Automating Response Evaluation for Franchising Questions on the 2017 Economic Census*

Andrew Baer† J. Bradford Jensen‡ Shawn Klimek§ Lisa Singh¶
Joseph Staudt‖ Yifang Wei**

March 18, 2019

Abstract

Between the 2007 and 2012 Economic Censuses (EC), the count of franchise-affiliated estblishments declined by 9.8%. One reason for this decline was a reduction in resources that the Census Bureau was able to dedicate to the manual evaluation of survey responses in the franchise section of the EC. Extensive manual evalution in 2007 resulted in many establishments, whose survey forms indicated they were not franchise-affiliated, being recoded as franchise-affiliated. No such evaulation could be undertaken in 2012. In this paper, we examine the potential of using external data harvested from the web in combination with machine learning methods to automate the process of evaluating responses to the franhise section of the 2017 EC. Our method allows us to quickly and accurately identify and recode establishments have been mistakenly classified as *not* being franchise-affiliated, increasing the unweighted number of franchise-affiliated establishments in the 2017 EC by 22%-42%.

JEL Classification Numbers: C81, L8

*We thank John Cuffe for help in getting the matching software working for this project. Thanks also to Emek Basker, Lucia Foster and seminar participants at the CRIW Pre-Conference: Big Data for 21st Century Economic Statistics for their useful comments. All remaining errors and omissions are those of the authors. The analysis, thoughts, opinions, and any errors presented here are solely those of the authors and do not necessarily reflect any official position of the U.S. Census Bureau or the International Monetary Fund (IMF). All results have been reviewed to ensure that no confidential information is disclosed. The Disclosure Review Board release number DRB #7598.

†International Monetary Fund (IMF), abaer@imf.org.
‡Georgetown University and NBER, jbj24@georgetown.edu.
§Center for Economic Studies (CES), U.S. Census Bureau, shawn.d.klimek@census.gov.
¶Department of Computer Science and MDI, Georgetown University, lisa.singh@georgetown.edu.
‖Center for Economic Studies (CES), U.S. Census Bureau, joseph.staudt@census.gov.
**Department of Computer Science, Georgetown University, yw255@georgetown.edu.

1 Introduction

The Economic Census (EC) is the most comprehensive collection of business activity data conducted by the U.S. Census Bureau. Every five years (those ending in 2 and 7), businesses are mandated to provide information including total sales, product sales, payroll, employment, and industry classification for each establishment that they operate.[1] In addition, businesses are asked to identify whether they are affiliated with a franchise, and if so, whether they are a franchisor or franchisee.[2] Data from the 2007 and 2012 Censuses indicated that, between the two time periods, the number of franchise-affiliated business establishments declined from 453,326 to 409,104, a 9.8 percent decrease. In contrast, comparable data produced by FRANdata, a research and advisory company and the strategic research partner of the International Franchise Association (IFA), showed a 4 percent *increase* in the number of franchise-affiliated outlets during this period.[3]

One possible reason for this discrepancy is the decline, between 2007 and 2012, in resources that the Census Bureau was able to dedicate to manually evaluating survey responses in the franchise section of the EC. After the 2007 EC, the Census Bureau exerted considerable effort by comparing survey responses to FRANdata and following-up with respondents over the phone. Through this process, a significant number of establishments who were not originally designated as franchise-affiliated based on their EC responses were recoded as franchise-affiliated. Unfortunately, in 2012, comparable resources were not available to conduct this extensive manual editing, contributing to the measured decline in franchise-affiliated establishments.[4]

[1] An establishment is defined as the smallest operating unit for which businesses maintain distinct records about inputs, such as payroll and operating expenses. In practice, establishments are typically individual business locations. https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf, Page 19.

[2] Cross-industry data on franchises were first published as part of the 2007 Economic Census. Franchise status had previously been collected only for restaurants. In 2007, businesses in 295 North American Industrial Classification System (NAICS) industries were asked to identify whether any of their establishments operated under a trademark authorized by a franchisor. The same question was asked to businesses in these same industries in 2012. In an attempt to reduce under-reporting, the franchise status question was modified for the 2017 Economic Census – it now asks whether the location operates under a trademark or brand authorized by a franchisor.

[3] FRANdata maintains a database of active franchise license agreements gathered from franchisors and franchisees.

[4] Another reason for the discrepancy between the Economic Census and FRANdata is that, for a variety of reasons, collecting data on franchise status is more challenging than most other business characteristics. For instance, franchise-affiliated establishments located in another retail outlet, such as a big-box store, are often not counted as a separate business establishment. In addition, multiple franchises are often operated out of a single location, such as a travel plaza. However, as the entity that fills out the survey form, the travel plaza only counts as a single-franchise-affiliated establishment. Finally, some franchises, such as colleges and universities, are out of scope for the EC. Growth in each of these types of franchise establishments likely contributes to the discrepancy in trends between the EC and FRANdata between 2007 and 2012.

The differences between the 2007 and 2012 Censuses show that, in order to ensure an accurate count of franchise-affiliated establishments, the quality of respondents' answers on the EC must be evaluated after collection. However, limited resources make it difficult to manually conduct such an evaluation. In this paper, we examine the potential of partially automating this process for the 2017 Economic Census. Specifically, we combine external data collected from the web with new machine learning algorithms designed for fuzzy name and address matching to quickly and accurately predict which establishments in the 2017 EC are likely to be franchise-affiliated and then compare our prediction to the responses (or non-responses) for these establishments' on the franchise section of the survey.[5]

To implement our procedure, we first obtain external data on franchise-affiliated establishments from two sources. First, we scrape information directly from franchise websites. This approach has the advantage of providing highly accurate and up-to-date information on a particular franchise's establishments. However, it also requires custom scripts to deal with the idiosyncrasies of each website. Second, we harvest data by querying Yelp's application programming interface (API).[6] This approach has the advantage of scalability – only a single script needs to be written and maintained. In addition, Yelp's API provides information not typically available elsewhere, such as establishment-level average customer ratings. Unfortunately, data harvested from Yelp's API is not always complete or timely.

