An Automated Industry Coding Application for New U.S. Business Establishments

Anne T. Kearney and Michael E. Kornbau
U.S. Census Bureau

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau.

1. Background

The U.S. Census Bureau receives industry codes from administrative sources such as the Social Security Administration (SSA) and other government agencies. The cost to SSA for generating these codes clerically for business births has increased over time. Since we currently receive electronic business descriptions and business names that often correspond to similar North American Industry Classification System (NAICS) codes, we developed an automated coding application. This paper summarizes the automated coding application originally proposed by author Michael Kornbau and augmented by Anne Kearney to assign industry codes to business births. Section 2 describes the NAICS codes. Section 3 reviews our history of receiving clerical industry codes from SSA. We review the history of automated industry coding at the Census Bureau in Section 4. Kornbau's automated coding application is presented in Section 5 followed by results in Section 6 and future research goals in Section 7.

The automated coding system is currently in use at two separate sites: the Census Bureau and the SSA. The two agencies agreed to keep the two systems as identical as possible. The original system does not use logistic regression for the weights, but rather uses the product of four weight components which lack a strong scientific foundation[1]. In an effort to make the weighting scheme more scientifically defensible, we developed the logistic regression weighting scheme described in Section 5. The new logistic regression weighting scheme also shows modest improvements in coding accuracy. However, SSA has delayed implementation due to resource constraints. We hope to have the logistic regression model in place soon.

2. What Is a NAICS Code?

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. There are 20 sectors in NAICS and 1,179 industries in NAICS United States [1]. 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 310000 indicates that the business is in the manufacturing sector. A code of 311000 indicates food manufacturing, 311300 indicates sugar and confectionary product manufacturing, and 311320 indicates chocolate and confectionery manufacturing from cacao beans. Depending on the amount of detail available to clerical coders, an establishment may receive a complete or partial (zero filled to the right) code.

3. SSA as a Source of NAICS Codes

3.1 Procedure for Receiving Industry Codes for Business Births

A new business that is planning to hire employees is required to obtain an Employer Identification Number (EIN) from the Internal Revenue Service (IRS). A business does this by filing Form SS-4, Application for Employer Identification Number, with the IRS. The SS-4 is a one-page application that includes several questions pertaining to the business name and type of business (see Figure 1 for the parts of Form SS-4 used by the autocoder).

An example (fictitious) of the input data could look something like this (see Figure 1). Item 1 may read JOE SMITH, LLC.; Item 2: JOE'S AUTO REPAIR; Item 8a may have CORPORATION box checked; Item 9 may have the STARTED NEW BUSINESS box checked; Item 14 may have the OTHER box checked with AUTO REPAIR written in the Please Specify field; and Item 15 may say TRANSMISSIONS.

The IRS has been keying the form since 2002. After assigning an EIN, the IRS provides the SSA with an electronic file of the Form SS-4 for geographic and industrial classification. The SSA has engaged in the industry classification of employer information received on the Form SS-4 for statistical purposes since 1936 when the SSA enumerated the first three million employers covered under the Federal Insurance Contributions Act. Although the IRS took over the processing of the form in 1950, the classification activities remain with SSA. The SSA has shared classifications with the Census Bureau since 1948 [2], [3].

3.2 Comparison of SSA NAICS Codes with Census/Survey Codes

The Economic Census and surveys collect detailed data used by the Census Bureau to assign codes. Therefore, they are able to assign more reliable codes to the businesses in their surveys as compared to SSA. SSA uses only the information on the Form SS-4 to assign a code. We rely on SSA codes, though, because they are the earliest source of

[1] Due to space considerations, the original weighting scheme is not described in this paper.

codes for new businesses.

Historically, the SSA NAICS codes compare to the census or survey codes as follows: 77% agreement at sector level; 67% agreement at 3-digit (subsector) level; 50% agreement at 4-digit (industry group) level; 41% agreement at 6-digit level. The autocoder described in Section 5 should assign codes equally as good or better than the clerks. The keyed clerical codes are used in the application but only the most frequently assigned business description-NAICS combinations and business name-NAICS combinations are used by the autocoder. This tends to remove the impact of human error.

4. History of Automated Coding at the Census Bureau

The Census Bureau has developed and used several automated coding applications since the mid-1960s to assign industry classifications, most importantly for the Population Census. We will not attempt to cover this history in detail, but to provide some key information about these systems and how they compare with the current coding application for business births. A more extensive review can be found in [7].

An automated coding system was developed prior to the 1967 Economic Census to assign Standard Industrial Classification (SIC) codes to small business establishments. This was known as the O'Reagan algorithm, named after Robert T. O'Reagan, formerly of the Census Bureau's Statistical Research Division (SRD), who developed the algorithm [5]. It involved the construction of a coding dictionary based on codes assigned to a set of around 10,000 historical business description responses. The stated production coding rates were excellent – between 75 to 80 percent with 90 percent accuracy. The success of the system relied upon the fact that relatively few distinct business descriptions (e.g., 'Restaurant') comprise the vast majority of business descriptions. Also, unlike NAICS there were a total of only 336 SIC codes reducing the complexity of assigning a code. Further modifications to the O'Reagan algorithm were incorporated over the next decade with little improvement.

