Automating Processes for the U.S. Census Business Register

Michael E. Kornbau
Business Register Research Staff
U.S. Census Bureau
Economic Statistical Methods Division
Michael.E.Kornbau@census.gov

Disclaimer: Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau

Outline of Processes

- Automated Industry Coding of New Businesses
- Automated Coding of Legal Form of Organization
- Automated Processing of Electronic Instrument Comments
- Quality Assurance of Administrative Records

Automated Industry Coding

- Administrative source of business name and description for new businesses applying for an Employer Identification Number (EIN)
- Training data for 4.3 million businesses
- Assign NAICS (North American Industry Classification System) codes

Automated Industry Coding

- Automated creation of coding dictionaries based on name and description tokens
  - One-word, two-word name tokens
  - One-word, two-word, full description tokens
- Entry requirement: Frequency of 20 and 40% assignment to one NAICS code

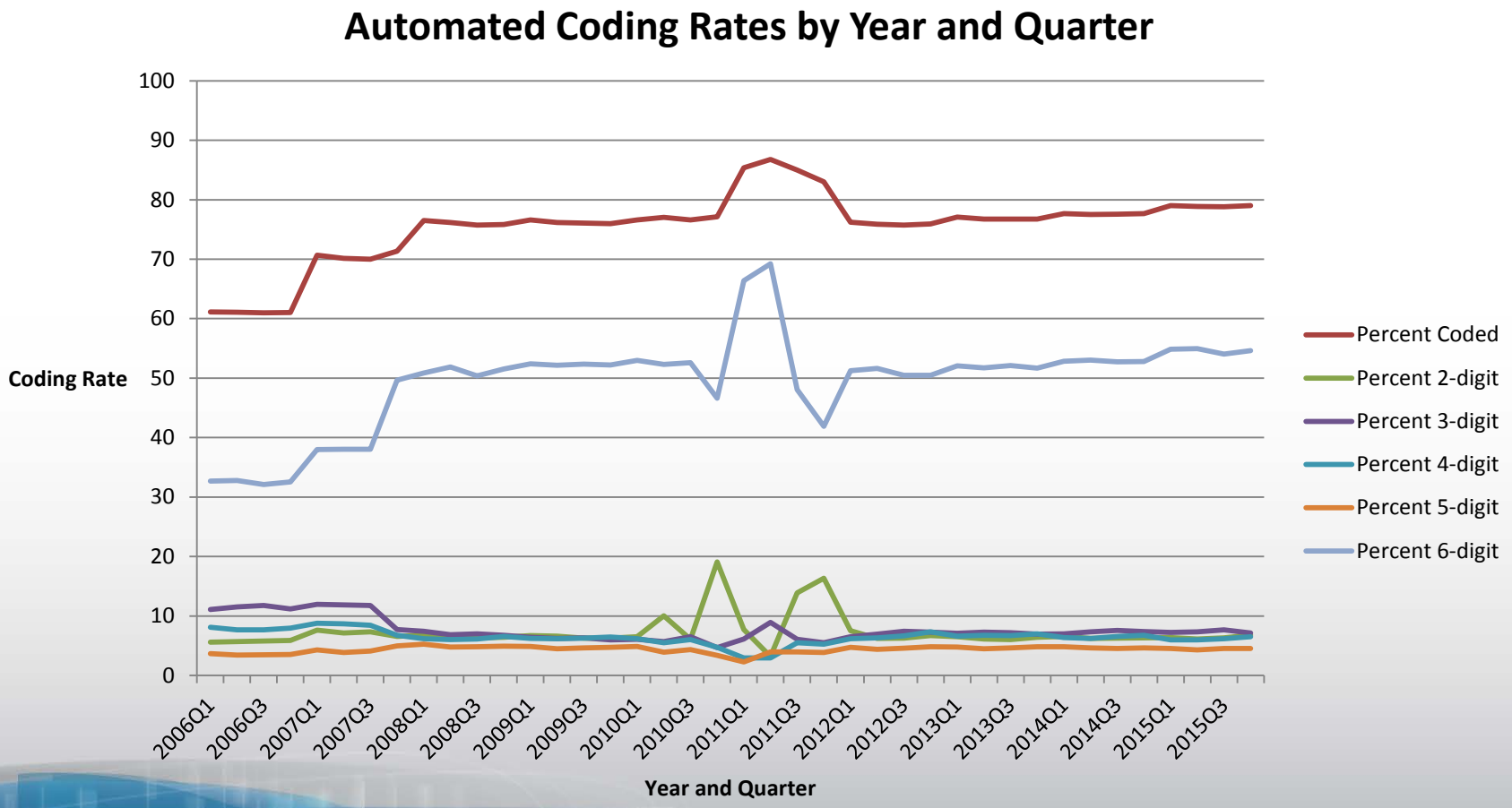
Automated Industry Coding

- Algorithm to assign NAICS code from multiple possible dictionary matches
  - Version 1: Four factor weighting
  - Version 2: Logistic regression model
- Score assigned for each match. Version 2 score is estimated probability that match is correct code. Select code with highest score.