After collecting the external data, we use new record-linking software developed at the U.S. Census Bureau (Cuffe and Goldschlag, 2018) to link external establishments (both web-scraped and Yelp-queried) to the Business Register (BR), a comprehensive list of all U.S. business establishments. The software – Multiple Algorithm Matching for Better Analytics (MAMBA) – constructs predictive features using name and address information, and feeds these features into a random forest, generating predicted probabilities of matches. In our case, for each external establishment, MAMBA identifies the establishments in the BR that are most likely to be a positive match, and thus likely to be franchise-affiliated. Finally, we link these matched establishments to the 2017 EC and compare MAMBA's predictions of franchise-affiliation to respondents' answers on the franchise section of the survey form.

Overall, we find that appoximately 70%-80% (depending on the source of external data) of establishments that MAMBA predicts to be franchise-affiliated and are in the 2017 EC (with processed forms) are identified as franchise-affiliated on the survey form – that is, MAMBA's

[5]The Economic Census (EC) is conducted at the firm-level, not the establishment-level. However, a surveyed firm gives information about each of its establishments. Thus, while a survey response may refer to a particular establishment, no one located at that etablishment may have actually filled out the survey form.

[6]Yelp is a search service that publishes crowd-sourced reviews of local business establishments. In addition to providing information on its website (yelp.com) and mobile app, Yelp provides information through an application programming interface (API).

prediction and the form responses are consistent. However, this implies that, for 20%-30% of establishments, MAMBA predicts them to be franchise-affiliated, but they are not identified as such on the survey form – that is, there is a discrepancy between MAMBA's prediction and form responses. Manual investigation of these discrepancies reveals that, in most cases, the establishments are, indeed, franchise-affiliated. That is, the MAMBA prediction is correct and the respondent mistakenly filled out the EC form.[7] Thus, we are able to identify, with a high degree of accuracy and minimal manual investigation, franchise-affiliated establishments that are mistakenly labeled as not being franchise-affiliated in the 2017 EC. Recoding these establishments increases the unweighted number of franchise-affiliated establishments in the 2017 EC by 22%-42%.

In sum, our approach of leveraging external data in combination with machine learning provides a way to reap the benefits of manually investigating the quality of 2017 EC responses to franchise questions, but in a mostly automated and cost-effective way. In particular, it allows us to identify a large set of establishments that are likely franchise-affiliated, but will not be counted as such if their 2017 EC survey forms are taken at face-value. Thus, for the 2017 EC, our approach should prove useful in avoiding the under-counting of franchise-affiliated establishments that ocurred in the 2012 EC and was only avoided in the 2007 EC by the dedication of substantial resources to manual curation.

The rest of this paper is organized as follows. The next section discusses the data – both external and restricted-use – that we use in our analyses. Section 3 discusses the linking of web-scraped and Yelp-queried establishments to the 2017 Business Register (BR) and the 2017 Economic Census (EC). Section 4 compares the MAMBA predictions of franchise-affiliation to survey form responses on the franchise section of the 2017 EC. Section 5 concludes.

2 Data

This project uses external data on franchise-affiliated establishments from two sources: 1) scraped directly from franchise websites ("web-scraped establishments") and 2) harvested from Yelp's API ("Yelp-queried establishments"). We also use franchise-level information from the *FranchiseTimes Top 200+* list and restricted-use data maintained by the U.S. Census Bureau, including the 2017 Business Register (BR) and the 2017 Economic Census (EC).

[7] In this context, a franchise-affiliated respondent can "mistakenly" fill out the EC form in two ways. First, they may not respond to the franchise section of the survey. Second, they may respond to the franchise section of the survey, but claim not to be franchise-affiliated.

2.1 FranchiseTimes

The *FranchiseTimes* is a trade publication that publishes news and data about franchising in the United States. Since 1999, it has published information on the largest U.S.-based franchises, and, in recent years, it has published information on the largest 500 franchises in its "Top 200+" list. Among other information, the list reports the number of U.S. establishments for each franchise. We use the Top 200+ list as a frame for franchises when querying Yelp's API (see Section 2.3) and as an independent source to validate the establishment counts obtained using external data (see Section 2.4).

2.2 Franchise Websites

We scrape establishment-level data directly from the websites of 12 different franchises: 7-Eleven, Ace Hardware, Burger King, Dunkin' Donuts, Great Clips, KFC, Marco's Pizza, McDonald's, Midas, Pizza Hut, Subway, and Wendy's. We refer to these 12 franchises as our "core" set of franchises. Though the list, like franchising generally, is restaurant-heavy, we made efforts to collect several non-restaurant franchises. Throughout 2017 – the reference period for the 2017 Economic Census (EC) – scripts were written and run to scrape establishment-level data using the "Find a Location" feature available on most franchise websites.[8] For a given franchise website, the script uses a zip code to submit a query for locations. By iteratively submitting a query for all U.S. zip codes, we are able to obtain an exhaustive list of establishments affiliated with the franchise. This process yielded information on 90,225 franchise-affiliated establishments.[9] Crucially for linking to the Business Register (BR), this information always includes the address of each establishment.