For Population Census industry and occupation coding, the O'Reagan algorithm and related systems were considered impractical. There were several reasons for this conclusion. One significant reason was the considerable expense in keying the historical data to build the coding dictionary. The Hellerman Industry and Occupation Coding System, later called Automated Industry and Occupation Coding System (AIOCS), was used for industry and occupation coding for the 1990 Population Census [6] and was also developed in SRD. It did not use historical response data but relied upon a comprehensive system of coding manuals (coding data base), a company name list, a lexicon and a list of synonyms in place of the coding dictionary. There were several algorithms involved in assigning a classification. An additional quality assurance plan was in place to improve upon the classifications, with a focus on special or new coding situations. For 1990 Population Census processing, AIOCS was credited with an overall coding rate of 58 percent with an error rate of 6 percent.

With the conversion from SIC to NAICS, the Census Bureau abandoned the use of AIOCS for the 2000 Population Census primarily because of a lack of resources to upgrade the system. Instead, attention was focused on automated coding applications from outside the Census Bureau. SRD evaluated around 20 potential systems. They selected the National Institute for Occupational Safety and Health's Standard Occupation and Industry Coder (SOIC) for 2000 Census industry coding [7]. The SOIC was originally developed to code death certificates.

Other related research by the Census Bureau on automated coding include [8], [9].

5. Description of the New Autocoder

The autocoder has two main components. The first is a set of five dictionaries used to store common business description-NAICS combinations and common business name-NAICS combinations. The second is the system of programs used to match the incoming files to the dictionaries.

5.1 Using Historical Data to Create Dictionaries

The IRS began keying the Form SS-4 in 2002. From that time through mid-2004, the SSA continued to clerically assign NAICS codes to these business births. Therefore we have about 2 ½ years of clerically coded EINs, or over four million clerical codes. This wealth of electronic and clerically coded data lends itself to building an automated coding procedure. A perusal of the electronic data shows that EINs with similar business descriptions or business names often map to the same NAICS codes. Therefore, we can use commonly recorded business descriptions and business names as dictionaries to look-up and automatically assign the NAICS codes.

To build the dictionaries from the clerically coded records, we concatenated the two business description fields (Items 14 and 15 in Figure 1) into one variable and under certain circumstances[2] concatenated the two business name fields (Items 1 and 2 in Figure 1) into one variable trying to avoid proper names. From these two variables we created five files. The first file contains the full description for each EIN. Next, we parse the business description and business name variables into one- and two-word tokens (or word strings) for the remaining four files. That is, we parse the business description variable into two-consecutive-word tokens and parse the same variable into one-word tokens. We do the same thing with the business name variable. The second file created is the two-word business description file. The third file created is the one-word business description file; the fourth is the two-word business name file, and finally the fifth is the one-word business name file.

Using the example in Section 3.1, we give an example of

[2] When Business Name (Item 1) ends with LLC or Trade Name (Item 2) ends with MBR, MEMBER, PTR, or TTEE we exclude that item from the business name variable. The items under these conditions will usually include a proper name.

the potential dictionary entries. The full description variable consists of the concatenation of Items 14 and 15 and would be AUTO REPAIR TRANSMISSIONS. The two-word tokens from the description would be: AUTO REPAIR and REPAIR TRANSMISSIONS. The one-word tokens would be AUTO; REPAIR; and TRANSMISSIONS. Item 1 (legal name) ends with the letters LLC and therefore would not be used in the business name variable. The business name variable would consist of Item 2, JOE'S AUTO REPAIR so the potential two-word token dictionary entries are JOE'S AUTO and AUTO REPAIR. The one-word tokens from the business name variable would be JOE'S; AUTO; and REPAIR.

Before the lists of tokens above (including the full-description) can be used as dictionaries, we need to select only the tokens (from the wealth of keyed data) that occur a minimum number of times and most often map to the same NAICS code. To accomplish this, we count the number of times each token occurs within each of the respective files. Then within token, we count the number of times a token maps to a NAICS code. If the token occurs at least 20 times, and it maps to the same NAICS code at least 40 percent of the time, then the token makes the dictionary. This method selects only the tokens with descriptive power and allows common misspellings, plurals, and common suffixes to make the dictionaries. After creating the preliminary dictionaries, they were cleaned up[3] and duplicate entries were adjudicated by expert coders.

