ADEP WORKING PAPER SERIES

Industry Self-Classification in the Economic Census

Brian Dumbacher
U.S. Census Bureau

Daniel Whitehead
U.S. Census Bureau

Working Paper 2024-04
June 2024

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

Industry Self-Classification in the Economic Census

Brian Dumbacher, U.S. Census Bureau, brian.dumbacher@census.gov
Daniel Whitehead, U.S. Census Bureau, daniel.whitehead@census.gov

ADEP Working Paper 2024-04
June 2024

Abstract

This paper describes the methodology behind BEACON – a tool that will be used by respondents to the 2022 Economic Census to self-designate their establishment's North American Industry Classification System (NAICS) code. BEACON, which stands for Business Establishment Automated Classification of NAICS, takes a respondent-provided business description as input and returns to the respondent a list of candidate NAICS codes from which to choose. BEACON is based on text analysis, machine learning, and information retrieval. The rich training dataset for BEACON contains over 3.7 million observations from sources such as past Economic Census responses and Internal Revenue Service data. It is shown how BEACON employs ensemble and hierarchical modeling techniques to propose relevant NAICS codes. This paper also discusses results from a recent Economic Census field test.

Keywords:

Economic Census, hierarchical modeling, information retrieval, machine learning, NAICS, text classification

JEL Classification Codes:

C. Mathematical and Quantitative Methods

- C1. Econometric and Statistical Methods and Methodology: General
    - C18. Methodological Issues: General

Industry Self-Classification in the Economic Census

Brian Dumbacher[†], Daniel Whitehead[†]
[†]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
brian.dumbacher@census.gov, daniel.whitehead@census.gov

Abstract

This paper describes the methodology behind BEACON – a tool that will be used by respondents to the 2022 Economic Census to self-designate their establishment's North American Industry Classification System (NAICS) code. BEACON, which stands for Business Establishment Automated Classification of NAICS, takes a respondent-provided business description as input and returns to the respondent a list of candidate NAICS codes from which to choose. BEACON is based on text analysis, machine learning, and information retrieval. The rich training dataset for BEACON contains over 3.7 million observations from sources such as past Economic Census responses and Internal Revenue Service data. It is shown how BEACON employs ensemble and hierarchical modeling techniques to propose relevant NAICS codes. This paper also discusses results from a recent Economic Census field test.

Key Words: Economic Census, hierarchical modeling, information retrieval, machine learning, NAICS, text classification

1. Background

1.1 North American Industry Classification System (NAICS)

The North American Industry Classification System (NAICS) was developed in conjunction with Canada and Mexico to facilitate economic analysis and was implemented in 1997. A key use of NAICS is to provide a consistent and uniform way to present summary statistics about the U.S. economy. The U.S. Census Bureau and other statistical agencies use NAICS throughout the survey life cycle including sample selection, data collection, editing, publication, and analysis. Consequently, proper NAICS classification of business establishments is important for the accuracy of official economic statistics. An establishment is defined as a physical location where business is conducted; companies and enterprises are comprised of one or more establishments. Establishments are classified based on their principal business or activity. NAICS uses a six-digit hierarchical coding scheme to identify business activity at different levels of detail. The first two digits of the NAICS code represent the broad economic sector. Additional non-zero digits add industry detail. Table 1 breaks down the structure of an example NAICS code.

Table 1. Structure of an Example NAICS Code: 721191

NAICS Level of Detail	NAICS Code	NAICS Description
Sector	72	Accommodation and Food Services
Subsector	721	Accommodation
Industry group	7211	Traveler Accommodation
NAICS industry	72119	Other Traveler Accommodation
National industry	721191	Bed-and-Breakfast Inns

Source: https://www.census.gov/naics/

NAICS code assignments are revised periodically. Revisions to the NAICS structure itself occur every five years and coincide with the Economic Census. The 2017 and 2022 vintages of NAICS identify 1,057 and 1,012 codes, respectively, at the 6-digit level (U.S. Census Bureau, 2017 and 2022). For additional information about NAICS, see the Census Bureau NAICS webpage at https://www.census.gov/naics/.

1.2 Economic Census

Every five years, for years ending in "2" and "7", the Census Bureau conducts the Economic Census, an extensive survey of approximately eight million establishments with paid employees that covers most industries[1] and all geographic areas of the United States, including U.S. territories. The Economic Census asks about half of the eight million establishments to complete questionnaires whereas the other half is accounted for through administrative records or imputation. The Economic Census provides a wealth of information to help policymakers, trade and business associations, individual businesses, and other federal agencies understand U.S. economic activity at a granular level. Key statistics include total number of establishments; total number of employees; value of sales, shipments, receipts, and revenue; and total annual payroll. Data products regarding establishments are broken down by industry, as classified by NAICS. For details about the design and methodology of the Economic Census, see https://www.census.gov/programs-surveys/economic-census/technical-documentation.html. The 2017 Economic Census was fully electronic, and respondents received an online questionnaire based on the most recent estimate of the establishment's NAICS code.

