This PDF is a selection from a published volume from the National Bureau of Economic Research

Volume Title: Big Data for Twenty-First-Century Economic Statistics

Volume Authors/Editors: Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-80125-4 (cloth),
978-0-226-80139-1 (electronic)

Volume URL:
https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics

Conference Date: March 15-16, 2019

Publication Date: Februrary 2022

Chapter Title: Automating Response Evaluation for Franchising Questions on the 2017 Economic Census

Chapter Author(s): Joseph Staudt, Yifang Wei, Lisa Singh, Shawn Klimek, J. Bradford Jensen, Andrew Baer

Chapter URL:
https://www.nber.org/books-and-chapters/big-data-twenty-first-century-economic-statistics/automating-response-evaluation-franchising-questions-2017-economic-census

Chapter pages in book: p. 209 – 227

7

Automating Response Evaluation for Franchising Questions on the 2017 Economic Census

Joseph Staudt, Yifang Wei, Lisa Singh, Shawn Klimek,
J. Bradford Jensen, and Andrew Baer

7.1 Introduction

The Economic Census (EC) is the most comprehensive collection of business activity data conducted by the US 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. In addition, businesses are asked to identify whether they are affiliated with a franchise, and if

Joseph Staudt is an economist at the Center for Economic Studies and the Center for Big Data Research and Applications at the US Census Bureau.

Yifang Wei is a postdoctoral researcher in the Department of Computer Science at Georgetown University.

Lisa Singh is a professor in the Department of Computer Science at Georgetown University.

Shawn Klimek is Assistant Center Chief for Business Research at the Center for Economic Studies at the US Census Bureau.

J. Bradford Jensen is McCrane/Shaker Chair in international business and professor of economics and international business at the McDonough School of Business at Georgetown University, and a research associate of the National Bureau of Economic Research.

Andrew Baer is currently a senior economist at the International Monetary Fund, but he was the Assistant Division Chief of the Services Sectors in the Economy-Wide Statistics Division at the US Census Bureau when this chapter was written.

We thank John Cuffe for help with the matching software. Thanks also to Emek Basker, Lucia Foster, Cathy Buffington, Ron Jarmin, 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 US 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 is DRB #7961. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see https://www.nber.org/books-and-chapters/big-data-21st-century-economic-statistics/automating-response-evaluation-franchising-questions-2017-economic-census.

so, whether they are a franchisor or franchisee. 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 derived from franchise license agreements and 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 establishments during this period.

One reason for this discrepancy was the decline, between 2007 and 2012, in resources the Census Bureau was able to dedicate to the manual evaluation of survey responses in the franchise section of the EC. After the 2007 EC, Census Bureau staff compared survey responses to FRANdata and followed up with respondents over the phone. Through this process, a significant number of establishments that 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.[1]

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 survey form 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 EC. 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 nonresponses) for these establishments on the franchise section of the survey.[2]

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 scraping scripts to deal with the idiosyncrasies of each website. Second, we harvest data by querying Yelp's application programming interface (API).[3] This approach has

[1] Another reason for the discrepancy, as discussed in section 7.2.6, was a growth in categories of franchise-affiliated establishments that were captured by FRANdata, but often missing from the EC data.

[2] 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 establishment necessarily filled out the survey form.

[3] Yelp is a search service that publishes crowdsourced 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).

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 US Census Bureau (Cuffe and Goldschlag 2018) to link external establishments (both web-scraped and Yelp-queried) to the US Census Bureau Business Register (BR), a comprehensive list of all US 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 approximately 70–80 percent (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 prediction and the form responses are consistent. However, this implies that for 20–30 percent 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.[4] 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 percent.

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 undercounting of franchise-affiliated establishments that occurred in the 2012

[4] 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—a nonresponse mistake. Second, they may respond to the franchise section of the survey but claim not to be franchise affiliated—an incorrect response mistake.

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. We also discuss possible alternative sources of external data on franchise-affiliated establishments that may overcome some of the shortcomings of the web-scraped and Yelp-queried data. Section 7.3 discusses the linking of web-scraped and Yelp-queried establishments to the 2017 BR and the 2017 EC. Section 7.4 compares the MAMBA predictions of franchise affiliation to survey form responses on the franchise section of the 2017 EC. Section 7.5 concludes.

7.2 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 US Census Bureau, including the 2017 BR and the 2017 EC.

7.2.1 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 US-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 US establishments for each franchise. We use the Top 200+ list as a frame for franchises when querying Yelp's API (see section 7.2.3) and as an independent source to validate the establishment counts obtained using external data (see section 7.2.4).

