

**Matching State Business Registration Records
to Census Business Data**

by

**J. Daniel Kim
U.S. Census Bureau**

**Kristin McCue
U.S. Census Bureau**

CES 20-03

January, 2020

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K038E, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov. To subscribe to the series, please click [here](#).

Abstract

We describe our methodology and results from matching state Business Registration Records (BRR) to Census business data. We use data from Massachusetts and California to develop methods and preliminary results that could be used to guide matching data for additional states. We obtain matches to Census business records for 45% of the Massachusetts BRR records and 40% of the California BRR records. We find higher match rates for incorporated businesses and businesses with higher startup-quality scores as assigned in Guzman and Stern (2018). Clerical reviews show that using relatively strict matching on address is important for match accuracy, while results are less sensitive to name matching strictness. Among matched BRR records, the modal timing of the first match to the BR is in the year in which the BRR record was filed. We use two sets of software to identify matches: SAS DQ Match and a machine-learning algorithm described in Cuffe and Goldschlag (2018). We find preliminary evidence that while the ML-based method yields more match results, SAS DQ tends to result in higher accuracy rates. To conclude, we provide suggestions on how to proceed with matching other states' data in light of our findings using these two states.

* Acknowledgements: We thank John Cuffe for his help in implementing the machine-learning algorithm, Jorge Guzman for help in understanding the BRR records, and Cristina Tello-Trillo and David Brown for helping us with use of the files they developed for name and address matching to the iLBD and BR.

Disclaimer: Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

1. Introduction

This memo describes our methods and results from matching state Business Registration Records (BRR) to Census business data and suggests how to proceed with matching other states' data based on our findings. These state registration records are of potential interest to Census as an additional source of data that includes information on the timing of business or establishment births. The BRR data are public records filed by businesses that are new to the state. The requirement to file is triggered by activity such as opening a bank account or leasing space, so the data include both employer and non-employer businesses, and businesses with a variety of legal forms, including corporations, LLCs, limited partnerships, and general partnerships (Guzman and Stern 2015a). While some businesses are not required to register (e.g., sole proprietorships), registration provides benefits such as limited liability and tax incentives. Guzman and Stern (2015b) argue that these benefits make registration a requirement for almost any new business that intends to grow.

2. BRR Data

The BRR files we currently have at Census are for Massachusetts and California. Both files include information on business name, the principal address of the business, whether a business is incorporated, and its jurisdiction. Jurisdiction is defined as either the state in which the business originated, or Delaware, if incorporated in Delaware. For example, records for new firms setting up businesses in Massachusetts will have a Massachusetts addresses in the BRR file, but may have jurisdiction in either Massachusetts or Delaware. If a business that was originally established in a different state expands into Massachusetts for the first time, the address in its BRR record will be in the state in which it already operates, rather than Massachusetts. As a result, in those cases we do not have a Massachusetts address to use in matching to Census data.¹ Additional information such as a list of names of business officers is available for some, but not all states.

¹ BRR records for established businesses that are expanding into the state of registration may have some utility for identifying states in which multi-units may have new establishments.

3. Census Data

We matched the BRR records to Census information on both employer and non-employer businesses. Data on employer businesses were sourced from annual end-of-year snapshot files from Census's Business Register (BR).² The BR covers all employer establishments in the non-farm private sector, and stores information on establishment characteristics such as industry, location, employment, payroll, and legal form of organization.

We sourced information on non-employer businesses from the Integrated Longitudinal Business Database (iLBD), which covers the universe of non-employer business units in the non-farm US private sector. It is built by linking businesses with administrative identifiers such as the owner's name, address, and a protected Identify Key (PIK) which is an internal person identifier, which in the iLBD is assigned based on social security numbers. The iLBD contains information on legal form (e.g., sole proprietorship, partnerships) and other characteristics of the business (Fairlie and Miranda, 2017). When a non-employer business in the iLBD hires its first paid employee, the business then appears in the employer universe. According to Fairlie and Miranda (2017), roughly 2.5% of US non-employer startups born in 1997 hired at least 1 employee within seven years of firm birth, while the vast majority eventually exited (i.e., firm death) without ever hiring an employee.

