25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

Michael E. Kornbau
U.S. Census Bureau

Session No. 5

Technology

Automating Processes for the U.S. Census Business Register

1. Introduction

With the goal of reducing costs and increasing productivity the U.S. Census Bureau, similar to many other statistical agencies, is attempting to automate processes that are typically manually intensive. The Census Bureau has automated several processes involving data from the Business Register, a database of all known single and multi-establishment employer companies maintained and updated by the U.S. Census Bureau. In one case, it was a large-scale collaborative effort with other federal administrative agencies that produced significant ongoing cost savings, in addition to quality improvements. In another case, an attempt at automation did not reveal any significant improvements. This paper will cover the successes that Census has had with automation as well as cases where success was not easily found.

2. Automated Industry Coding for New Businesses

The Census Bureau has established a long-term relationship with the U.S. Social Security Administration (SSA) to receive information on new businesses that file for an Employer Identification Number (EIN). The EIN is primarily used for tax purposes and the application for the EIN is managed by the U.S. Internal Revenue Service (IRS). The traditional history is that the IRS provided the EIN information to the SSA for the clerical assignment of industry and geography codes. With several million EIN applications filed annually, this involved significant manual coding and keying operations at the SSA. After these efforts were completed, the SSA would provide the data back to the IRS and to the Census Bureau with the industry codes. The Census Bureau uses the SSA-provided data as initial information about a business for inclusion in the Business Register.

The SSA has been assigning industry codes using the North American Industry Classification System (NAICS) since 1999. The NAICS is an industry classification system that groups establishments into industries based on the activities in which they are primarily engaged. It is a comprehensive system covering the entire field of economic activities, producing and nonproducing. The current version (2012) of NAICS has 20 sectors and 1,065 industries as implemented in the United States. The NAICS industry

codes are made up of 6 numerical digits. The first two digits indicate the industry sector and subsequent nonzero digits add more and more detail to the description of the business. For example, a code of 440000 indicates that the business is in the retail trade sector. A code of 445000 indicates food and beverage stores, 445200 indicates specialty food stores, 445290 indicates other specialty food stores and 445291 is the code for baked good stores. Depending on the amount of detail available from the EIN application, an establishment may receive a complete or partial (zero filled to the right) NAICS code.

At one point in 2002, the IRS started to electronically capture data from the EIN application, and when combined with SSA manual coding, it became possible to readily build a significant file of training data for automated coding. In less than two years, Census staff accumulated 4.3 million individual business names and descriptions combined with a NAICS code. From the training data, Census staff built automated coding dictionaries based on one-word and two-word business name tokens, and one-word, two-word and full business description tokens that occurred frequently (20 or more occurrences) and that mapped over 40 percent of the time to one particular NAICS code. An example of a coding dictionary entry is:

AUTO BODY REPAIR	811121	168	193	0.8705

This entry is based on the term 'AUTO BODY REPAIR' occurring in the training data (business description field) on 193 records. For 168 records, the assigned code was 811121, for a frequency percentage of 0.8705.

By matching new incoming records based on name and description tokens to the coding dictionaries, it is possible to generate multiple matches to different NAICS codes. Version I of the automated industry-coding program, known as the Autocoder, used a four-factor weighting algorithm to assign one code from among multiple possible codes. The frequency percentage was the most influential measure. A score was assigned to each coded record, with a higher score indicating a higher-quality code. A score cutoff was determined through analyst review, with a goal of maintaining overall coding accuracy at the same level as 100% manual coding. Version I was run in production from summer 2004 to February 2006, when Census implemented Version II. The automated coding rate was initially 60 percent. The Autocoder is run at both the Census Bureau and the SSA. Records not assigned a NAICS code by the Autocoder continue to be manually coded at the SSA, and these codes are shared with the Census Bureau.

Version II included a logistic regression model with 89 independent variables to estimate a probability that the dictionary-matched code is correct. The primary independent variable in the model is the frequency percentage from the coding dictionary for the dictionary token. Other variables include indicator variables for other EIN application data such as type of entity and reason for applying. There are also interaction terms and indicator variables to represent the number of words in the description and the number of words in the name. With Version II, the score cutoff was set to 0.534, indicating at least a 53.4% probability that the assigned code would agree with what a coding clerk would assign. It was discovered in testing that this estimated match probability is highly correlated with coding accuracy. A large proportion of scores are above 0.90, indicating a high level of accurate assignments. Eventually the score cutoff was decreased to 0.414 for a majority of industries in order to increase the coding rate and thereby reduce the number of cases requiring manual coding.