7.2.2 Franchise Websites

We scrape establishment-level data directly from the websites of 12 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 nonrestaurant franchises. Throughout 2017—the reference period for the 2017 EC—scripts were written and run to scrape establishment-level data using the "Find a Location" feature available on most franchise websites.[5] For a given franchise website, the script uses a zip code to submit

[5] 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 EC and run all scripts from within the Census Bureau's IT system.

a query for locations. By iteratively submitting a query for all US 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.[6] Crucially for linking to the 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 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 "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 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 may make this extra cost worthwhile.

[6] 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.

7.2.3 Yelp API

Yelp is a search service that publishes crowdsourced 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.[7] This process took place in 2017 and resulted in a harvest of 220,064 establishments affiliated with at least one of the 500 queried franchises and 63,395 establishments affiliated with one of the 12 franchises for which we have web-scraped data (again, we refer to these 12 as "core" franchises). From the list of 500 franchises, 496 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 uniformity of the Yelp data across all establishments, and thus its comparability across franchises. In particular, all establishments across all franchises have address information—which, as noted, is crucial for linking to the BR.

The main disadvantage is that Yelp data are generated through user reviews and are inevitably incomplete. For a given franchise, this incompleteness likely decreases the number of establishments in the BR that we can identify as being affiliated with the franchise. 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

[7] Here is the section of the Yelp API *terms of use* that allows for the bulk download of data for noncommercial 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 emphasis] (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.

Table 7.1 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 (488 non-Core)	.	156,669	284,716
Total (12 Core)	90,213	63,395	91,363
Total (All 500)	90,213	220,064	376,079

Notes: We used the *FranchiseTimes* list to avoid disclosure risk from using confidential Census Bureau or IRS data. All external data were harvested from outside the Census Bureau's IT system.

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).

7.2.4 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 7.1. 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 the other 488 (noncore) 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 are usually less comprehensive than the web-scraped data—they do not contain all establishments for all franchises. Indeed, for all but two franchises (Pizza Hut and Midas), the number of web-scraped establishments exceeds the number of Yelp-queried establishments.

7.2.5 Business Register (BR)

The BR is a comprehensive list of US businesses, containing information on approximately 1.8 million establishments affiliated with 160,000 multiunit firms, 5 million single-unit firms, and 21 million nonemployer firms (DeSalvo, Limehouse, and Klimek 2016). It is updated continuously and serves as the frame for most business surveys conducted at the Census Bureau—including the EC. Since we scraped data from franchise websites and queried Yelp during 2017, we linked these external establishments to the 2017 BR.

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

7.2.6 Economic Census

The EC 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 US Census Bureau. Businesses are mandated to provide information including total sales, product sales, payroll, employment, and industry classification for each establishment that they operate.[8] In addition, businesses are asked whether they are affiliated with a franchise, and if so, whether they are a franchisor or franchisee.[9] Prior to the 2007 EC, franchise status was collected only for restaurants. In the 2007 and 2012 Censuses, businesses across 295 North American Industrial Classification System (NAICS) industries were asked whether any of their establishments operated under a trademark authorized by a franchisor. In an attempt to reduce underreporting, the 2017 EC franchise status question was modified to ask whether an establishment operates under a trademark *or brand* authorized by a franchisor.

As noted in the introduction, FRANdata, a research/advisory company and the strategic research partner of the International Franchise Association (IFA), uses active franchise license agreements to construct a database on franchise-affiliated establishments. In contrast to EC data, which indicates a decline from 453,326 to 409,104 in the number of franchise-affiliated establishments between 2007 and 2012, comparable FRANdata indicates a 4 percent increase in franchise-affiliated establishments. After the release of the 2012 EC, Census Bureau staff, in collaboration with representatives from IFA and FRANdata, set out to identify the reasons for this discrepancy.

The first main reason for the discrepancy was a growth in categories of franchise-affiliated establishments that were captured by FRANdata but were often missing from the EC 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 the EC. In addi-

[8] 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. See: https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf, page 19.

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

tion, multiple franchises are often operated out of a single location, such as a travel plaza. However, as the entity that fills out the EC survey form, the travel plaza only counts as a single franchise-affiliated establishment. Finally, some franchises are owned by institutions that are out of scope to the EC, such as colleges and universities and government agencies.