Obtaining establishment-level information directly from franchise websites has several advantages. First, it yields data close to "ground truth" – since a franchise has a strong incentive to maintain a complete and up-to-date list of locations on its website, we are unlikely to find a more accurate source of information about the existence of individual franchise establishments. Second, there is no ambiguity regarding the franchise with which an establishment is affiliated – if an establishment is returned from a query of Franchise A's website, we can be confident that the establishment is, in fact, affiliated with franchise A (as noted below, this is not always true for Yelp-queried establishments).

Lack of scalability is a disadvantage of obtaining information directly from franchise

[8] All scripts were run from outside the Census Bureau's IT system and the data were then transferred to Census. However, the goal is to formalize this process for the 2022 Economic Census and run all scripts from within the Census Bureau's IT system.

[9] For this paper, we collected a one-time snapshot of 2017 establishments. We did not continuously scrape information from franchise websites over the course of the year.

websites. Since each website has its own peculiarities, a custom script must be written and maintained for each franchise. Moreover, franchise websites often change, making the task of maintaining working scripts more difficult.

Another disadvantage is ambiguity regarding the *terms of use* for franchise websites (as noted below, no such ambiguity exists for Yelp's API). One franchise website explicitly allows accessing the site as long as scripts do not do so in a "manner that sends more request messages to the...servers in a given period of time than a human can reasonably produce in the same period by using a conventional online Web browser". We scraped the data using Python's *selenium* package – this allows a script to interact with a website in a point-and-click fashion, which significantly reduces the load on servers hosting franchise websites and which we initially believed was consistent with the *terms of use* for these websites. However, further review of the core franchise websites indicates that there is typically standard language prohibiting data collection without caveat. A representative example of prohibited activity includes the "Use or launch any unauthorized technology or automated system to access the online services or extract content from the online services, including but not limited to spiders, robots, screen scrapers, or offline readers..." In the future, the Census Bureau can follow the lead of the The Bureau of Labor Statistics, which obtains permission from each company to scrape their websites for price data. This would increase the cost of collecting location information directly from franchise websites, but the high quality of the data would likely make this extra cost worthwhile.

2.3 Yelp API

Yelp is a search service that publishes crowd-sourced reviews of local business establishments. In addition to providing information on its website (yelp.com) and mobile app, Yelp provides information through an application programming interface (API). We obtained the Yelp data by repeatedly querying its API using the names of the 500 franchises listed in the 2017 *FranchiseTimes Top 200+* and approximately 3,000 county names.[10] This process took

[10]Here is the section of the Yelp API *terms of use* that allows for the bulk download of data for non-commercial use: "You agree that you will not, and will not assist or enable others to: a) cache, record, pre-fetch, or otherwise store any portion of the Yelp Content for a period longer than twenty-four (24) hours from receipt of the Yelp Content, or attempt or provide a means to execute any "bulk download" operations, **with the exception of using the Yelp Content to perform non-commercial analysis (our empahsis)** (as further explained below) or storing Yelp business IDs which you may use solely for back-end matching purposes...Notwithstanding the foregoing, you may use the Yelp Content to perform certain analysis for non-commercial uses only, such as creating rich visualizations or exploring trends and correlations over time, so long as the underlying Yelp Content is only displayed in the aggregate as an analytical output, and not individually...'Non-commercial use' means any use of the Yelp Content which does not generate promotional or monetary value for the creator or the user, or such use does not gain economic value from the use of our content for the creator or user, i.e. you." See: https://www.yelp.com/developers/api_terms.

place in 2017 and resulted in a harvest of 2,551,273 unique establishments. Of these, we determined that 220,064 (8.6%) are affiliated with at least one of the 500 queried franchises and 63,395 (2.5%) are affiliated with one of the 12 franchises for which we have web-scraped data (again, we refer to these as "core" franchises). Moreover, 496 franchises in the list have at least one establishment in Yelp.

The primary advantage of using the Yelp API is scalability – a single script can be used to obtain establishment-level data on any franchise. Another advantage is the existence of establishment-level information, such as average review scores, that is not typically obtainable from franchise websites or economic surveys. Finally, the data Yelp provides is uniform across all establishments, and so is comparable across franchises. In particular, all establishments across all franchises have address information, which, as noted, is crucial for linking to the Business Register (BR).

The main disadvantage is that Yelp data are generated through user reviews and are inevitably incomplete. In addition, Yelp may be slow to expunge establishments that no longer exist. A second disadvantage is ambiguity regarding the franchise with which an establishment is affiliated. When a franchise name is used to query Yelp's API, not all harvested establishments are actually affiliated with the queried franchise. For instance, a query for "Franchise A" might yield several establishments affiliated with that franchise, but might also yield other nearby establishments affiliated with "Franchise B" (or nearby establishments not affiliated with any franchise). Thus, it is crucial to identify which establishments harvested from a query for a franchise are actually affiliated with that franchise. We are able to effectively address this issue by taking advantage of the structure of Yelp URLs, which typically contain franchise name information (see Appendix A for details).

2.4 Comparing External Data

In this section, we compare establishment counts from the *FranchiseTimes* and our two sources of external data. We display these counts in Table 1. As noted, across the 12 core franchises, we harvested 90,213 web-scraped establishments and 63,395 Yelp-queried establishments. The *FranchiseTimes* indicates that there are 91,363 establishments affiliated with these 12 franchises. There are an additional 156,669 Yelp-queried establishments affiliated with other (non-core) franchises. The *FranchiseTimes* indicates that there are 284,716 establishments affiliated with these other franchises.