The Census Bureau implemented a quality control (QC) process to ensure the continuation of coding accuracy and to avoid deterioration in coding dictionaries and in the model. The QC process involves selecting a sample every three months of records coded by the Autocoder into 42 different code categories. The code categories represent different NAICS detail levels. The overall sample size is around 8,000 records.

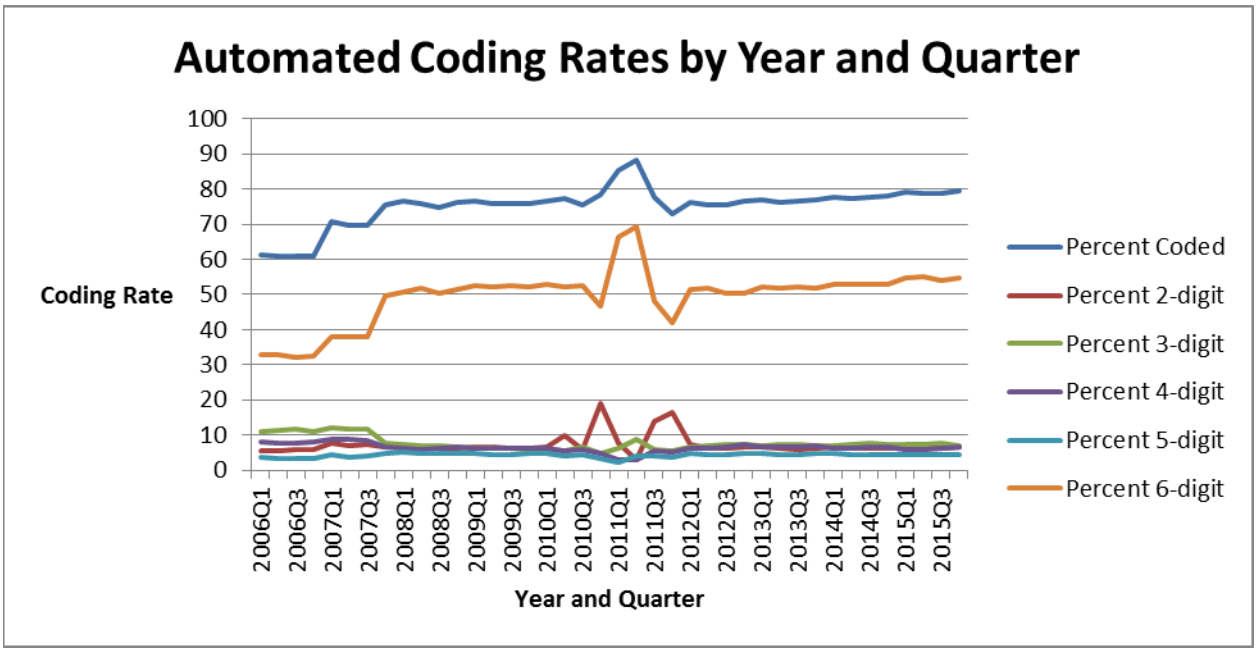
Two expert coders review the sample independently and assign their own NAICS codes based on available EIN application information. Any differences between the two expert coders are adjudicated by a third coder. Adjudication rates range from 20% to 25%, revealing the difficulty in assigning and agreeing upon one code. An error rate is computed by code category. The error rate is compared against tolerances. The tolerances are ±5% from an initial sample error rate . If two consecutive samples are above the upper tolerance, a special sample is pulled for review to determine what changes are necessary to bring the error rate back into tolerance. The usual solution is to identify additions or modifications to the coding dictionaries. The most extreme response is to increase the score cutoff (lower automated coding rate) to improve the quality of the codes and to return to meeting the tolerances. This has never been necessary.

The year 2007 saw an initial increase in the automated coding rate to 70 percent (by lowering the score cutoff) for most code categories and the IRS introduction of a new online EIN application. This new online EIN application included questions about the industry that were specifically designed with input by the Census Bureau to improve automated coding. The IRS allowed up to four screens of questions to drill down to a detailed NAICS code. In some cases, such as in food services, this involves a radio button list of NAICS descriptions. In other cases, such as in manufacturing, it requests a write-in description about the products or goods used.