The second main reason for the discrepancy is that in 2007, a Census Bureau staff member spent approximately three months evaluating EC survey responses, comparing them to FRANdata and following up with respondents over the phone. Through this process, a significant number of establishments owned by firms that did not fill out the franchise section on the EC form (i.e., item nonresponse) were recoded to franchise affiliated. In addition, a smaller number of establishments owned by firms that claimed not to be franchise affiliated were recoded as franchise affiliated (i.e., incorrect response). In 2012, comparable resources were not available to conduct this extensive manual editing, contributing to a *measured* decrease in the number of franchise-affiliated establishments. The substantial number of labor hours needed to fully validate and correct the franchise section on the EC form served as motivation in this paper to pursue alternative methods that could be used to quickly and accurately identify (and when necessary, reclassify) franchise-affiliated establishments in the 2017 EC.

7.2.7 Other Possible Sources of External Data

Though franchise websites are an attractive source for harvesting establishment-level franchise data, as noted earlier, this approach has some serious disadvantages. In particular, it is difficult to scale—both because many scraping scripts must be written and maintained and because prohibitions on scraping in websites' *terms of use* requires obtaining permission from each company. The use of Yelp's 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*.

7.2.7.1 Search Engine Location Services

One possible alternative approach relies on location services provided by search engine companies. For example, Google provides the Google Places API and Microsoft's Bing provides the Bing Maps Locations API. A user can submit a franchise name and location information (e.g., the zip code or a county/city/state combination) and addresses of the franchise-affiliated establishments in that location are returned. The main advantages of this approach are that Google and Bing continually curate and maintain an up-to-date list of business addresses, ensuring high-quality and timely data, and that only a single script needs to be written to query an API, ensuring scalability. The main disadvantage of this approach is cost. For instance, to ensure comprehensive coverage of 500 franchises across 3,141 counties, we would need to submit over 1.5 million queries to an API, which would cost over $7,500 using Google and over $4,500 using Bing.

7.2.7.2 State Government Websites

The offer and sale of a franchise requires compliance with federal and state franchise laws. While federal law provides a franchise regulatory framework, some states have enacted supplemental franchise laws. In particular, 14 states known as "franchise registration states" require the registration of franchisors' Franchise Disclosure Documents (FDDs), which are another possible source of establishment-level franchise data.[10] One major advantage of this source is the avoidance of *terms of use* violations. Indeed, Census Bureau policy currently allows the scraping of government websites, and the Scraping Assisted by Learning (SABLE) software (Dumbacher and Diamond 2018), which has built-in checks to ensure compliance with a website's *terms of use*, is already used for this purpose. An additional advantage is that FDDs list franchisees, allowing us to distinguish between franchisee- and franchisor-owned establishments within each brand.

7.3 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 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.

7.3.1 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, because all our external establishments are affiliated with a franchise, MAMBA essentially identifies a subset of BR establishments that are likely to be franchise affiliated.

[10] These states include California, Hawaii, Illinois, Indiana, Maryland, Michigan, Minnesota, New York, North Dakota, Rhode Island, South Dakota, Virginia, Washington, and Wisconsin. On 3/11/2019, a review of active FDDs for Wisconsin suggested the existence of 1,401 active franchises—well in excess of the 500 contained in the *FranchiseTimes Top 200+*. See https://www.wdfi.org/apps/franchiseefiling/activeFilings.aspx for the current list.

Table 7.2 **Match of external establishments to Business Register (BR)**

	Web-scraped	Yelp (Core)	Yelp (non-Core)
External establishments	90,213	63,395	156,669
Any match	65,000	47,500	93,000
1-to-1 match	57,500	44,500	89,000

Notes: The counts in the "External establishments" row are exact and the counts in the "Any match" and "1-to-1 match" rows are rounded. All counts are unweighted.

The results of this linking exercise are displayed in table 7.2.[11] 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 noncore Yelp-queried establishments. The row titled "Any Match" shows that approximately 65,000 (72 percent), 47,500 (75 percent), and 93,000 (59 percent) of these are matched to a BR establishment. Thus, it is clear that establishments affiliated with a core franchise are much more likely than those affiliated with a noncore franchise to match to a BR establishment.

Note that in the "Any Match" row, a given BR establishment may be matched to more than one external establishment.[12] The next row, titled "1-to-1 Match," shows that approximately 57,500 (64 percent) web-scraped, 44,500 (70 percent) core Yelp-queried, and 89,000 (57 percent) noncore 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 external establishments are affiliated with a franchise, these 1-to-1 matches can be treated as BR establishments that MAMBA predicts to be franchise affiliated.

7.3.2 Linking 1-to-1 Matches to the Economic Census

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 EC.[13] This allows us to examine whether MAMBA's predictions are consistent with whether an establishment is characterized as franchise affiliated on their EC form.

[11] 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 noncore Yelp-queried establishments are matched at the same time).

[12] 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 multiple 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.

[13] We use EC files captured in May 2019.