4. Matching Protocols

We restricted our Massachusetts BRR sample to businesses with filings in the Massachusetts BRR file in years 1975-2013 with usable name and address fields, and jurisdiction in either Massachusetts or Delaware. We imposed the same restrictions for the California BRR sample, except to require jurisdiction in either California or Delaware. Our first step was to standardize the name and address fields in the BRR data using the code used to standardize the BR and iLBD files that we match to. We then assigned SAS DQMatch codes to each business name and

² For years 1978-2001, the Census employer data was instead drawn from the Standard Statistical Establishment List (SSEL) which was the predecessor to the BR. DeSalvo, Limehouse and Klimek (2016) provide some additional background on the BR.

address.³ Using a series of algorithms, we then try to match the set of names and addresses in the state BRR file to BR records for businesses in that state. We carry out the match for each available year of BR and iLBD data so that we can compare the timing of entry in Census and BRR sources. We use all BRR records with usable name and address information for each year of Business Register (BR) data from 1978-2014, and to non-employer data from the Integrated Longitudinal Business Database (iLBD) files for years 1977, 1982, 1987, 1992, and 1994-2013.⁴

The DQMatch software identifies pairs of records that meet its criteria for a match given the specified parameters. For Massachusetts, we ran a series of matches in which we varied the parameters to gradually reduce the stringency of the match criteria. We drew random samples from the set of pairs identified as matches based on the DQMatch outcomes, and coded each pair as a correct or incorrect match based on the clerical reviews. We used the share of matches that we deemed were correct as an estimate of match quality for each pass. In matching the California data, we used the Massachusetts results to identify a more parsimonious sequence of matches and then applied those to the California BRR data to confirm that the matching results appear similar for different states. While we found somewhat lower match rates for California, Our conclusion is that the approach is reasonably generalizable.

5. DQ Match Results

Table 1 Panel A shows the share of Massachusetts records matched, by level of sensitivity (as defined by DQMatch), and separately for the employers and non-employer. As expected, blocking at the Zip-3 or Zip-5 level attenuates the match rate, but the differences were small, so we blocked only at the state level in later rounds of matching. Because we were interested in the timing of BRR matches to Census businesses, we included the full set of Massachusetts records in matching to each year of Census data. Many of the matched BRR records met the DQMatch criteria for matching in more than one year, so in calculating the BRR match rates we count a BRR record as matched if it matches in at least one year. Here we focus

³ We match to a shared set of general purpose files that CES staff have created for name and address matching to Census business data using SAS DQMatch software.

⁴ Prior to 1994, Census only has non-employer data for economic census years, which are years ending in 2 or 7.

on whether DQMatch produces any matches for the BRR records. We then examine the share that appear accurate based on our clerical reviews.

Table 1: Overall BRR Match Rates with Standard Thresholds

Panel A: Massachusetts BRR (N=736,000)

Panel A: Massachusetts BRR (N=736,000)

Matching Threshold	Location Blocking	Nb. Of Records Matched	Match Rate (%)	Accuracy Rate (%)
Employer Business Register				
Name 80, address 80	None	221,000	30.0%	95%+
Name 65, address 80	None	246,000	33.4%	95%+
Non-Employer Business Register				
Name 80, address 80	None	119,000	16.2%	95%+
Name 65, address 80	None	134,000	18.2%	95%+

Panel B: California BRR (N=3,862,000)

Matching Threshold	Location Blocking	Nb. Of Records Matched	Match Rate (%)	Accuracy Rate (%)
Employer Business Register				
Name 80, address 80	None	1,000,000	25.9%	95%+
Name 65, address 80	None	1,108,000	28.7%	95%+
Non-Employer Business Register				
Name 80, address 80	None	516,000	13.4%	95%+
Name 65, address 80	None	587,000	15.2%	95%+

Notes: For disclosure purposes, accuracy rates higher than 95% are topcoded as 95%+.

