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 15, 2019

Disclaimer

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.

Discrepancies in the Data

- Between 2007 and 2012 Economic Censuses (EC), tabulated number of franchise-affiliated establishments declined by 9.8%.
- Data derived from active franchise license agreements (FRANdata) suggested a 4% increase over this period.
- After 2012 EC, Census employees set out to identify reasons for discrepancy.

Discrepancies in the Data

- Between 2007 and 2012 Economic Censuses (EC), tabulated number of franchise-affiliated establishments declined by 9.8%.
- Data derived from active franchise license agreements (FRANdata) suggested a 4% increase over this period.
- After 2012 EC, Census employees set out to identify reasons for discrepancy.

Discrepancies in the Data

- Between 2007 and 2012 Economic Censuses (EC), tabulated number of franchise-affiliated establishments declined by 9.8%.
- Data derived from active franchise license agreements (FRANdata) suggested a 4% increase over this period.
- After 2012 EC, Census employees set out to identify reasons for discrepancy.

Why the Discrepancies?

- One reason for discrepancy: reduction in resources dedicated to manual evaluation of survey responses.
- 2007
  - Census staff compared EC responses to FRANdata
  - Followed up with respondents over the phone
  - Many establishments recoded to franchise-affiliated
- 2012
  - Resources not available for manual evaluation
  - No recoding takes place

Why the Discrepancies?

- One reason for discrepancy: reduction in resources dedicated to manual evaluation of survey responses.
- 2007
  - Census staff compared EC responses to FRANdata
  - Followed up with respondents over the phone
  - Many establishments recoded to franchise-affiliated
- 2012
  - Resources not available for manual evaluation
  - No recoding takes place

Why the Discrepancies?

- One reason for discrepancy: reduction in resources dedicated to manual evaluation of survey responses.
- 2007
  - Census staff compared EC responses to FRANdata
  - Followed up with respondents over the phone
  - Many establishments recoded to franchise-affiliated
- 2012
  - Resources not available for manual evaluation
  - No recoding takes place

Response Evaluation: Manual and Automated

- Obtaining accurate counts of franchise-affiliated establishments requires evaluation of EC survey responses
- But manual evaluation is costly...
- So we combine external data and machine learning to partially automate this process

Response Evaluation: Manual and Automated

- Obtaining accurate counts of franchise-affiliated establishments requires evaluation of EC survey responses
- But manual evaluation is costly...
- So we combine external data and machine learning to partially automate this process

Response Evaluation: Manual and Automated

- Obtaining accurate counts of franchise-affiliated establishments requires evaluation of EC survey responses
- But manual evaluation is costly...
- So we combine external data and machine learning to partially automate this process

Automation Helps

- Our method quickly and accurately identifies establishments mistakenly classified as not being franchise-affiliated on their 2017 EC form.
- Recoding these establishments increases unweighted count of franchise-affiliated establishments by 22-42%.

Automating the Evaluation of 2017 EC Responses

- Harvest external data from the web on franchise-affiliated estabs.
- Use machine learning to link external data to the Business Register (BR)
    - Identifies BR establishments that are likely franchise-affiliated
- Link external data to 2017 Economic Census
- Evaluate answers to franchise questions on EC form
    - Compare ML algorithm prediction to survey responses
    - Identify and manually evaluate discrepancies
    - Estimate proportion of establishments that can be recoded to franchise-affiliated

Automating the Evaluation of 2017 EC Responses

- Harvest external data from the web on franchise-affiliated estabs.
- Use machine learning to link external data to the Business Register (BR)
  - Identifies BR establishments that are likely franchise-affiliated
- Link external data to 2017 Economic Census
- Evaluate answers to franchise questions on EC form
  - Compare ML algorithm prediction to survey responses
  - Identify and manually evaluate discrepancies
  - Estimate proportion of establishments that can be recoded to franchise-affiliated

Automating the Evaluation of 2017 EC Responses

- Harvest external data from the web on franchise-affiliated estabs.
- Use machine learning to link external data to the Business Register (BR)
  - Identifies BR establishments that are likely franchise-affiliated
- Link external data to 2017 Economic Census
- Evaluate answers to franchise questions on EC form
  - Compare ML algorithm prediction to survey responses
  - Identify and manually evaluate discrepancies
  - Estimate proportion of establishments that can be recoded to franchise-affiliated

