Industry Self-Classification in the Economic Census

Joint Statistical Meetings

Washington, DC

August 10, 2022

Brian Dumbacher and Daniel Whitehead

U.S. Census Bureau

United States® Census Bureau

North American Industry Classification System (NAICS)

- Establishments are physical locations where business is conducted
- U.S. Census Bureau classifies establishments by NAICS industry based on primary business or activity
- NAICS is utilized throughout the survey life cycle
  - Sample selection
  - Data collection
  - Publication
- Hierarchical 6-digit coding structure
  - First two digits represent the economic sector (52 – Finance and Insurance)
  - Additional non-zero digits add industry detail (522291 – Consumer Lending)

United States® Census Bureau

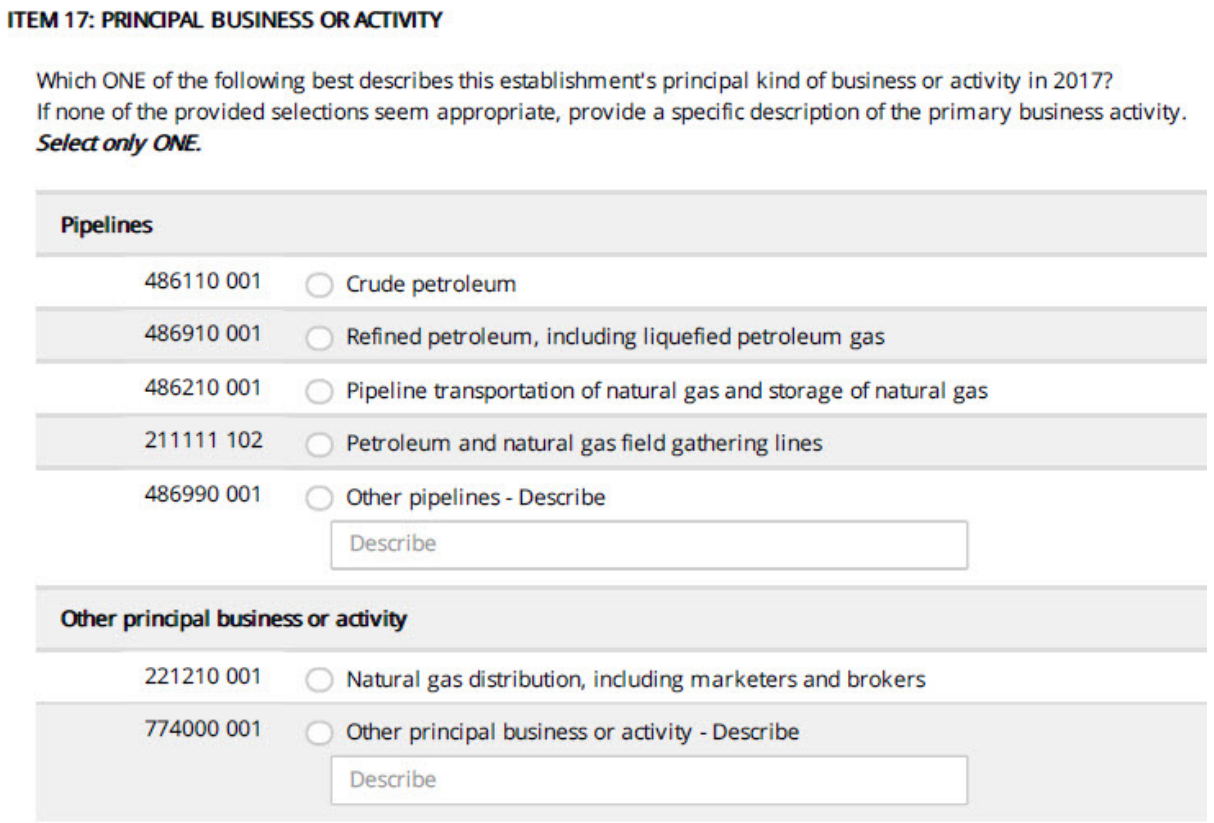
Economic Census (EC)