Overall, when using our baseline matching thresholds (sensitivity of 80 for both name and address) and not blocking on zipcode, the match rates for Massachusetts were 30% for the employer BR and 16% for the non-employer iLBD. Some businesses matched to both the employer and non-employer lists, so overall 46% of the Massachusetts BRR records matched to at least one of the 2 lists.

Panel B shows match results for California. Our findings for Massachusetts data indicated that blocking on zipcode made little difference, so for brevity, we used only our

baseline name-address thresholds of 80-80 and thresholds of 65-80 in matching the California data. Consistent with the results for the Massachusetts, match rates to the employer BR are substantially higher than those to the non-employer BR. Overall, we find somewhat lower match rates using the California BRR records appear to have modestly lower match rates. Based on the baseline name-address threshold of 80-80, the match rates for employer BR and non-employer BR are 26% and 13%, respectively (compared to 30% and 16% for Massachusetts).

Moreover, we report measures of the quality of the matches we found using name and address thresholds of 80-80 and 65-80, and not blocking on zipcode. The quality measures are the share of reviewed records that we labeled as correct during our clerical reviews. The last column in Table 1 shows that almost all randomly selected matched records for both the employer and non-employer sides appear to be accurate matches. We find consistent results for the California BRR records. Even when using less strict matching thresholds of Name-65 and Address-80, the rate of false positives is low. Therefore, it appears that a sufficiently high threshold on address – even if the name is set to a lower threshold – identifies matched samples with relatively few cases that appear incorrect.

Next, we explore the heterogeneity in match rates by various firm characteristics. Table 2 documents the differences in match rates associated with the firm traits available for the BRR records, using DQ match results based on sensitivity levels of 65 for name and 80 for address. We find that match rates to employer firms are systematically higher for incorporated businesses and businesses with higher quality scores as measured in Guzman and Stern (2015). This appears consistent with their finding that higher quality startups (c.f., Guzman and Stern 2012) are more likely to hire and to have other business activity including fundraising. Surprisingly, match rates do not differ substantially depending on whether or not the business was incorporated in Delaware, which Guzman and Stern (2015) identify as a predictor of high-growth startup firms.

Table 2: Heterogeneity in Match Rates (Massachusetts BRR)

Dimension	Nb. Records	Match Rate	
		Employer BR	Non-Emp. BR
Overall Match Rate	736,000	33%	18%
Corporation Status			
No	217,000	16%	21%
Yes	518,000	41%	17%
Startup Quality*			
Quartile 1 (lowest)	127,000	23%	22%
Quartile 2	129,000	30%	20%
Quartile 3	118,000	38%	18%
Quartile 4 (highest)	145,000	40%	23%
Top 1%	6,000	54%	18%
<i>Missing quality score</i>	217,000	34%	12%
Delaware Jurisdiction			
No	705,000	33%	18%
Yes	31,000	36%	25%

These patterns do not generally hold with matching to the iLBD. This result is consistent with the view that initial firm characteristics – which are correlated with measures of firm success and growth – are also related to the decision to hire employees.

As mentioned above, we also applied a machine learning approach to the match between the Massachusetts BRR and the BR, using the MAMBA algorithm described in Cuffe and Goldschlag (2017). Comparing the matches identified by the two sets of software allows us to examine how the outcomes differed. Table 3 shows a comparison of the match rates, and Table 4 compares the accuracy rate of the two sets of matches based on our clerical reviews.

Table 3: Comparison of Match Rates from Alternative Software

<i>DQ Match Sensitivity</i> Match Category	<u>Name-80 Address-80</u>		<u>Name-65 Address-80</u>	
	Count	%	Count	%
Full BRR Sample	736,000	100%	736,000	100%
No Match	383,000	52%	371,000	50%
Both Match	142,000	19%	155,000	21%
Partial Match				
only SAS DQ	79,000	11%	91,000	12%
only ML-approach	132,000	18%	119,000	16%

Overall, 71% of the BRR records have the same outcome using DQ and MAMBA – i.e. either they identify the same match, or neither of them finds a match. For the other 29% of the BRR records, the MAMBA software finds a larger number of matches than does DQ when using 80 for both thresholds. Among that residual set, the ML-based approach accounts for roughly 60% of the additional matches. Lowering the DQ name-address thresholds to 65 and 80 the match coverage premium for the ML-based approach is attenuated to 4 percentage points.

