A Large Scale, High Quality U.S. Occupational Database: Results from Merged ACS and IRS Write-Ins

Julia Beckhusen, Victoria L. Bryant, David B. Grusky, Thomas N. Hertz, Michael Hout, Liana Christin Landivar, Lynda Laughlin, Ananda Martin-Caughey, Javier Miranda, Kevin Pierce, Carl Sanders

FCSM, 24 October 2024

Purpose & motivation

- Worker occupation is a key driver in economic growth (Violante 2008), career progression (Yamaguchi 2011), and cross-sectional and intergenerational inequality (Card and DiNardo 2002, Long and Ferrie 2013).
- Universe-level occupation data available in some countries (e.g. Denmark), but administrative and data collection difficulties in the U.S.
- Census: American Community Survey
- IRS: Form 1040 “Occupation” field

Contribution

- Create near-universe dataset of coded worker occupations
    - Match e-filed Form 1040s and 1-Year ACS
- Evaluate quality of matched IRS/ACS write-ins
    - Token similarity
    - Semantic similarity
- Create a Large Language Model-based autocoder mapping text write-ins to Census 2018 occupation codes.
- (Preliminary) Evaluate cross-sectional and longitudinal accuracy of IRS occupational distribution

Data

- American Community Survey 2019 1-Year Microdata (ACS) write-ins
- IRS Tax Year 2018 Form 1040 write-ins

ACS and IRS Occupation Prompts

ACS

e. What was this person's main occupation? *(For example: 4th grade teacher, entry-level plumber)*

Driver

f. Describe this person's most important activities or duties. *(For example: instruct and evaluate students and create lesson plans, assemble and install pipe sections and review building plans for work details)*

Pick people up in my car,
drive them where they need to
go, and drop them off

F1040

Sign Here

Joint return? See instructions. Keep a copy for your records.

Under penalties of perjury, I declare that I have examined this return and accompanying schedules and statements, a
belief, they are true, correct, and complete. Declaration of preparer (other than taxpayer) is based on all information o

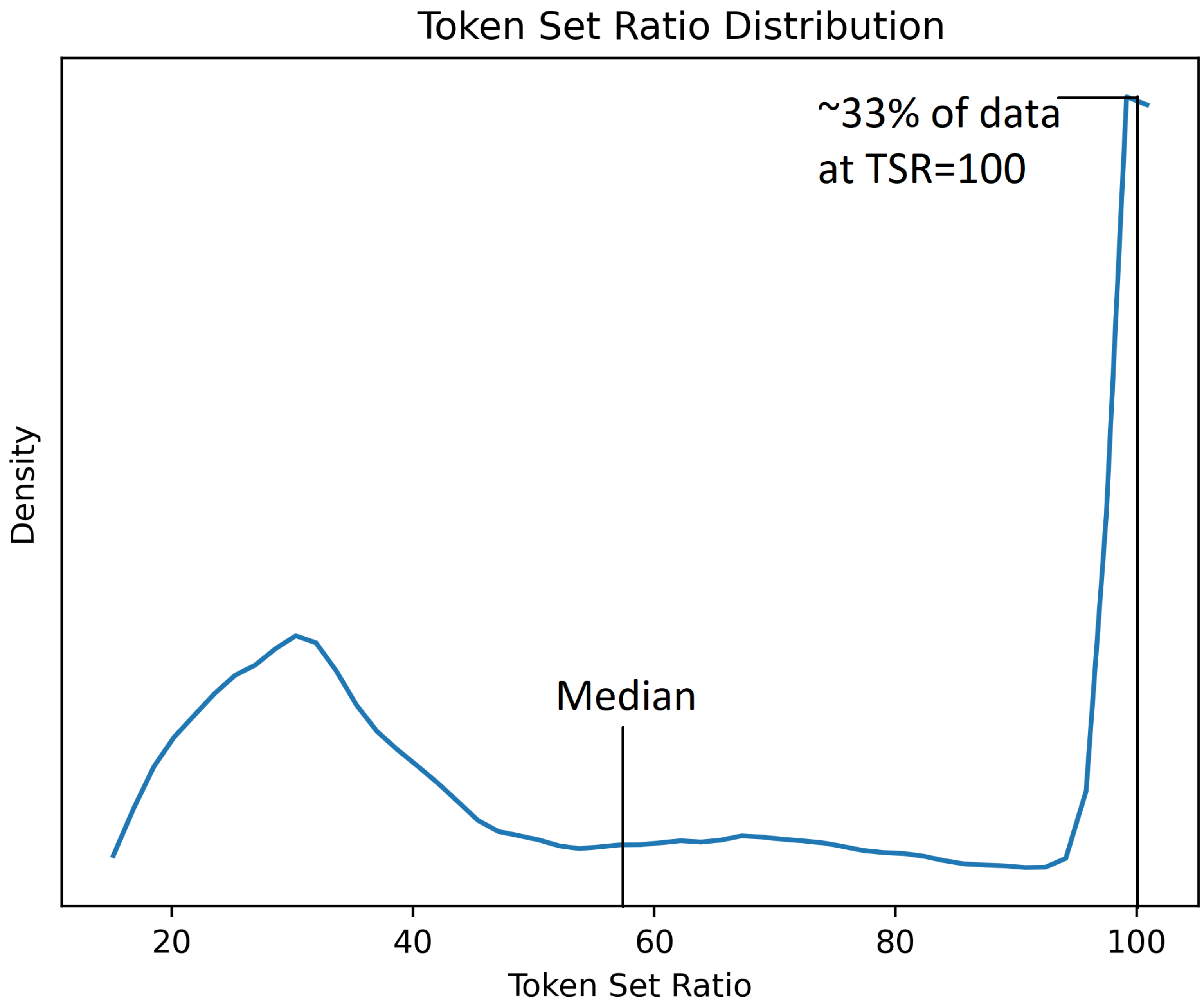
Your signature	Date	Your occupation
Spouse's signature. If a joint return, **both** must sign.	Date	Spouse's occupation