Table 7.3 **Match of 1-to-1 establishments to Economic Census (EC)**

	Web-scraped	Yelp (Core)	Yelp (non-Core)
1-to-1 match with BR	57,500	44,500	89,000
Surveyed in 2017 EC	52,500	40,500	78,500
EC form processed	29,000	21,500	41,000

Note: All counts are rounded and all are unweighted.

Once an external establishment is linked to the BR, it is straightforward to link it to the EC using an internal establishment identifier. Table 7.3 summarizes this link. The row labeled "1-to-1 Match with BR" shows that, as in table 7.2, there are approximately 57,500 web-scraped, 44,500 core Yelp-queried, and 89,000 noncore 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 52,500 (91 percent), 40,500 (91 percent), and 78,500 (88 percent) of these are included in the 2017 EC. Since the processing of the 2017 EC is still ongoing, 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 29,000 (55 percent) web-scraped, 21,500 (53 percent) core Yelp-queried, and 41,000 (52 percent) noncore Yelp-queried establishments.

For most of the remainder of the paper, we focus on these 29,000 web-scraped and 62,500 Yelp-queried (21,500 core and 41,000 noncore) establishments. These are the subset of establishments that MAMBA predicts to be franchise affiliated, for whom we can also examine survey responses (or nonresponses) about their franchise status on the 2017 EC form.

7.4 Evaluating Responses on the 2017 Economic Census

As noted in the previous section, we have 29,000 web-scraped, 21,500 core Yelp-queried, and 41,000 noncore 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 7.4 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 21,500 (74 percent) web-scraped, 16,500 (77 percent) core Yelp-queried, and 28,500 (70 percent) noncore Yelp-queried establishments are identified as franchise

Table 7.4 **Responses to franchise questions for 1-to-1 establishments with processed forms**

	Web-scraped	Yelp (Core)	Yelp (non-Core)
EC form processed	29,000	21,500	41,000
Franchisor or franchisee	21,500	16,500	28,500
Not affiliated or not answered	7,400	5,000	12,500

Note: All counts are rounded and all are unweighted.

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 somewhat 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—7,400 (26 percent), 5,000 (23 percent), and 12,500 (30 percent)—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. Both these groups are classified as not being franchise affiliated because they would be classified as such if their EC forms were taken at face value. Overall, table 7.4 shows that a substantial portion of establishments have conflicting information.

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 7,400 web-scraped and 17,500 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 7.5.

As in table 7.4, there are approximately 7,400 web-scraped, 5,000 core Yelp-queried, and 12,500 noncore 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 98.4 percent of web-scraped establishments whose survey form con-

Table 7.5 **MAMBA's predictions vs. EC form responses**

	Web-scraped	Yelp (Core)	Yelp (non-Core)
Not affiliated or not answered	7,400	5,000	12,500
MAMBA prediction correct (est.)	98.4%	95.5%	93.5%

Notes: All counts are rounded and all are unweighted. The estimates for the percent of establishments that MAMBA correctly predicts to be franchise-affiliated is based on random samples of size 300 from each category.

tradicts MAMBA's prediction are, in fact, franchise affiliated. Similarly, we estimate that the percentages are 95.5 percent and 93.5 percent for core and noncore Yelp-queried establishments. Thus, it appears that, as was also found in the 2007 EC, a substantial fraction of respondents either incorrectly filled out the franchise section on their 2017 EC survey form or did not fill it out at all.

These results suggest that we can conservatively recode the responses of 90 percent 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 7,282 web-scraped, 4,755 core Yelp-queried, and 11,688 noncore Yelp-queried franchise-affiliated establishments,[14] which is an increase of 34 percent, 29 percent, and 41 percent, respectively, relative to the counts obtained from the 2017 EC form alone.[15]

As noted above, 26 percent of web-scraped, 23 percent of core Yelp-queried and 30 percent of noncore Yelp-queried establishments whose 2017 EC forms have been processed are classified by respondents as not being franchise affiliated (see table 7.4). If these proportions hold, once all 52,500 web-scraped, 40,500 core Yelp-queried, and 78,500 noncore Yelp-queried establishments' EC survey forms are processed (see table 7.3), we can expect 13,650 (= 52,500 * 0.26), 9,315 (= 40,500 * 0.23), and 23,550 (= 78,500 * 0.30) to be classified as not being franchise affiliated on the basis of their EC form. If we conservatively reclassified 90 percent of these as franchise affiliated, we would obtain an extra 12,285 web-scraped, 8,384 core Yelp-queried, and 21,195 noncore Yelp-queried franchise-affiliated establishments than would be suggested by the EC form alone.