Table 4 compares the accuracy of the matches from the two sets of software, in the cases where they identify different matches.⁵ We clerically review a random sample 300 records where the matches differ, and label the match assigned by each as correct or incorrect.

Table 4: Match Accuracy using SAS DQ vs. Machine Learning-Based Approach (Massachusetts BRR)

Method	<u>% Match Accuracy</u>
SAS DQ	93%
ML-Approach	70%

We find that where the two approaches identify different matches, the DQ matches are on average more accurate, with 93% of the DQ-matched observations labelled accurate, compared

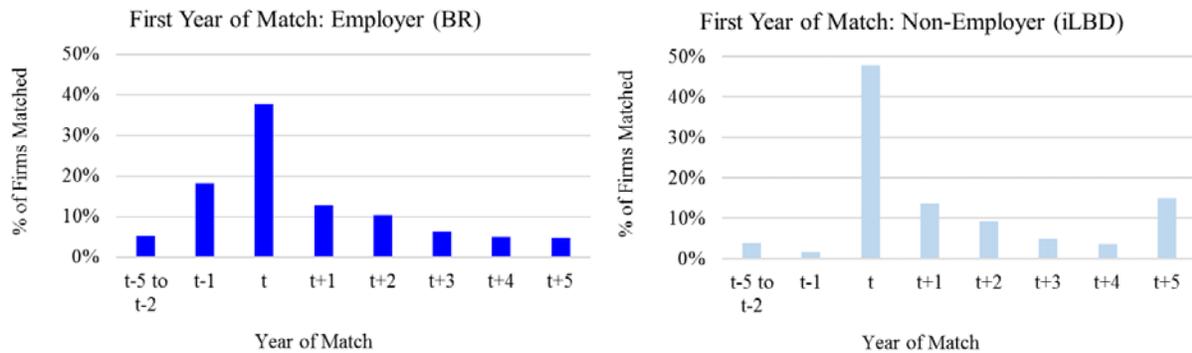
⁵ The SAS DQ match was run using Name and Address thresholds of 65 and 80, respectively.

to 70% of the MAMBA matches.⁶ Qualitatively, the leading driver of lower match accuracy for the ML-based approach was the looser strictness on address. Many of the inaccurate matches were businesses with similar (but not same) names that had different addresses.

Our final analysis explores the timing of the matches between the BRR and Census records. First, we assess the timing differences in matching of BRR records to the employer versus non-employer side of Census records. For BRR records that are registered in year t , we examine the years in which the record is matched to Census employer and non-employer businesses. For brevity, we begin with the Massachusetts BRR businesses that register in 2005, and restrict the time window of match to $t-5$ and $t+5$ (i.e., between 2000 and 2010).⁷ In this analysis, we explore both (1) the first year of match and (2) all years of match.