Overall, these counts make it clear that the Yelp-queried data is usually less comprehensive than the web-scraped data – it does not contain all establishments for all franchises. Indeed, for all but two franchises, Pizza Hut and Midas, the number of web-scraped establish-

ments exceeds the number of Yelp-queried establishments. The higher count of Yelp-queried establishments for these two franchises appears to be driven, at least in part, by a large number of store closures during 2016-2017 which is quickly reflected on the franchise websites but not on Yelp. Thus, in addition to being less comprehensive than web-scraped data, the Yelp-queried data tend to be less up-to-date.

Table 1: Establishment Counts for External Data.

Franchise	Web-Scraped	Yelp-Queried	FranchiseTimes
Subway	27,085	13,556	26,741
McDonald's	14,153	12,060	14,153
Burger King	7,139	6,223	7,156
Pizza Hut	6,022	6,116	7,667
Wendy's	5,721	5,535	5,739
Marco's Pizza	838	789	770
KFC	4,193	3,871	4,167
Dunkin' Donuts	8,839	4,697	8,431
7-Eleven	7,624	4,067	7,008
Great Clips	3,702	3,163	3,945
Midas	1,081	1,258	1,125
Ace Hardware	3,816	2,060	4,461
Other	.	156,669	284,716
Total (12 Core)	90,213	63,395	91,363
Total (All 500)	90,213	220,064	376,079

Notes – These 500 franchises represent roughly 83-92% of the 2007-2012 franchise total. We used the external list from the *FranchiseTimes* to avoid the risk of disclosure from using confidential data from Census or IRS. All external data was obtained from outside the Census Bureau's IT system.

2.5 Business Register (BR)

The Business Register (BR) is a comprehensive list of U.S. business establishments, containing information on approximately 1.8 million establishments affiliated with 160,00 multi-unit firms, 5 million single-unit firms, and 21 million non-employer firms (DeSalvo et al., 2016). It is updated continuously, and serves as the frame for most business surveys conducted

at the Census Bureau – including the Economic Census (EC). Since we scraped data from franchise websites and queried Yelp during 2017, we link these external establishments to the 2017 November Monthly BR file.[11]

The BR contains a wide range of information on each establishment, including industry, legal form of organization, payroll, and employment. Crucially for linking, like our external data, it also contains information on the name and address of each establishment.

2.6 Economic Census (EC)

The Economic Census is a quinquennial survey (conducted in years ending in 2 and 7), and is the most comprehensive collection of business activity data conducted by the U.S. Census Bureau. Businesses are mandated to provide information including total sales, product sales, payroll, employment, and industry classification for each establishment that they operate.[12] In addition, businesses are asked whether they are affiliated with a franchise, and if so, whether they are a franchisor or franchisee.

Cross-industry data on franchise businesses was first published as part of Economic Census in 2007.[13] Franchise status had previously been collected only for restaurants. In 2007, businesses in 295 North American Industrial Classification System (NAICS) industries were asked to identify whether any of their establishments operated under a trademark authorized by a franchisor. The same question was asked to businesses in these same industries in 2012.

As noted in the introduction, data on franchise businesses is also produced by FRANdata, a research and advisory company and the strategic research partner of the International Franchise Association (IFA). FRANdata is derived from a database of active franchise license agreements gathered from franchisors and franchisees. While Economic Census data indicated that the number of franchise businesses establishments in the 295 queried industries declined from 453,326 in 2007 to 409,104 in 2012, comparable FRANdata showed a 4 percent increase in the number of these franchise outlets during this period. Once the 2012 Economic Census data was finalized and released, Census Bureau staff set out to research the discrepancy and identify steps to improve data collection of franchise status. As part of this process, meetings were conducted with representatives from IFA and FRANdata. They provided a detailed list of all US locations for the 6 largest fast food franchises from the FRANdata files

[11] The final version of this paper will match external establishments to the final version of the 2017 BR and the final collected version of the 2017 EC.

[12] An establishment is defined as the smallest operating unit for which businesses maintain distinct records about inputs, such as payroll and operating expenses. In practice, establishments are typically individual business locations.

[13] Franchise data is also collected as part of the Survey of Business Owners (SBO) and the Annual Survey of Entrepreneurs (ASE).

as of the end of 2012 that was compared to the 2012 Economic Census results.

In comparing the files, a few categories of franchise locations were identified as missing from the Economic Census data. For instance, franchise-affiliated establishments located in another retail outlet, such as a big-box store, are often not counted as a separate business establishment. In addition, multiple franchises are often operated out of a single location, such as a travel plaza. However, as the entity that fills out the survey form, the travel plaza only counts as a single-franchise-affiliated establishment. Finally, some franchises, such as colleges and universities, are out of scope for the EC. Growth in each of these types of franchise establishments likely contributes to the discrepancy in trends between the EC and FRANdata between 2007 and 2012.

Of the locations that both identified as franchises in the FRANdata files and provided a response to the franchise status question on the 2012 Economic Census, 5.6% answered that they were not operating under a trademark authorized by a franchisor. In 2007, a Census Bureau staff member spent approximately 3 months evaluating Census responses, comparing them to FRANdata, and following-up with respondents over the phone. Through this process, a significant number of franchise status non-responses and a smaller number of "no" responses were changed to "yes". In 2012, comparable resources were not available to conduct this extensive manual editing. This was also a factor in the decrease of tabulated franchised establishments between 2007 and 2012. The franchise status question was modified for the 2017 Economic Census in an attempt to reduce under-reporting. The question now asks whether the location operates under a trademark or brand authorized by a franchisor.

