 people). The map for 2006 indicates a high degree of variation across states in the applications per capita. Many states in the west, as well as those in the East Coast, tend to have high levels of applications per capita, whereas states in the middle exhibit lower levels of application activity. In particular, Nevada, Florida and Delaware stand out with more than 2 high-propensity applications per capita. In contrast, West Virginia and many states in the Midwest have less than one application per capita.

The 2010 map in Figure 12 shows that the number of high-propensity applications per capita declined broadly after the Great Recession. The decline persists into 2014, with particular weakness in application activity in the middle of the country. While the 2016 map indicates slight recovery for some states, many states still have much lower levels of high-propensity applications per capita in 2016 compared to the pre-recession levels in 2006.

Overall, Figure 12 shows that the volume of application activity per capita varies significantly across states and responds to the changing economic conditions brought about by the Great Recession. What about the success rate of applications in becoming employer businesses? How does it vary across states and over time? Figure 13 presents heat maps that illustrate the variation across states in the number of business formations (within a 4-quarter window) per high-propensity business application made in a given quarter. This measure can be interpreted as the average success rate of a high-propensity business application in turning into an employer business. The maps for 2006 and 2010 use actual business formations, whereas the maps for 2014 and 2016 use model-based projected business formations.[31]

As in the case of applications per capita, there is considerable variation across states in business formations resulting from high-propensity applications. For instance, in the pre-Great Recession year of 2006, Florida and Nevada had average success rates less than 1 business formation for every 5 high-propensity business applications. These rates contrast with the relatively better performance of these two states in terms of high-propensity applications per capita in Figure 12. On the higher end, Idaho, Montana, North and South Dakota, and Vermont had success rates that exceed 2 business formations for every 5 high-propensity applications.

[31] Footnote 16 and Appendix A discuss how the 4-quarter formation window is implemented.

Figure 13 also shows that the number of business formations per high-propensity application declined after the Great Recession. From 2006 to 2014, many states experienced a drop in the average success rate, though the decline does not appear to be as broad and pronounced as in the case of high-propensity business applications per capita. Projections of business formations based on applications in 2016 suggest an increase in some states' success rates since 2014. At the same time, the relative ranking of states' success rates does not change as much over the years, which hints at the possibility that success rates reflect highly persistent state-specific factors.[32] The variation in success rates may stem from differences across states in the distribution of entrepreneurial ability, the degree of competition entrepreneurs face for each business opportunity, population density, and the types of business activity specific to a state. We plan to explore further the patterns in Figures 12 and 13 in future work to better understand the sources and implications of the variation across states in the application activity and the success rate.

6.2 County-Level Analysis

The exercises in this section exploit the county-coded applications data. Our initial analysis focuses broadly on whether counties in the United States that experience above average economic activity also experience above average application activity. To measure county-level economic activity, we classify counties into employment growth quintiles over the recent recovery, 2010-2016. Each county is placed in a growth quintile based on its employment growth from 2010 to 2016 using first quarter employment from the Local Area Unemployment Statistics program of the BLS. The application data are then grouped by county based on the employment growth quintiles. The application series that we examine are business applications (BA) and high propensity applications (HBA), using only the first quarter applications. Figure 14a shows the plot of the business application series disaggregated into the five county employment growth categories. Application activity fell less in high growth counties during the recession and expanded at a greater pace during the recovery. Figure 14b shows the same chart for the high propensity application group. Since, the end of the recession in 2009, high propensity applications have rebounded more in higher growth counties.

A second analysis looks at application growth for counties that have been most involved

[32] The correlations of 2010, 2014, and 2016 state rankings with the 2006 ranking are 0.94, 0.94 and 0.92, respectively.

in the recent shale oil and gas boom. In this exercise we divide the 3,141 U.S. counties into 4 categories. Counties in states with no shale oil and gas activity (2,366 counties), counties in states with shale oil and gas activity but not within a shale gas/oil field (490 counties), counties in states with shale oil and gas activity and within a shale gas/oil field but not in a core drilling area (217 counties), and counties in states with shale oil and gas activity and with core drilling activity (68 counties). Shale oil and gas activity is identified using information on fields from the Energy Information Agency (EIA) and on drilling statistics from state agencies. Figures 15a and 15b show the business applications and high propensity series from 2010 to 2016, during the period of rapid expansion in shale oil and gas activity. Business application in core shale oil and gas counties are consistently higher than application counts for the other three county groupings, though the relative gaps start to diminish in 2015 as application counts outside the core areas pick up. High propensity application activity for core counties peaks in 2015:Q1 and then experiences a decline in 2016:Q1. This is in-line with the decline in overall drilling activity that began at the end of 2014 and continued through the start of 2016.

6.3 CBSA-Level Analysis

The next analysis examines the correlations between application activity and house price growth. A simple empirical model is estimated that regresses the growth in the number of applications of a particular type on the growth in house prices and the growth in unemployment. The geographic unit of observation is the CBSA, as the Federal Housing Administration (FHA) provides information on house prices at that level. We also include the unemployment rate and year effects to control for overall economic activity. The models are estimated in log-difference form using annual data from 2006 through 2015, and they include estimated fixed effects at the CBSA level. The sample includes data on applications, house prices, and unemployment rates for 401 CBSAs.

Table 10 shows the results of this analysis. Each column of the table reports the results for a different application type: overall business applications, high propensity applications, corporate applications and wage applications. The final column presents the results of a similar regression using the growth in the 4-quarter window births as the dependent variable. The results are quite consistent across the specifications – there is a positive correlation between house price growth and the growth of applications and a negative correlation between the growth in unemployment and application growth. The last column

shows that the growth in 4-quarter window births are negatively correlated with the growth in the unemployment rate and positively correlated with the growth in home prices. Figure 16 provides further information on the patterns between house prices and application activity. The chart depicts the time series of high propensity applications broken into house price change quintiles. The quintiles are based on house price growth in CBSAs over the period 2006-2010, the period of the housing price collapse in the United States. The chart shows that CBSAs with greatest (lowest) house price declines experienced substantially higher (lower) cumulative reductions in application activity from 2006 through 2010.

We perform a final check on the link between applications and regional activity by examining how application activity is correlated with business formation using data from the Business Dynamics Statistics (BDS) program. The release of the BDS covering years 1977-2014 includes statistics on the annual number of new firms births that appear in a CBSA. Figure 17 shows the relationship between the number of new firms per capita in a CBSA and the number of high propensity applications per capita. Each dot represents a CBSA, along with the linear fit line. The underlying data are from 2013 and represent the 355 CBSAs that the BDS reports on. The chart depicts a strong positive relationship between applications per capita and business formations per capita. The correlation is .765, indicating that high-propensity application activity provides solid information about business formation activity at a relatively disaggregated geographic level.

On balance, our initial analysis of the relationship between county and CBSA level variables and application volumes is suggestive that application activity may provide useful information about local economic conditions. The current analysis is only descriptive and more work needs to be performed to establish the timing of the relationships, but our sense is that business applications and projected business formations may act as leading variables in gauging local economic conditions. Work is under way to analyze the properties of various business application and formation series as potential economic indicators.

7 Public Use Files

A main output of this research project is the development of the public use data, Business Formation Statistics (BFS), on application filings and business formation at a regional level. The BFS are available at the Census Bureau's website and include a set of applica-

tion and business formation series at the state and U.S. levels at a quarterly frequency.[33] The data are presented both non-seasonally adjusted and seasonally adjusted based on the Census Bureau's X-13ARIMA-SEATS utility.[34] The application data series released include the following series, which are a subset of those shown in Figure 4: business applications (BA), high-propensity business applications (HBA), business applications with planned wages (WBA), business applications from a corporation (CBA). The latter two groups, WBA and CBA, are proper, but not mutually exclusive, subsets of the more comprehensive group HBA. The business formation series released include: business formations within either 4 or 8 quarters (BF4Q and BF8Q), projected business formations within either 4 or 8 quarters (PBF4Q and PBF8Q), and two series that splice together the actual and projected business formations within either 4 or 8 quarters (SBF4Q and SBF8Q). The series, projected business formations within 4 quarters, is based on a probit model that uses the 2012-2014 sample with full interactions, estimated separately for each state. Similarly, the series, projected business formations within 8 quarters, is based on a probit model that uses the 2011-2013 sample with full interactions, estimated separately for each state. Finally, two series are provided to give information on the delay in business formation: Average duration between business application and formation, conditional on the application turning into an employer business within either 4 or 8 quarters (DUR4Q and DUR8Q). Appendix B provides a full description of the data included in the BFS.

8 Concluding Remarks

Business applications data provide a novel, timely and granular source for tracking new business activity. Using the information contained in an application, we find that we can generate accurate and timely indicators of business startup activity at the national and local levels. The public-use data, Business Formation Statistics, include the projection series, as well as various measures of business application volumes, so that users can develop their own approach and interpretation of these novel series.

We regard our approach to developing timely and granular quarterly series as being very promising, but further progress can be made to improve and refine the methodology and extend the analysis. In particular, we would like to assess whether application series can be produced at a monthly frequency and explore whether public use files at the county

[33] Visit the BFS website https://www.census.gov/programs-surveys/bfs.html.

[34] For more on X-13ARIMA-SEATS, see https://www.census.gov/srd/www/x13as/.

level can be released. We have also begun to examine whether applications can be used to project the total number of new businesses formed in a given time period, as opposed to the forward-looking window approach described above.

While not reported here, we have also begun to examine how application activity is related to local demographics and the economic structure of neighborhoods. At the tract level, application activity appears strongly related to measures of workforce skill, population demographics, and the level of employment activity in a neighborhood. Finally, we plan to explore: (1) the link between EIN applications and the non-employer universe; (2) EIN application submissions by mature businesses and how they relate to local economic conditions and to applications by new businesses; (3) whether business applications are a leading indicator of other types of economic activity, such as employment, housing starts, manufacturing and retail sales at the national and local levels; and (4) the potential use of machine learning techniques to filter applications and to model business formations in a more flexible way.

References

[1] Adelino, Manuel, Antoinette Schoar and Felipe Servino, 2015, "House Prices, Collateral and Self Employment," Journal of Financial Economics, 117(2): 288-306.

[2] Davis, Steven, John Haltiwanger, Ron Jarmin, C.J. Krizan, Javier Miranda, Alfred Nucci, and Kristin Sandusky, 2009,"Measuring the Dynamics of Young and Small Businesses: Integrating Employer and Non-Employer Businesses," in Producer Dynamics: New Evidence from Micro Data, Dunne, Jensen, and Roberts (eds.), NBER/University of Chicago Press, 329-366.

[3] Decker, Ryan, John Haltiwanger, Ron Jarmin and Javier Miranda, 2015,"Changing Business Dynamism and Productivity: Shocks vs. Responsiveness," NBER Working Paper 24236.

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

[5] Dunne, Timothy, J. Bradford Jensen, and Mark Roberts, 2009, Producer Dynamics: New Evidence from Micro Data, NBER/University of Chicago Press.

[6] Fairlie, Robert, 2005, "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models," Journal of Economic and Social Measurement, 30(4): 305-316.

[7] Fort, Teresa, John Haltiwanger, Ron Jarmin and Javier Miranda, 2013, "How firms respond to business cycles: The role of firm age and firm size," IMF Economic Review 61 (3), 520-559.

[8] Glaeser, Edward, Stuart Rosenthal, and William Strange, 2010,"Urban Economics and Entrepreneurship," Journal of Urban Economics, 67(1): 1-14.

[9] Glaeser, Edward, and William Kerr, 2009,"Local Industrial Conditions and Entrepreneurship: How Much of the Spatial Distribution Can We Explain," Journal of Economics and Management Strategy, 18(3): 623-663.

[10] Guzman, Jorge and Scott Stern, 2015, "Nowcasting and Placecasting Entrepreneurial Quality and Performance," NBER Working Paper No. 20954.

[11] Guzman, Jorge and Scott Stern, 2016, "The State of American Entrepreneurship: New Estimates of the Quality and Quantity of Entrepreneurship for 15 States, 1988-2014," NBER Working Paper No. 22095.

[12] Haltiwanger, John, Ron Jarmin and Javier Miranda, 2013,"Who Creates Jobs? Small vs. Large vs. Young," Review of Economics and Statistics, 95(2): 347-361.

Table 1. Type of Entity

Entity	Applications (%)	4-Qtr Window Birth (%)	8-Qtr Window Birth (%)
Sole Proprietorship	24.2	8.8	9.9
Partnership	8.3	9.2	10.8
Corporation	24.2	24.8	28.9
Personal Service Corp	0.4	25.9	29.6
Church Related	1.0	6.7	8.1
Nonprofit	3.2	2.8	3.8
Other	38.6	11.0	13.2
Miscellaneous	0.2	8.4	9.4
Total Applications	100	13.6	15.9

Source: EIN Applications Files, Census Bureau.

Table 2. Reason for Applying

Reason	Applications (%)	4-Qtr Window Birth (%)	8-Qtr Window Birth (%)
Start New Business	88.1	12.4	14.7
Hiring Employees	1.5	56.5	58.8
IRS Compliance	0.2	10.7	12.6
Banking Purposes	5.8	3.9	5.2
Changed Org. Type	2.4	32.5	36.7
Purchased Business	1.3	44.4	47.9
Other	0.8	9.7	11.4
Total Applications	100.0	13.6	15.9

Source: EIN Applications Files, Census Bureau.