- Conducted every five years for years ending in “2” or “7”
- Represents approximately eight million establishments, covering most industries and all geographic areas of the U.S.
- Key statistics
  - Total number of establishments
  - Total number of employees
  - Value of sales, shipments, receipts, and revenue
  - Total annual payroll
- Data products are presented by NAICS and geography

United States® Census Bureau

Principal Business or Activity Question from the 2017 EC

- Question asks respondents to describe their business
- There are prelisted descriptions corresponding to an estimated NAICS code, but the respondent can also provide a description
- Clerical analysis of write-in text is resource-intensive



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

Source: 2017 Economic Census

United States® Census Bureau

BEACON Overview

- Business Establishment Automated Classification of NAICS
- General idea
  - The respondent inputs a business description
  - BEACON returns a ranked list of 6-digit NAICS codes with industry descriptions
- Goals
  - Help respondents self-designate their NAICS code
  - Send respondents down correct EC questionnaire path
  - Reduce clerical work associated with write-ins
- Methodology is based on machine learning, natural language processing, and information retrieval (e.g., internet search)

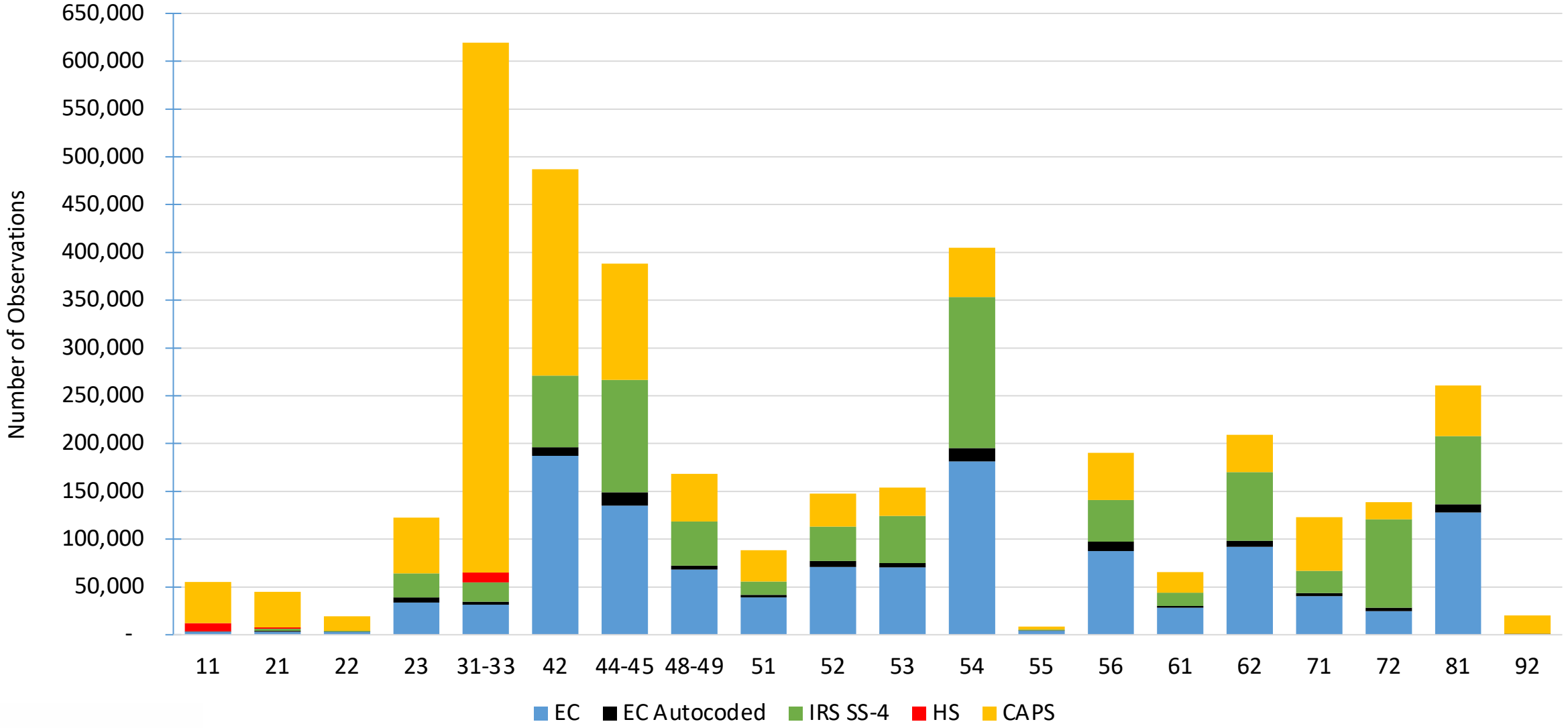
United States® Census Bureau

Training Data

- Historical write-in responses to the EC
- Frequent write-in text that was autocoded during 2017 EC
- Business descriptions from Internal Revenue Service (IRS) SS-4 forms
- Classification Analytical Processing System (CAPS) items
- Harmonized System (HS) commodity descriptions
- Variables
  - Business description text
  - Corresponding NAICS code

United States® Census Bureau

Training Data Breakdown by Sector and Source



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

Text Cleaning

- Convert to lowercase and account for numbers and punctuation
- Remove common “stop” words
- Stem words to reduce the number of word variations
- Correct common misspellings

Input Text	Clean Text
This is a convenence store.	conveni store
automobile mfg	car manufactur
We rapair watches & jewelry.	repair watch jewelri