The full business description (AUTO REPAIR TRANSMISSIONS) and the two-word tokens from description in our example would probably often map to the same NAICS code. So, if they occurred often enough they would make the dictionary. The business name tokens that include the word JOE'S probably would map to a variety of NAICS codes, and therefore would not make the dictionary. Similarly, most of the one-word tokens for name and description would not make the dictionaries.

5.2 Autocoding New Form SS-4 Files Using Dictionaries

The incoming Form SS-4 files from the IRS do not yet have a NAICS code assigned. We use the autocoder dictionaries to assign a code to EINs with common business descriptions or business names. We accomplish this with the incoming Form SS-4 records by following a procedure similar to that of building the dictionaries. That is, we parse the business description and name variables into five files.

We take each file above and match to the corresponding dictionary. This procedure can generate many potential NAICS codes for each EIN. For example, a different automobile repair shop with a business description on the incoming file of AUTO REPAIR SHOP and a (fictional) business name of ANNE'S AUTO REPAIR can match to several dictionaries. The business description variable may match to the full description dictionary. AUTO REPAIR and/or REPAIR SHOP may also match to the two-word token description dictionary. Finally, AUTO REPAIR may match to the two-word token name dictionary. AUTO and/or REPAIR would probably not be on the one-word token name or the one-word token description dictionary since they could easily map to a variety of NAICS codes. So, the EIN ANNE'S AUTO REPAIR could possibly generate a total of four matches.

The matching dictionary entries and corresponding NAICS codes may look like this:

Full Descrip:	Auto Repair Shop	811100
Two-word Descrip:	Auto Repair	811100
Two-word Descrip:	Repair Shop	811000
Two-word Name:	Auto Repair	811100

Only strings that compare exactly on each character are considered matches.

To select the "best" NAICS code choice for the EIN, we use a logistic regression procedure to assign a probability to each choice. The code with the highest probability is selected. After selecting the code with the highest probability, detail is added to the selected code when another code choice has more detail and agrees at the partial level. The logistic regression procedure is described in the next section.

5.3 Selecting the "Best" NAICS Code Using Logistic Regression

There were a total of approximately 4.3 million electronic EIN records with a clerical code since the beginning of 2002. This provides an excess of records with which to build the model. Since we believe the clerical coding accuracy has improved over time, we excluded 2002 data from the model. We also excluded data from three cycles (essentially 3 months of data) in year 2004 which were used as validation data sets. From the remaining EIN records, we selected a systematic sample of one million records.

The file input into the logistic regression model is not at the EIN level, rather there is one record for each match to the dictionary. The autocoder system handles an essentially unlimited number of dictionary matches per EIN (limited only by the number of words in the business description and business name and the number of dictionary entries). We used SAS for the stepwise logistic regression procedure (proc logistic) which assumes independence between the input records. Our model does not have complete independence between the input records, but through validation we believe that the model achieves the goal of selecting the "best" NAICS code choice within the EIN. We used the three excluded cycles of data as validation data sets. As you will see in Section 6, our results with the validation data sets are promising.

[3] We need to review the dictionary contents because sometimes tokens that meet the 20 occurrence 40 percent rule can not be relied upon to uniquely map an EIN to a NAICS code. For example, depending on the goods sold or services provided, the token MARKETING corresponds to several sectors, and therefore even if a particular sector makes the dictionaries, the token should be removed from the dictionaries to avoid misclassification. Also, the token FINANCE was rolled up to 520000 because it is not descriptive enough to assign more detail.

The model definition looks like this: $y = \sum_{i=0}^{n} \beta_i X_i$. The dependent variable in the logistic regression model is as follows: $y = 1$ if the autocoder choice equals the clerical code, $y = 0$ otherwise, where each dictionary match is compared to the clerical code for the EIN.

There are 89 independent variables entered into the model which include 37 interaction terms. Of these, seven were not selected into the model through the stepwise procedure. So, i in the model statement goes from 0 to 82 after the stepwise procedure. We used a significance level of 0.30 for a variable to enter the model and 0.35 for a variable to remain in the model as new variables were added. There was only one continuous independent variable; the remainder were recoded dummy variables. The independent variables were taken primarily from the dictionaries and from the Form SS-4 data. The variables taken from the dictionaries include the type of dictionary, the number of times a word-token occurred when building the dictionary, and the percent of times the NAICS code mapped to that word-token (this percent is Frequency Pct in Table 1 below). The dummy variables created from Form SS-4 include Items 8a, 9, and 14 (see Figure 1).

We also calculated some values like number of words in the business description and business name, number of matches to the dictionaries for the EIN (GroupCnt), and of the matches to the dictionaries the number of times each NAICS code occurred (CodeCnt). These last two were used in ratio form as seen in rows two, six, and ten[4] of Table 1. After analyzing the predictive power of these calculated values, we recoded them into dummy variables.