Table 3. Industry of Application

Industry	Applications (%)	4-Qtr Window Birth (%)	8-Qtr Window Birth (%)
Agriculture, Mining, Util.	0.5	12.1	13.7
Construction	10.6	15.3	18.1
Manufacturing	1.9	17.2	20.5
Wholesale, Retail Trade	21.6	12.5	14.9
Financial, Business Serv.	32.2	11.1	13.3
Education, Health Serv.	7.6	17.2	20.6
Leisure, Hospitality	8.8	24.2	27.5
Personal Services	8.5	12.1	14.3
Missing Industry Code	8.3	9.8	11.4
Total Applications	100.0	13.6	15.9

Source: EIN Applications Files, Census Bureau.

Table 4. Presence of Wage Date

Wage Date Presence	Applications (%)	4-Qtr Window Birth (%)	8-Qtr Window Birth (%)
No Wage Date Given	76.4	5.4	7.0
Wage Date Given	23.6	36.8	40.2
Total Applications	100.0	13.6	15.9

Source: EIN Applications Files, Census Bureau.

Table 5. High Propensity Applications

Propensity	Applications (%)	4-Qtr Window Birth (%)	8-Qtr Window Birth (%)
Not High Propensity	50.7	2.8	3.8
High Propensity	49.3	23.7	27.0
Total Applications	100.0	13.6	15.9

Source: EIN Applications Files, Census Bureau.

Table 6. Model Fit: RMSDs

Est. Sample	Specification	2013(U)	2013(W)
2006-2012	LPM, By State, Full Inter.	5.94	5.10
2008-2012	LPM, By State, Full Inter.	5.34	4.24
2010-2012	LPM, By State, Full Inter.	5.29	4.04
2011-2012	LPM, By State, Full Inter.	5.77	3.96
2010, 2011, 2012	LPM, By State, Full Inter.	5.35	4.09
2008-2010	LPM, By State, Full Inter.	5.79	4.99
2009-2011	LPM, By State, Full Inter.	5.52	4.53
2005-2007	LPM, By State, Full Inter.	14.71	13.97
2006-2008	LPM, By State, Full Inter.	8.21	7.04
	Alternative Specifications		
2010-2012	LPM, Pooled, No Inter.	5.58	4.27
2010-2012	LPM, Pooled, Full Inter.	5.56	4.30
2010-2012	LPM, By State, No Inter.	5.44	4.08
2010-2012	Probit, By State, Full Inter.	5.15	4.03

Note: 2013 RMSD are out-of-sample.

Table 7. Projection Model: Year Effects

Year	Parameter
2005	.0084 (.0004)
2006	- .0137 (.0004)
2007	- .0238 (.0004)
2008	- .0251 (.0004)
2009	-.0284 (.0004)
2010	- .0256 (.0004)
2011	- .0274 (.0004)
2012	- .0280 (.0004)
2013	- .0282 (.0004)

Table 8. Decomposition Analysis: 2006-2013

Decomposition	$\Delta Startups$	ΔP (%)	ΔN (%)
Startups	-146,000	92.0	8.0
	ΔP	ΔZ (%)	$\Delta \beta$ (%)
Propensity (2006-2008 Base Parameters)	-.034	78.6	21.4
Propensity (2011-2013 Base Parameters)	-.034	78.3	21.7

Table 9. Startup Employment Size by Selected Application Characteristics

Application Type	Average Startup Size (Employees)	Std. Error of Avg.
No Wage Date	6.0	0.039
Wage Date Given	8.2	0.044
Sole Proprietorship	5.9	0.028
Partnership	9.1	0.226
Corporation	6.6	0.042
Not an LLC	6.1	0.034
Single Member LLC	7.9	0.077
Multi Member LLC	10.5	0.078
Not High Propensity	5.4	0.068
High Propensity	7.6	0.034
Manufacturing	11.6	0.349
Retail and Wholesale Trade	5.3	0.044
Hospitality and Leisure	14.2	0.083
Personal Services	4.9	0.058

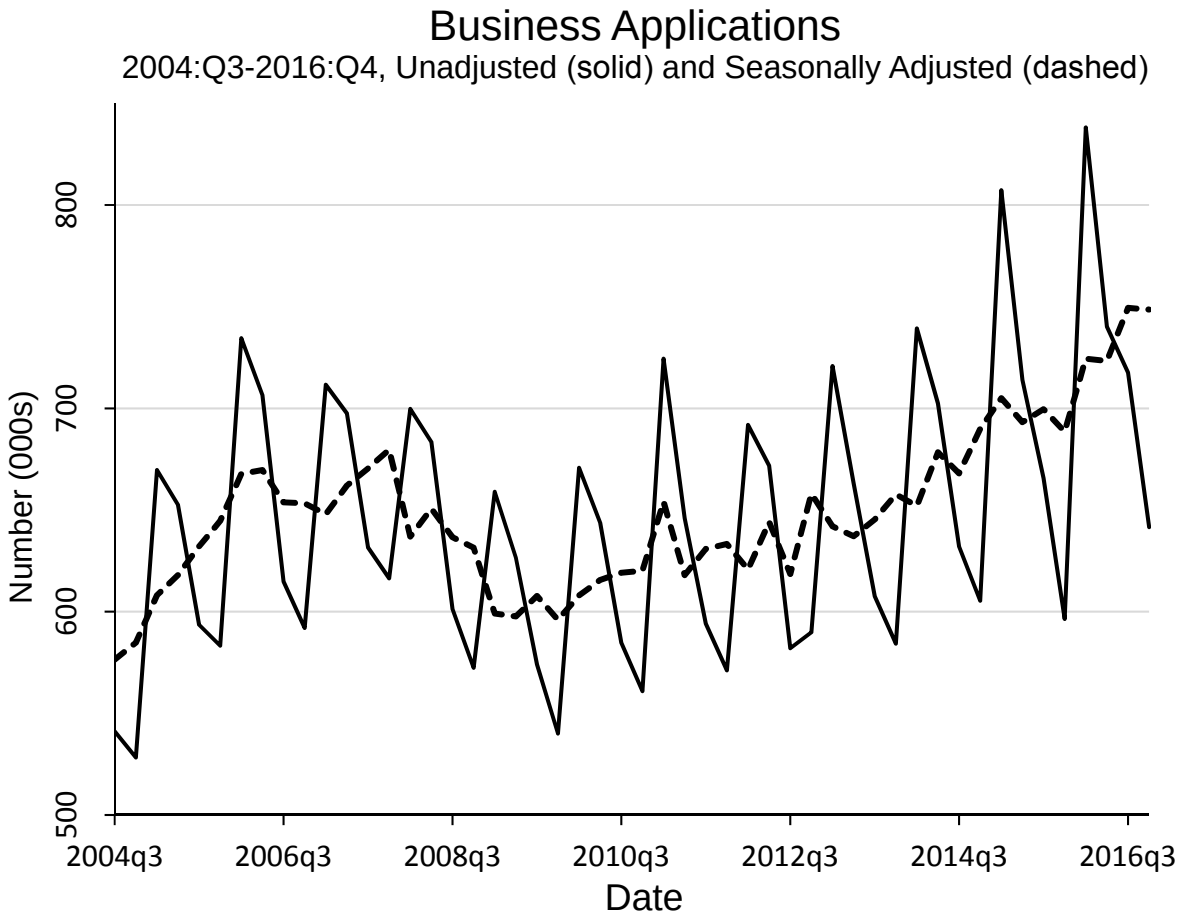
Note: Data include only the 2012 and 2013 employer business startups born in the 8-quarter window.

Table 10. House Prices and Application Activity

Variable	Business	High Propensity	Corporate	Wage	Window Births
House Prices	0.238	0.192	0.150	0.212	0.085
	(.024)	(.025)	(.040)	(.031)	(.043)
Unemployment	-0.035	-0.069	-0.076	-0.119	-0.090
	(.020)	(.021)	(.032)	(.025)	(.029)
N	4010	4010	4010	4010	3208
R^2	0.327	0.423	0.318	0.467	0.283

Note: Models include CBSA and year fixed effects and std. errors clustered at the CBSA level.

Figure 1



Source: EIN Applications File, US Census Bureau

Figure 2

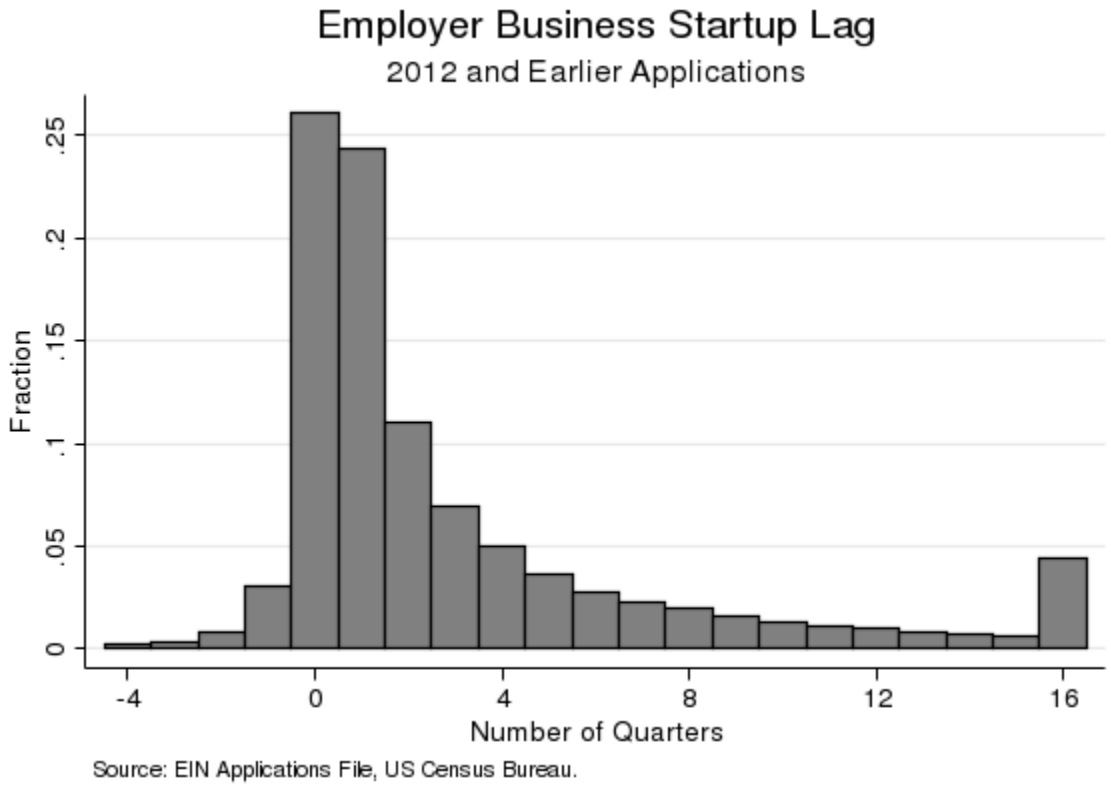

Employer Business Startup Lag
2012 and Earlier Applications
Fraction
0
.05
.1
.15
.2
.25
-4
0
4
8
12
16
Number of Quarters
Source: EIN Applications File, US Census Bureau.