Each of these factors show that collecting data on franchise status is more challenging than most other business characteristics. In addition, Internal Revenue Service administrative files typically used to validate or impute collected responses do not indicate whether an establishment belongs to a franchise. As a result, the labor hours that are needed to fully validate and correct Economic Census data is substantial. This served as motivation to pursue alternative methods that could be used to accurately identify franchise locations – namely, obtaining external data on franchise-affiliated establishments from the web and using machine learning algorithms to link this data to the 2017 EC.

2.7 Other Possible Sources of External Data

Though franchise websites seem like the obvious place to begin harvesting franchise data, as noted earlier, this approach also has some serious disadvantages. In particular, it is difficult to scale – both because many scripts must be written and maintained and because

prohibitions on scraping in websites' *terms of use* will require obtaining permission from each company. The use of location services via an API is more promising with regard to *terms of use*, but, as noted, coverage is incomplete. In this section, we discuss two alternative sources of establishment-level data on franchises that may allow us to achieve comprehensive coverage without violating websites' *terms of use*.

2.7.1 Search Engine Location Services

This approach relies on Location Services that search engine companies make available. For example, Google provides the Google Places Application Program Interface (API) which returns information about a given location. A user can submit the name of a franchiser, the zip code or a county/city/state combination. The addresses of the franchisees in the zip code are returned. For this option, we wrote a script that submitted the names and zip codes of different franchisors and stored the returned addresses of the affiliated establishments in a database. The main advantage of this approach is that search engines like Google and Bing continually curate and maintain an up-to-date list of business addresses. This is important for general search and phone-based location services. The other advantage is that only a single script needs to be written. Search engines have very clear APIs that can be used to collect these data. The disadvantage of this approach is the cost. For 500 franchises in 3,141 counties (counties are larger than zip codes), our script would need to submit approximately 1.5 million queries to the API. This would cost over $7,500 using Google search and over $4,500 using Microsoft's Bing search.

2.7.2 State Government Websites

The offer and sale of a franchise requires compliance with federal and state franchise laws. While federal law provides an overriding franchise regulatory framework, some states have enacted their own franchise laws that supplement and add additional regulations. Fourteen of the states, known as the "franchise registration states", require the registration of the franchisor's Franchise Disclosure Document (FDD).[14] These states include: California, Hawaii, Illinois, Indiana, Maryland, Michigan, Minnesota, New York, North Dakota, Rhode Island, South Dakota, Virginia, Washington and Wisconsin.[15]

Obtaining data from these FDDs has the advantage of avoiding *terms of use* violations. According to Census Bureau policy, it is acceptable to scrape government websites – indeed,

[14]For example, see the 2018 McDonald's Franchise Disclosure Document.

[15]On 3/11/2019, a review of active FDDs for Wisconsin reveals 1,401 active franchises – well in excess of the 500 that we queried from Yelp's API. See https://www.wdfi.org/apps/franchiseefiling/activeFilings.aspx for the current list.

Census currently collects data from goverment websites. The Scraping Assisted by Learning (SABLE) tool (Dumbacher and Capps, 2016; Dumbacher and Diamond, 2018) was designed to scrape government websites and has built in checks to ensure it does not violate a website's *terms of use*. Thus, using SABLE to harvest establishment-level franchise data from the FDDs of the fourteen franchise registration states is a potentially fruitful approach that avoids the complication of restrictions on harvesting data directly from franchise websites. An added advantage is that these documents list franchisees, allowing us to distinguish between franchisee and franchisor-owned establishments within each brand.

3 Linking the Data

We link the external establishments scraped from franchise websites and queried from Yelp's API to the 2017 EC in two steps. First, we use MAMBA to link the external establishments to establishments in the November 2017 BR. Second, the subset of external establishments that are successfully matched to the BR are then linked to establishments in the 2017 EC. These steps are described, in detail, in the rest of this section.

3.1 Linking External Establishments to BR Establishments

MAMBA, developed by Cuffe and Goldschlag (2018), is specialized software designed to link business establishments from external data sources to establishments in the BR. It does this by constructing predictive features using name and address information, and then feeding these features into a random forest, which generates predicted probabilities of matches. In our case, for each external establishment (web-scraped or Yelp-queried), MAMBA identifies the establishments in the BR that are most likely to be positive matches. In this context, since all of our external establishments are affiliated with a franchise, MAMBA essentially identifies a subset of BR establishments that are likely to be franchise-affiliated.

The results of this linking exercise are displayed in Table 2.[16] The row titled "External Estabs" shows that, as discussed, there are 90,213 web-scraped establishments, 63,395 core Yelp-queried establishments, and 156,669 non-core Yelp-queried establishments. The row titled "Any Match" shows that approximately 83,000 (92%), 48,500 (77%), and 95,000 (61%) of these are matched to a BR establishment. Thus, it is clear that web-scraped establishments

[16]Since core Yelp-queried establishments are affiliated with the same 12 franchises as the web-scraped establishments, there is substantial overlap between the two data sources (see Appendix B), and so combining them will create duplicate establishments. To prevent this, web-scraped and Yelp-queried establishments are separately matched to the BR (though core and non-core Yelp-queried establishments are matched at the same time).

are much more likely than Yelp-queried establishments to match to a BR establishment and that establishments affiliated with a core franchise are much more likely than those affiliated with a non-core franchise to match to a BR establishment.

Table 2: Match of External Establishments to Business Register (BR).