Automating the Evaluation of 2017 EC Responses

- Harvest external data from the web on franchise-affiliated estabs.
- Use machine learning to link external data to the Business Register (BR)
    - Identifies BR establishments that are likely franchise-affiliated
- Link external data to 2017 Economic Census
- Evaluate answers to franchise questions on EC form
    - Compare ML algorithm prediction to survey responses
    - Identify and manually evaluate discrepancies
    - Estimate proportion of establishments that can be recoded to franchise-affiliated

Harvesting External Data

- Web-scraping individual franchise websites
  - 12 “core” franchises
- Querying Yelp’s API
  - 12 “core” and 488 “non-core” franchises
  - *FranchiseTimes* is frame for top 500 U.S. franchises

Harvesting External Data

- Web-scraping individual franchise websites
  - 12 “core” franchises
- Querying Yelp’s API
  - 12 “core” and 488 “non-core” franchises
  - *FranchiseTimes* is frame for top 500 U.S. franchises

Establishment Counts for External Data

Franchises	Web-Scraped	Yelp-Queried
12 Core	90,213	63,395
488 Non-Core	N/A	156,669
All 500	90,213	220,064

Establishment Counts for External Data

Franchises	Web-Scraped	Yelp-Queried
12 Core	90,213	63,395
488 Non-Core	N/A	156,669
All 500	90,213	220,064

Establishment Counts for External Data

Franchises	Web-Scraped	Yelp-Queried
12 Core	90,213	63,395
488 Non-Core	N/A	156,669
All 500	90,213	220,064

Establishment Counts for External Data

Franchises	Web-Scraped	Yelp-Queried
12 Core	90,213	63,395
488 Non-Core	N/A	156,669
All 500	90,213	220,064

Linking External Data to the Business Register

- Multiple Algorithm Matching for Better Analytics (MAMBA)
    - Cuffe and Goldschlag (2018)
- Specialized software designed to link business estabs. from external data to estabs. in the BR.
    - Constructs predictive features using name and address info
    - Feeds features into a random forest
    - Generates predicted probabilities of matches
- In our context, MAMBA identifies a subset of BR estabs. likely to be franchise-affiliated.

Linking External Data to the Business Register

- Multiple Algorithm Matching for Better Analytics (MAMBA)
  - Cuffe and Goldschlag (2018)
- Specialized software designed to link business estabs. from external data to estabs. in the BR.
  - Constructs predictive features using name and address info
  - Feeds features into a random forest
  - Generates predicted probabilities of matches
- In our context, MAMBA identifies a subset of BR estabs. likely to be franchise-affiliated.

Linking External Data to the Business Register

- Multiple Algorithm Matching for Better Analytics (MAMBA)
  - Cuffe and Goldschlag (2018)
- Specialized software designed to link business estabs. from external data to estabs. in the BR.
  - Constructs predictive features using name and address info
  - Feeds features into a random forest
  - Generates predicted probabilities of matches
- In our context, MAMBA identifies a subset of BR estabs. likely to be franchise-affiliated.

Linking External Data to the Business Register

- Multiple Algorithm Matching for Better Analytics (MAMBA)
  - Cuffe and Goldschlag (2018)
- Specialized software designed to link business estabs. from external data to estabs. in the BR.
  - Constructs predictive features using name and address info
  - Feeds features into a random forest
  - Generates predicted probabilities of matches
- In our context, MAMBA identifies a subset of BR estabs. likely to be franchise-affiliated.

Match of External Establishments to Business Register

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
External Estabs	90,213	63,395	156,669
1-to-1 Match	60,500	44,500	90,000
External Estabs	90,213	63,395	156,669
1-to-1 Match	67.1%	70.2%	57.4%

Match of External Establishments to Business Register

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
External Estabs	90,213	63,395	156,669
1-to-1 Match	60,500	44,500	90,000
External Estabs	90,213	63,395	156,669
1-to-1 Match	67.1%	70.2%	57.4%

Match of External Establishments to Business Register

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

External Estabs	90,213	63,395	156,669
1-to-1 Match	67.1%	70.2%	57.4%

Linking External Data to 2017 Economic Census

- Once linked to the BR, it is straightforward to link external establishments to 2017 EC.
- We use EC forms captured as of November 2018 – final paper will use fully processed EC.