Figure 3

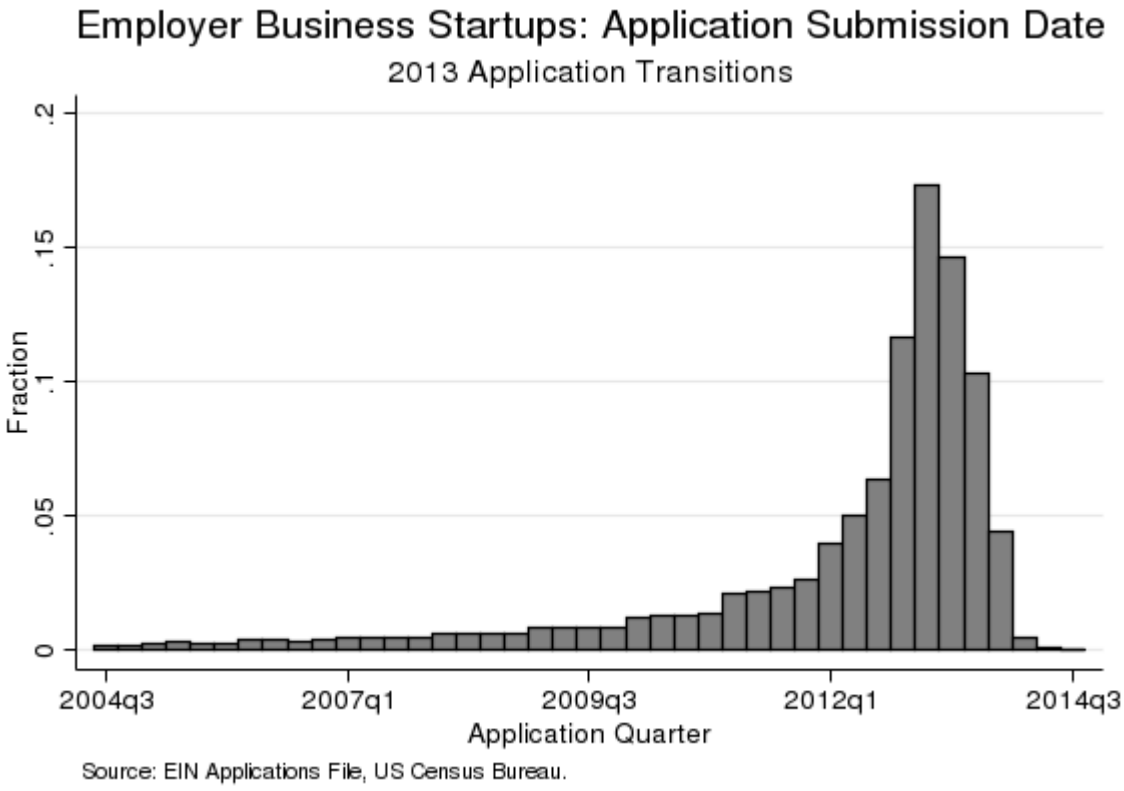
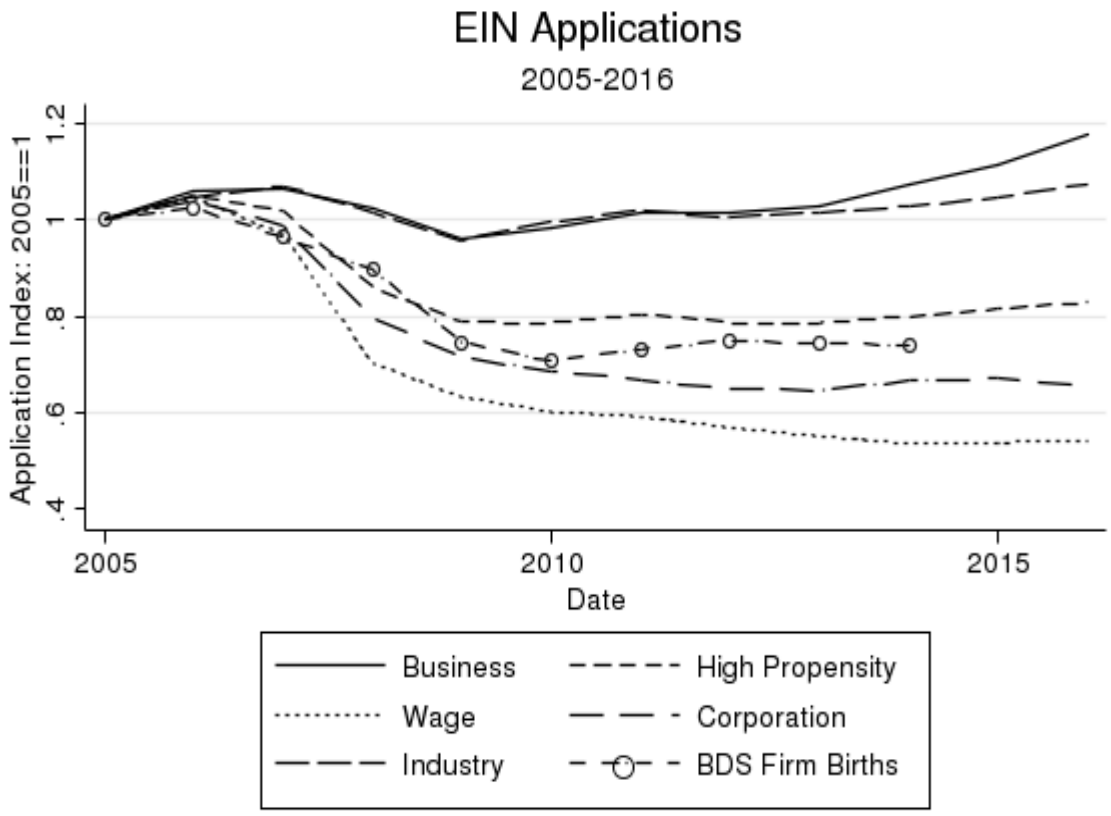

Employer Business Startups: Application Submission Date
2013 Application Transitions
Fraction
0
.05
.1
.15
.2
2004q3
2007q1
2009q3
2012q1
2014q3
Application Quarter
Source: EIN Applications File, US Census Bureau.

Figure 4



Source: EIN Applications File, US Census Bureau.

Figure 5

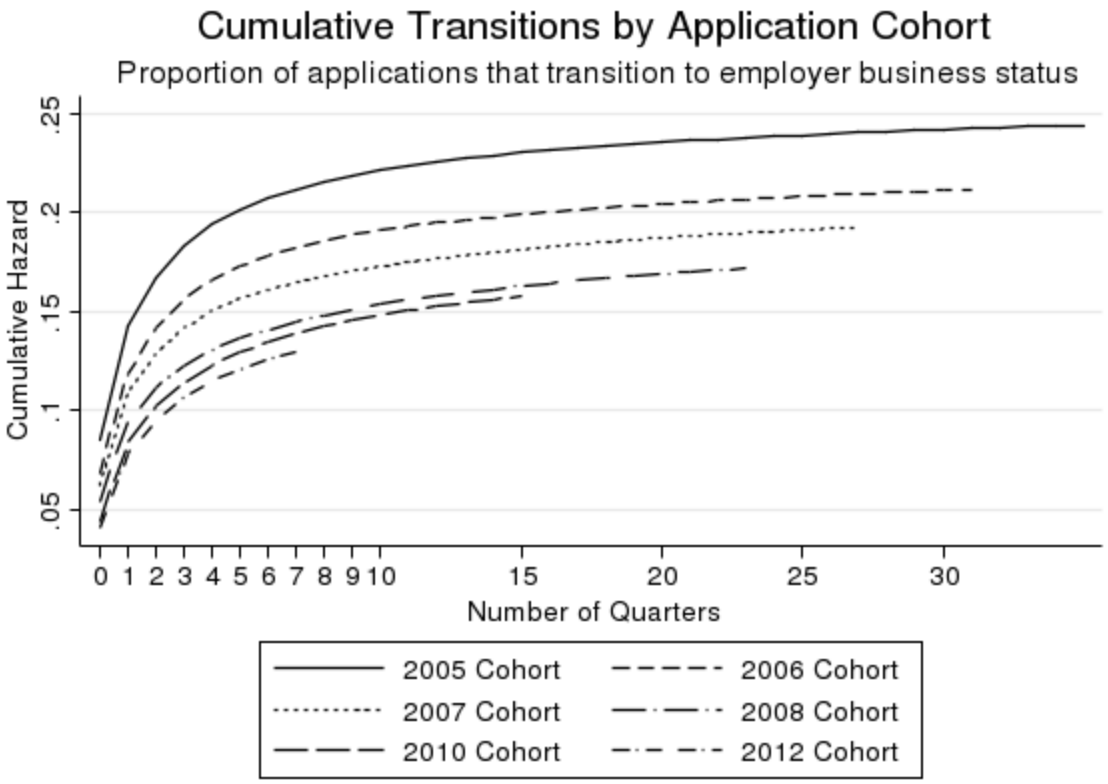
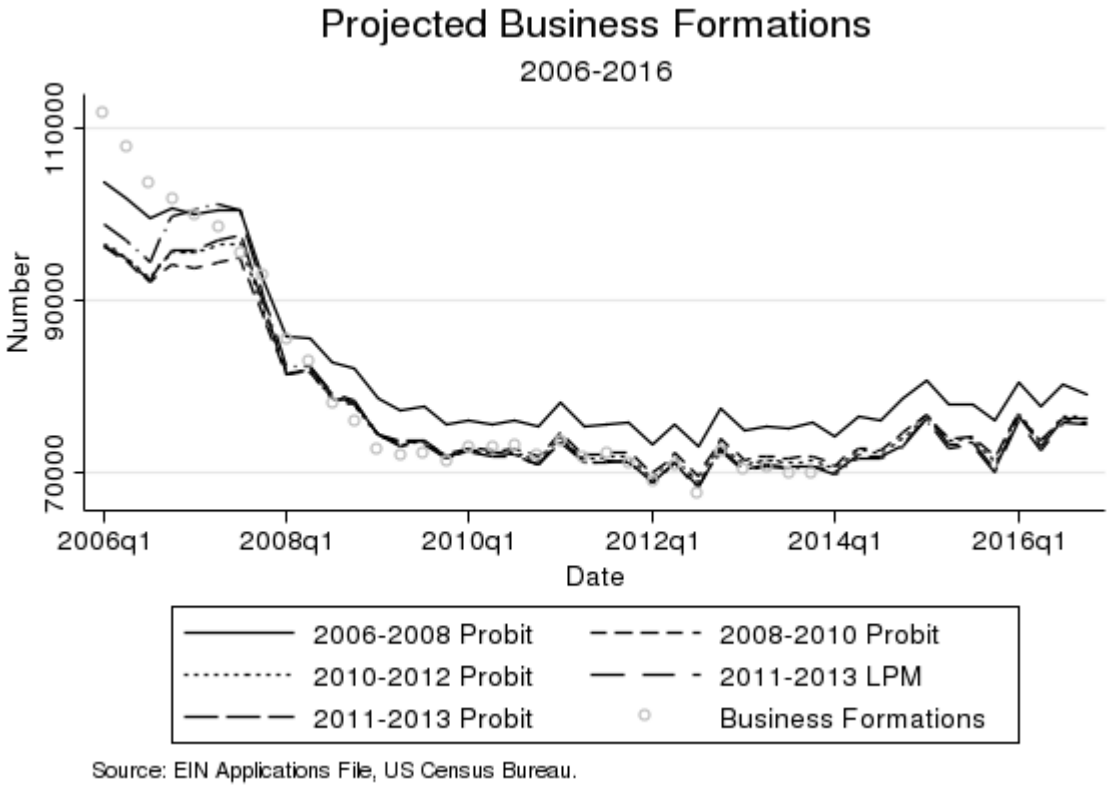
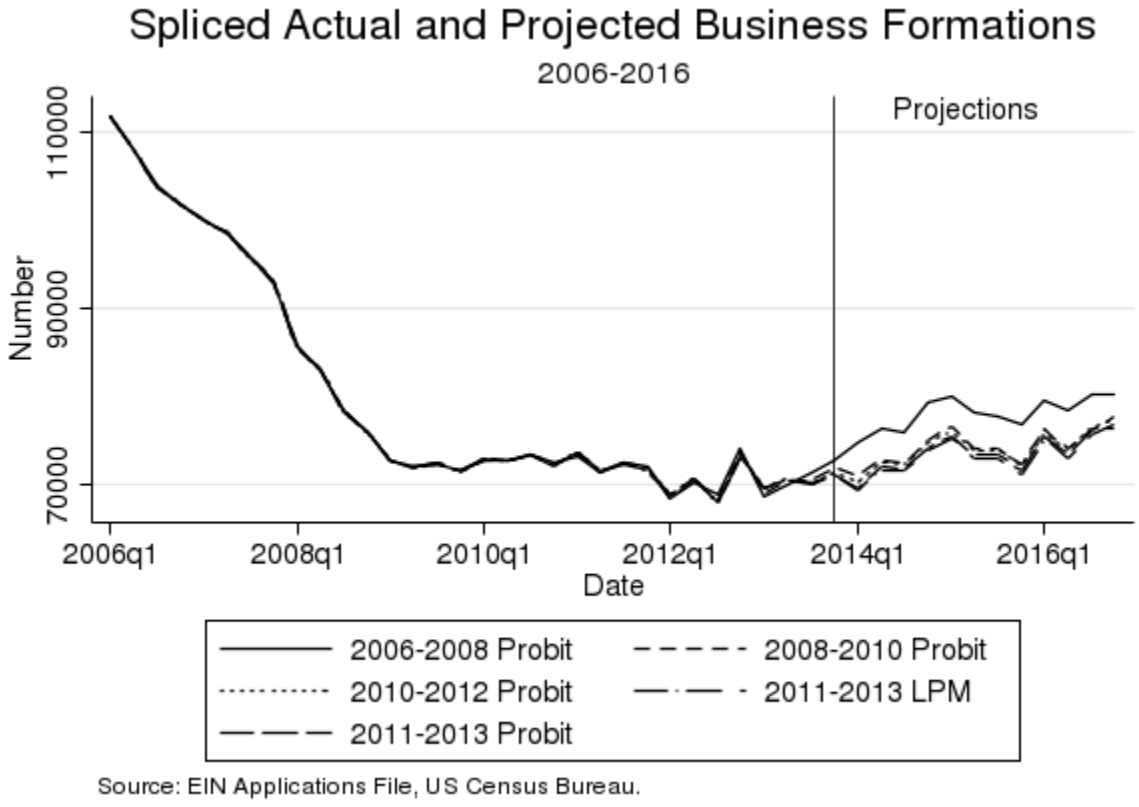