Figure 1: Timing of Match between Massachusetts BRR and Census Data

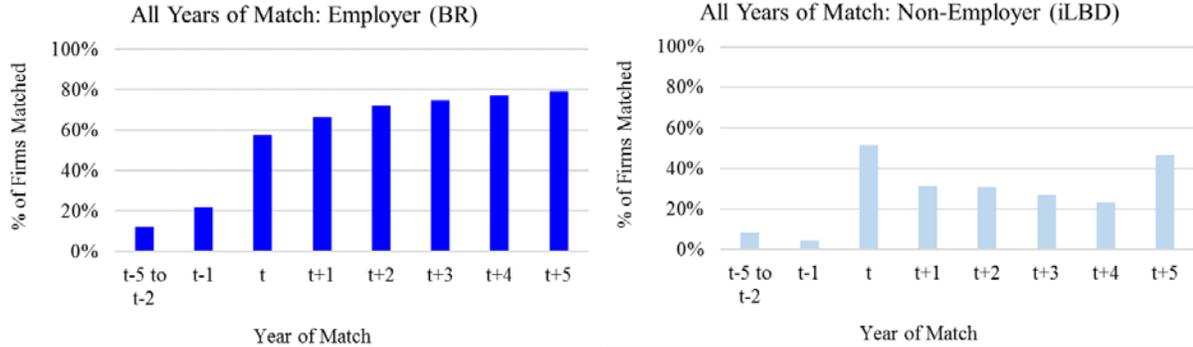
Panel A: First Year of Match



⁶ A key strength of MAMBA is the ability to enhance the training data with clerically reviewed match results, thereby increasing the accuracy of the matches. Therefore, the reported match accuracy from MAMBA is likely closer to a lower bound.

⁷ As a robustness check, all analyses were repeated with the 1995 Massachusetts registration cohort with consistent results.

Panel B: All Years of Match



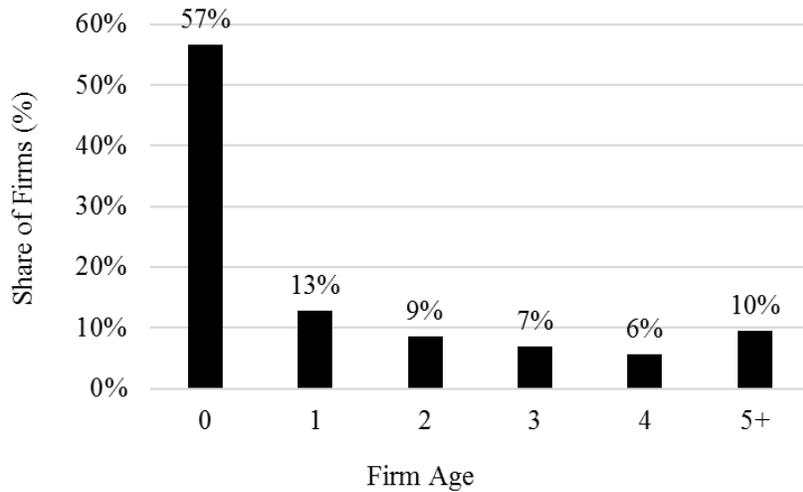
Notes: Figure 1 is based on businesses that first register in Massachusetts in 2005, conditional on matching to Census records at least once in the years between 2000 and 2010. Panel A reports year in which the business is first matched to Census records, meaning each business is counted only once. Panel B reports all years in which the business is matched to Census records, meaning each business may be counted multiple times.

Figure 1 displays the distribution of time coverage of the matches between BRR records and Census data. Panel A shows the first year of match, separately for employer and non-employer businesses. Expectedly, year t is the most frequent year in which businesses are first matched to Census data. Moreover, especially for employer businesses, most of the mass is concentrated around year t .

Similarly, Panel B demonstrates that records are disproportionately likely to be matched in year t . In the subsequent years, businesses are increasingly likely to match to the employer side of Census data, although this trend is absent for the non-employer universe. Though untested, a reasonable explanation is the differences in the underlying survival rates of employer versus non-employer businesses.

Moreover, we test the timing of BRR-Census matching with respect to the lifecycle of the firms. In particular, for the BRR records that match to an employer business in the Census data, we document the distribution of firm age in the year of the first match. Consistent with the analyses in Figure 1, we use the Massachusetts BRR businesses that register in 2005, and in particular those that match to employer-side of Census at least once between 2000 and 2010.

Figure 2: Massachusetts BRR Match Timing and Firm Age



Notes: Figure 2 is based on businesses that first register in Massachusetts in 2005 and match to employer-side Census data (Business Register) at least once between 2000 and 2010. Firm age, which is based on the year of hiring the first employee, is sourced from the Longitudinal Business Database.

Figure 2 shows that almost 60% of the businesses are new (age 0) firms in the first year of BRR-Census match. Furthermore, 85% of the initial match occurs when the firm is young (ages 0-3). Therefore, as expected, initial BRR-Census match typically occurs in the early stage of the firm’s lifecycle.