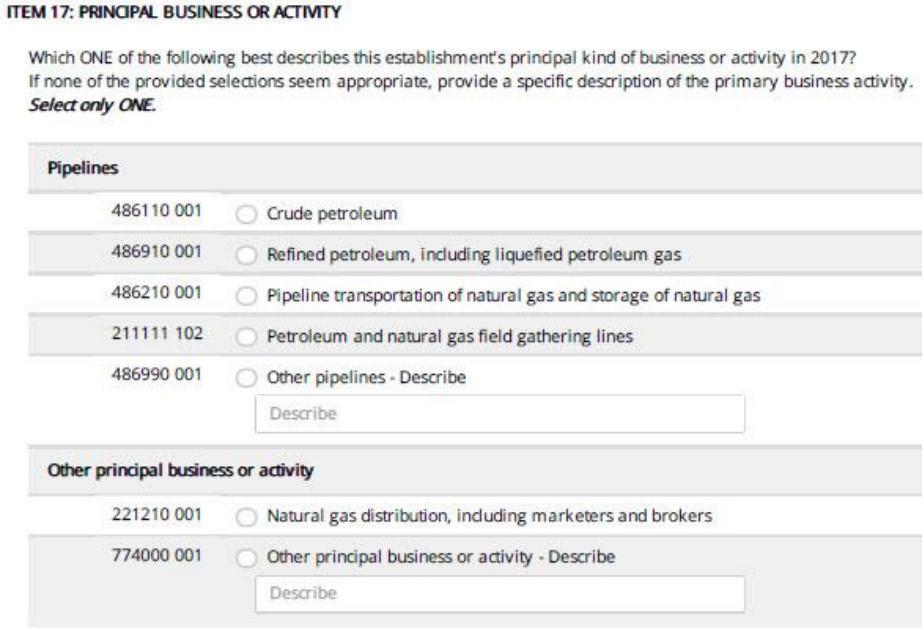
1.3 Business Description Write-Ins

One question in the Economic Census, the Principal Business or Activity (PBA) question[2], asks respondents to describe their business. Answers to the PBA question help keep NAICS code assignments up to date. This question displays a list of descriptions related to the current classification, and the respondent is asked to select one. The respondent also

[1] NAICS 92 (Public Administration) and most of NAICS 11 (Agriculture, Forestry, Fishing and Hunting) along with a few industries in other sectors are out of the scope of the Economic Census. For more information on industry coverage, see the "Target Population" section of the Economic Census methodology.

[2] Before 2017, this question was known as the self-designated kind of business question.

has the option of typing in a description if the listed descriptions do not seem accurate. To illustrate, Figure 1 is a screenshot of the PBA question from a 2017 questionnaire.



ITEM 17: PRINCIPAL BUSINESS OR ACTIVITY

Which ONE of the following best describes this establishment's principal kind of business or activity in 2017?
If none of the provided selections seem appropriate, provide a specific description of the primary business activity.
Select only ONE.

Pipelines		
	486110 001	Crude petroleum
	486910 001	Refined petroleum, including liquefied petroleum gas
	486210 001	Pipeline transportation of natural gas and storage of natural gas
	211111 102	Petroleum and natural gas field gathering lines
	486990 001	Other pipelines - Describe
		Describe
Other principal business or activity		
	221210 001	Natural gas distribution, including marketers and brokers
	774000 001	Other principal business or activity - Describe
		Describe

Figure 1. Principal Business or Activity (PBA) question from the 2017 Economic Census pipelines questionnaire (TW-48600). Example write-ins include "pipeline terminal" and "field office". Source: https://bhs.econ.census.gov/ombpdfs/export/TW-48600_su.pdf

Over the years, there have been hundreds of thousands of these so-called "write-in" responses to the PBA question. During the 2017 Economic Census alone, there were over 400,000 write-ins. For the most part, clerks and analysts assign NAICS codes to these cases manually. According to Snijkers *et al*. (2013, p. 478), manual coding is expensive, time-consuming, and subjective. Using more automated methods can help address these disadvantages. For example, Kornbau (2016) and Kearney and Kornbau (2005) describe a successful NAICS autocoder for classifying new businesses.

In addition to reducing clerical work associated with write-ins, autocoding can also make the Economic Census questionnaire more dynamic. Along these lines, the Census Bureau has developed a tool called BEACON (Business Establishment Automated Classification of NAICS) for use in the upcoming 2022 Economic Census. For respondents who provide a write-in, BEACON will help them self-designate their 6-digit NAICS code in real time. BEACON takes the write-in as input, applies a text classification model, and returns a ranked list of candidate 6-digit NAICS codes for the respondent to choose from. The subsequent questions on the questionnaire depend on the respondent's NAICS code, so using BEACON represents an important step in the survey.

1.4 Outline of Paper

The rest of the paper is organized as follows. Section 2 provides an overview of BEACON. Methodological aspects such as the training data, text cleaning algorithm, model ensemble,

and hierarchical model structure are detailed in Section 3. Section 4 presents results from an Economic Census field test used to assess BEACON with respondents. Lastly, Section 5 lays out ideas for future work.