7.5 Conclusion

In this paper, we develop a method to mostly automate the evaluation of responses to the franchise section of the 2017 EC. The method combines external data on franchise-affiliated establishments with machine learning

[14] These were computed using information in table 7.5: 7282 = 7400 * 0.984, 4755 = 5000 * 0.955, and 11688 = 12500 * 0.935.

[15] These were computed using information from tables 7.4 and 7.5: 0.339 = 7282/21500, 0.288 = 4755/16500, and 0.410 = 11688/28500.

algorithms to predict which establishments in the BR are franchise affiliated, 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 US.

If a version of our process is used to augment the production of official franchising statistics, several improvements can be made. First, since we only collect data on 12 core and 488 noncore franchises, it will be crucial to obtain a much more comprehensive external list of franchise-affiliated establishments. We believe the most promising sources for this comprehensive data are search engine location services and franchise disclosure documents from state government websites, both of which are discussed in section 7.2.7. Our process allowed us to reclassify enough establishments to increase (relative to taking the EC form at face value) the franchise-affiliated count by 34 percent (web-scraped) and 29 percent (Yelp-queried) for the 12 core franchises and by 41 percent (Yelp-queried) for the 488 noncore franchises. Since additional franchises from an expanded list are likely to more closely resemble the 488 noncore franchises, we may expect a higher reclassification *rate* for EC establishments matched to establishments affiliated with the newly acquired franchises. However, since the newly acquired franchises will tend to have fewer affiliated establishments, the impact of adding these franchises to the total *count* of reclassified establishments may be modest.

Second, it will be important to improve MAMBA's predictions. More comprehensive data will help with this. In addition, MAMBA enables users to manually create bespoke training data tailored for a specific use case. Though the creation of these training data will require extensive manual labeling of true and false matches, the probability of significantly improving match rates between the external data and the BR is likely to make it worthwhile.

Finally, in this paper we only manually examine discrepancy cases in which MAMBA predicts that an establishment is franchise affiliated, but its EC form indicates otherwise. It will also be crucial to examine discrepancy cases in which an establishment's EC form indicates it is franchise affiliated, but MAMBA fails to predict it as such. To get a truly accurate franchise count, some of these establishments may need to be reclassified as not franchise affiliated.

Appendix 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 suggests 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 the second two URLs suggests 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.

Appendix 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 preprocessing and a matching step—along with a series of sub-steps:[16]

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

- Step 1: Address and Name Preprocessing
  –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 lowercase, eliminating nonalphanumeric characters, etc. Step 1B enables matching separately on different address components (e.g., zip code, street number, street name), 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, located in the same zip code, and have the

[16] 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.

Table 7A.1 **Match of web-scraped establishments to Yelp-queried establishments**

	Web-to-Yelp	Yelp-to-web
External establishments	90,213	63,395
Any match	51,144	51,642
1-to-1 match	50,255	50,255

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.[17] We then compute the mean similarity score and identify the subset of establishment combinations that have the highest score.

Table 7A.1 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 7.1, there are a total of 90,213 web-scraped and 63,395 Yelp-queried establishments across the 12 core franchises.

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 percent) web-scraped establishments match to a Yelp-queried establishment and 51,642 (81.4 percent) 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 percent of web-scraped establishments and 79.3 percent of Yelp-queried establishments.

In sum, there is a large number of web-scraped establishments (43.3 percent) that are unmatched to a Yelp-queried establishment and substantially fewer Yelp-queried establishments (18.5 percent) that are unmatched to a web-scraped establishment. Conversely, about 79.3 percent of Yelp-queried establishments are 1-to-1 matches with a web-scraped establishment, but only 55.7 percent 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

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

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.

References

Cuffe, John, and Nathan Goldschlag. 2018. "Squeezing More Out of Your Data: Business Record Linkage with Python." Center for Economic Studies Working Paper 18-46, US Census Bureau, Washington, DC.

DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek. 2016. "Documenting the Business Register and Related Economic Business Data." Center for Economic Studies Working Paper 16-17, US Census Bureau, Washington, DC.

Dumbacher, Brian, and Cavan Capps. 2016. "Big Data Methods for Scraping Government Tax Revenue from the Web." In *Proceedings of the American Statistical Association*, Section on Statistical Learning and Data Science, 2940–54. Alexandria, VA: American Statistical Association.

Dumbacher, Brian, and L. K. Diamond. 2018. "SABLE: Tools for Web Crawling, Web Scraping, and Text Classification." Federal Committee on Statistical Methodology Research Conference, March 7, 2018. https://nces.ed.gov/FCSM/2018_ResearchPolicyConference.asp.