Here are the five largest and five smallest odds ratios which were obtained from the parameter estimates (β_i) by computing e^{β_i} . The odds ratios might better be described as adjusted odds ratios because they control for other variables in the model [10]. Odds ratios greater then 1.0 indicate a positive parameter estimate, and ratios less than 1.0 indicate a negative parameter estimate. Negative parameter estimates decrease the estimated match probability.

Table 1. Five Largest and Five Smallest Odds Ratios

Parameter	O_i	Confidence Limits Lower	Upper
Frequency Pct	59.7	58.4	61.0
GroupCnt/CodeCnt = 1	4.8	4.8	4.9
Full Description Dict * Mfg	3.6	3.1	4.2
Full Description Dict * Whsle	3.4	3.1	3.8
Full Description Dictionary	3.1	2.4	4.2
GroupCnt/CodeCnt > 3	0.2	0.2	0.2
Wholesale Sector (Whsle)	0.2	0.2	0.2
Manufacturing Sector (Mfg)	0.2	0.2	0.3
Full Descr Dict * number of words in Description ≤ 2	0.3	0.2	0.4
2 < GroupCnt/CodeCnt < 3	0.4	0.4	0.4

As you can see by the size of the odds ratio, Frequency Pct dominates the model. If the frequency is close to 1.0 then the predicted probability is likely to be high. In rows two, six, and ten, when GroupCnt/CodeCnt is > 1 then the word token mapped to more than one NAICS code. The larger this ratio (GroupCnt/CodeCnt) is the smaller the odds ratio. For Wholesale Sector (row 7), the EIN applicant checked one of the two wholesale boxes, and for Manufacturing Sector (row 8) the applicant checked the manufacturing box. These main effects lower the predicted value. Refer to Item 14 in Figure 1 for the rows labeled Wholesale and Manufacturing . Rows three, four, and nine are interaction terms. Each interaction term involves the Full Description Dictionary variable. The first interaction term equals one when the EIN matches the full description word token dictionary, and the EIN checked the Manufacturing box in Item 14. The second is the same but one of the Wholesale boxes were checked. Unlike the main effects for wholesale and manufacturing, the interaction terms increase the predicted value. This is because the sectors alone are too broad to easily code (each have many product lines within them), but when the products involved are narrowed down through the description field we can get a good code for them. The final interaction term equals one when the EIN matches the full description dictionary and there were less than or equal to two words in the description.

In the final model there were only four non-significant variables at the 0.10 level remaining after the stepwise procedure. We left them in the model for two reasons: 1) they are from small subgroups and the effects may be important to those subgroups, and 2) we are more interested in the predicted probabilities than the parameter estimates, and the model is predicting well.

6. Results

Due to space constraints, results from only the February 2004 validation data set are shown here. The other two validation data sets produced similar results.

The validation input file contains 305,365 EINs (used in Table 2). After parsing the business description and business name and matching to the dictionaries, there were 770,579 records input into the logistic regression model (used in Table 3). The autocoder was able to assign a NAICS code to 78.8 percent of the EINs. The remaining 21.2 percent of the EINs had business descriptions and business names that did not match to a dictionary and therefore remained uncoded. (See Section 7.2 for a discussion of why some EINs remain uncoded.) Using the 770,579 dictionary matches we estimated the probabilities of match for the validation file using the parameters estimated from the one million EINs.

6.1 Agreement Rates by Coding Level

Table 2 shows the agreement rates by coding level and is based on the final automated coder NAICS code choice. To create this table, we sorted the EINs by descending estimated probability of match. The column labels in Table 2 represent the coding rate. For example, the column labeled 10% shows the results if we accepted the 10 percent of the EINs with the

[4] In rows two, six, and ten, GroupCnt is always greater than 2.

largest estimated probabilities of match. Likewise, the column labeled 70% shows what we could expect in coding quality if we accepted the autocode for 70 percent of the file.

In the 10% column, the autocode agrees exactly with the clerical code 96.6 percent of the time. The agreement rate declines as the coding rate goes up. In the 70% column, the agreement rate is only 72.6 percent. Likewise the complete disagreement rate goes from 1.4 percent in the 10% column to 14.2 percent in the 70% column. Other rows show levels of partial agreement. To determine the acceptable level of coding, we enlisted the help of expert coders at the Census Bureau.

6.2 Using Expert Coders to Determine an Acceptable Level of Coding

In this section we address the question: "How many of the autocoded EINs are assigned a code of high enough quality to use in production, and what portion of the EINs should continue to be clerically coded?" In an operation performed in early 2003, we selected a sample of 232 EINs where the autocode disagreed with the clerical code, and provided these EINs to the expert coders at the Census Bureau. We did not disclose either the autocode or the clerical code to the experts. While the experts worked, a measurement of agreement was produced to measure the comparability of the autocode and the clerical code with the expert code taking into account partial agreement. The algorithm assigned a value that ranged from +15 to -15 to each code depending on its agreement level with the expert code. The results indicated that if we used a coding level of 60 percent that the autocoder would be just as accurate as the SSA clerks in assigning codes. This meant that approximately 40 percent of the EINs would be coded manually. The choice of coding level was further supported by a match of autocodes and clerical codes to the Business Sample Revision (BSR) codes. The BSR is a survey of new businesses at the Census Bureau that obtains NAICS codes of higher quality than the SSA clerical codes. A subsequent expert coding operation of 300 EINs further supported the choice and suggested that we might be able to code at an even higher rate.