2. BEACON Overview

2.1 Key Characteristics

BEACON is an example of a ranked text classifier. Instead of predicting a single NAICS code, BEACON returns multiple codes for the respondent to choose from, ranked from most to least relevant. Text classification is interdisciplinary and lies at the intersection of machine learning (Aggarwal, 2018), natural language processing (Jurafsky and Martin, 2009), and information retrieval (Goswami, 2014). As such, BEACON's text classification methodology has the following key characteristics:

Rich training data – The training data come from various sources that complement one another well: historical write-ins to the Economic Census, business descriptions from the Internal Revenue Service, business descriptions from an internal Census Bureau system used by analysts in their industry classification work, and publicly available commodity descriptions.

Extensive text cleaning algorithm – BEACON cleans business descriptions to prepare the text for modeling. This process involves addressing punctuation and spacing issues, removing common words, stemming words (removing suffixes to reduce the number of word variations), and correcting common misspellings. A hybrid approach is employed that is based on contextual rules, a modified version of a well-known stemming algorithm, and various word lookup lists.

Representation of text based on word combinations – Business descriptions are represented by the occurrence of individual words, 2-word combinations, and 3-word combinations. To emphasize, these are combinations of words and not sequences of words (commonly known as n-grams). Unlike n-grams, word combinations do not place any restrictions on the order of words or distances between them in the text. Write-ins from the Economic Census are examples of short text documents that contain very few words. Consequently, using word combinations is both appropriate and feasible in this setting.

Optimized model ensemble – BEACON is composed of multiple sub-models that use the representation of text in different ways. Each sub-model can produce a ranked list of NAICS codes. The ensemble methodology involves calculating a weighted average of output from the sub-models, where the weights are optimized using machine learning.

Hierarchical model structure – The model structure reflects the hierarchical structure of NAICS and borrows strength across industries within the same

economic sector. This helps boost the sample size for underrepresented industries in the training data.

2.2 BEACON as a Search Engine

It is helpful to view BEACON as a search tool like an internet search engine. Given a query, a search engine uses a "score function" to assign relevance scores to the websites in its database. The score is typically higher if there is greater agreement between the text in the query and the text on the website[3]. The search engine ranks websites according to the relevance score and then presents the highest-scoring websites to the user.

There are many similarities between this process and how BEACON operates. Table 2 lists the internet search engine concepts described above and their BEACON analogues. Given a respondent-provided business description, BEACON assigns a relevance score to each 6-digit NAICS code. BEACON's score function is based on how words and word combinations are distributed across NAICS codes in the training data. The more highly associated the words in the business description are with the NAICS code, the higher the score. Finally, BEACON ranks the NAICS codes according to relevance score and presents the highest-scoring NAICS codes to the respondent.

Table 2. Search Engine Concepts and BEACON Analogues

Internet Search Engine Concept	Analogous BEACON Concept
Search query	Business description
Websites	NAICS codes
Past searches	Historical write-ins
User	Respondent

3. Methodology

3.1 Training Data

The training data for BEACON come from five sources with different characteristics. These data sources complement one another well and combine to form a rich dataset for modeling. Table 3 summarizes the data sources' advantages and disadvantages.

EC – Write-in business descriptions from the 2002, 2007, 2012, and 2017 Economic Census provided by single-unit (one physical location) establishments. This data source is the most representative of the target population.

IRS SS-4 – Write-in business descriptions from the Internal Revenue Service's SS-4 form, which is used by businesses to apply for an Employer Identification Number. The relevant question on the SS-4 form asks for the "principal line of

[3] Internet search engines also consider website linkages to measure importance and search trends to measure recency and popularity, but these concepts do not carry over to BEACON.

merchandise sold, specific construction work done, products produced, or services provided." As with the EC data source, these descriptions were provided by single-unit establishments. This data source covers 2002-2016.

CAPS – The Classification Analytical Processing System (CAPS) is used by Census Bureau analysts in their industry classification work. CAPS contains many business descriptions for every NAICS code, including the index descriptions found in the NAICS manual (U.S. Census Bureau, 2017 and 2022). This data source provides a rich vocabulary for modeling. Variations and duplicates of original observations are added. BEACON's methodology does not use sampling weights, so the duplicates give the original observations more weight in the training data.

EC Autocoded – During the 2017 Economic Census, an exact-match autocoder was used to assign NAICS codes to frequently occurring write-in text. This data source consists of frequently occurring business descriptions and the corresponding NAICS codes that were assigned automatically. Many of these descriptions can be found in the EC data source, but including this source further helps associate them with the correct NAICS code. As with the CAPS data source, this data source includes duplicates of original observations.

HS – The Harmonized System is an internationally standardized system of commodity descriptions and codes used to classify traded products. Lists of commodity descriptions are publicly available[4] that also contain corresponding NAICS codes.