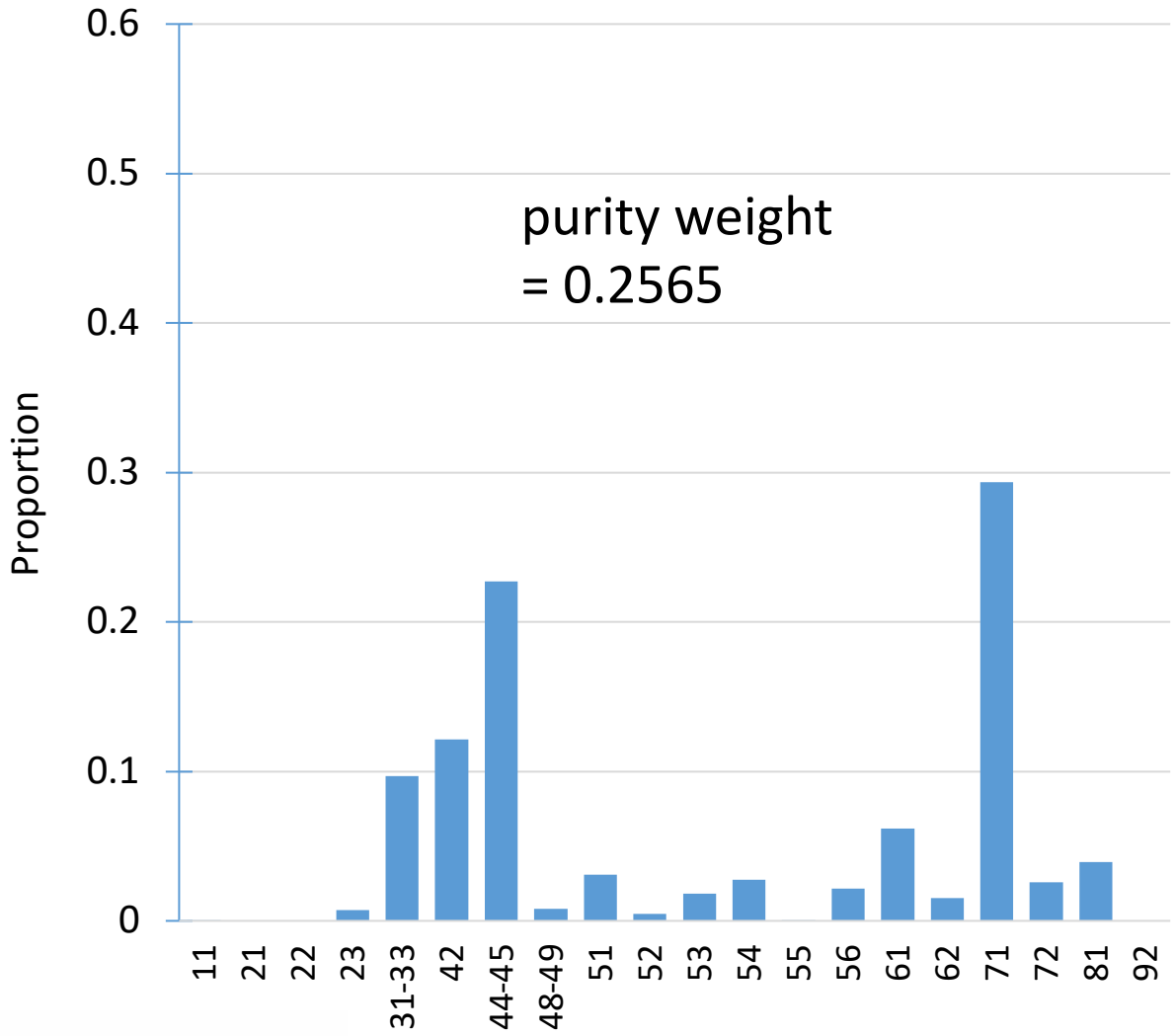
United States® Census Bureau

Dictionary

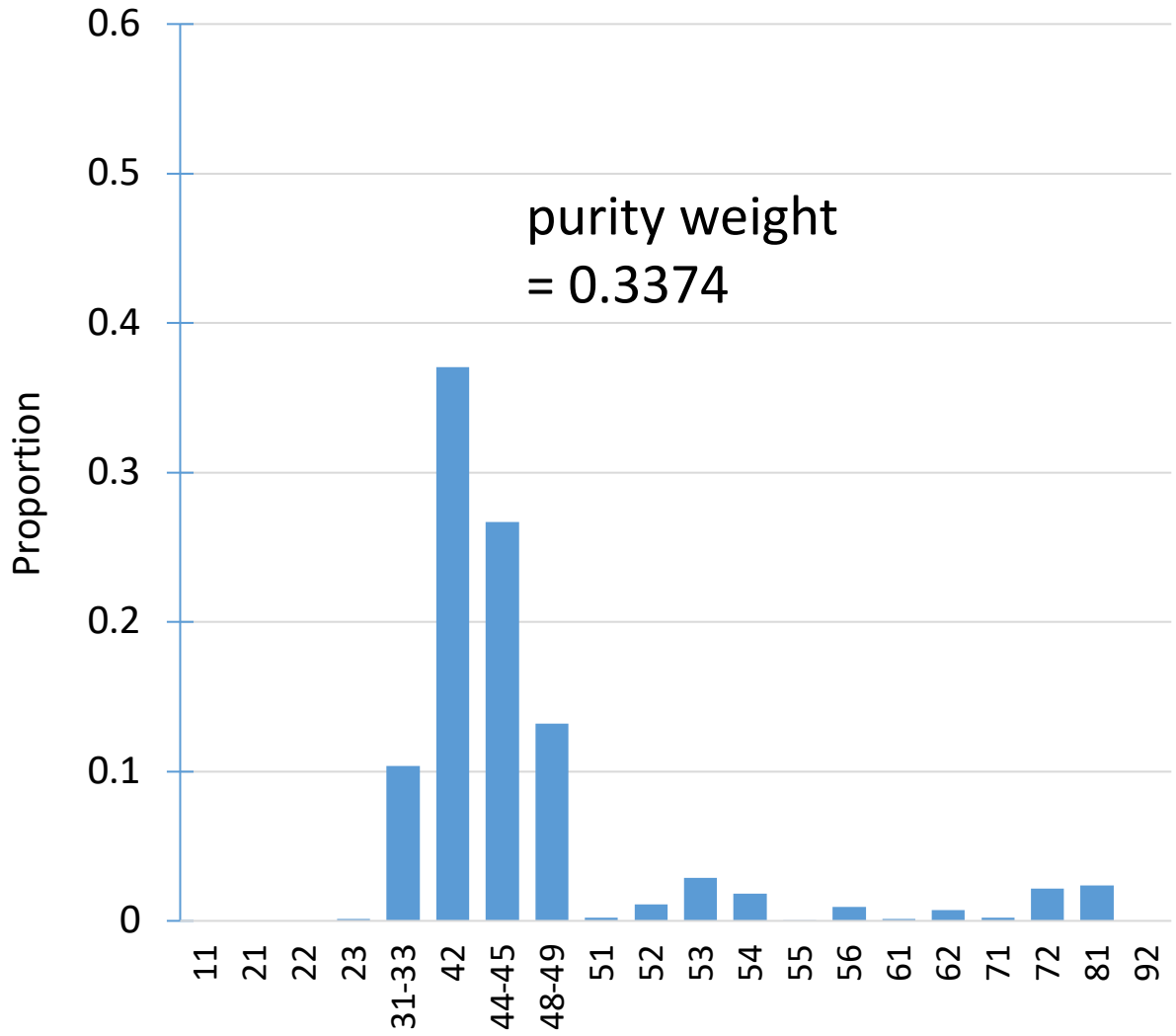
- Underlying BEACON is a dictionary of text that occurs frequently in the cleaned training data
  - Words
  - Word combinations
  - Full-length/exact descriptions
- Dictionary size is currently 455,380
- These pieces of text are the model features
- Dictionary stores features' NAICS distributions and associated purity weights that measure how concentrated, or pure, the distribution is