Referring back to Table 2, this means that at the 60 percent level we have exact agreement 77.7 percent of the time. In addition, row three says both agree at the coded level but the autocode has more detail for another 4.1 percentage points. Row four says both agree but SSA NAICS has more detail another 5.3 percentage points, and row five says SSA NAICS is unclassified when the autocoder is able to classify another 1.6 percentage points. This means the autocoder produces an "acceptable" code 88.1 percent of the time at the 60 percent coding level.

6.3 Comparison of Estimated Match Probability with Actual Match Rate

Table 3 uses the same validation data set used in Table 2 but the input to Table 3 contains all the dictionary matches. In Table 3 we show how the logistic regression's estimated probability of match compares with the actual agreement level between the autocoder NAICS and the clerical NAICS. In all but the smallest two estimated probability categories, the agreement level falls within the range of the estimated match probability. This table validates the model and confirms that the lack of independence between the observations did not severely diminish the ability of the model to predict the match. For example, when the estimated match probability was in the range of 0.95 to 1.00, you can see in column three that the autocoder NAICS matched the clerically coded NAICS 96.8 percent of the time. There was at least partial agreement between the two codes another 1.6 percent of the time as is shown in column four. By partial agreement we mean that the codes agreed at the partial level, but one of the codes had more detail where the other code was right filled with zeros. The last column shows the percentage of time where the two codes disagreed at the coded level. This could be complete disagreement (i.e. disagreement at the sector level), or disagreement at some subsequent level of detail (i.e. disagreement in the third through sixth digit).

6.4 Quality Control for the Autocoded NAICS Codes Before Implementing Logistic Regression

As mentioned in Section 1, the original autocoder does not use logistic regression, but rather a system of four weight components. The logistic regression weighting scheme produces modest improvements in the accuracy of the NAICS code choice. We have not implemented the logistic regression scheme yet, and therefore have not yet submitted the codes to our quality control (QC) program. We expect the results to be similar to the original system, so we present the QC results of the original system here. The QC process is designed to monitor and prevent autocoder deterioration.

Because NAICS codes are used in sampling for economic surveys we consulted classification experts and administrators of the surveys for help in determining which code categories to focus on in the QC effort. They recommended 41 code categories which included each sector and some key subsectors (3 digit code) and industry groups (4 digit codes).

We used the nonmatch rate from the match of the autocode to the BSR NAICS code as initial error rates ($p_{i.}$; $q_i = p_i - 1$). We obtained the batch sizes (B_i) from a frequency by code category of the cycles entering the QC. In addition we used a 95 percent confidence level ($t = 1.96$) and a margin of 0.05 ($d_i = p_i + 0.05$). The margin represents the additive additional amount of error we could live with above the initial error rate.

With this information, we used the following formulas to get the necessary sample size for the i^th code category where i goes from 1 to 41: $n_{0i} = \frac{t^2 * p_i q_i}{d_i^2}$ and

$$n_i = \frac{n_{0i}}{1+\frac{n_{0i}-1}{B}} .$$

After selecting the initial sample of 9,476 EINs, the expert coders at the Census Bureau assigned each EIN a NAICS code. Error rates were calculated by comparing the expert codes to the autocodes, and some dictionary flaws were discovered and fixed[5].

With the parameters in place, quarterly samples are drawn for the production QC operation. This operation takes place at the Census Bureau's National Processing Center (NPC) in Jeffersonville, IN. Each quarter the QC coders at the NPC assign a NAICS code to each EIN in the sample without knowledge of the autocode.

In the first quarter QC effort, 3 of the 41 code catergories had error rates above our cut-off for that category. These code catergories are categories 22 (the utilites sector), 61 (the educational services sector), and 5417 (the scientific research and development industry group). Their error rates were 0.750, 0.150, and 0.206 respectively. For code category 22, there were only 4 EINs in the batch (and sample). The errors were reviewed by our expert coders at headquarters, and they determined that the autocode was actually the correct code. Therefore no action was necessary. In code category 61, we found that for 15 percent of the 153 sampled cases the autocode did not match the QC code. After our expert coders reviewed the failures, we decided that action was not necessary in this quarter because the current error rate was less than 2 percentage points above the upper bound (d_i). If the category failed in the second quarter of 2005, then we would look into this category further. In code category 5417, we found that for 21 percent of the 63 EINs the autocode did not match the QC code. The expert coders found that one of the dictionaries needed to be updated, and therefore an entry was added to the dictionary to correct this problem in the future.