Table 3. Summary of Data Source Advantages and Disadvantages

Data source	Advantages	Disadvantages
EC	• Represents target population • Reflects natural language	• Descriptions not classified perfectly • Descriptions contain misspellings
IRS SS-4	• Provides timely data • Reflects natural language	• Descriptions not classified perfectly • Descriptions contain misspellings
CAPS	• Provides a rich vocabulary • Descriptions are classified correctly	• Text does not always reflect natural language
EC Autocoded	• Improves consistency with the 2017 EC autocoding effort • Descriptions are classified correctly	• Relatively small data source
HS	• Increases sample sizes for sectors not represented well in other data sources	• Relatively small data source

[4] See the import and export concordance files at https://www.census.gov/foreign-trade/reference/codes/index.html

	• Descriptions are classified correctly	• Covers only three sectors (11, 21, 31-33)

Source: 2002, 2007, 2012, 2017 Economic Census; Internal Revenue Service; and Classification Analytical Processing System

Figure 2 breaks down the training data by economic sector and data source. The contributions of the data sources vary by sector. For example, for sector 31-33 (Manufacturing), most observations come from the CAPS data source. For most sectors there is a good mix of observations from multiple data sources. In total, the training data contain over 3.7 million observations.

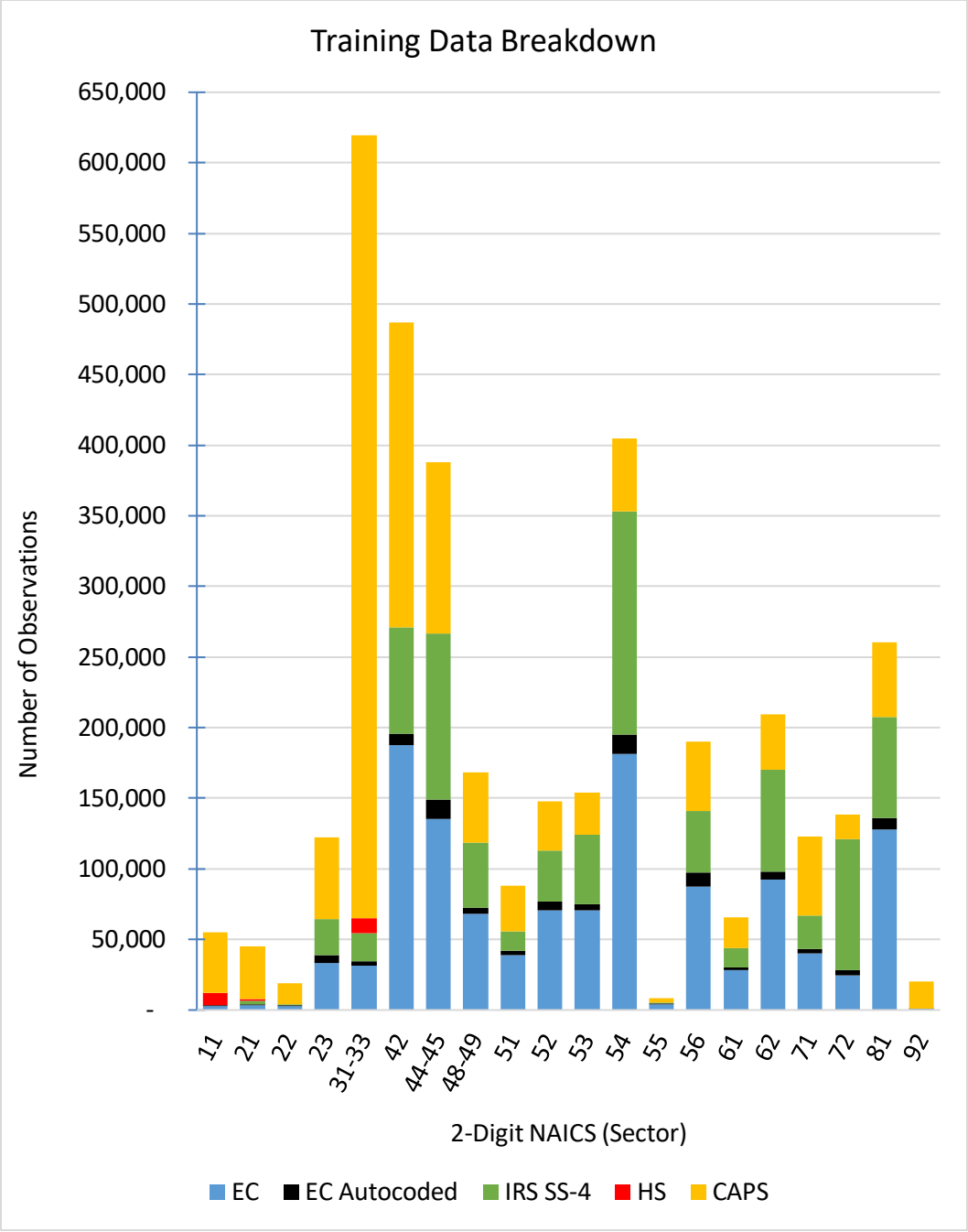


Figure 2. Breakdown of the training data by economic sector and data source. Source: 2002, 2007, 2012, 2017 Economic Census; Internal Revenue Service; and Classification Analytical Processing System

3.2 Text Cleaning

BEACON applies an extensive text cleaning algorithm to prepare business descriptions for modeling. The algorithm is implemented using rules called regular expressions and various word lookup lists. The steps include the following:

- Convert text to lowercase
- Address numbers in various ways depending on the context (e.g., remove or convert to a word)
- Standardize common compound words (e.g., "fast-food" and "barbershop")
- Remove filler phrases (e.g., "all of the above" and "types of")
- Apply basic negation handling rules (to handle phrases such as "not a laundromat")
- Remove common "stop" words such as "the" and "or" that are not expected to be predictive of NAICS
- Apply a modified version of the Porter 2/Snowball stemming algorithm (Porter, 2001) to strip suffixes and reduce the number of word variations (e.g., stem the words "manufacturer" and "manufacturing" to "manufactur")
- Associate synonyms with the same stem to further reduce the number of word variations (e.g., "automotive" with "car" and "lady" with "woman")
- Correct common misspellings (e.g., map the stem "manifactur" to "manufactur")

The output of the text cleaning process is a string of words (stems) separated by spaces. Table 4 lists illustrative examples. Misspelled words and typographical errors in the raw business descriptions are intentional.

Table 4. Illustrative Examples of the Text Cleaning Process

Business Description	Clean Business Description
Automotive rapair & car-wash.	car repair carwash
This is a convenence store.	conveni store
motel with 30 rooms	motel room
seller of 2nd hand ussed goods	sell secondhand used good
ammunition (used for sport)	ammo sport
wholesaler ladies' clothi ng	wholesal woman cloth
mfg of frzn cakes	manufactur frozen cake
We do fin. and ins.	financ insur
This co. does constructi on	compani construct
Consulting svcs	consult service
mini warehouse rntl & storing	miniwarehous rent storag

3.3 Dictionary

Underlying BEACON is a dictionary of words, word combinations, and full-length/exact business descriptions that occur frequently in the cleaned training data and are recognized by the model. In machine learning terminology, these pieces of text are the model features. Table 5 presents some dictionary metrics.

Table 5. BEACON Dictionary Metrics

Metric	Value
Number of words	10,668
Number of 2-word combinations	141,649

Number of 3-word combinations	247,546
Number of full-length/exact business descriptions	55,517
Total	455,380

The dictionary stores the features' distributions across NAICS codes in the training data. Associated with each feature is a "purity weight" that measures how concentrated, or pure, its distribution is. The purity weight reflects the feature's predictive ability. Values range from 0 (evenly distributed across NAICS codes) to 1 (occurring in only one NAICS code). As an example, Figure 3 displays the 2-digit (sector) distribution of the word "bakeri" (the stemming algorithm converts the "y" to "i"). "Bakeri" is associated with sectors 31-33 (Manufacturing), 42 (Wholesale Trade), 44-45 (Retail Trade), and 72 (Accommodation and Food Services). It has a moderately large purity weight of 0.4509.