Linking 1-to-1 Matches to 2017 EC

Franchises	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
1-to-1 Match with BR	60,500	44,500	90,000
Surveyed in 2017 EC	90.1%	91.0%	87.2%
2017 EC Form Processed	40.5%	41.6%	40.6%

Linking 1-to-1 Matches to 2017 EC

Franchises	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
1-to-1 Match with BR	60,500	44,500	90,000
Surveyed in 2017 EC	90.1%	91.0%	87.2%
2017 EC Form Processed	40.5%	41.6%	40.6%

Linking 1-to-1 Matches to 2017 EC

Franchises	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
1-to-1 Match with BR	60,500	44,500	90,000
Surveyed in 2017 EC	90.1%	91.0%	87.2%
2017 EC Form Processed	40.5%	41.6%	40.6%

Linking 1-to-1 Matches to 2017 EC

Franchises	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
1-to-1 Match with BR	60,500	44,500	90,000
Surveyed in 2017 EC	90.1%	91.0%	87.2%
2017 EC Form Processed	40.5%	41.6%	40.6%

Responses to 2017 EC Franchise Questions

- Franchise-affiliated
- Not franchise-affiliated

Responses to 2017 EC Franchise Questions

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	19,500	15,000	25,000
Not franchise-affiliated	5,200	3,500	11,000
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	79.6%	81.1%	68.5%
Not franchise-affiliated	21.2%	18.9%	30.1%

Responses to 2017 EC Franchise Questions

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	19,500	15,000	25,000
Not franchise-affiliated	5,200	3,500	11,000
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	79.6%	81.1%	68.5%
Not franchise-affiliated	21.2%	18.9%	30.1%

Responses to 2017 EC Franchise Questions

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	19,500	15,000	25,000
Not franchise-affiliated	5,200	3,500	11,000
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	79.6%	81.1%	68.5%
Not franchise-affiliated	21.2%	18.9%	30.1%

Responses to 2017 EC Franchise Questions

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	19,500	15,000	25,000
Not franchise-affiliated	5,200	3,500	11,000
2017 EC Form Processed	24,500	18,500	36,500
Franchise-affiliated	79.6%	81.1%	68.5%
Not franchise-affiliated	21.2%	18.9%	30.1%

Evaluating 2017 EC Responses

- Is MAMBA correct or is the survey form correct?
- Take random samples of discrepancies and manually examine whether MAMBA or survey form is correct
- Estimate the fraction of estabs. that can be recoded to franchise-affiliated

Evaluating 2017 EC Responses

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Not franchise-affiliated	5,200	3,500	11,000
MAMBA Prediction Correct (est.)	91.3%	93.0%	94.5%
Number Recoded (est.)	4,748	3,255	10,395

Evaluating 2017 EC Responses

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Not franchise-affiliated	5,200	3,500	11,000
MAMBA Prediction Correct (est.)	91.3%	93.0%	94.5%
Number Recoded (est.)	4,748	3,255	10,395

Evaluating 2017 EC Responses

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Not franchise-affiliated	5,200	3,500	11,000
MAMBA Prediction Correct (est.)	91.3%	93.0%	94.5%
Number Recoded (est.)	4,748	3,255	10,395

Evaluating 2017 EC Responses

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Not franchise-affiliated	5,200	3,500	11,000
MAMBA Prediction Correct (est.)	91.3%	93.0%	94.5%
Number Recoded (est.)	4,748	3,255	10,395

Implications

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Franchise-Affiliated (EC Form)	19,500	15,000	25,000
Recoded (est.)	4,748	3,255	10,395
Franchise-Affiliated (EC Form + Recoded)	24,248	18,255	35,395
Percent Increase	24%	22%	42%

Implications

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Franchise-Affiliated (EC Form)	19,500	15,000	25,000
Recoded (est.)	4,748	3,255	10,395
Franchise-Affiliated (EC Form + Recoded)	24,248	18,255	35,395
Percent Increase	24%	22%	42%

Implications

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Franchise-Affiliated (EC Form)	19,500	15,000	25,000
Recoded (est.)	4,748	3,255	10,395
Franchise-Affiliated (EC Form + Recoded)	24,248	18,255	35,395
Percent Increase	24%	22%	42%

Implications

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Franchise-Affiliated (EC Form)	19,500	15,000	25,000
Recoded (est.)	4,748	3,255	10,395
Franchise-Affiliated (EC Form + Recoded)	24,248	18,255	35,395
Percent Increase	24%	22%	42%

Implications

Franchises	Web-Scraped	Yelp (Core)	Yelp (Non-Core)
Franchise-Affiliated (EC Form)	19,500	15,000	25,000
Recoded (est.)	4,748	3,255	10,395
Franchise-Affiliated (EC Form + Recoded)	24,248	18,255	35,395
Percent Increase	24%	22%	42%

Moving Forward

- Alternative sources of external data
    - Search engine location services (e.g. Google, Bing)
    - Franchise Disclosure Documents (FDDs)
- Improving MAMBA through clerical review