The rollout to the new IRS online EIN application occurred in September 2007. After a review period where the Autocoder was not run in production, it was determined that no changes were needed for the regression model to account for the new data collection instrument. There were some coding dictionary changes. The most significant change was that with the use of radio buttons for specific NAICS codes it was now possible to assign some codes directly, bypassing the logistic regression model. In fact, the direct assignment of codes is approximately 20 percent. The overall coding rate improved to over 75%, with a significant increase in coding to more detailed NAICS levels, while staying within tolerances.

The results for the Autocoder in 2015 show a coding rate of 79% for 3.6 million records. This amounts to over 2.8 million records coded with the Autocoder. Approximately 69% of the coded records are classified to a complete 6-digit NAICS code level. The good news is that with continual improvements and a quality control process, there was no deterioration in the automated coding rate and quality, but instead an improvement was realized.

The following chart shows the coding rates of the Version II Autcoder for the ten-year period from 2006 through 2015. The overall automated coding rate is shown, in addition to the component detail-level automated coding rates of the assigned NAICS code. Since 2006, more than 50% of the NAICS codes assigned are full 6-digit codes.



The chart shows an anomaly from late 2010 to early 2012. This is not due to an Autocoder problem, but resulted from swings in the number and type of EIN applications. The chart shows the overall improvement in coding rate and detail level over time.

The Autocoder methodology has not changed significantly in Version II since 2007. There has been an interest in expanding the use of the methodology to other programs. The American Community Survey (ACS) is a large ongoing population survey conducted by the Census Bureau that replaced the use of a long form in the U.S. decennial population census and provides important information about the U.S. and its people. In 2012, the Census Bureau implemented automated industry and occupation (I&O) coding using the Autocoder methodology. The automated coding rates are significantly lower (56% for industry and 43% for occupation), which are a reflection of the limited detail provided by the respondent to meet desired accuracy levels for automated coding. But cost savings are substantial as compared to 100% manual coding.

3. Automated Coding of Legal Form of Organization

A number of Census Bureau business statistics programs publish data by legal form of organization (LFO). The most common legal forms of organization include:

- Corporation (C or S)
- Partnership
- Sole Proprietorship
- Tax-exempt Corporation
- Tax-exempt Other
- Government

The Census Bureau uses data from administrative records to assign an LFO to individual EIN records. However, these data are sometimes insufficient and an EIN is left with an 'Unknown' designation for LFO. In the BR, the LFO codes assigned to EINs are typically carried to the records of single-establishment companies. In 2014, there were over 300,000 of these single-unit establishments with payroll on the Business Register with an unknown LFO.

In order to assign an LFO code to the unknown establishments, the names and LFOs of successfully classified establishments were extracted to build coding dictionaries based on one-word and two-word tokens in the business names. To do this, each business name (which is actually comprised of two separate variables or fields) was parsed into one-word and two-word tokens. A token was included in a coding dictionary if it met the following criteria:

First Name Field

- The word token occurred at least 20 times AND
- The word token was associated with a particular LFO code at least 60% of the time.

Second Name Field

- The word token occurred at least 30 times AND
- The word token was associated with a particular LFO code at least 75% of the time.

There are four coding dictionaries – two for the first name field (one-word token and two-word token) and two for the second name field (one-word token and two-word token). The initial creation of the dictionaries was in 2008. The criteria for dictionary inclusion was modified for 2014. The following table shows the size of the coding dictionaries:

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

To assign an LFO code to an unknown company, the business name is parsed into one- and two-word tokens and matched to the corresponding dictionaries. Each word token may have a potential match – more than one per company is possible. Once all possible LFO codes are obtained from the dictionaries, the LFO code with the highest frequency rate from the dictionaries is assigned to each establishment.

One general rule is applied to the first name field: Sole proprietorships are typically shown with the owner's name in the first name field. So, if the first name field has the format of a person's name, then the LFO code is set to sole proprietorship.

For 2014, the automated LFO coding rate was 72%.

LFO	2014 Frequency Assigned	2014 Percentage
Corporation (C)	11,237	1.6
Partnership	125,687	17.6
S-Corporation	219,123	30.7
Sole Proprietorship	142,378	20.0
Tax-exempt Corporation	11,049	1.6
Other	3,014	0.4
Unknown	200,415	28.1

4. Automated Processing of Electronic Instrument Comments

With survey data collection shifting away from paper questionnaires and moving more towards online electronic reporting, it is easier to capture and analyze text responses and respondent comments. It is also apparent that respondents are more likely to provide comments in electronic data collection instruments than they were when using paper. With this increase in free-form, text-based data, it became important to determine whether or not this information could be processed automatically, thereby bypassing the need for manual review.