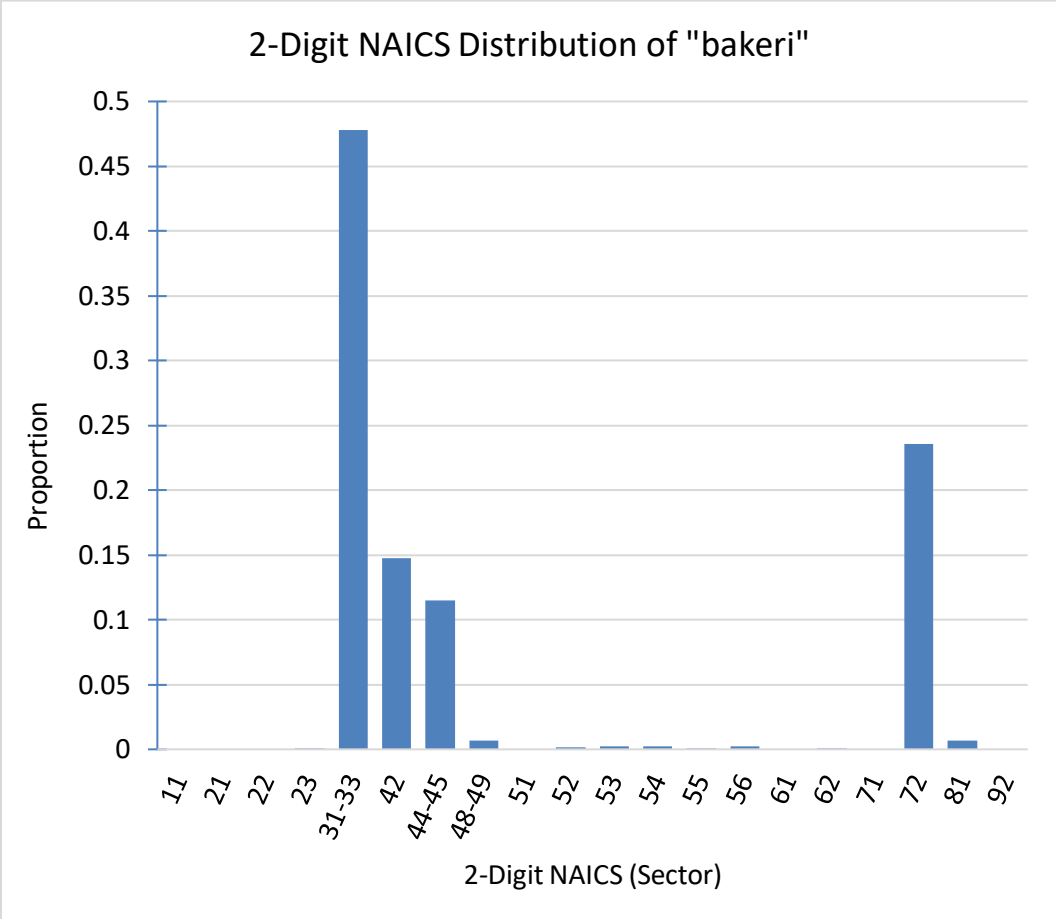


Figure 3. 2-digit NAICS distribution of "bakeri". Purity weight = 0.4509. Source: 2002, 2007, 2012, 2017 Economic Census; Internal Revenue Service; and Classification Analytical Processing System

3.4 Model Ensemble

Given the large number of model features and 6-digit NAICS codes, there exist many relationships in the training data that a model must recognize in order to be effective. To

this end, BEACON employs an ensemble methodology to take advantage of different approaches. BEACON's model ensemble is composed of three information retrieval sub-models, which are referred to as "all", "umbrella", and "exact". The sub-models use different features, but each can assign relevance scores to NAICS codes.

All – This sub-model considers all words and combinations in the cleaned business description.

Umbrella – Similar to the "all" sub-model, the "umbrella" sub-model considers all words and combinations in the cleaned business description except for those that are subsets of other combinations. Keeping just the "umbrella" words and combinations allows this sub-model to focus on the details of the text.

Exact – This sub-model considers observations in the training data that consist of, and only of, the exact same words as the business description. Arguably, these observations are most like the respondent-provided business description.

The sub-models consider how the relevant features are distributed across NAICS codes. These distributions are pre-computed; BEACON just looks them up in its dictionary. For the "all" and "umbrella" sub-models, which consider the NAICS distributions of possibly multiple features, the distributions are averaged using the purity weights. This gives more influence to the NAICS distributions of features that are more predictive.

The final relevance score for a particular NAICS code is a weighted average of the scores from the three sub-models, where the "ensemble weights" have been optimized using the holdout method (Tan *et al.,* 2019). The holdout method is a training/testing paradigm commonly used in machine learning. A large fraction of the training data was randomly selected and used to fit model ensembles for various combinations of ensemble weights (multiples of 0.1 constrained to sum to 1). The different models were then applied to the held-out fraction of the data – the test set. The combination of weights yielding the best results was used. This process determined the optimal ensemble weights to be 0.1 for "all", 0.6 for "umbrella", and 0.3 for "exact".

The following example outlines how the model ensemble works at the 2-digit level. Suppose a respondent provides the description "This is a retail bakery." This description gets cleaned to "retail bakeri", where the features "retail", "bakeri", {"retail", "bakeri"}, and exact{"retail", "bakeri"} are all in BEACON's dictionary. The following are the steps for applying the three sub-models and model ensemble:

- All
    - Look up NAICS distribution of "retail"
    - Look up NAICS distribution of "bakeri"
    - Look up NAICS distribution of {"retail", "bakeri"}

    - To determine relevance scores, calculate a weighted average of the NAICS distributions using the features' purity weights
- Umbrella
    - The words "retail" and "bakeri" are subsets of {"retail", "bakeri"}, so they are excluded from this sub-model
    - To determine relevance scores, look up the NAICS distribution of {"retail", "bakeri"} [no need to calculate a weighted average of NAICS distributions in this case because there is only one distribution]
- Exact
    - To determine relevance scores, look up the NAICS distribution of exact{"retail", "bakeri"}
- Ensemble
    - To determine the final relevance scores, calculate a weighted average of the relevance scores from the "all", "umbrella", and "exact" sub-models using the ensemble weights

3.5 Model Hierarchy

Assigning relevance scores directly at the 6-digit level is challenging. The approach used by BEACON takes advantage of the hierarchical structure of NAICS and helps boost the sample size for underrepresented industries in the training data. BEACON uses the model ensemble to first assign scores at the 2-digit (sector) level. It is verified that these scores sum to 1. For each sector SS, BEACON uses the model ensemble to assign sector-conditional scores to the constituent 6-digit NAICS codes. The conditional score for NAICS code SS####[5] can be interpreted as this industry's relevance score, given or assuming that the true sector is SS. It is verified that these conditional scores for sector SS also sum to 1. To calculate the unconditional 6-digit scores, BEACON combines the 2-digit score and 6-digit sector-conditional scores using a formula that resembles the conditional probability formula:

$$score(SS\#\#\#\#) = score(SS) \times score(SS\#\#\#\#|SS)$$

This step essentially allocates the 2-digit score among the constituent 6-digit NAICS codes. In summary, the model ensemble is used 21 times in the hierarchy – first to assign scores at the 2-digit level, and 20 more times, once for each sector, to assign sector-conditional scores at the 6-digit level.