If a code category has an error rate above d_i, and a fix could not be implemented, we would draw a second sample. If this produces the same results, we would reevealute the category and may discontinue accepting autocodes for that code category and instead clerically code them.

7. Future Research Goals

With changes in the types and the nature of businesses, industry code revisions, and technological changes, the automated coding program will not continue with its full coding potential without maintenance and improvements. In particular, the coding dictionaries need to be maintained and updated as businesses change, when there is a NAICS revision, or a conversion to a totally new coding structure. These types of changes may also erode the effectiveness of the algorithm that selects the best potential code. In addition, there may be pressure to further reduce manual coding, requiring work to improve the automated coding program. For all these reasons, serious thought and planning needs to be given to how the automated coding program will be maintained and improved under the environment of change.

7.1 Autocoder Deterioration

There are several potential causes for deterioration in the quality of automated coding with the current program: NAICS revision; Form SS-4 revision; change in types of businesses (e.g. new trends); and change in quality of IRS description collection and keying.

All of these causes need to be tracked. At this point, the NAICS 2007 revision is not expected to have a major impact. The development of the autocoder effectively handled changes from the NAICS 2002 revision, which was a more significant revision. A difficult adjustment will occur with a major change to the coding structure, as occurred with the change from SIC to NAICS.

The change in types of businesses and new trends will probably best be handled through the quality control program and periodic evaluations of clerical coding to pick up new common business trends. A revision to Form SS-4 or a change in the quality of the business descriptions are more problematic. These have the potential of changing descriptions and word combinations to phrases that will no longer match the coding dictionaries, leading to a lower coding rate, or incorrect codes using the new description entries. The Census Bureau will need to work closely with the IRS and the SSA to monitor changes to electronic business name and description entries.

At this point, it is unknown how quickly or slowly the program will deteriorate with little or no maintenance. The Census Bureau is planning to track the accuracy of the autocoder through the quality control program plan covered in section 6.4, and through evaluations against codes obtained from survey sources or other administrative records sources.

7.2 Improvements to the Autocoder

The current coding rate for the autocoder is set at 60 percent to obtain quality consistent with clerical coding. A review of Form SS-4 data not included in the coded 60 percent reveals no prominent type of improvement that would significantly raise the coding rate and maintain at least comparable quality. The following reasons for not getting a classification were discovered during this investigation: a vague description or a description with insufficient information for any classification; a name or description with information not meeting reliability criteria for classification; unique or detailed business description not normally encountered; messy business description with extra or missing words; and misspellings or abbreviations that result in a lack of coding

Through this review, four insights can be gained: 1. an algorithm to handle misspellings would improve coding, but not by a substantial amount; 2. for some cases it will be impossible to assign a decent code, because the business name and description are inadequate for coding; 3. a significant portion (almost half) is not assigned a code because the information in the name and description are either uncommon or too detailed to be found in the coding dictionaries. For example, “INSTALL BLINDS” is one description that has no match to the coding dictionaries; and 4. around 30 percent are assigned a good code but the score is insufficient to allow a classification.

Cases from insight number 3 will require a more intelligent method that simulates the steps taken by a clerk, where key words are used to narrow down a lookup list to a specific code. Even with this approach it is unlikely to code close to half of the current uncoded cases, because description wording can be very messy. For example, “TEL COM

[5] After fixing the dictionary flaws, we calculated new error rates by rerunning the autocoder with the updated dictionaries. These became the new error rates (p_i) that would be used in the production QC process to calculate the sample size for the code categories.

SERVICES" can be interpreted by a person, but due to spelling variations this would be more difficult for the autocoder. An attempt was made to improve classification by expanding the criteria for dictionary inclusion. In addition to the current 20 occurrence, 40 percent rule, we tried including a 5 occurrence, 100 percent to the same code rule (and we also included a 10 occurrence, 80 percent rule) – but this only increased the coding rate by about 0.5 percent. The substantial increase in the size of the coding dictionaries with the lowered criteria made it unacceptable to the classification experts who review the dictionaries.

With the advent of the online Form SS-4, it is recommended that we investigate producing an application to assist Form SS-4 filers in assigning their own NAICS codes. This would have a big payoff to agencies using the NAICS codes, and would help ensure that the proper forms (including Economic Census forms) are sent to the business. It will need to be coordinated with the IRS, who develops and owns the form. This type of improvement would preclude the need for an autocoder except for offline filing and EINs left uncoded through the internet application. The Form SS-4 autocoder is a product of the opportunity made available with substantial electronic name and description information. Future improvements should also be based on available opportunities.

8. Acknowledgments

The authors would like to thank Don Malec, Bill Winkler and Rob Creecy of the Census Bureau for their helpful comments and advice. Don was especially helpful with building a better logistic regression model. We would also like to thank Shelley Vile of the Census Bureau who worked on the QC operation and contributed to Section 6.4, and Brian Gibson of SSA who suggested a method for measuring the agreement between the NAICS code choices.

