Making Quality Improvements to an Automated Industry Coding Application for U.S. Business Establishments

Michael Kornbau, U.S. Bureau of the Census,
Julie Bouffard, U.S. Bureau of the Census,
Michelle Vile, U.S. Bureau of the Census

Abstract

In 2003 and 2004, the U.S. Census Bureau developed a methodology for the automated assignment of industry classifications to new business establishments based on common name, business description, and NAICS patterns discovered from clerical coding of EIN applications at the Social Security Administration (SSA). The Census Bureau and the SSA placed an automated coding application into production to assign a partial or complete NAICS code to at least 60 percent of new business births. The application consists of five coding dictionaries and an algorithm to match electronic name and description from new businesses against the coding dictionaries to assign the NAICS code. This paper presents Census Bureau results and experience from the first two years of using the auto coder, including revisions to the original application and a description of a quality control procedure.

Keywords: business establishments, automated industry coding, NAICS

Disclaimer: This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). The views expressed on (statistical, methodological, technical, or operational) issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1.Background

New business owners use the Form SS-4 from the Internal Revenue Service (IRS) to apply for an Employer Identification Number (EIN). The IRS assigned over 3.1 million EINs in 2006. The Form SS-4 requests the business name and a description of the business in addition to several other items used by the IRS and the Social Security Administration (SSA). The IRS has provided Form SS-4 data to the SSA since 1936 for the purpose of assigning geographic and industrial classification. Both the IRS and the SSA use the industry classification for statistical purposes. Since 1948 the SSA has shared its SS-4 industry assignments with the Census Bureau [1], [2]. This SS-4 data was always kept in paper form, with the EIN and industry code supplied by SSA in electronic format until 2002, when IRS supplied all of the SS-4 data to SSA in electronic format. This data was also passed on to the Census Bureau. Prior to mid-2004, when the Census Bureau and the SSA implemented an automated coding application, industry codes were assigned primarily by clerical coders at SSA. For a period of over two years between 2002 and 2004, the Census Bureau received electronic business name and description information along with the clerically assigned NAICS code from the SSA for over 4.3 million Form SS-4 filers. This provided a wealth of data for developing an automated coding application using common business name-description-NAICS code patterns. The Census Bureau, in cooperation with the SSA, developed and placed into production in 2004 an automated coding application to assign NAICS codes for approximately 60 percent of businesses filing Form SS-4. The remainder continued to be assigned by SSA clerical coders. At the initial production phase in 2004, the Census Bureau maintained a SAS version of the application, also known as the Autocoder, while the SSA maintained a similar version of the program in MS SQL Server. Both agencies ran their own versions of the program on SS-4 data files received from the IRS, resulting in slight coding differences, until a joint Service Level Agreement was signed by the Census Bureau and the SSA in April 2006 that ensured both agencies would assign the same codes using the same SAS version of the Autocoder.

This paper will provide an overview on the Autocoder including its development at the Census Bureau and also cover earlier versions of automated coding applications. The paper will review the development and maintenance of coding dictionaries, the algorithm for assigning a NAICS code, and a quality assurance system in place to ensure continued quality and improvements to the Autocoder and prevent deterioration. It will also cover challenges faced by the Census Bureau in the first couple years of running the Autocoder in production, consider potential improvements and enhancements, and suggest the future direction for assigning industry codes to new businesses applying for an EIN.

This paper follows a report by Anne T. Kearney and Michael E. Kornbau entitled "An Automated Industry Coding Application for New U.S. Business Establishments" [3], presented at the 2005 Joint Statistical Meetings (JSM) in Minneapolis, MN. The report appears in the ASA Proceedings. Kearney and Kornbau cover aspects that will not be covered in this paper, including a brief history of automated coding at the Census Bureau, the structure of NAICS codes, and a detailed description of developing the coding dictionaries and the logistic regression model that are the basis for the current version of the Autocoder. This current research paper will take on a different perspective – by reviewing how the current version of the Autocoder evolved from earlier versions of automated coding. It will document the experiences of running the Autocoder in production, while making improvements to continue its coding efficiency. Readers are encouraged to review Kearney and Kornbau (2005) [3] for additional details on the Autocoder.

2. Predecessors to the Form SS-4 Autocoder

The O'Reagan algorithm developed in the 1960's and described in [4] and briefly in [3], was largely forgotten by the Economic Directorate at the Census Bureau when the need developed for an SS-4 automated coding application in 2002. Instead, the earliest known version of an automated coding application was a simple procedure to assign IRS Principle Business Activity codes to unclassified sole proprietorships filing an income tax return with a supplied business description. The IRS keyed and provided a 20-character business description to the Census Bureau for these unclassified businesses every five years corresponding with the Census Bureau's Economic Census cycle, to save costs of mailing classification forms. In other years, the keying costs were prohibitive, and the Census Bureau did not request the business descriptions.