Automated Industry Coding

- Score Cutoff: Goal is same level accuracy as 100% manual coding
- Score > cutoff → Assign NAICS code
- Score < cutoff (or no match) → Manually code
- 60% to 80%+ coding rate from 2004 to 2016

Automated Industry Coding



Automated Industry Coding

- Quality Control Process:
  - Quarterly sample of around 8,000 records in 42 NAICS code categories
  - Two independent coders with adjudication
  - Track error rates by code category
- New online EIN application – example of designing questions to increase coding rate and accuracy

Legal Form of Organization

- Most common LFOs:
  - Corporation (C or S)
  - Partnership
  - Sole Proprietorship
  - Tax-exempt Corporation or Other
- Commonly assigned through administrative record data

Legal Form of Organization

- Use classified records to automatically create coding dictionaries
  - One-word, two-word tokens from first name field
  - One-word, two-word tokens from second name field
  - Different entry requirements between first and second name fields

Legal Form of Organization

Dictionary Description	2008	2014
	Number of Records	Number of Records
1st Name Field – one-word token	15,039	17,313
1st Name Field – two-word token	45,598	62,166
2nd Name Field – one-word token	3,301	3,523
2nd Name Field – two-word token	9,905	10,021

Legal Form of Organization

- Algorithm to assign LFO: Match tokens from unclassified business name to dictionaries. Assign the matched LFO with the highest frequency rate.
- Automated coding rate was 72% in 2014. Remainder coded as 'unknown'.

Electronic Comments

- Move to electronic reporting → easier to capture and analyze text responses
- Reviewed general comments provided by electronic reporters to 2012 Economic Census
- Over 158,000 single-unit comments and over 344,000 multi-unit comments

Electronic Comments

- Use SAS Text Miner to identify clusters of similar comments
  - Parses text data into a term-by-document frequency matrix
  - Singular-value decomposition to reduce dimensionality
  - Expectation-maximization algorithm to separate documents into clusters

Electronic Comments

- Types of Clusters
  - Industry classification (what they do)
  - Operational status: closures, openings or acquisitions
  - Feedback on the questionnaire
  - Meaningless (i.e., ‘None’) comments
  - Business location

Electronic Comments

- Effective at grouping similar comments
- Attempt at status changes from comments
  - Questionnaire already has operational status question
  - No value added from comments
- Possible uses by looking at specific clusters

Administrative Records Quality Assurance

- Administrative record extracts from other government agencies:
  - New businesses applying for an EIN
  - Quarterly and annual payroll tax data
  - Annual business income tax data
  - Industry data from several sources
  - Address and geographic data from several sources

Administrative Records Quality Assurance

- Objectives:
  - Identify potential errors in data received
  - Ensure agreement between data requested and data received
  - Provide record counts for administrative reports
  - Review distributions of different data items
  - Track problems

Administrative Records Quality Assurance

- Significant revisions to SAS QA programs in 2014
- New programs:
  - Metadata specified in Excel spreadsheets
  - Generic SAS programs
  - Output in Excel spreadsheets for analytical review
  - Standards – Flagging system

Administrative Records Quality Assurance

- Example of flags:

	201507	18		CHECK INVALID DATASET, 18 instances of non-numeric data in continuous or ID fields
Beginning of Year Inventory	201505	41,310		95P = 63,221 Reference=(80,000 to 300,000)
Total Income	201505	41,310		95P = 1,051,063 Reference=(1,200,000 to 3,000,000)
Interest Income	201503	23,222		Values not appearing when expected for intr and for form 03
Gross Receipts or Sales less Returns and Allowances	201505	41,310		Q3 = 384,883 Reference=(400,000 to 750,000)
	201503	12,422		Values appearing when not expected for depr and for form 04

Conclusion

- Account for automation in data collection → May significantly improve results
- Census Bureau looking into uses of Big Data