Cumulative Transitions by Application Cohort
Proportion of applications that transition to employer business status
Cumulative Hazard
.05
.1
.15
.2
.25
0 1 2 3 4 5 6 7 8 9 10
15
20
25
30
Number of Quarters
2005 Cohort
2006 Cohort
2007 Cohort
2008 Cohort
2010 Cohort
2012 Cohort

Figure 6



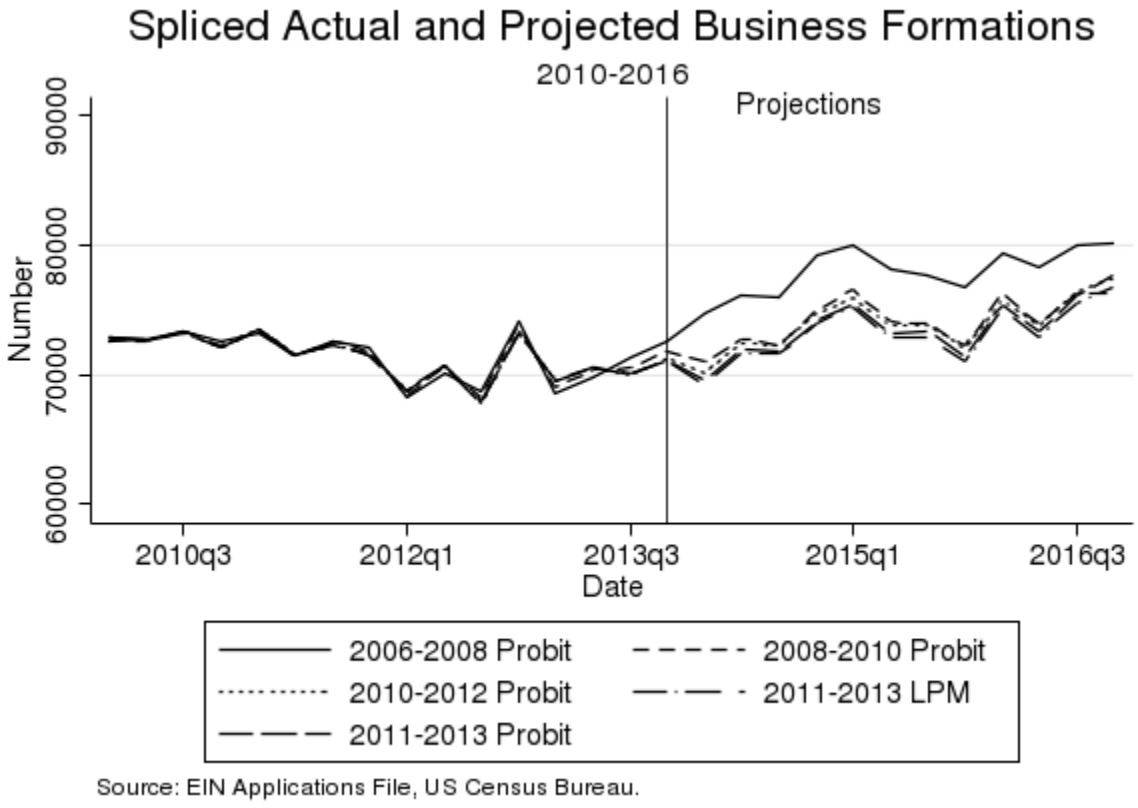
Source: EIN Applications File, US Census Bureau.

Figure 7a



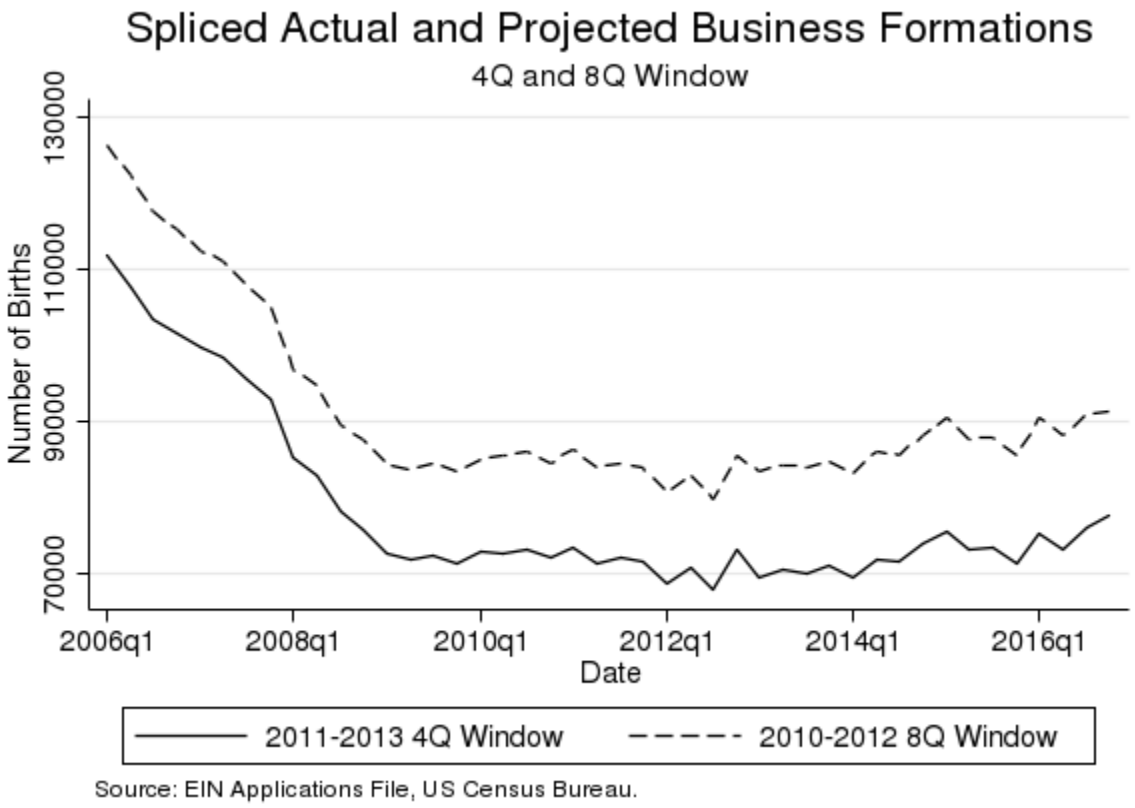
Source: EIN Applications File, US Census Bureau.

Figure 7b



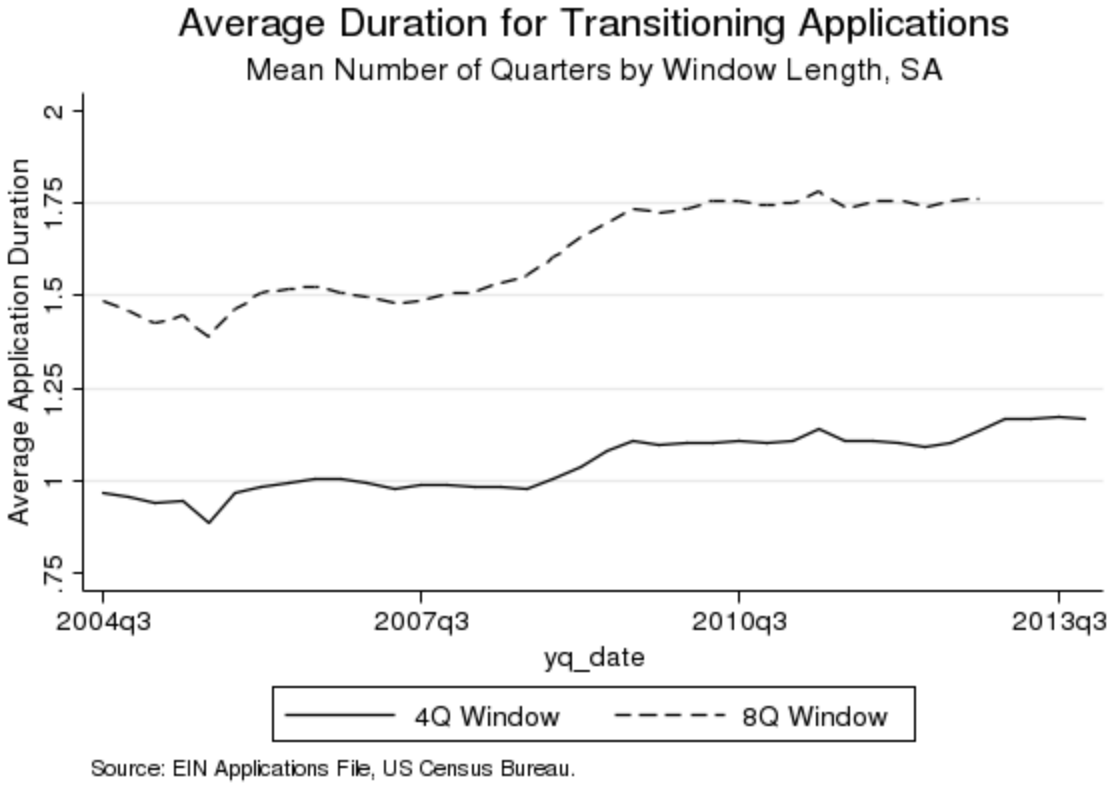
Source: EIN Applications File, US Census Bureau.

Figure 8



Source: EIN Applications File, US Census Bureau.

Figure 9



Source: EIN Applications File, US Census Bureau.

Figure 10

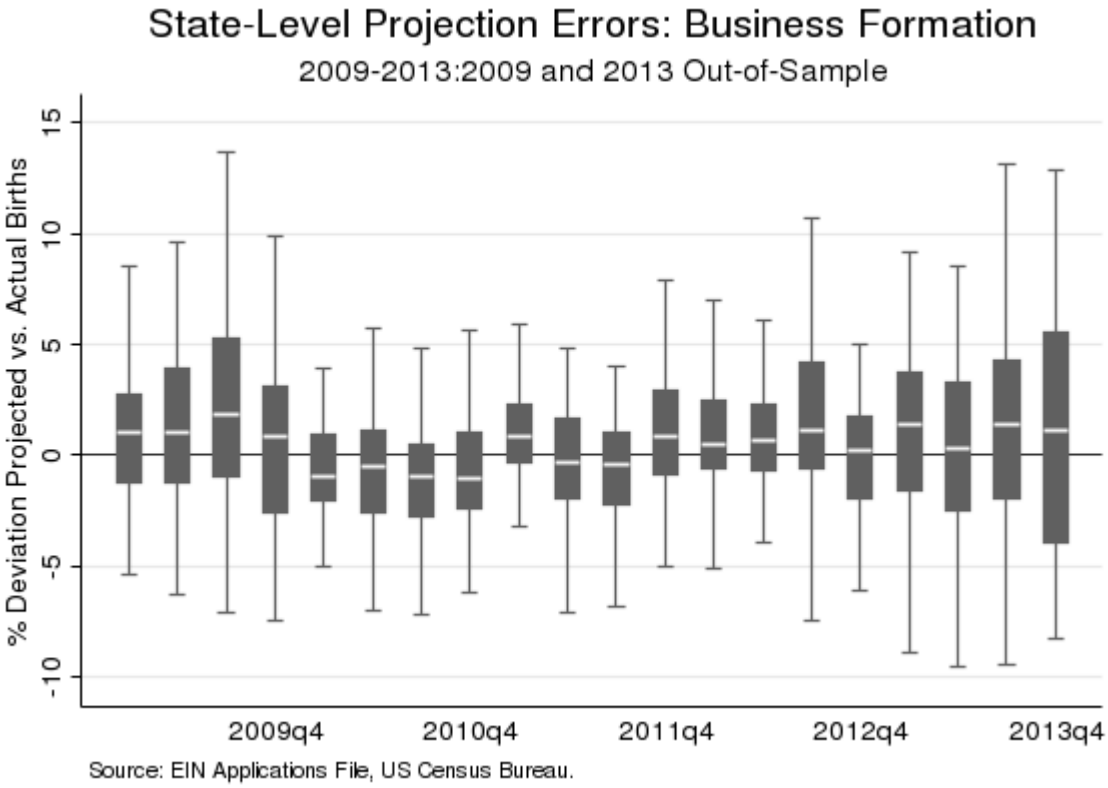

State-Level Projection Errors: Business Formation
2009-2013:2009 and 2013 Out-of-Sample
% Deviation Projected vs. Actual Births
-10
-5
0
5
10
15
2009q4
2010q4
2011q4
2012q4
2013q4
Source: EIN Applications File, US Census Bureau.