The assignment process started with a clerk assigning an industry code to each unique 20-character business description. The description and the assigned code became part of a coding dictionary so that subsequent entries with the same description were automatically assigned the identical code. The business name was unavailable. Only the 20-character description was useful in assigning a classification. Many of the supplied descriptions were vague and didn't allow for assigning a reliable industry code. The coding rates using this process ranged from 65 percent in 1992 to 62 percent in 1997 and 59 percent in 2002.

Another earlier version was an application developed in 2001 that assigned NAICS codes to unclassified businesses on the Census Bureau's Business Register[1] based on business name. The developers automated the creation of one-word and two-word name token dictionaries from businesses assigned a NAICS code from the 1997 Economic Census, based on common name token – NAICS patterns. For example, the two-word name token 'Carpet Cleaning' appeared in the name of 746 businesses, and 82 percent of those businesses had an assigned NAICS code of '561740'. At the time, the 1997 Economic Census NAICS code was considered the most reliable industry code source because the codes were based on detailed data collected from census forms. A one-word or two-word business name token was added to the coding dictionary whenever it appeared in at least 40 business names and was coded to a particular NAICS code at least 75 percent of the time. This rule kept out names of individuals such as Jones, Robert, or Smith while including descriptive words such as 'Restaurant', 'Pharmacy', or 'Auto Repair'. In some cases, the rule could not be met at the 6-digit NAICS level, but could be met at fewer digits such as 3-, 4-, or 5-digit levels. In these cases, the token was added to the coding dictionary with the most detailed partial code satisfying the rule. This effort led to the assignment of NAICS codes to approximately 32 percent of unclassified businesses on the Business Register in 2001, saving additional cost of mailing classification forms to obtain NAICS codes.

With the receipt of electronic SS-4 name and description in 2002, the Census Bureau and the SSA made an effort to design an automated coding application that would relieve increasing clerical coding costs at the SSA. Initial attempts included an automated assignment based on the appearance of particular business descriptions. This led to a 10 to 15 percent coding rate, which helped to some extent but was not significant when considering the rising number of filed SS-4 forms. A review of business descriptions found that a small number of unique business descriptions occur frequently, such as 'Construction' or 'Real Estate', but the frequency is not to the level that would lend itself to an effective coding application. An application was necessary that

[1] The Business Register is a current and comprehensive database of U.S. business establishments and companies for statistical program use at the Census Bureau. The Business Register covers more than 180,000 multi-unit companies, representing 1.7 million affiliated establishments, 5.8 million single-establishment companies, and 19.5 million non-employer businesses.

combined business name, similar to name coding on the Business Register, with the business description, making use of tokens of each component. Using the same approach as the Business Register name coding, it would be necessary to create coding dictionaries, but also an algorithm to decide on potential codes from the business name and business description that may conflict.

Considering that potential codes could come from either the business name or description, five coding dictionaries were developed from one- and two-word business name tokens, one- and two-word business description tokens, and a full business description. It was thought that the full business name and tokens of three or more words would not be of much additional benefit. The criteria for dictionary inclusion was lowered to 20 occurrences with 40 percent mapped to the same NAICS code based on codes assigned to SS-4 filers by the SSA clerical coding staff. This rule was somewhat arbitrary, but designed to keep out individual names or non-descriptive words, while avoiding duplication of entries with more than one NAICS code. The usage of five dictionaries with loose requirements leads to the potential of more than one possible code, so a scoring method would be necessary to decide on the best NAICS code.

To derive a score, several factors were considered that might affect the accuracy of a potential code. The first, and most important, was the percentage of all occurrences that a particular token was associated with a particular NAICS code. If 'Restaurant' is associated with '722110' at a rate of 95 percent, it's expected that assigning '722110' whenever 'Restaurant' appears, would lead to being accurate around 95 percent of the time. If the percentage were 40 percent, the accuracy rate would be much lower. Another factor was the actual number of occurrences – if only the minimum number was reached (20) for the token to make the dictionary, it would not be as significant as a higher occurrence such as 1,000. A third factor was the agreement with other potential codes. If a business generated five matches from the coding dictionaries and each yielded the same code, the code should be considered more highly than a match of five codes with three potential codes. The fourth, and last, factor considered in assigning a score to a potential code, was the coding dictionary the token matched. A match to the full description dictionary was seen as more significant than a match of one word to the one-word description dictionary. Also, business description was given more weight than business name. With these factors, the formulation of a score was based on a methodology using personal judgment and evaluation with arbitrarily assigned parameters to produce a score for potential codes.