6. Conclusion

State business registration records provide an important marker of business creation. While entrepreneurs may start new businesses without necessarily registering them with the state, the benefits associated with registering a business result in individuals systematically opting into registering their company (Guzman and Stern 2015b). Therefore, state business registration data are a promising source of data – especially when coupled with the business data at Census – in advancing our understanding of new firm creation and business dynamism.

We begin to integrate the two important data sources by matching these records based on the business name and address. We demonstrate a large portion of the state business registration data from Massachusetts and California can be accurately matched to Census business data. From these efforts, several key results emerge:

- (1) Address appears to be more important than name in terms of improving matching accuracy. Therefore, a sufficiently high threshold on addresses ensures that the resulting matches are generally accurate. Address threshold below 70 is the breaking point after which the additionally matched records seem to be of poor quality.
- (2) Lower name threshold allows for new matches that would otherwise be rejected as false negative due to differences from name abbreviations and trivial typos. With these data sources, a name threshold below 65 is the breaking point after which the additionally matched records seem to be of poor quality.
- (3) SAS DQ and MAMBA, which is the machine learning-based matching algorithm, demonstrate noticeable differences in matching performance.
 - a. **Match Coverage:** MAMBA matches more BRR records than does SAS DQ.
 - b. **Match Accuracy:** SAS DQ matches are more likely to be correct than the MAMBA matches.
- (4) Massachusetts and California BRR records produce consistent results for both rates of match coverage and match accuracy. However, match rates are modestly lower for California.
- (5) Timing of the match to Census data concentrated around the year in which the business incorporates in the state.

We conclude by discussing future research questions that can be answered by leveraging the two data sources. First, the timing of entrepreneurial entry can be more accurately investigated. While the modal new entry appears in both data sources at approximately the same time, a sizable number of employer businesses appear to file BRR records prior to appearing on Census's BR. It may be the case that there is gradual process of launching a business that begins with registering with the state, and progresses to applying for an federal EIN (see Bayard et al. 2018) to hiring the first employee (Fairlie and Miranda 2017). Combining these data sets may be useful for closely tracking the lifecycle of a young business. Future research can shed light on the evolution of young businesses by leveraging such organizational milestones. Second, linking the state BRR data from states with information on the executive officers of the business maybe useful for studying business founder.⁸ While the listed officers are not necessarily the original founders of a firm, these individuals may represent a large share of the founding team when considering nascent ventures. In light of the empirical challenges in accurately identifying the founders of a startup (Azoulay et al. 2018), state BRR data appears to be a promising source of information on the key personnel.

⁸ Massachusetts is one such state.

References

- Azoulay, P., Jones, B., Kim, J. D., & Miranda, J. (2018). *Age and High-Growth Entrepreneurship* (No. w24489). National Bureau of Economic Research.
- Bayard, K., Dinlersoz, E. M., Dunne, T., Haltiwanger, J., Miranda, J., & Stevens, J. J. (2018). *Measuring Early-Stage Business Formation* (No. 2018-03-07). Board of Governors of the Federal Reserve System (US).
- DeSalvo, B., Limehouse, F., & Klimek, S. (2016). "Documenting the Business Register and Related Economic Business Data," Working Papers 16-17, Center for Economic Studies, U.S. Census Bureau. <https://ideas.repec.org/p/cen/wpaper/16-17.html>
- Fairlie, R. W., & Miranda, J. (2017). Taking the leap: The determinants of entrepreneurs hiring their first employee. *Journal of Economics & Management Strategy*, 26(1), 3-34.
- John Cuffe & Nathan Goldschlag, 2018.
"Squeezing More Out of Your Data: Business Record Linkage with Python,"
Working Papers 18-46, Center for Economic Studies, U.S. Census Bureau.
<https://ideas.repec.org/p/cen/wpaper/18-46.html>
- Guzman, J., & Stern, S. (2015a). Where is Silicon Valley?. *Science*, 347(6222), 606-609.
- Guzman, J., & Stern, S. (2015b). *Nowcasting and placecasting entrepreneurial quality and performance* (No. w20954). National Bureau of Economic Research.