Figure 11

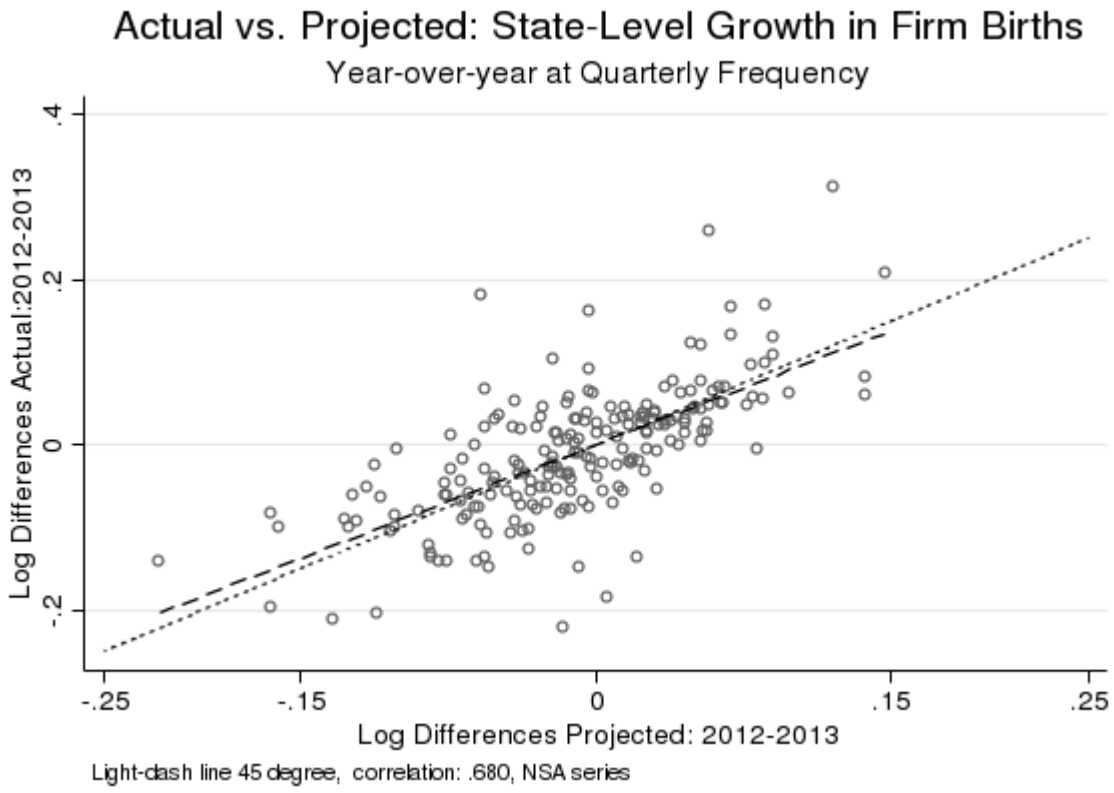

Actual vs. Projected: State-Level Growth in Firm Births
Year-over-year at Quarterly Frequency
Log Differences Actual:2012-2013
.4
.2
0
-.2
-.25
-.15
0
.15
.25
Log Differences Projected: 2012-2013
Light-dash line 45 degree, correlation: .680, NSA series

Figure 12. High-Propensity Business Applications per 1,000 People

Notes: Average of non-seasonally adjusted data across all quarters in a year by state; population estimates as of July 1.

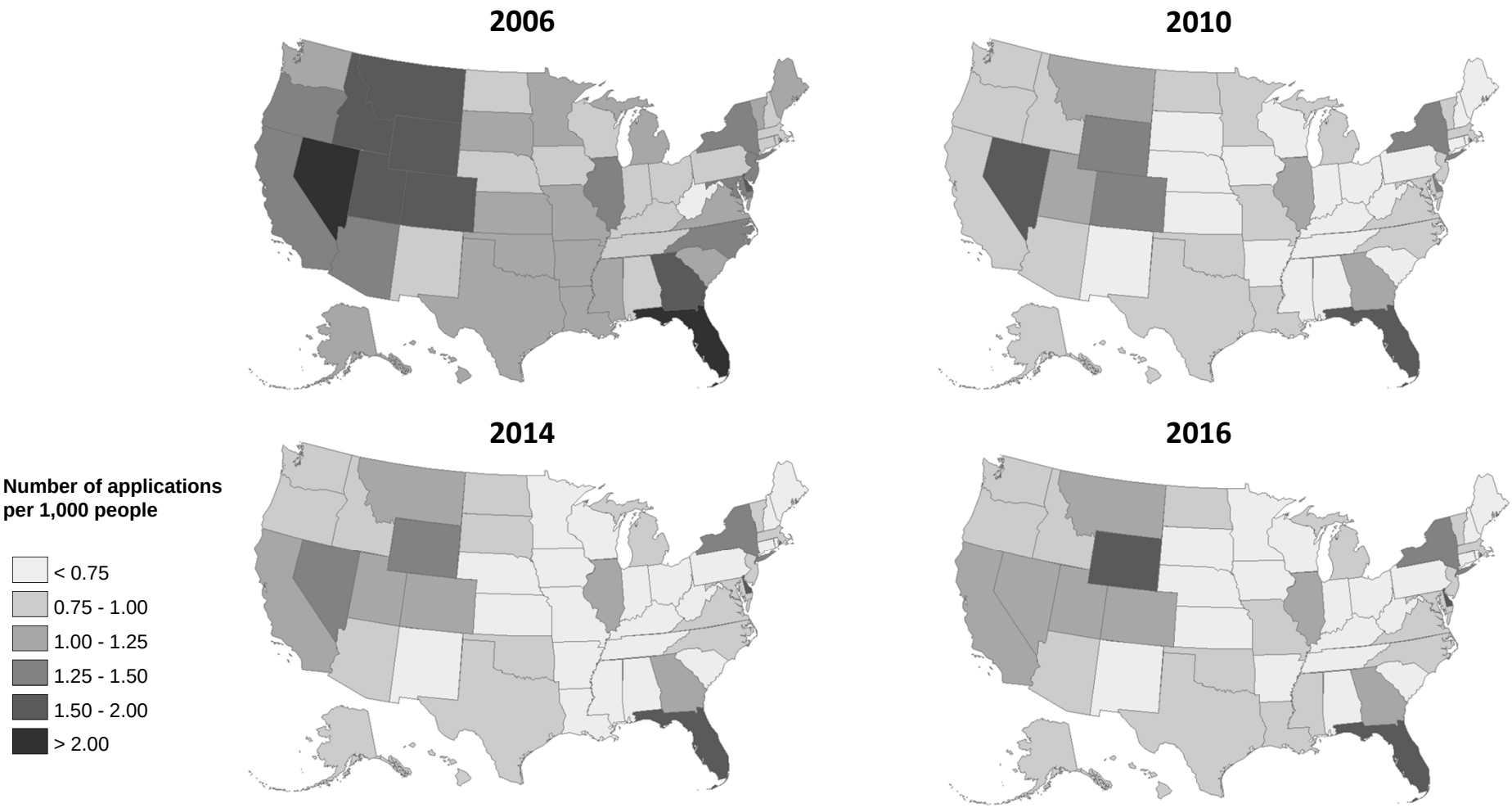


Figure 13. Business Formations (within 4 quarters) per High-Propensity Business Application

Notes: Average of non-seasonally adjusted data across all quarters in a year by state; actual formations for applications in 2006 and 2010, and projected formations for applications in 2014 and 2016.

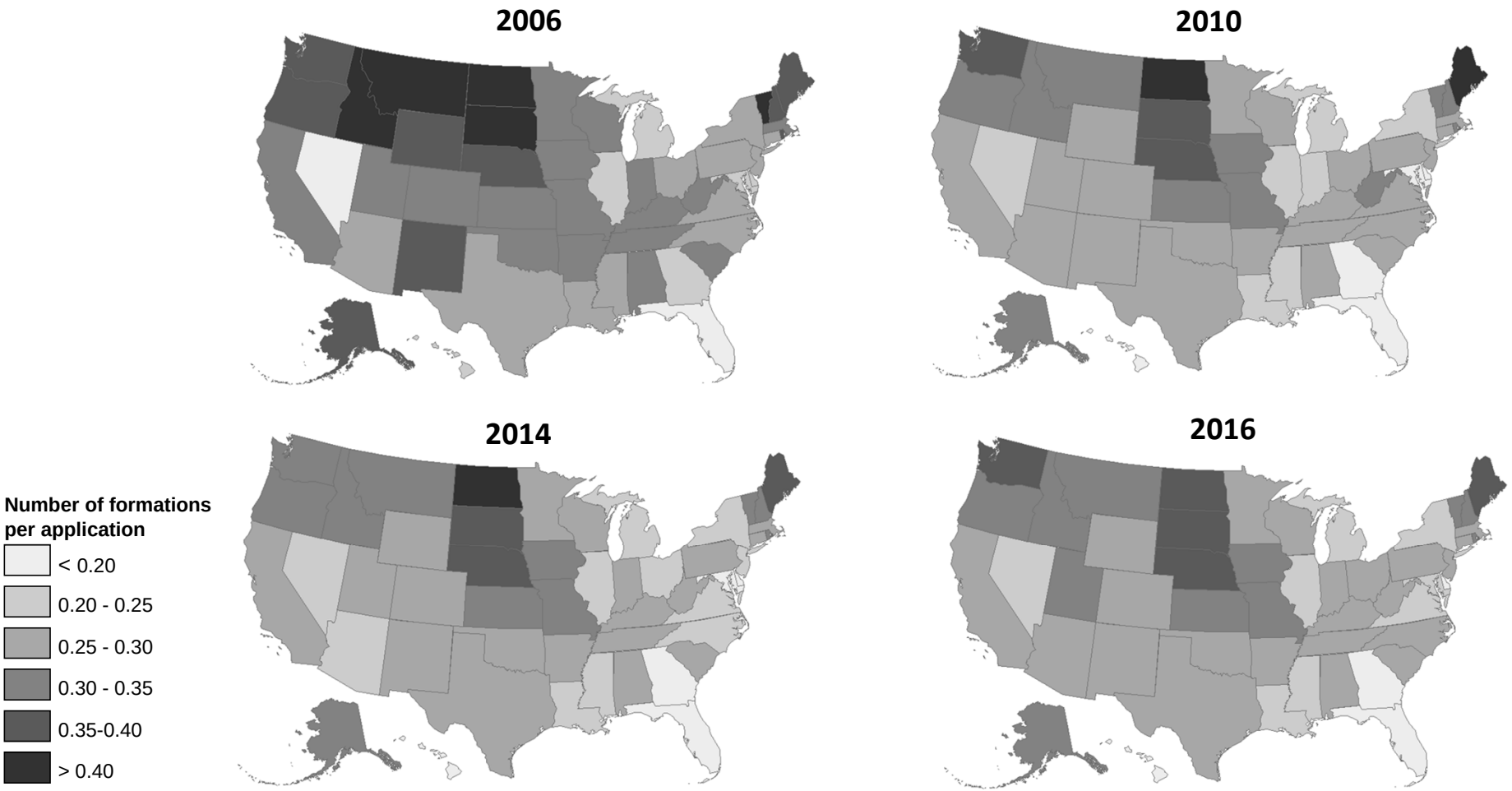
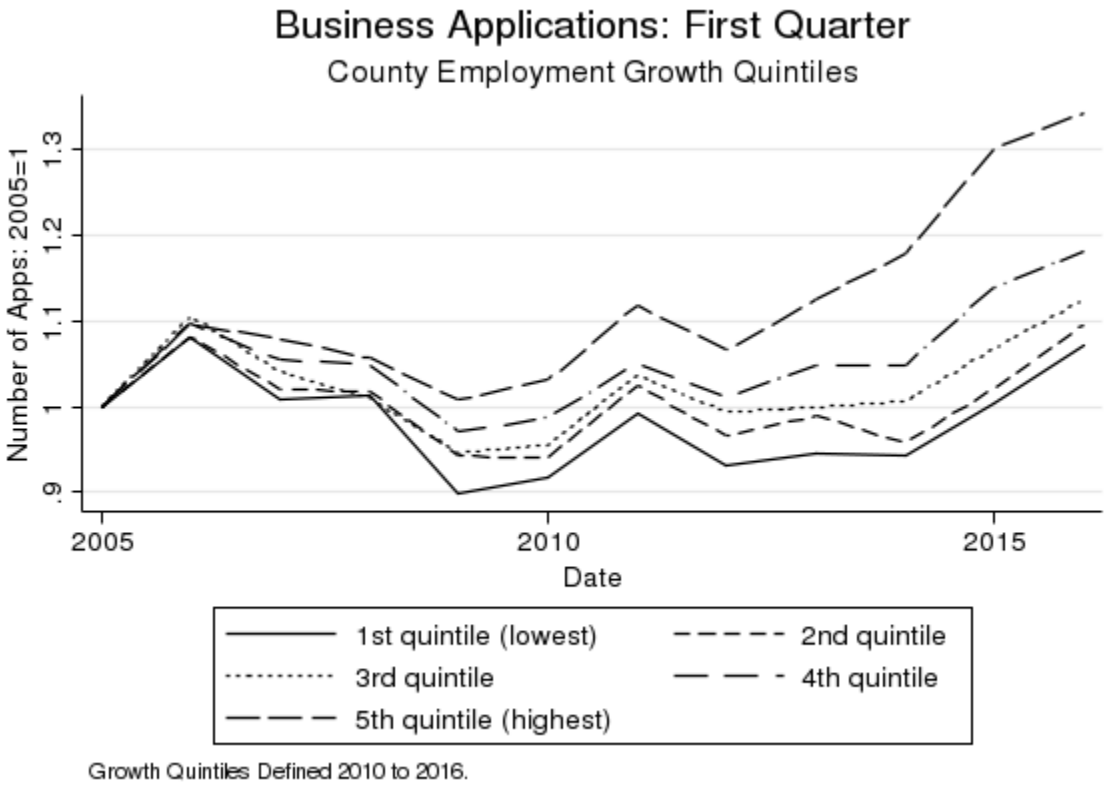
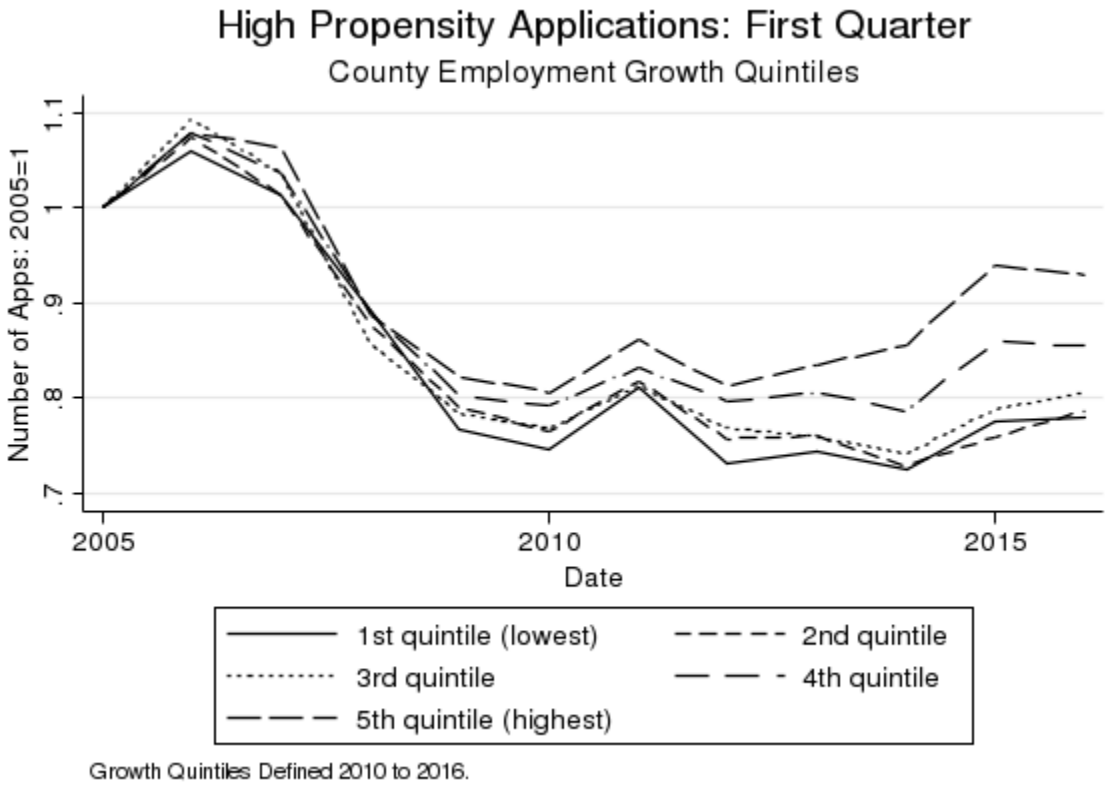