	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
External Estabs	90,213	63,395	156,669
Any Match	83,000	48,500	95,000
1-to-1 Match	60,500	44,500	90,000

Notes – The counts in the "External Estabs" row are exact and the counts in the "Any Match" and "1-to-1 Match" rows are rounded. All counts are un-weighted.

Note that, in the "Any Match" row, a given BR establishment may be matched to more than one external establishment.[17] The next row, titled "1-to-1 Match", shows that approximately 60,500 (67%) web-scraped, 44,500 (70%) core Yelp-queried, and 90,000 (57%) non-core Yelp-queried establishments are 1-to-1 matches with a BR establishment – that is, an external establishment uniquely matches to a BR establishment and the BR establishment matches uniquely back to the external establishment. Since we know all of these external establishments are affiliated with a franchise, these 1-to-1 matches can essentially be treated as BR establishments that MAMBA predicts to be franchise-affiliated.

3.2 Linking 1-to-1 Matches to the Economic Census (EC)

Our next step is to link the BR establishments that MAMBA predicts as being franchise-affiliated (i.e., external establishments that are 1-to-1 matches with a BR establishment) to the 2017 Economic Census (EC).[18] This will allow us to examine whether MAMBA's predictions are consistent with whether or not an establishment is characterized as franchise-affiliated on their EC form.

Once an external establishment is linked to the BR, it is straightforward to link it to the EC using an internal establishment identifier. Table 3 summarizes this link. The row

[17]Since web-scraped and Yelp-queried establishments are separately matched to the BR, these multiple matches are not driven by the fact that some web-scraped establishments correspond with establishments in the Yelp-queried data and vice versa. Indeed, these multple matches occur even *within* each source of external data – that is a BR establishment may match to multiple web-scraped establishments or multiple Yelp-queried establishments.

[18]We use EC files captured in November of 2018. Collection of 2017 EC survey forms will continue until March 29, 2019. Once collection is complete the final 2017 version of the BR will become available.

labeled "1-to-1 Match with BR" shows that, as in Table 2, there are approximately 60,500 web-scraped, 44,500 core Yelp-queried, and 90,000 non-core Yelp-queried establishments that MAMBA identifies as 1-to-1 matches with a BR establishment. The row labeled "Surveyed in 2017 EC" shows that, approximately 54,500 (90%), 40,500 (91%), and 78,500 (87%) of these are included in the 2017 Economic Census. Since the 2017 EC is still being produced, not all survey forms have been collected. Thus, the row labeled "2017 EC Form Processed" reports the number of 1-to-1 matches that are included in the 2017 EC whose forms have been processed – approximately 24,500 (40%) web-scraped, 18,500 (42%) core Yelp-queried, and 36,500 (41%) non-core Yelp-queried establishments.

Table 3: Match of 1-to-1 Establishments to Economic Census (EC).

	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
1-to-1 Match with BR	60,500	44,500	90,000
Surveyed in 2017 EC	54,500	40,500	78,500
2017 EC Form Processed	24,500	18,500	36,500

Notes – All counts are rounded and all are un-weighted.

For most of the remainder of the paper, we focus on these 24,500 web-scraped and 55,000 Yelp-queried (18,500 core and 36,500 non-core) establishments. These are the subset of establishments that MAMBA predicts to be franchise-affiliated for whom we can also examine survey responses (or non-responses) about their franchise status on the 2017 EC form.

4 Evaluating Responses on the 2017 Economic Census

As noted in the previous section, we have 24,500 web-scraped, 18,500 core Yelp-queried, and 36,500 non-core Yelp-queried establishments that are both predicted to be franchise affiliated by MAMBA and have had their survey forms processed for the 2017 EC. This gives us a unique opportunity to examine whether survey responses about the establishments are consistent with MAMBA's predictions, and if they are inconsistent, examine which is correct.

Table 4 examines these responses to the 2017 EC survey form. The row titled "Franchisor or Franchisee" shows the number of establishments that respondents claim to be franchise-affiliated. As the row name suggests, an establishment is classified as franchise-affiliated if the respondent claimed to be either a franchisor or franchisee on its EC survey form. We see

that 19,500 (80%) web-scraped, 15,000 (81%) core Yelp-queried, and 25,000 (68%) non-core Yelp-queried establishments are identified as franchise-affiliated by respondents, consistent with MAMBA's prediction. Thus, for a majority of establishments, the MAMBA prediction and EC form agree that the establishment is franchise-affiliated, with higher proportions for establishments affiliated with our 12 core franchises. However, the row labeled "Not Affiliated or Not Answered" shows that this leaves a substantial number of establishments – 5,200 (21%), 3,500 (19%), and 11,000 (30%) – that respondents claim not to be franchise-affiliated, contradicting MAMBA's prediction. An establishment is classified as not being franchise-affiliated if the respondent either did not fill out the franchise portion of its EC survey form or did fill it out, but claimed that the establishment was not franchise-affiliated. Overall, Table 4 shows that a substantial portion of establishments have conflicting information.

Table 4: Responses to Franchise Questions for 1-to-1 Establishments with Processed Forms.

	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
2017 EC Form Processed	24,500	18,500	36,500
Franchisor or Franchisee	19,500	15,000	25,000
Not Affiliated or Not Answered	5,200	3,500	11,000

Notes – All counts are rounded and all are un-weighted.

These conflicts raise a crucial question: for how many establishments is MAMBA's prediction correct and for how many establishments is the EC survey form correct? To the extent that MAMBA correctly identifies franchise-affiliated establishments that respondents mistakenly label as not being franchise-affiliated, this information can be used to recode incorrect EC forms and improve the accuracy of the count of franchise-affiliated establishments in the 2017 EC.