4. Economic Census Field Test

[5] The notation SS and SS#### is convenient for representing the NAICS hierarchy, but it might suggest that 6-digit NAICS codes with different leading digits SS always belong to different sectors. The 2-digit codes 31, 32, and 33 collectively represent the Manufacturing sector. Similarly, 44 and 45 represent the Retail Trade sector, and 48 and 49 represent the Transportation and Warehousing sector.

4.1 Overview

From October 2021 – February 2022, the Census Bureau conducted the Economic Census 2021 Industry Classification Report. This survey is usually sent out the year before the Economic Census to obtain classification for cases without a reliable NAICS code. It was repurposed as a field test to assess new questionnaire features with live respondents in a production environment. This test allowed an evaluation of BEACON's ease of use and ability to provide relevant NAICS codes to respondents. In this section, we discuss the performance of BEACON during the field test.

There were two inputs to BEACON: the respondent's description of the establishment's principal business or activity and an optional sector selection from a drop-down menu (for a list of all 20 sectors, see Appendix A). BEACON accounted for the respondent's selection by ranking results from the selected sector above those of other sectors. Given that respondents tend to select from the top of the NAICS results, the ability of BEACON to assist the respondent in choosing the correct NAICS code was intertwined with the respondent's choice of sector. Because data collection for the 2022 Economic Census will still be in terms of the 2017 NAICS vintage, BEACON returned 2017 NAICS codes to the field test respondents. For each candidate NAICS code, a short industry description and a link to the industry's webpage on https://www.census.gov/naics/ were also provided.

4.2 Word Recognition

In total, BEACON received more than 20,000 descriptions, over 15,000 of which contained text. The remainder mostly contained NAICS codes, which respondents can enter if known[6]. We applied BEACON's text cleaning algorithm to the descriptions containing text and observed that the cleaned descriptions contained more than 40,000 word instances. BEACON recognized over 99% of these instances and thus was able to return results for a large majority of descriptions.

4.3 Sector Selection

About 93.9% of respondents selected a sector from the drop-down menu before using BEACON. Of these cases, there were 7,797 "truth deck" cases. Truth deck cases are those for which the existing NAICS code from past collections was considered reliable for evaluation purposes. Using these cases, we examined how often the correct sector was selected. This analysis is summarized by the confusion matrix in Figure 4. The rows correspond to the correct sector, and the columns correspond to the sector selected from the drop-down menu. For example, many respondents incorrectly selected sector 81 [Other Services (except Public Administration)]. This can be seen most clearly when examining the truth deck cases for which the correct sector was sector 56 (Administrative and Support and Waste Management and Remediation Services). Of the 484 truth deck cases belonging to sector 56, 272 selected sector 81. The overall sector selection accuracy was 62.4%.

[6] BEACON recognizes NAICS codes in the business description and returns appropriate results.

Correct Sector	Sector Selected from Drop-Down Menu																			
	11	21	22	23	31-33	42	44-45	48-49	51	52	53	54	55	56	61	62	71	72	81	92
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	1	159	1	21	6	0	1	8	2	0	1	10	3	0	0	0	0	0	72	0
22	12	2	290	9	1	0	1	1	0	0	1	5	1	0	0	0	0	0	27	1
23	1	0	3	444	9	1	10	0	0	0	1	5	0	0	0	0	0	0	35	0
31-33	6	1	0	26	477	8	10	2	0	0	0	5	0	1	0	4	0	2	19	0
42	30	2	1	15	70	256	44	12	0	0	0	10	2	1	0	3	1	1	57	1
44-45	8	0	7	20	10	16	350	6	0	0	0	4	0	0	1	45	4	6	43	1
48-49	3	0	0	5	14	2	11	145	1	0	0	17	0	3	2	0	17	3	133	0
51	1	0	12	1	8	2	3	0	95	3	1	53	1	2	8	5	53	0	136	0
52	1	0	0	0	0	0	0	0	0	470	0	4	1	1	0	2	0	0	23	0
53	2	0	0	10	0	2	15	12	0	2	335	1	20	1	1	3	15	6	59	0
54	9	1	2	11	5	1	3	1	14	25	3	267	4	2	2	36	1	0	152	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
56	3	1	8	24	2	1	10	12	5	6	2	73	15	33	0	3	7	3	272	4
61	1	0	0	0	0	0	4	0	1	0	0	14	1	0	163	3	226	0	49	1
62	1	0	0	1	0	0	0	0	0	0	1	21	0	0	3	474	1	2	47	1
71	6	0	0	1	3	0	12	1	1	0	2	2	0	0	11	11	298	11	87	0
72	0	0	0	1	0	0	5	0	0	0	16	0	2	0	1	0	7	277	35	2
81	3	0	0	5	8	4	49	18	0	1	10	35	1	0	6	14	10	7	332	5
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4. Confusion matrix of 7,797 truth deck cases with a sector selection. Source: Economic Census 2021 Industry Classification Report

Subject matter experts concluded that the sector name "Other Services (except Public Administration)" was confusing. If a respondent does not know what sector to select, he/she may be inclined to select the sector beginning with the word "Other". As a result, it was decided to combine sectors 51, 54, 56, 61, 62, 71, and 81 into an "Other Services" category, which will be used in the sector drop-down menu for the 2022 Economic Census. It is believed this new category will reduce misclassification within the services trade.