Figure 14a



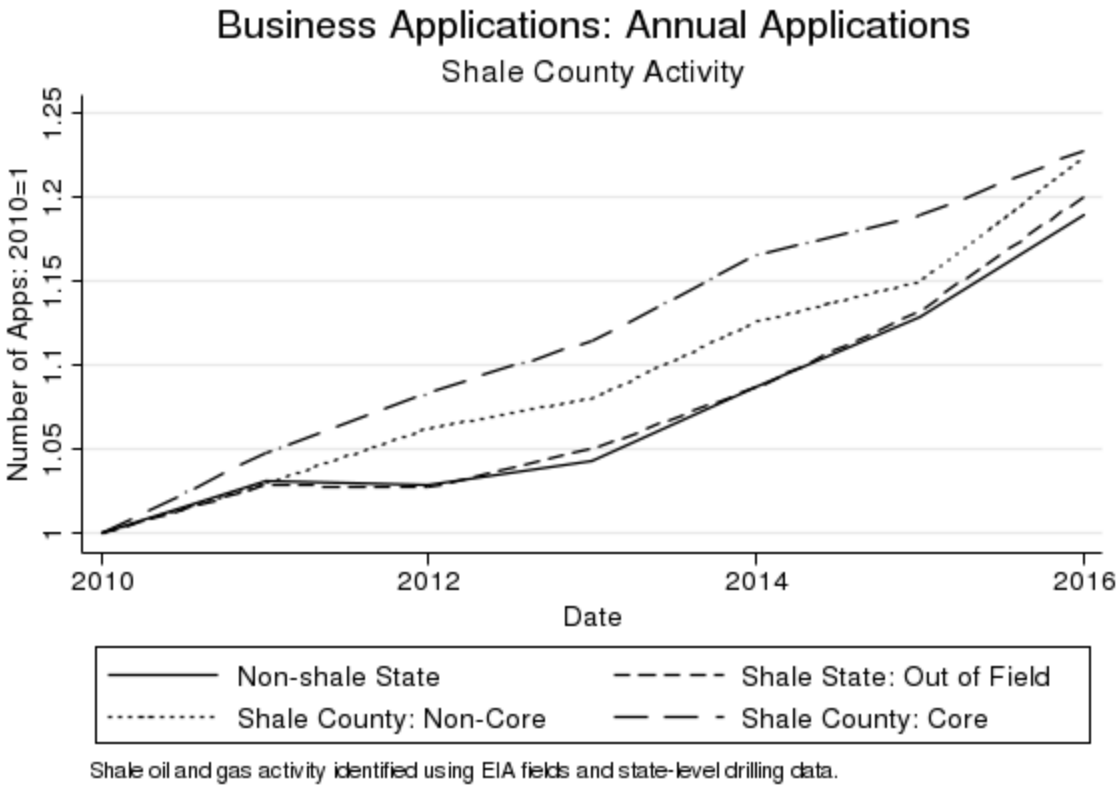
Growth Quintiles Defined 2010 to 2016.

Figure 14b



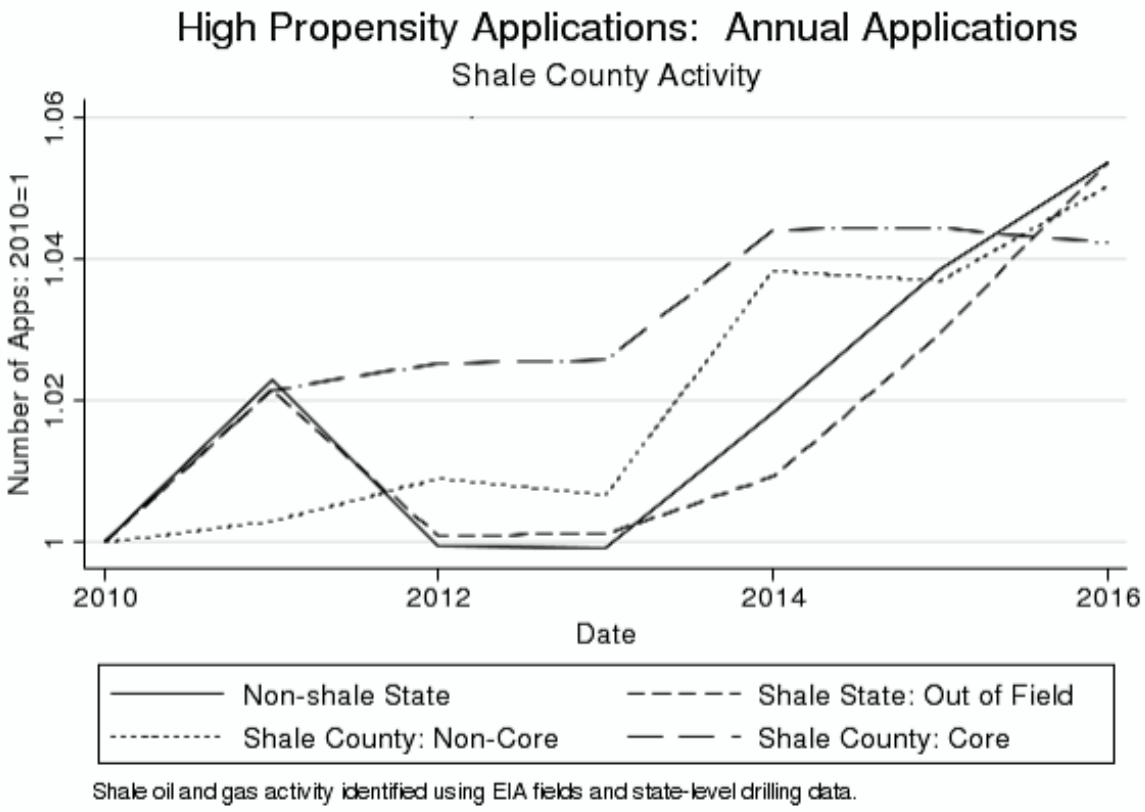
Growth Quintiles Defined 2010 to 2016.

Figure 15a



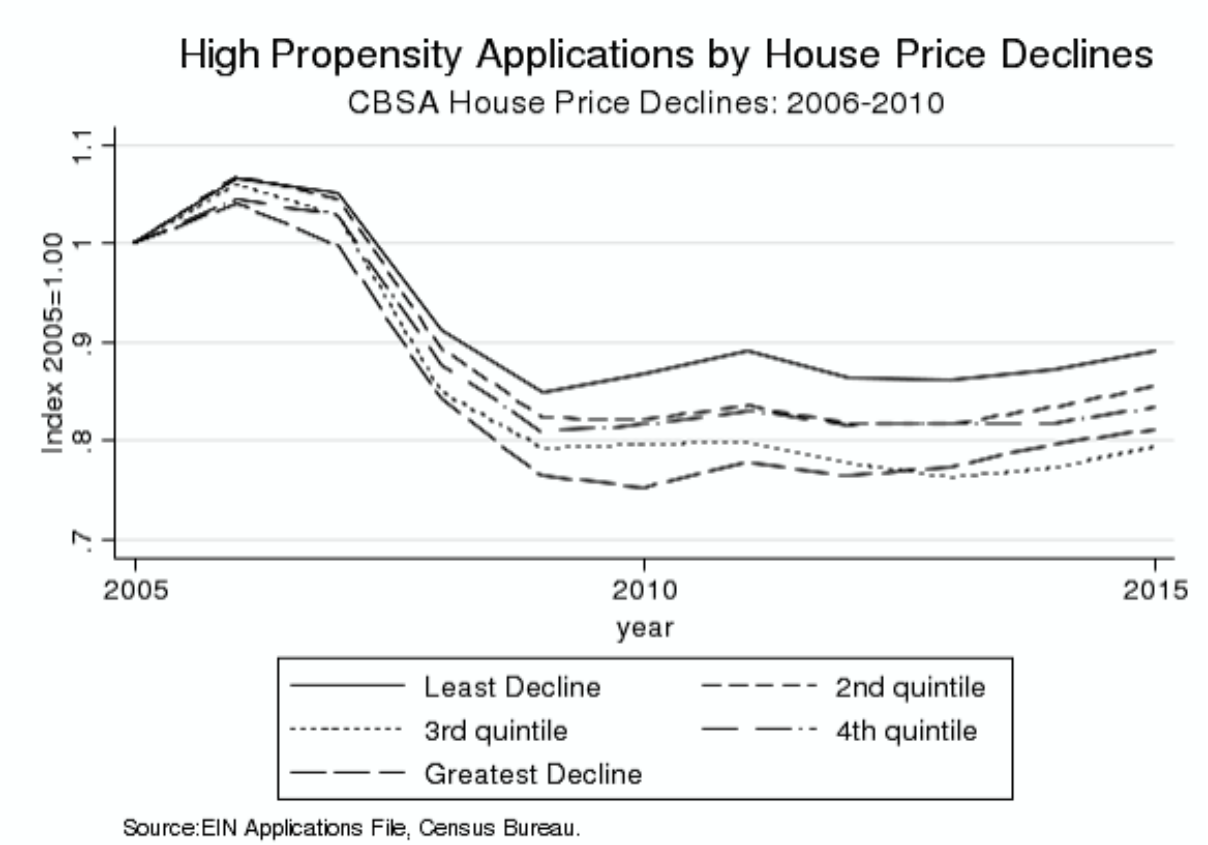
Shale oil and gas activity identified using EIA fields and state-level drilling data.