United States® Census Bureau

Sector Distribution of “sport”

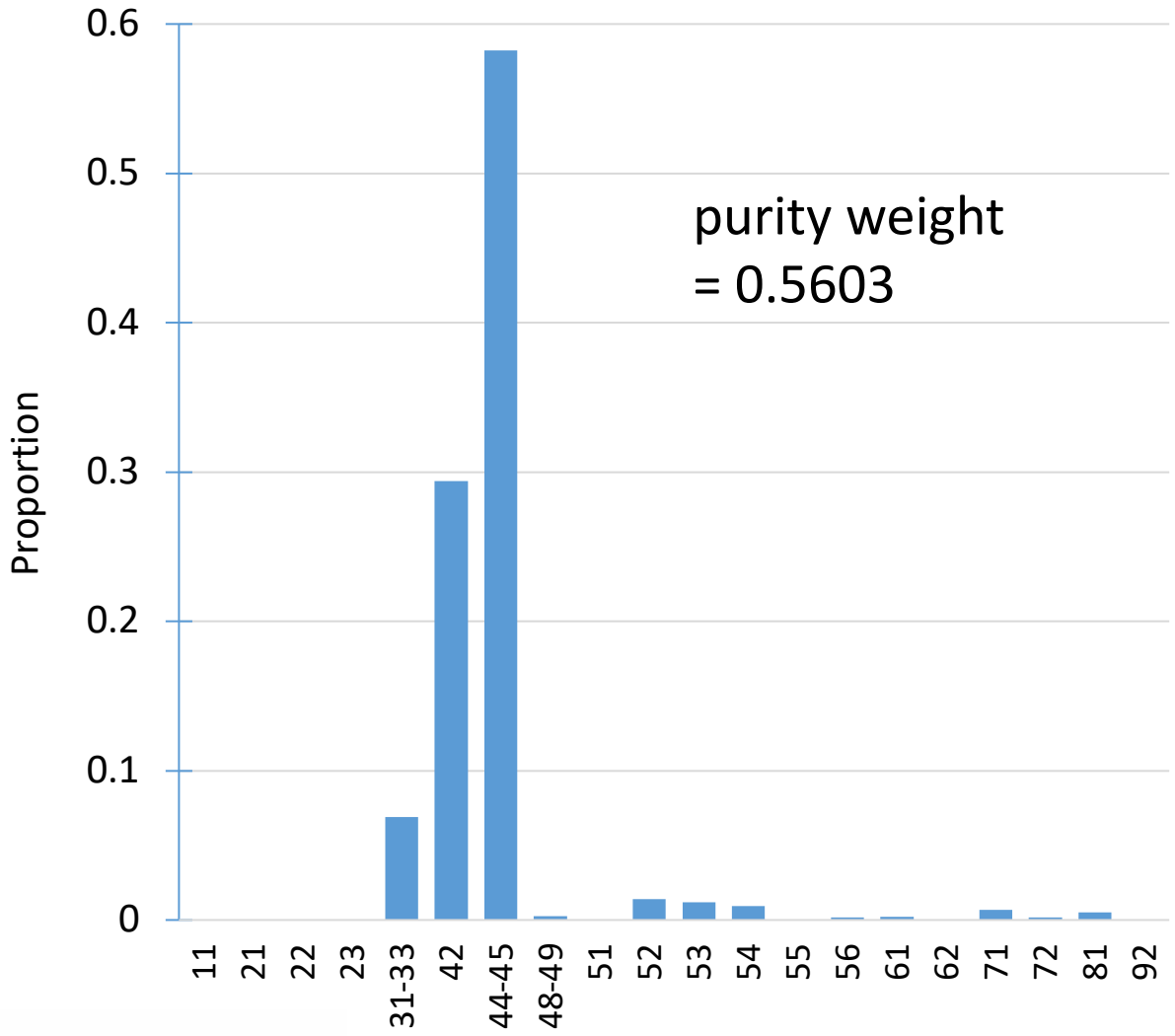


Sector Distribution of “good”