4.4 Overall Accuracy and Ease of Use

A correct NAICS self-classification depends on (1) BEACON returning the correct NAICS code as a candidate and (2) the respondent understanding the industry descriptions well enough to select the correct code. Of the 7,797 truth deck cases, there were 7,005 cases where the respondent made a valid NAICS selection (i.e., the selected NAICS code was neither blank nor "none of these"). For these cases, BEACON returned the correct NAICS code 90.1% of the time, which is consistent with past evaluations using just the training data. Given that the correct NAICS code was returned, respondents selected it 83.7% of the time, resulting in an overall self-classification accuracy of 75.5%. Lastly, based on web probe questions at the end of the field test questionnaire, 65.3% of respondents found BEACON very easy or somewhat easy to use.

5. Future Work

There are many methodological components to BEACON and, consequently, many directions in which to extend the research. Regarding the text cleaning algorithm, refinement is an ongoing process. BEACON already recognizes a large percentage of word instances, but we continue to correct additional misspellings, add words to the dictionary via synthetic training data observations, and research new rules in the form of regular expressions.

We are also on the lookout for additional data sources, both internal and publicly available. Before the 2022 Economic Census begins, we will incorporate the descriptions from the field test into BEACON's training data. As described in Section 4.4, there are only 7,005 descriptions from the field test with a reliable NAICS code. However, they are representative of the target population and are, therefore, important to include. The next big boost to BEACON's training data will come after the hundreds of thousands of write-ins expected from the 2022 Economic Census are processed and incorporated. This data source will provide timely observations and possibly many new words and combinations from which BEACON can learn.

The current score functions have proven successful in picking up on textual details and returning relevant NAICS codes. We may investigate a more advanced model based on word embeddings, for example Google's word2vec (Tomas *et al.*, 2013a; Tomas *et al.*, 2013b). An ensemble framework for incorporating additional sub-models already exists, so a word embedding sub-model could join the "all", "umbrella", and "exact" sub-models. New optimal ensemble weights would be calculated using the holdout method. Model stacking is another ensemble modeling technique we may investigate (Güneş, 2017). More sophisticated than simply averaging scores from multiple sub-models, stacking involves developing a second-stage, or meta, model such as logistic regression that uses the sub-models' scores as input.

Finally, future work also involves improving the Economic Census questionnaire to make it easier for respondents to use BEACON. One idea is to implement a text autocomplete feature. This feature would recommend commonly provided business descriptions to the respondent as he/she types in the text field. Autocomplete would help standardize descriptions further and reduce the number of misspellings and extraneous words.

Acknowledgments

The authors would like to thank Justin Nguyen, Amy Newman-Smith, Matthew Thompson, William Davie Jr., and Theresa Riddle of the U.S. Census Bureau for reviewing drafts of this paper and providing helpful comments.

References

Aggarwal, C.C. (2018). *Machine Learning for Text*. Cham, Switzerland: Springer International Publishing.

Goswami, P. (2014). *Learning Information Retrieval Functions and Parameters on Labeled Collections*. Ph.D. Dissertation. Université Joseph Fourier.

Güneş, F. (2017). Why do stacked ensemble models win data science competitions? *The SAS Data Science Blog*. May 18, 2017. <https://blogs.sas.com/content/subconsciousmusings/2017/05/18/stacked-ensemble-models-win-data-science-competitions/>. Accessed April 21, 2022.

Jurafsky, D. and Martin, J.H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second Edition). Upper Saddle River, NJ: Pearson Education, Inc.

Kearney, A.T. and Kornbau, M.E. (2005). An Automated Industry Coding Application for New U.S. Business Establishments. *2005 Proceedings of the American Statistical Association, Business and Economic Statistics Section*. Alexandria, VA: American Statistical Association, 867–874.

Kornbau, M.E. (2016). Automating Processes for the U.S. Census Business Register. *25th Meeting of the Wiesbaden Group on Business Registers*.

Porter, M.F. (2001). Snowball: a language for stemming algorithms. <http://snowball.tartarus.org/texts/introduction.html>. Accessed April 18, 2022.

Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D.K. (2013). *Designing and Conducting Business Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. (2019). *Introduction to data mining (2nd edition)*. New York: Pearson Education, Inc.

Tomas, M., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv*:1301.3781.

Tomas, M., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv*:1310.4546.

U.S. Census Bureau. (2017). 2017 North American Industry Classification System Manual. <https://www.census.gov/naics/reference_files_tools/2017_NAICS_Manual.pdf>. Accessed May 2, 2022.

U.S. Census Bureau. (2022). 2022 North American Industry Classification System Manual. <https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf>. Accessed May 2, 2022.

Appendix A: List of Economic Sectors

The first two digits of the NAICS code represent the economic sector. Table A-1 lists all 20 sectors and their corresponding 2-digit NAICS codes.

Table A-1. Economic Sectors as Defined by NAICS

2-Digit NAICS	Sector
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

Source: U.S. Census Bureau (2017 and 2022)