Figure 15b



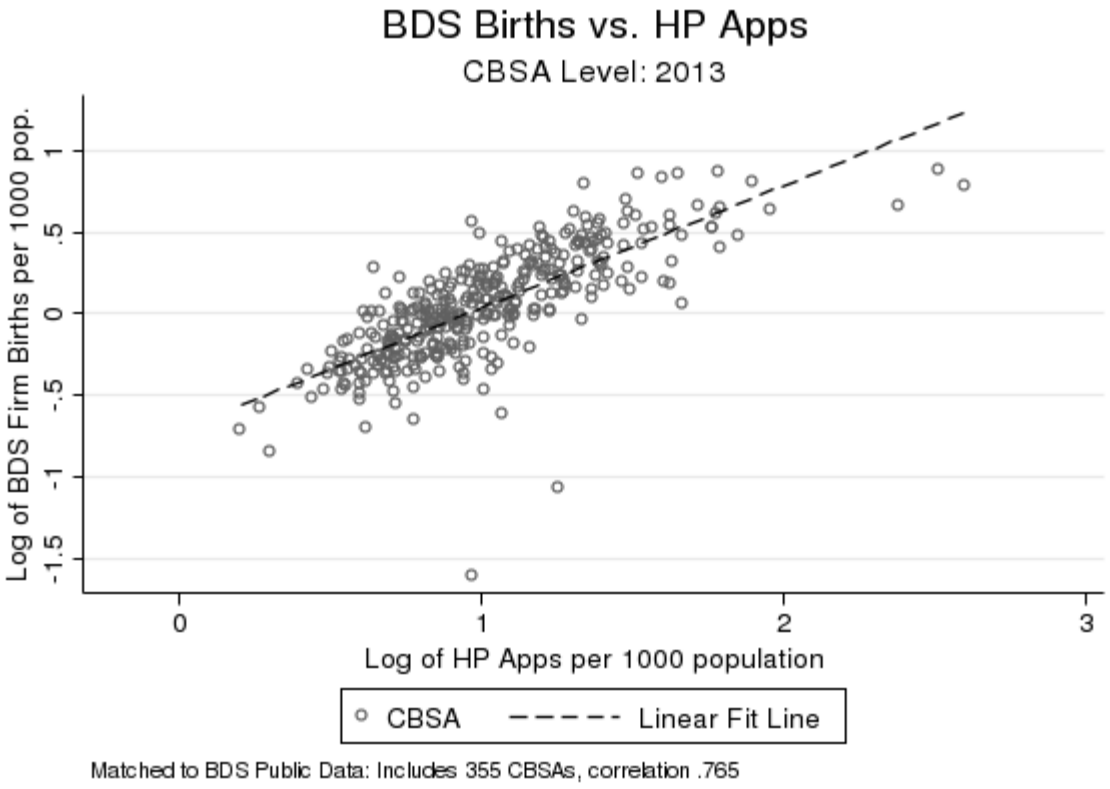
Shale oil and gas activity identified using EIA fields and state-level drilling data.

Figure 16



Source:EIN Applications File, Census Bureau.

Figure 17


BDS Births vs. HP Apps
CBSA Level: 2013
Log of BDS Firm Births per 1000 pop.
-1.5
-1
-.5
0
.5
1
0
1
2
3
Log of HP Apps per 1000 population
CBSA
Linear Fit Line
Matched to BDS Public Data: Includes 355 CBSAs, correlation .765

A Appendix

This appendix provides additional details on the SS-4 applications data and variable descriptions.

A.1 Data

The SS-4 or EIN applications data set is constructed from IRS filings that are transmitted on a weekly basis to the Census Bureau. Each week is referred to as a cycle and there are 52 or 53 cycles per year, numbered 1 through 53. The file structure has been nearly uniform over the period 2004:Q3 to 2016:Q4. Each application record has 31 variables. Table A.1 provides a complete list of the variables (responses) that come from the SS-4 form, along with the associated item number and description on the form. Not all items on the SS-4 form are transmitted to the Census Bureau. In particular, the Census Bureau does not receive the responses to items numbers 7, 9b, and 14. The dataset also includes a cycle variable and a derived NAICS industry code, autocoded by the Census Bureau.

Table A.2 contains the number of EIN filings received by Census on a yearly basis broken out into 5 main categories: all filings, tax liens filings, trusts and estates filings, filtered-out filings, and the estimation sample. Tax liens are applications typically associated with real estate transactions. These types of filings are quite volatile, likely represent uneven coverage, and spike in the data during the years of the Great Recession. There are a total of 2.3 million tax lien filings in the data. The number of such applications have fallen since 2011, as state auctions for tax liens have declined and multiple filings for EINs for such auctions have been restricted. Filings associated with a type of entity that self identifies as a trust and estate fell sharply in 2010. This pattern may be attributable to changes in estate tax laws during this period. The trusts and estate restriction omits 2.5 million applications. A number of applications are filtered out prior to the analysis. The applications that are contained in the filtered out group are one of three cases: (1) applications that do not report a geographic location or report a geographic location outside the 50 states and DC; (2) applications that are identified as coming from a concentrated spike; and (3) applications from a set of industries that are typically associated with a low rate of transition to employer businesses. With regard to (2), we identified applications as a part of a concentrated spike based on the number of filings that came from the same zip code, in the same week, and in the same 6-digit NAICS code. If the number of filings from such a cell exceeded 25,

the applications from the cell are identified as a concentrated filing and omitted from the final estimation sample. Such concentrated filings are typically associated with financial filings, as opposed to filings associated with starting a new business. With regard to (3), we omitted applications from a set of NAICS industry codes: 110000, 112000, 525100, 525900, 531100, 531110, 531120, 813200, 813410, 813990, 814110 and 900000. These industries are in agriculture, financial services, personal services and government services. These overall data restrictions generate an estimation sample that includes 70 percent of the original applications. We refer to this subset of applications as the set of business applications and use it to form our overall business applications series, as depicted in Figure 1 of the paper.

A.2 Application Microdata Files

Applications File: The raw SS-4 data are contained in a set of SAS data files that cover an aggregation of weekly cycles of various time intervals. A master data file was constructed by taking the 2007-2011 applications from the initial panel file and appending the incoming applications for 2012-2016 and the recovered files from 2004:Q3-2006:Q4. The incoming application files are stored as annual files that include unduplicated annual applications. These files are appended together to construct a master application file. The current data file contains the full set of unduplicated applications. The file is set up to be appended with incoming applications data with little processing involved (duplicates check, variable names, type and length checks). No data cleaning is done at this stage.

Geography File: All applications data are sent to the Census Bureau's geocoding operations to be geocoded to the state, county, tract, block group, and block levels. Both the mail address and the business address (if available) are coded. Currently 99 percent of applications are coded to the state and county levels, with 88 percent coded to the tract, block group and block level. An auxiliary file is constructed that contains the EIN, State FIPS code, and County FIPS code used to process the main applications file. The file will be appended with incoming geocoded applications data on an applications flow basis.

In processing the applications, if a business address is present that is the information used as geographic codes for the application. In all other cases, the mail address is used. When a business address is reported, the business address almost always has the same generated state and county codes as the mail address – 96.5 percent of such cases.

Clerical Industry Coding File: A file that contains the EIN and the NAICS code assigned by

the SSA (Social Security Administration) is available starting in 2010. Industry codes for a subset of applications are coded by a clerical operation at SSA. The SSA codes are used to revise the original NAICS code and to replace missing NAICS codes on the application files. As new application files come in, this file needs to be updated on an applications flow basis. Updates to the industry codes typically take two to three months to complete. The updates to the industry codes will cause revisions in past data releases, as industry codes for recent applications may be changed.

Business Register Quarterly Panel: This panel is a set of application (EIN) level records matched to the business register (BR). The BR match data are contained currently in several files. These files include all applications, along with business register match information from 2003-2014. The data identify the quarter that an application becomes payroll active. An auxiliary file is constructed that contains EIN and a variable that identifies the quarter of transition to an employer business for an application EIN. This information is updated every year once the BR/LBD processing for the year is complete.

Business Register Employment Files: These files consist of the subset of applications that match to the business register (Firm Age=0 cases). The files contain a small set of BR variables and up to 12 quarters of employment data. The files are stored as annual files based on year of transition to employer business status. The Employment Files are used in supplementary analysis. The BR quarterly employment data are only available beginning in 2005.

Business Applications File: The Applications File is matched to the auxiliary files derived from: (1) Geography File; (2) Clerical Industry Coding File; and (3) Business Register Quarter Panel. The geography and industry files are used to augment and create final geography codes (FIPS state and county) and NAICS industry codes. The BR file is used to identify the quarter of transition to employer status. The incorporation of the information from the three files allows for the cleaning and filtering of the applications data. The resulting file is the Business Applications File.

Model Sample Estimation File: This file uses the Business Applications File as input. Some additional cleaning is done at this stage in preparation for the programs that estimate the projection series. In addition, this file is used to generate the state- and national-level application counts that are contained in the public data sets.

A.3 Variable Construction

Industry Codes: The applications files contain a NAICS code variable. In the original applications file, the code is missing for 17 percent of the cases. The frequency of missing values in the applications file is relatively constant over time. While the industry code is stored in a 6-digit field, the code itself may represent a higher level of aggregation. It is quite common to have a code that identifies a 3- or 4-digit NAICS code within the 6-digit field. The NAICS code is based on three response items on the applications: primary business activity, primary business (other), and primary merchandising line.

Industry codes are augmented in two ways. For applications with missing codes, a check on the response to the primary business activity inquiry is made. If a response is available, the application is assigned to the primary 1-digit NAICS sector identified on the application form. The second main edit is through the use of the clerical industry code files. Clerical codes are only available from 2010 onward and only for a subset of applications. If an EIN has a valid clerical industry code record, the code on the application is replaced by the clerical industry code for both missing/non-missing application industry codes. Once the industry-coding replacement process has been completed, industry codes are put through a filtering process. This process fixes a set of incorrect codes and takes codes with low frequency codes and assigns them to a higher level code. The resulting industry code reflects a mix of 1-digit, 2-digit, ... , 6-digit NAICS codes. Roughly 8 percent of applications are not coded and are assigned a code of "000000". The number of uncoded applications shifts down in 2010 by 5 percent when the clerical-industry coding file becomes available.

Geography Codes: The applications data come with address field information including mailing address, executors address (if applicable), a business county and business state. Both the mail and business address information are sent to the Census Bureau's geocoding operations. The geocoder places FIPS state and county codes and census tract, block group and block codes on the application records for the mail and business addresses, separately. For almost all applications, we obtain FIPS state and county codes for the mail addresses. For the business address, this is coded on roughly 15 percent of applications. In almost all cases the geocoded state and observed state mail code is the same.

Census tract, block group and block codes are available on 88 percent of application records. The geocoder codes both the mail and business addresses (if available). Longitude and latitude are also included on the geocoded records for both mail and business addresses.

For applications outside the U.S. (usually U.S. territories) or with invalid codes, the final state code is assigned a "00" and these applications are omitted from the analysis and application statistics.

Cycle and Quarter Variables: Each application is identified by the cycle week the application was received. The cycle variable runs from 427 (2004:week 27) to 1652 (2016: week 52). Each year contains either 52 or 53 cycles. Each cycle is also assigned to a variable that identifies the overall quarter the application is submitted in. The quarter variable is a sequence number initiated as 3 (in 2004:Q3) and incremented by one each quarter. The only modification made to the cycle and quarter variables was that 30 percent of applications in cycles 540 and 541 were reassigned to cycle 538 and 539 to smooth out what is likely a trough-spike in the data due to processing issues. This re-assignment is done at the end of processing.

Year and Quarter Variable: Each cycle is assigned to a year and quarter. Cycles are assigned to quarters based on the week of submission (1 through 53): Q1:1-13, Q2:13-26, Q3:27-39, Q4:40-53.

Type of Entity: The type of entity represents the self-reported legal status of the entity assigned the EIN. The main types of entity are: (1) Sole Proprietorship; (2) Partnership; (3) Corporation; (4) Personal Service Corporation; (5) Church; (6) Non-Profit; (7) Other; (8) Estates; (9) Plan Administrator; (10) Trusts; (11)-(16) special cases and governments. There is noise in the type of entity field and one does observe response information that lies outside the form values. This information looks to be unusable for the most part and affects only a small number of cases. The key treatments of the type of entity variables include: (1) drop estates and trusts – these are inconsistently reported over time with a small number of applications reported 2004 and 2006 and from 2011-onwards, and a very large number of applications between 2007 and 2010; (2) groups (9) and (11) through (16) into one "99" category that covers a very small number of observations. The main trend in the variable is that the category (7) Other has expanded sharply while traditional LFOs (Partnerships and Corporations) have declined as a share of applications. The reason is likely due to the growth of LLCs – both single and multi member. If the LLC variable available from mid-2007 onward indicates an S or an M, then the type of entity reported is usually recorded as (7) Other. LLC's are pass-through entities, where income taxes paid are usually reported on the member's individual tax returns. The usage of this legal

business entity has expanded sharply over time according to the Statistics of Income (+60 percent 2004-2012), while traditional C-Corp and Partnerships have declined (-20 percent: 2004-2012). The (7) Other category is not limited to LLC's and contains other business entities, as well. The Other category also has a lower propensity to become an employer business than the Corporate entity.