We answer this question by taking random samples of the 5,200 web-scraped and 14,500 Yelp-queried establishments for which the MAMBA prediction and EC form are inconsistent, manually comparing the name and address information from the BR to the franchise name and address information from the external data, and determining whether the establishments are, in fact, true matches. Note that this is the only manual part of our process. The results of this manual validation are displayed in Table 5.

Table 5: MAMBA's Predictions vs. EC Form Responses.

	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Not Affiliated or Not Answered	5,200	3,500	11,000
MAMBA Prediction Correct (est.)	91.3%	93.0%	94.5%

Notes – All counts are rounded and all are un-weighted. The estimates for the percent of establishments that MAMBA correctly predicts to be franchise-affiliated is based on random samples of size 400 from each category.

As in Table 4, there are approximately 5,200 web-scraped, 3,500 core Yelp-queried, and 11,000 non-core Yelp-queried establishments that EC respondents report are *not* franchise-affiliated, but that MAMBA predicts to be franchise-affiliated. Manual investigation reveals that, in most cases, MAMBA's prediction of franchise-affiliation is correct. Indeed, we estimate that 91.3% of web-scraped establishments whose survey form contradicts MAMBA's prediction are, in fact, franchise-affiliated. Similarly, we estimate that the percentages are 93.0% and 94.5% for core and non-core Yelp-queried establishments. Thus, it appears that, as was also found in the 2007 EC, a substantial fraction of respondents incorrectly filled out the franchise section on their 2017 EC survey form.

These results suggest that we can conservatively recode the responses of 90% or more of establishments that MAMBA predicts are franchise-affiliated but that respondents report are *not* franchise-affiliated. In our data, this translates into an additional 4,748 web-scraped, 3,255 core Yelp-queried, and 10,395 non-core Yelp-queried franchise-affiliated establishments[19] which is an increase of 24%, 22%, and 42% relative to the counts obtained from the 2017 EC form alone.[20]

As noted above, 21% of web-scraped, 19% of core Yelp-queried and 30% of non-core Yelp-queried establishments whose 2017 EC forms have been processed are classified, by respondents, as not being franchise-affiliated (see Table 4). If these proportions hold, once all 54,500 web-scraped, 40,500 core Yelp-queried, and 78,500 non-core Yelp-queried establishments' EC survey forms are processed (see Table 3), we can expect 11,445, 7,695, and 23,550 to be classified as not being franchise-affiliated. If 90% are, in fact, franchise-affiliated, this means we can estimate that there exist an extra 10,301 web-scraped, 6,926 core Yelp-queried, and 21,195 non-core Yelp-queried franchise-affiliated establishments than would be suggested by the EC form alone.

[19]These were computed using information in Table 5: $4748 = 5200 * 0.913$, $3255 = 3500 * 0.930$, and $10395 = 11000 * 0.945$.

[20]These were computed using information from Tables 4 and 5: $0.243 = 4748/19500$, $0.217 = 3255/15000$, and $0.416 = 10395/25000$.

5 Conclusion

In this paper, we develop a method to partially automate the evaluation of responses to the franchise section of the 2017 Economic Census (EC). The method combines external data on franchise-affiliated establishments with machine learning algorithms to predict which establishments in the Business Register (BR) are franchise-affiliated, then links these establishments to the 2017 EC, and then examines whether respondents also characterize the establishment as franchise-affiliated.

We find that, while the predictions and survey forms agree for a majority of establishments, there are a substantial minority of cases in which an establishment is predicted to be franchise-affiliated, but the survey form does not characterize the establishment as such. The only manual part of our approach is the examination of a random sample of these discrepancies, which reveals that the predictions of franchise-affiliation are typically correct, and the form is filled out incorrectly. Recoding these establishments substantially increases the count of franchise-affiliated establishments in the 2017 EC. Thus, we find that our method provides a cost effective way to evaluate responses to the franchise section of the 2017 EC and, in turn, to potentially improve the count of franchise-affiliated establishments in the U.S.

References

Cuffe, John and 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.

DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek (2016), "Documenting the Business Register and related economic business data." Working Papers 16-17, Center for Economic Studies, U.S. Census Bureau.

Dumbacher, Brian and Cavan Capps (2016), "Big data methods for scraping government tax revenue from the web." Proceedings of the American Statistical Association, Section on Statistical Learning and Data Science 2940–2954.

Dumbacher, Brian and L.K. Diamond (2018), "SABLE: Tools for web crawling, web scraping, and text classification." Federal committee on statistical methodology research conference.

A Identifying Franchise-Affiliated Yelp-Queried Establishments

One of the disadvantages of the Yelp-queried data is ambiguity regarding the franchise with which an establishment is affiliated. Unfortunately, when a franchise name is used to query Yelp's API, not all harvested establishments are actually affiliated with the queried franchise. For instance, a query for "Franchise A" might yield several establishments affiliated with that franchise, but might also yield other nearby establishments affiliated with "Franchise B" (or nearby establishments not affiliated with any franchise). Thus, it is crucial to identify which establishments harvested from a query for a franchise are actually affiliated with that franchise.

We address this issue by taking advantage of the fact that Yelp URLs typically embed the name of the franchise with which each establishment is affiliated. Moreover, each URL is augmented with information that distinguishes the establishment from other establishments affiliated with the same franchise. This allows us to identify, with a fairly high level of confidence, all establishments in the Yelp database that are affiliated with a given franchise. To illustrate, consider the Yelp URLs listed below.