Electronic reporters in the 2012 Economic Census were given the opportunity to provide comments at the completion of the data collection instrument. The big question is: What information can be gleaned from these comments? Possible uses include:

- Obtaining Feedback on the Questionnaire
- Status Changes (new, closed or sold establishments)
- Sentiment Analysis

Each comment could have up to 1,000 characters. For the 2012 Economic Census, comments were retrieved for 158,739 single-unit establishments and 344,457 multi-unit company establishments[1]. Comments varied significantly between the single-unit establishments and multi-unit establishments. Due to the design of the multi-unit data collection instrument, there was significant repetition in multi-unit comments that adversely affected the analysis of multi-unit text data and the assignment to meaningful clusters. Therefore, it was decided to focus primarily on single-unit comments.

The first step was to use SAS Text Miner software to identify clusters of similar comments. The dataset was divided into three partitions – a training set, a cross-validation set and a test set. SAS Text Miner

[1] A multi-unit company operates at more than one physical location. Each location is referred to as an establishment. A single-unit company operates at one physical location, or establishment.

parses the text data in the training set into a term-by-document frequency matrix, uses singular-value decomposition (SVD) to reduce the dimensionality of the data, and uses an expectation-maximization algorithm to separate the documents into clusters. The final number of clusters can change significantly with different options, but it will typically run in the 10 to 30-cluster range. Characteristics of the clusters include:

1. Industry classification: the respondent clarifies what they do and provides information about their type of business.
2. Status changes: Clusters set up by business closures, openings or acquisitions.
3. Comments about the questionnaire itself.
4. Meaningless (i.e., "None") comments.
5. Reporting period (fiscal vs. calendar year) comments.
6. Business location comments.

This clustering was effective in placing the comments into batches, but the question is what can be done from here.

One possibility that was explored was initiating specific BR update actions for respondents with comments in the status change cluster. The questionnaire already does have an item designed to capture operational status changes – i.e., whether the establishment is new, sold, seasonal or closed. But, is it possible to find status updates from the comments alone? After some exploration, the answer is essentially:"No". It turns out that most respondents that give a comment about status change also answered the operational status item. If they did not provide a positive response to the operational status question, the comments typically reference something else – commonly a change outside of the survey reference period – or a false positive for a status change.

Despite these findings, clustering did identify groups that may warrant further investigation. These include changes to business location, feedback that can be useful for improving instrument design and data collection, and NAICS code changes not identifiable through other reported data elements. Each of these applications has their specific audience. With the move to electronic-only reporting for the 2017 Economic Census and other surveys, this information would be very useful to improve survey processing.

5. Quality Assurance of Administrative Records

The Census Bureau receives multiple administrative record extracts from three other government agencies which contain data to build and support the Business Register. The extracts include:

- Data about new businesses filing for an EIN, including industry classification
- Quarterly and annual payroll tax data
- Annual business income tax data
- Industry data from multiple sources
- Address and geographic information from multiple sources

The Census Bureau receives these extracts on a weekly, monthly, quarterly or an annual basis, depending on the extract. For reference purposes, the term 'cycle' is used to designate the week of the year when the data arrives at Census. For example, cycle 201628 represents the 28th week of 2016.

The administrative records data is generally of excellent quality. However, data quality issues do appear periodically. In addition, staff must ensure that Census is receiving the correct data items from the different agencies that supply them.

Quality assurance (QA) programs are necessary to monitor the quality of incoming administrative records data before they can be applied to the BR. The objectives of the QA programs are as follows:

- To identify potential errors in data received.
- Ensure agreement between data requested and data received.
- Track the number of records received for administrative reports.
- Provide a first glance of the distributions of different data items.
- Track problems as identified in the QA output.

Staff members engaged in a series of meetings in early 2014 to evaluate and review output from the QA programs that were in use at that time. These programs were developed using SAS. The purpose of the meetings was to get a current assessment of the QA process. This included determining if the QA output was effective in detecting problems with incoming data, eliminating unnecessary output and suggesting improvements. The result was a consensus that the programs were in need of an overhaul, in part because of changes to incoming administrative records data, the age of the programs and the ability of staff to understand how to make updates to the programs. Analysts were also interested in seeing output in a spreadsheet format (Excel) instead of looking at PDF files, which were the output of the old system.