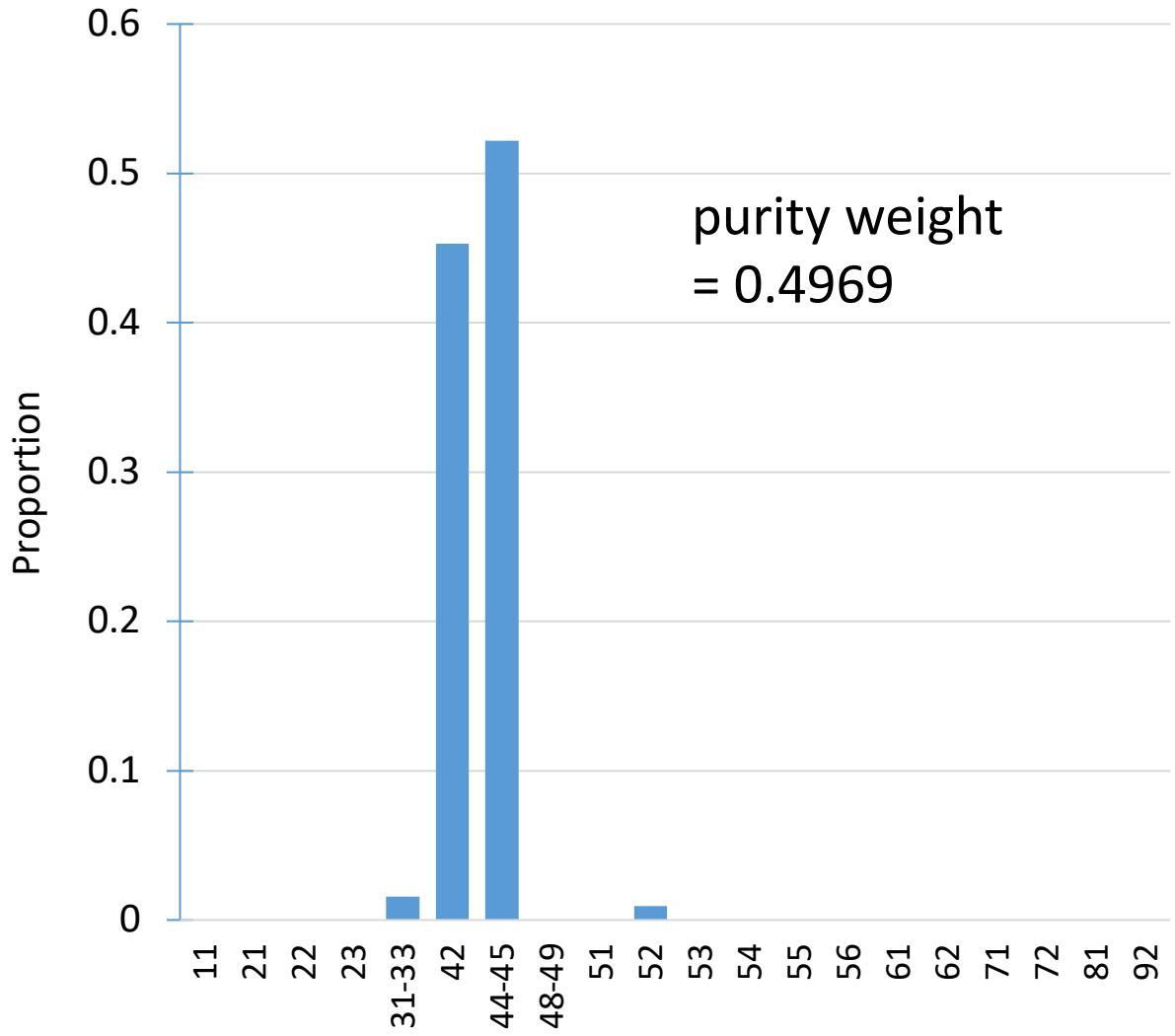


Source: 2002, 2007, 2012, 2017 Economic Census; Internal Revenue Service; and Classification Analytical Processing System

Sector Distribution of {“sport”, “good”}



Sector Distr. of exact{“sport”, “good”}



United States® Census Bureau

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

Model Ensemble

- Three information retrieval sub-models use different features
- Example: description “Sporting goods” with clean text “sport good”

Sub-Model	Description	Features in Example
All	All words and word combinations	Word “sport” Word “good” Two-word combination {“sport”, “good”}
Umbrella	Words and word combinations that are not subsets of others	Two-word combination {“sport”, “good”}
Exact	Full-length/exact description	Full-length description exact{“sport”, “good”}

United States® Census Bureau

Model Ensemble (cont.)

- BEACON assigns relevance scores to NAICS codes for ranking
- For each sub-model, the NAICS distributions of the relevant features are averaged using purity weights to calculate relevance scores
- The model ensemble calculates a weighted average of scores from the three sub-models, where the weights are optimized using machine learning

0.1 (All) + 0.6 (Umbrella) + 0.3 (Exact)

United States® Census Bureau

Hierarchical Model Structure

- Assigning relevance scores directly at the 6-digit level is challenging
- Model ensemble is applied 21 times
  - 1x to assign scores at the 2-digit level