United States® Census Bureau

Token Similarities

- Token Set Ratio: 0-100 score of similarity of two strings

- TSR(“Lawyer”, “Lawyer”) = 100
- TSR(“Clown”, “Teacher”) = 17
- TSR(“Lawyer”, “Attorney”) = 29
- TSR(“Paralegal”, “Paramedic”) = 56



United States® Census Bureau

Transformer-based Autocoder

- BERT (Bidirectional Encoder Representations from Transformers) architecture for Large Language Modeling
  - Open Source LLM, pretrained on Wikipedia and the Toronto BookCorpus (3.3 billion words)
  - Maps a text string to a numerical vector representation (“encoding”).
- Occupational coding problem estimated as a Multinomial Logit with 565 choices
- Inputs: text writein -> BERT encoding, industry category
- Target: assigned 2018 Census occupational code (565 categories).

United States® Census Bureau

Estimation Results

Model	Match Rate	Top 2	Top 10
ACS LLM Text + Industry	**0.81**	**0.90**	**0.97**
IRS LLM Text + Industry	**0.42**	**0.54**	**0.77**

United States® Census Bureau

Source: U.S. Census Bureau, 2019 American Community Survey 1-year and IRS Form 1040 Tax Year 2018

Semantic Similarity

- The ACS and IRS model each predict a probability distribution
- **Total Variation Distance** (TVD) between them measures prediction disagreement
- Results from TVD broadly agree with results from token-based analysis
- Approx. 50% paired entries semantically similar, approx. 33% high quality semantic matches

Agency Benefits

- IRS:
  - Fully coded occupational field
  - Response quality control via ACS comparisons

- Census:
  - Show feasibility of Open Source, Machine Learning-based occupation coding
  - Improved imputes for missing records

United States® Census Bureau

Conclusion

- Creating a near-universe file of coded occupations from Form 1040 write-ins is feasible when combined with ACS data.
- Economically significant information in IRS write-ins, but measurement challenges remain.
- Next steps: aggregation; years 2011-2018.

Funding: Russell Sage Foundation [Hout & Grusky]

Thank you!

Carl Sanders

Carl.E.Sanders@Census.gov

United States® Census Bureau