The number of dictionary matches for any one EIN (one SS-4 form) could range from zero to well over 10 matches. The Census Bureau's version limited the number of matches considered for the assigned code to five, while the SSA version of the Autocoder looked at all potential matches. This difference in handling multiple dictionary matches led to a difference in the third weighting factor between the Census Bureau and the SSA, resulting in some differences in the assigned code. The approach taken by the Census Bureau in computing the third weight factor became considerably lengthier with more matches because it considered all possible combinations of agreement, while the SSA computed two variables – one based on the total number of matches and the other on the number of unique codes. The SSA approach could handle an indefinite number of codes without adding length to the program.

After weight factors were assigned for each match to the coding dictionaries for each EIN, a weight was computed for each match by multiplying the four factors. The weights were compared and the code associated with the highest weight was the assigned code.

The five coding dictionaries combined with the four-factor weighting approach constituted Version 1 of the Form SS-4 Autocoder. It was placed into production at both the Census Bureau and the SSA in the summer of 2004 and remained in production at both agencies until Version 2 was implemented in February 2006. The Autocoder assigned a code to approximately 80 percent of new businesses, but the low requirement for coding dictionary entries led to some codes of inferior quality. To assure comparable quality with clerical coding, the coding rate for Version 1 was set at 60 percent at the Census Bureau using a score cut-off, although SSA's coding rate showed more variation. An overall 60 percent coding rate was set after several evaluations revealed that a 60 percent coding rate would yield NAICS codes with equivalent quality as clerical coding. A description of these evaluations is presented in Kearney and Kornbau (2005) [1].

The SSA version of the Autocoder rolled out in 2004 was slightly different than the Census Bureau version, as previously mentioned, which led to assigned code differences of around 8% of total EINs. The Census Bureau coding rate was consistently around 60 percent and only fluctuated by one to two percent, while the SSA rate was slightly higher but with a greater fluctuation.

3. Version 2 of the Form SS-4 Autocoder

Kearney and Kornbau (2005) [3] covers the development of the second version of the Autocoder, which uses the same coding dictionaries, but replaces the four-factor weighting approach with a more defensible logistic regression weighting model. This new model lends itself better to revision when improvements are introduced. The model initially developed consisted of 82 independent variables, including 37 interaction terms. It can be defined as

$$y = \sum_{i=0}^{n} \beta_i X_i$$

where y=1 if the Autocoder choice equals the clerical code, y=0 otherwise, and where each dictionary match is compared to the clerical code for the EIN. The model variables include the frequency percentage from the coding dictionary, the coding dictionary that was matched, and the agreement among codes that were part of the four-factor weighting. However, the model also includes many other variables such as the type of entity (corporation, sole proprietorship, etc.), reason for applying, geographic location, number of words in the description, and the number of words in the business name.

With the new model, the score takes on a value between 0 and 1, and represents the probability that the assigned code would agree exactly with a code assigned by a clerical coder. The score is derived from the predicted value from the logistic regression model that was developed from a training set of 1 million SS-4 records. Table 1 shows how the assigned score compares with the agreement rate between the Autocoder and the clerically assigned code for a test sample from 2004 of over 770,000 records assigned a code independently by the Autocoder and a coding clerk. In general, the results show that the score is a good indicator of the probability that the clerically assigned code would be identical. This table doesn't take into account partial agreement, which often occurs, or changes to methods and interpretation by human coders that can lead to differences.

4. Implementation at the Census Bureau and the SSA

The move to Version 2 of the Autocoder in February 2006 reduced the coding differences to virtually zero, and also kept the coding rate at both agencies consistently at 60 to 61 percent. Differences continue to be monitored to ensure programs at both agencies are assigning high quality codes.

The SSA and the Census Bureau both signed a service level agreement in April 2006 for one year. Terms of this agreement include:

- The Autocoder program is to be run using SAS version 9.1.
- The Census Bureau will be responsible for the programming of the SAS Autocoder and the accompanying dictionaries.
- A goal of 70 percent coding is to be set, with a move to that level contingent upon acceptable quality to both agencies.
- Updates and modifications to the program will be implemented on an "as needed" basis and when IRS data quality issues arise. The Census Bureau will implement major enhancements no more than once per year. Dictionary changes will occur no more than four times per year. Members of a change control board that includes Census Bureau and SSA staff members must approve all changes.
- Both the Census Bureau and the SSA will be responsible for validating the quality of the assigned codes to meet their agency goals.

By signing a service level agreement, both agencies define and agree to their specific roles. It also encourages inter-agency cooperation and avoidance of unnecessary changes while pursuing improvement in quality and coding rates.