9. References

[1] North American Industry Classification System manual, United States, (2002), pp. 16 http://www.census.gov/naics

[2] Konschnik, C.A., Hanczaryk, P.S. Kornbau, M.E., (2000), "The Transition of the U.S. Business Register to NAICS," Proceedings of the Second International Conference on Establishment Surveys (CD-ROM)

[3] U.S. Department of Commerce, Census Bureau and Social Security Administration, Memorandum of Understanding Between the Census Bureau and the SSA for Sustaining Employer (SS-4) Coding Within SSA, dated July 27, 2004.

[4] Gillman, D.W. (2002) "Decision Criteria for Using Automated Coding in Survey Processing," Proceedings of the American Statistical Association, Section on Survey Research Methods, 2002 (CD-ROM)

[5] O'Reagan, R.T. (1967) Internal Memorandum to Mr.. James L. O'Brien From Robert T. O'Reagan Subject Project to Study Potential for SIC Coding of Establishments on the Computer - Project 1330, dated November 7, 1967.

[6] Appel, M.V. and Hellerman, E. (1983). "Census Bureau Experience with Automated Industry and Occupation Coding," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 32-40

[7] Gillman, D.W., Appel, M.V., (1999). "Developing an Automated Industry and Occupation Coding System for Census 2000," Proceedings of the American Statistical Association, Social Statistics Section, 1999.

[8] Chen, B., Creecy, R. H., and Appel, M. (1993), "On Error Control of Automated Industry and Occupation Coding", *Journal of Official Statistics*, Vol. 9, No. 4, pp. 729-745.

[9] Creecy, R. H., Masand, B. M., Smith, S. J., and Waltz, D. L. (1992), "Trading MIPS and Memory for Knowledge Engineering", Communications of the ACM, Vol. 35, No.8, 48-64.

[10] Allison, Paul, D. *Logistic Regression Using the SAS System: Theory and Application*, Cary, NC: SAS Institute Inc., 1999, pp. 29

Table 2. Percentage Agreement between SSA Clerical NAICS and Auto NAICS Using Logistic Regression for Weights (SSA0402: 305,365 records)

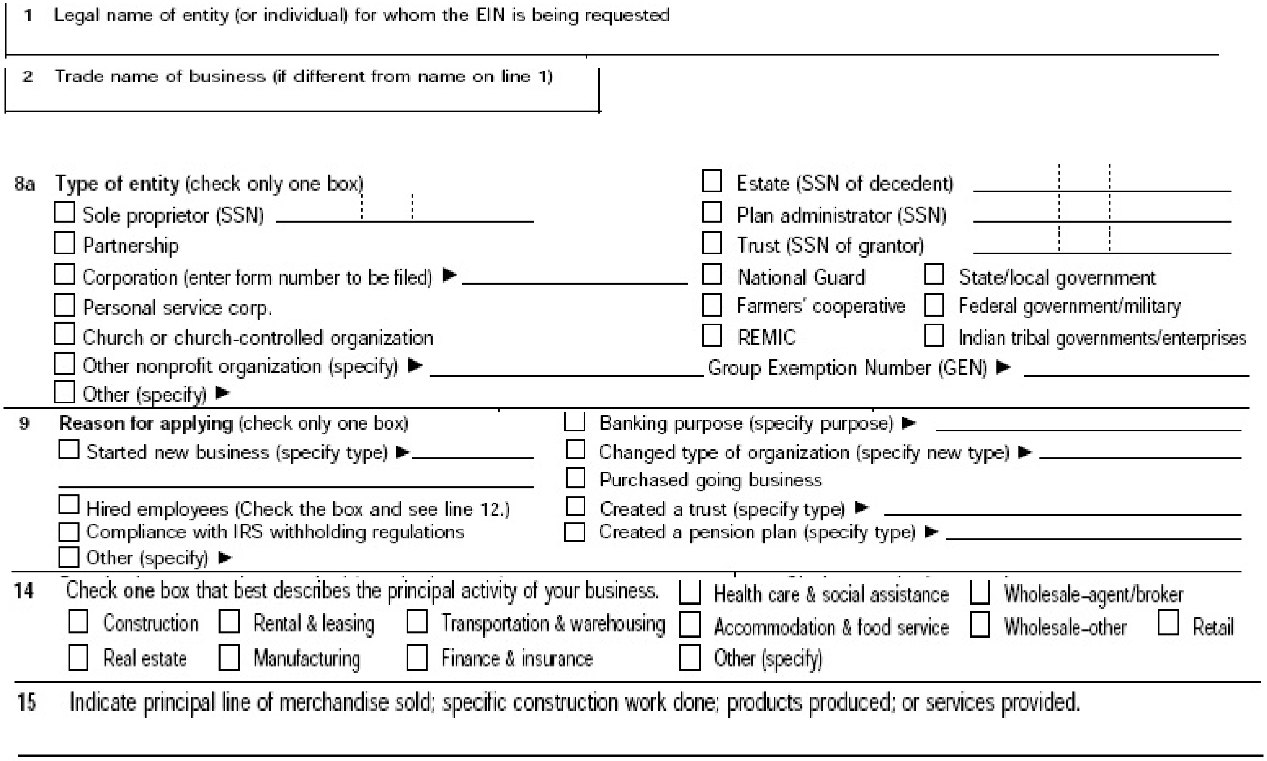
	Level of Coding						
	10%	20%	30%	40%	50%	**60%***	70%
SSA NAICS = Auto NAICS	96.6	92.5	89.9	86.2	82.2	**77.7**	72.6
Both agree, but Auto NAICS has more detail	0.7	1.5	2.2	2.8	3.5	**4.1**	4.5
Both agree, but SSA NAICS has more detail	1.0	2.4	2.9	4.2	4.9	**5.3**	6.3
No agreement at coded level	1.4	3.0	4.4	6.0	8.4	**11.2**	14.2
Sector level disagreement	0.8	1.8	2.6	3.8	5.4	**7.3**	9.6
SSA NAICS is unclassified, Auto NAICS is classified	0.3	0.5	0.6	0.8	1.1	**1.6**	2.4