The new QA programs continue to be in based on SAS but also create Excel files for output. They are designed to be shorter, more flexible and easier to update than the old QA programs that were used in production up through 2014. Input to the programs include the following Excel spreadsheets which provide a great deal of useful metadata on the administrative records:

- A main data dictionary, which specifies the data items we expect to receive for each extract, along with some basic characteristics of each item.

- A validation spreadsheet that lists value domains (i.e., expected values) for different variables. For example, if the variable is "tax period", valid values are expected to be in a certain range, such as 2012 to 2016. Unexpected values, such as 3016 for "tax period", are tagged as 'Invalid', and will show up in QA output as invalid.

- Standards for data items. They are separated into continuous and class variables. Standards for continuous variables may include a mean, percentile or sum being in a certain range. For class variables, standards are defined by count or percentage. For the "tax period" example, we may specify a standard that the invalid count be equal to zero. This would alert an analyst whenever unexpected values are found in administrative records data.

The old QA programs had much of this information hard-coded into the SAS code. In the new QA system, separating the metadata from the program code makes it easier to add new variables, modify value domains and to change standards.

The QA output includes the following:

- Descriptive statistics of all continuous, numeric variables that typically hold dollar values or counts, such as employment. The statistics include percentiles and quartiles (1st and 3rd quartiles, median, 95th percentile, 99th percentile) and minimum and maximum values. The output also includes negative-value, zero-value and positive-value counts and the total value. Through the standards, the QA programs checks if any of these values are outside of expected ranges. The programs also checks for non-numeric data and produces a 'red' flag if any are discovered.

- Descriptive statistics of categorical data such as NAICS codes, tax periods or checkboxes. There is also derived categorical data such as the number of characters in a text field or income categories. The QA programs calculate counts and percentages of specific values out of the total number of records. Similar to continuous variables, standards make it possible to flag items if any counts or percentages are outside of expected ranges. The programs also catch any invalid values.

  The QA programs check addresses by using SAS Data Quality Server to evaluate if an address is of the correct format. The programs cannot validate if the company actually resides at the address, or if a street number is in the valid range for the street, but rather checks the structure of the address. In some cases, a business name may appear in the address field, and this would be flagged as an invalid address. The result of this address check is a derived categorical variable. There is also a check to ensure that the ZIP code is valid for the state provided in an address.

- A system to mark data items not meeting standards with a 'yellow' or 'red' flag. A yellow flag represents "caution" and implies that a problem may exist. A red flag represents "take action" and implies a more serious problem that should lead to follow-up. The standards determine how these flags are set.

  The table below is an example of the flagging system.

	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

In this example, there are three 'yellow' warnings for percentiles being out-of-range for specific variables. There is one red flag for non-numeric data, a yellow flag for not finding any positive values for a variable and a red flag for finding positive values in a field where a positive value is not expected. The second column is the cycle number of the extract and the third column is a record count.

The QA programs create a set of spreadsheets for the above statistics and flags for each incoming administrative records extract. It also creates a cumulative spreadsheet that allows an analyst to compare the incoming extract against those from earlier cycles in order to check on trends in the data. There is also a year-to-date spreadsheet that holds totals, ranges and overall counts and percentages for the year. The year-to-date spreadsheet includes comparisons against prior year or, if appropriate, prior quarter to determine if there are any significant differences.

Other features of the QA include:

- 100-record listings to quickly review incoming data
- A check for duplicate records
- Two-dimensional tables of counts and percentages
- Record Counts

All of these features enable the Census Bureau to meet the QA objectives.

The new QA programs were first used in production at the start of 2015. For a period of two years the old programs will be run concurrently with the new programs to allow for a comparison between the two and to ensure a successful transition to the new QA process. The QA programs require an annual maintenance process to update the import control spreadsheets with any new data items or changes to categorical data values. Also, standards are reviewed and updated. This maintenance work is facilitated by having the metadata in Excel and separate from the SAS program code.

6. Conclusion

The preceding processes are four examples of efforts to speed up the processing and review of large amounts of administrative or survey-collected data through automation. It will be possible to increase automation and further improve efficiency through a good usage of machine learning combined with smart data collection. The Census Bureau is investigating uses of Big Data, which may further reduce respondent burden and improve economic statistics.