- https://www.yelp.com/biz/**franchise-a**-*boston-downtown-seaport-boston-2*
- https://www.yelp.com/biz/**franchise-a**-*boston-back-bay-fenway-boston*
- https://www.yelp.com/biz/**franchise-b**-*atlanta-ne-atlanta-2*
- https://www.yelp.com/biz/**franchise-b**-*austin-austin*
- https://www.yelp.com/biz/**nonfranchise-establishment-1**-*boulder-longmont*
- https://www.yelp.com/biz/**nonfranchise-establishment-2**-*brooklyn-queens-queens*

The bold fragments of each URL indicate the name of the establishment. The italicized fragments give information on the location of the establishment, which differentiates URLs affiliated with different establishments but the same franchise. For instance, the bold fragment of the first two URLs suggest that the establishments are affiliated with Franchise A, and the italicized fragment suggests the establishments are located in different neighborhoods in Boston. The bold fragment of second two URLs suggest that the establishments are affiliated with Franchise B, and the italicized fragment suggests that one establishment is located Atlanta and the other in Austin. Finally, the bold fragment of the last two URLs suggests that the establishments are not affiliated with any franchise on the *FranchiseTimes 200+* list.

B Linking Web-Scraped Establishments to Yelp-queried Establishments

In this section, we use franchise names and establishment addresses to link web-scraped establishments to Yelp-queried establishments, which allows us to examine the extent of

overlap between the two data sources. To do this, we use a deterministic rule-based algorithm to link establishments, which we show to be highly accurate in this context – less than 1 percent of matches are false positives.

The deterministic rule-based algorithm we use to link web-scraped and Yelp-queried establishments can be broken down into two broad steps – a pre-preprocessing and a matching step – along with a series of sub-steps:[21]

Web-Scraped / Yelp-queried (W-Y) Establishment Matching Algorithm

- Step 1: Address and Name Pre-Processing
  - A: Clean and standardize franchise names and addresses in both the web-scraped and Yelp-queried data.
  - B: Parse addresses into component parts.
- Step 2: Matching
  - A: Exact match using street number, zip code, and franchise name.
  - B: Fuzzy match on street name.

W-Y Step 1 involves preparing the web-scraped and Yelp-queried data for matching. W-Y Step 1A involves organizing the data scraped from the 12 franchise websites and data scraped from Yelp into the same format. It also involves standardization operations such as trimming of whitespace, converting all text to lower-case, eliminating non alpha-numeric characters, etc. Step 1B enables matching separately on different address components (e.g. zip code, street number, street name, etc.), rather than matching based on the entire unparsed address string.

W-Y Step 2 implements the matching process using the standardized data produced in the previous step. In W-Y Step 2A, we identify all pairwise combinations of web-scraped and Yelp-queried establishments that are affiliated with the same franchise, are located in the same zip code, and have the same street number. Notice that the street name plays no role in the match process at Step 2A. However at W-Y Step 2B the street address is used to narrow the number of possible matches. Specifically, we use 26 different string comparators to compute 26 similarity scores between the street names for each pairwise combination of establishments identified in the previous step.[22] We then compute the mean similarity score and identify the subset of establishment combinations that have the highest score.

Table A1 gives an overview of the results this algorithm produces. The column titled "Web-to-Yelp" examines links of web-scraped establishments to Yelp-queried establishments. The column titled "Yelp-to-Web" examines the results for matching in the reverse direction – Yelp-queried establishments to web-scraped establishments. As also shown in Table 1,

[21] For this linking exercise, since we scrape data from 12 franchise websites, we only retain Yelp-queried establishments belonging to these same 12 franchises. When we link scraped establishments to the BR, we use Yelp-queried establishments from all 496 franchises in the *FranchiseTimes 200+* list.

[22] We use Stata's *matchit* command to compute the similarity scores.

there are a total of 90,213 web-scraped and 63,395 Yelp-queried establishments across the 12 core franchises.

Table 6: Match of Web-Scraped Estabs to Yelp-Queried Estabs.

	Web-to-Yelp	Yelp-to-Web
External Estabs	90,213	63,395
Any Match	51,144	51,642
1-to-1 Match	50,255	50,255

Notes –

The row titled "Any" indicates the count of establishments from one source that match to at least one establishment from the other source. We see that 51,144 (56.7%) web-scraped establishments match to a Yelp-queried establishment and 51,642 (81.4%) Yelp-queried establishments match to a web-scraped establishment. The row titled "1-to-1 Match" indicates the count of establishments from one source that are uniquely matched to an establishment in the other source and vice versa. By definition, this count must be the same whether we are matching Web-to-Yelp or Yelp-to-Web. We see that 50,225 external establishments are uniquely matched across the two data sources, which is 55.7% of web-scraped establishments and 79.3% of Yelp-queried establishments.

In sum, there are a large number of web-scraped establishments (43.3%) that are unmatched to a Yelp-queried establishment and substantially fewer Yelp-queried establishments (18.5%) that are unmatched to a web-scraped establishment. Conversely, about 79.3% of Yelp-queried establishments are 1-to-1 matches with a web-scraped establishment, but only 55.7% of web-scraped establishments are 1-to-1 matches with a Yelp-queried establishment. These patterns reflect the less comprehensive coverage of the Yelp data.

It is important to note that, just because a web-scraped establishment and a Yelp-queried establishment are designated as a 1-to-1 match, does not mean the match is correct. Thus, to examine the accuracy of the deterministic rule-based algorithm, we manually examine random samples of the 50,225 1-to-1 matches. This exercise leads us to conclude that the algorithm is highly accurate in this context – indeed, we estimate a false positive match rate of less than 1 percent.