*** This is what we expect to achieve in production. We currently (under the old weighting scheme) accept 60 percent of the codes, and will likely continue to accept 60 percent when the logistic regression weighting scheme is implemented.**

Table 3. Logistic Regression Estimated Match Probability by Agreement Level Between the Autocoder NAICS and the Clerically Coded NAICS

Estimated Match Probability	Number of Records	Agreement Rate Between Autocoder and Clerical Code (%)		
		Exact Agreement	Partial Agreement	Some Level of Disagreement
0.95 to 1.00	103675	96.8	1.6	1.4
0.90 to 0.95	95051	92.0	4.1	3.4
0.85 to 0.90	63580	87.5	6.3	5.4
0.80 to 0.85	59596	82.8	9.1	7.1
0.75 to 0.80	51866	75.4	13.4	10.0
0.70 to 0.75	37114	72.4	13.0	13.0
0.65 to 0.70	27823	68.4	12.9	16.8
0.60 to 0.65	23824	62.6	15.3	19.7
0.55 to 0.60	25055	59.2	15.7	22.4
0.50 to 0.55	30024	53.9	19.9	23.3
0.45 to 0.50	35047	46.6	26.2	24.0
0.40 to 0.45	36546	41.6	25.2	29.0
0.35 to 0.40	32005	37.4	27.1	32.4
0.30 to 0.35	27265	32.7	28.1	37.0
0.25 to 0.30	18801	29.2	26.1	43.3
0.20 to 0.25	14803	24.7	29.0	45.5
0.15 to 0.20	19905	18.3	33.6	47.3
0.10 to 0.15	40842	12.7	39.0	47.3
0.05 to 0.10	26572	10.5	31.2	56.9
0.00 to 0.05	1185	9.5	20.2	69.8
Total	770579			

Figure 1. Part of IRS Form SS-4 (see http://www.irs.gov/pub/irs-pdf/fss4.pdf for entire form)



1 Legal name of entity (or individual) for whom the EIN is being requested

2 Trade name of business (if different from name on line 1)

8a **Type of entity** (check only one box)
□ Sole proprietor (SSN)
□ Partnership
□ Corporation (enter form number to be filed) ▶
□ Personal service corp.
□ Church or church-controlled organization
□ Other nonprofit organization (specify) ▶
□ Other (specify) ▶
□ Estate (SSN of decedent)
□ Plan administrator (SSN)
□ Trust (SSN of grantor)
□ National Guard □ State/local government
□ Farmers' cooperative □ Federal government/military
□ REMIC □ Indian tribal governments/enterprises
Group Exemption Number (GEN) ▶

9 **Reason for applying** (check only one box)
□ Started new business (specify type) ▶
□ Hired employees (Check the box and see line 12.)
□ Compliance with IRS withholding regulations
□ Other (specify) ▶
□ Banking purpose (specify purpose) ▶
□ Changed type of organization (specify new type) ▶
□ Purchased going business
□ Created a trust (specify type) ▶
□ Created a pension plan (specify type) ▶

14 Check **one** box that best describes the principal activity of your business. □ Health care & social assistance □ Wholesale–agent/broker
□ Construction □ Rental & leasing □ Transportation & warehousing □ Accommodation & food service □ Wholesale–other □ Retail
□ Real estate □ Manufacturing □ Finance & insurance □ Other (specify)

15 Indicate principal line of merchandise sold; specific construction work done; products produced; or services provided.