Reason for Applying: The reason for applying is a self-reported reason for requesting an EIN. The checkbox options for the reason for applying are: (0/1) Started a New Business; (2) Hiring Employees (3) IRS Compliance; (4) Other; (5) Banking Purposes; (6) Change in Organization Type; (7) Purchase a Business; (8) Create a Trust; (9) Create a Pension Plan. Most of the responses fall in the first category (88 percent). There are two codes (0) and (1) initially that correspond to the Started a New Business. These collapse into (1) in the middle of 2010-onward. Since, (8) and (9) are very small groups, we include them with (4) Other. One notable difference in the Reason for Applying trends is that the response Banking Purposes rises somewhat sharply (still a relatively small group) between 2010 and 2011.

Business Start Date: The business start date is a month-year variable that indicates when a business was started or acquired. A variable is created that codes whether the reported start date is missing, precedes the application date, is in the same quarter as the application date, or is after the application date. There are a significant number of missing business start dates.

Wage Date: The wage date variable reflects the first month-year that wages or annuities are paid. A variable is created that codes whether the reported wage date is missing, precedes the application date, is in the same quarter as the application date, or is after the application date. There are a significant number of wage dates that are missing, as this field only pertains to applicants paying or planning to pay wages. A second indicator variable is constructed that equals 1 if a wage date appears, 0 otherwise.

Response Indicator Variables: Several variables are constructed based on whether a response was received in a particular application item. Four indicator variables are constructed based on whether a response was given for: (1) Trade Name; (2) Executor's Name (3) Business Address; (4) Prior EIN.

LLC Variable: Post 2007, an LLC categorical variable is available. The response takes on one of three values – missing, S, or M, with S indicating a single or sole member LLC and M indicating a multi-member LLC.

Quarter of Employer Birth: This variable identifies the quarter of transition to employer status. It is generated in the match to the Business Register data. Note that quarter of transition can precede the application quarter, as a set of applications have employment data recorded that occurs before the application quarter.

Employer Status: This is an indicator variable denoting whether an application is matched to an employer record in the Business Register.

Four-Quarter Window Business Startup Indicator: This variable is an indicator variable that identifies whether an application becomes an employer birth within 4-quarters of the application. The variable is constructed by using the quarter of payroll birth from the Business Register and constructing the difference between the quarter of birth and the quarter of application. All applications where the difference is less than or equal to 4 quarters are denoted as employment births within the window. Since applications come in at different times within a quarter, we adjust the window for applications based on which week in the quarter the application was submitted. An application that comes in week $n = 1, 2, .., 13$ of a quarter receives a probability of $n/13$ to look ahead an additional quarter for potential transition. For example, for applications that enter the third week of a quarter, 3/13's of the applications (randomly selected) will be allowed to look ahead an additional quarter to measure whether they become a business birth in the fifth quarter. For applications that enter in the last week of the quarter, we allow all these applications to look ahead a full additional quarter to measure employer birth status. In this way, application timing within a quarter will not affect the average length of the birth window utilized in measuring transition to employer birth.

Eight-Quarter Window Business Startup Indicator: This variable is an indicator variable that identifies whether an application becomes an employer birth within 8-quarters of the application. The variable is constructed by using the quarter of payroll birth from the Business Register and constructing the difference between the quarter of birth and the quarter of application. All applications where the difference is less than or equal to 8 quarters are denoted as employment births within the window. No adjustment for window

length is made for these applications, as opposed to the four-quarter window case. This is because an 8-quarter window is large enough to include most transitions to employer status over time.

Table A.1 Application Data Record

Item Number	Variable Name Description	Availability
1	Name of Legal Entity	Y
2	Trade Name	Y
3	Executor or Trustee Name	Y
4a	Mail Address	Y
4b	Mail City	Y
4b	Mail State	Y
4b	Mail Zip Code (9 digit: +4 rarely used)	Y
5a	Executor Address	Y
5a	Executor City (often contains state, zip code)	Y
6	Business County	Y
6	Business State	Y
7	Responsible Party	N
8	LLC Designation – Single, Multi	2008-
9a	Type of Entity	Y
9a	Type of Entity: Remarks	Y
9b	Incorporation Location	N
10	Reason for Applying	Y
11	Business Start Date	Y
12	Closing Month Accounting Year	Y
13	Expected Max. Number of Employees (12 months)	2015-
14	Employment Tax Liability less than 1000	N
15	First Wages Paid Date	Y
16	Principal Activity of Business	Y
17	Principal Line of Merchandising	Y
18	Previous EIN	Y

Source: EIN Application Files 2007-2016, Census Bureau

Table A.2. Application Activity by Year

Year	Number	Tax Liens	Trusts-Estates	Filtered	Sample
2004	1.3		0.0	0.2	1.1
2005	3.0		0.0	0.5	2.5
2006	3.2		0.0	0.5	2.6
2007	4.1		0.8	0.5	2.7
2008	4.1		0.9	0.5	2.6
2009	3.7		0.7	0.5	2.5
2010	3.3	0.1	0.0	0.7	2.5
2011	5.2	2.0	0.0	0.7	2.5
2012	3.1	0.0	0.0	0.6	2.5
2013	3.2	0.0	0.0	0.6	2.6
2014	3.3	0.0	0.0	0.7	2.7
2015	3.6	0.0	0.0	0.8	2.8
2016	4.7	0.1	0.5	1.2	3.0

Source: EIN Application Files 2007-2016, Census Bureau

B Public Use Data Documentation

This section provides documentation for the public-use files made available through the Business Formation Statistics (BFS) program.[35] The BFS uses information from the IRS SS-4 Form on EIN applications to construct measures of business applications and formation at the national and regional levels. The national and state statistics files each contain business application and business formation series. The application series provide the number of applications in each of the following categories: business applications, high-propensity business applications, business applications with planned wages, and corporations. The formation series include the number of actual and projected business formations (employer births) within either 4 or 8 quarters of application. Two other series give information on the average duration between business application and formation, conditional on the application turning into an employer business within either 4 or 8 quarters of application. All series are presented both non-seasonally adjusted and seasonally adjusted. The data in all files are reported at a quarterly frequency. The variables on which each series is based are described in detail below.

Variable Definitions:

Business Applications (BA): The number of business applications received. This variable incorporates a set of exclusion restrictions that filter out a set of EIN applications associated with tax lien, trust, estate and a subset of financial and industry-specific filings.

High Propensity Business Applications (HBA): The number of high propensity business applications received. This variable is a subset of business applications (BA) that are identified as applications that have a higher propensity to become an employer business based on specific responses to the industry, type of entity, reason for application and wage date inquiries on the SS-4 application.

Business Applications with Planned Wages (WBA): The number of business applications received that provide a planned date to pay wages - a subset of high propensity business applications (HBA). This variable is the number of applications received that indicated a payment date for wages (Item 15) on the SS-4 application.

[35] Visit the BFS website https://www.census.gov/programs-surveys/bfs.html.

Business Applications from Corporations (CBA): The number of business applications received from corporate entities - a subset of high propensity business applications (HBA). This variable is a count of all filings from applications that indicated they were a corporation or personal service corporation (Item 9a).

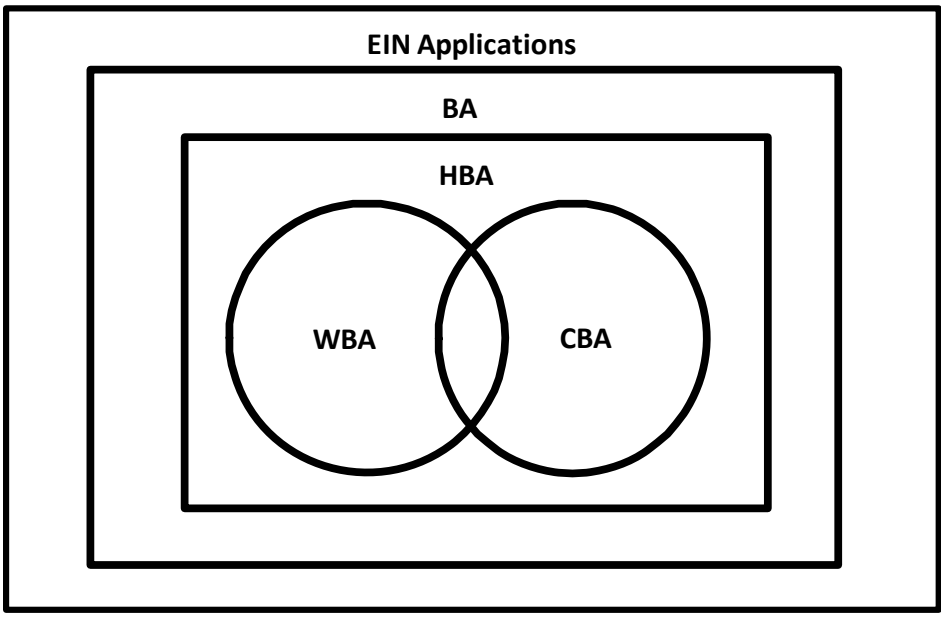


Figure B.1: The relationship between EIN applications and various business applications series

Business Formations within 4-Quarters (BF4Q): The number of applications that become an employer business within four quarters of application submission. This variable is constructed using application information linked to the Census Bureau's Business Register. It is a forward-looking measure of business formation based on incoming applications received. The end-point of this series is determined by the availability of quarterly data on payroll and employment from the business register. The interpretation of the count is the number of applications received in a quarter that become an employer business over the next 4 quarters.

Projected Business Formations within 4-Quarters (PBF4Q): A model-based projection of the BF4Q variable that extends the series beyond the available business register-to-application matched data. This variable uses application data and model parameters to

construct a projection of BF4Q starting from the period following the last available period of business-register-to application matched data. Combining the projected series with BF4Q provides an up-to-date, forward-looking business formation series.

Spliced Business Formations within 4 Quarters (SBF4Q): This series combines (splices) BF4Q and PBF4Q to provide the entire time series for the actual and projected business formations within 4 quarters. The series BF4Q and PBF4Q are connected starting with the last quarter for which BF4Q is available.

Business Formations within 8-Quarters (BF8Q): The number of applications that become employer businesses within eight quarters of application submission. This variable is constructed similar to BF4Q above. The end-point of this series is determined by the availability of quarterly data on payroll and employment from the business register. The interpretation of the count is the number of applications received in a quarter that become an employer business over the next 8 quarters.

Projected Business Formations within 8-Quarters (PBF8Q): A model-based projection of the BF8Q variable that extends the series beyond the available business register-to-application matched data. This variable uses application data and model parameters to construct a projection of BF8Q starting from the period following the last available period of business-register-to application matched data. Combining the projected series with BF8Q provides an up-to-date, forward-looking business formation series.

Spliced Business Formations within 8 Quarters (SBF8Q): This series combines (splices) BF8Q and PBF8Q to provide the entire time series for the actual and projected business formations within 8 quarters. The series BF8Q and PBF8Q are connected starting with the last quarter for which BF8Q is available.

Average Duration (in Quarters) from Business Application to Formation within 4 Quarters (DUR4Q): A measure of delay between business application and formation, defined as the average duration (in quarters) between the quarter of business application and the quarter of business formation, conditional on business formation within four quarters. This series by definition span the same time period as BF4Q.

Average Duration (in Quarters) from Business Application to Formation within 8 Quarters (DUR8Q): A measure of delay between business application and formation, defined as the

average duration (in quarters) between the quarter of business application and the quarter of business formation, conditional on business formation within eight quarters. This series by definition span the same time period as BF8Q.

Public-Use Data Files:

The following table gives a list of publicly available data files. Each file contains both the non-seasonally adjusted and seasonally-adjusted series for the corresponding variable as described above. Note that while the analysis in this paper runs up to 2016:Q4, the public-use files in the BFS website contain the most recent updates to the business applications and formations series.

Table B.1. Public-use Data Files

Series Name	Acronym	National File	States File
Business Application Series			
Business Applications	BA	BA_US.xlsx	BA_ST.xlsx
High-Propensity Business Applications	HBA	HBA_US.xlsx	HBA_ST.xlsx
Business Applications with Planned Wages	WBA	WBA_US.xlsx	WBA_ST.xlsx
Business Applications from Corporations	CBA	CBA_US.xlsx	CBA_ST.xlsx
Business Application Series			
Business Formations within 4 Quarters	BF4Q	BF4Q_US.xlsx	BF4Q_ST.xlsx
Projected Business Formations within 4 Quarters	PBF4Q	PBF4Q_US.xlsx	PBF4Q_ST.xlsx
Spliced Business Formations within 4 Quarters	SBF4Q	SBF4Q_US.xlsx	SBF4Q_ST.xlsx
Business Formations within 8 Quarters	BF8Q	BF8Q_US.xlsx	BF8Q_ST.xlsx
Projected Business Formations within 8 Quarters	PBF8Q	PBF8Q_US.xlsx	PBF8Q_ST.xlsx
Spliced Business Formations within 8 Quarters	SBF8Q	SBF8Q_US.xlsx	SBF8Q_ST.xlsx
Avg. Duration from Business Application to Formation (4 Qtrs)	DUR4Q	DUR4Q_US.xlsx	DUR4Q_ST.xlsx
Avg. Duration from Business Application to Formation (8 Qtrs)	DUR8Q	DUR8Q_US.xlsx	DUR8Q_ST.xlsx