  - 20x to assign sector-conditional scores at the 6-digit level
- Scores are calculated using the conditional probability formula

$$\text{score}(52) \times \text{score}(522291|52) = \text{score}(522291)$$

United States® Census Bureau

Economic Census Field Test

- October 2021 – February 2022
- Test usability of new EC questionnaire features with respondents
- Sample consisted of 37,000 single-unit establishments
- By design, a third of the sample had a reliable NAICS code on record
  - “Truth deck”
  - Used to evaluate accuracy of BEACON
- Received approximately 20,000 descriptions

United States® Census Bureau

Economic Census Field Test (cont.)

- For a successful NAICS self-classification, two actions need to occur
  - BEACON needs to return the correct NAICS code as one of its search results
  - The respondent then needs to select the correct NAICS code
- Probability estimates based on truth deck

90.1%	×	83.7%	=	75.5%
BEACON returns correct NAICS code		Respondent selects correct NAICS code given that it is a search result		Successful NAICS self-classification

United States® Census Bureau

Future Work

- Refine text cleaning algorithm
- Incorporate new data sources such as write-ins from the 2022 EC
- Research more advanced models
  - Word embeddings
  - Model stacking – applying a second-stage, or meta, model that uses the sub-models' scores as input

United States® Census Bureau

Contact Information

- Brian.Dumbacher@census.gov
- Daniel.Whitehead@census.gov

United States® Census Bureau