5. Quality Control (QC) Process for the Autocoded NAICS Codes

One of the primary concerns of an automated coding tool is the deterioration of its ability to assign quality codes at a consistent level over time. Maintenance and updates are constantly needed. To ensure the quality of the Autocoder, the Census Bureau developed a quality control process to measure coding error rates on a quarterly basis. Based on needs and uses of industry codes at the Census Bureau, Census Bureau staff selected 42 code categories for measuring coding error. These code categories are at the 2-digit, 3-digit and 4-digit NAICS level (sector to industry group level). Each quarter, an appropriate sample is selected for each of these code categories, using the code assigned by the Autocoder. The total sample size covering all code categories is close to 10,000 EINs. Initial estimates of error rates came from two sources - comparisons of autocoded NAICS codes with codes assigned through the Census Bureau Business and

Professional Classification Survey[2] and by an initial sample of approximately 9,500 EINs coded by the Autocoder and by Census Bureau "expert" coders, who are trained in industry coding and in business descriptions encountered on the Form SS-4.

The quarterly process of assigning an expert code through the QC program begins with two trained Census Bureau coders assigning a NAICS code to each sampled case independently. If a disagreement occurs between the two coders at any level, a separate adjudicator decides between the two potential codes. The autocoded NAICS is not revealed to the two coders or adjudicator, and batches are randomly organized to avoid clusters of similar types of businesses that may indicate comparable codes. Adjudication rates are usually between 20 to 25 percent, and indicate the difficulty in assigning one correct NAICS code to a business establishment. There is usually some partial agreement – for example, one coder may assign a 4-digit partial NAICS code while the other coder may believe the SS-4 information warrants a complete 6-digit code, but within the same 4-digit industry group assigned by the first coder.

The goal of the QC program is to assure that each code category falls within the acceptable range of error. The acceptable range varies by code category and was set with the goal of keeping error rates within ±5 percent of the initial error rate. The QC error rates are determined by dividing the number of disagreements between the adjudicated codes and the autocoded codes, by the total number of cases in the particular code category. If a certain code category has an error rate above its assigned upper tolerance, our practice was to take note and wait for the review of a second sample. If the error rate for a second sample is also above tolerance, Census Bureau staff investigates the code category to determine how to bring the error rate back within tolerance. This usually involves modifying several dictionary entries to ensure correct coding. The final, most drastic step is to reduce the coding rate for the specific code category so the error rate is acceptable until a solution can be found. This step was never taken in the first two years of production.

[2] The Business and Professional Classification Survey is a quarterly mail-out survey of a sample of new business births to determine potential inclusion in several Census Bureau economic surveys covering retail trade, wholesale trade and service industry sectors. The survey requests information to assist in assigning a NAICS code.

We discovered early in the process that for a couple code categories, the initial error rate was either too low or too high. Therefore, we later updated the expected error rates and tolerances so as to represent results from more than one sample.

6. Making Adjustments to the Autocoder

Incoming SS-4 data and the assignment of NAICS codes is not static in its nature. Data quality can change over time, businesses change, and how we assign NAICS codes also changes. Steps are necessary to keep the Autocoder program current with the changes. After rolling out Version 1 of the Autocoder, the Census Bureau and the SSA enacted some small improvements:

- Edited the name field for sole proprietorship records to remove personal names that could possibly lead to inaccurate code assignments.
- Defined list of special characters to remove for the automated coding process.
- Revised or added dictionary entries.
- Handled IRS changes to data input.

A special challenge was encountered with revising or adding dictionary entries. The original dictionary entries were created through an automated process based on clerical coding by the SSA of over 4 million records. To make a change, we would be basically modifying the work of the coding staff. However, this is acceptable considering that the perception of types of businesses and their descriptions may change from the original coding by the SSA. This still leaves the challenge of how to update the counts and frequency percentages that are an important part of the coding dictionaries.

To modify the dictionaries, Census Bureau staff identified potential name or description tokens to include in the dictionary revisions, and selected a sample of EINs containing the token. This sample was provided to expert coders to decide if the revised code was agreeable for the EIN using a 'Yes'/'No' flag for each sampled EIN. The percentage of EINs assigned a 'Yes' flag for a particular token or dictionary revision became the new frequency percentage for the new entry. If the percentage was below 40 percent, the Census Bureau did not make the revision. This happened for several potential revisions, including several entries related to 'Investments'. Census Bureau staff wanted to map 'Investments' to '523000' and not to the dictionary entry of '531000', which

contains real estate investments. In selecting the sample and assigning codes it was discovered that most occurrences of the token 'Investments' are associated with real estate investments, so an update was not advised. A similar situation occurred with 'Hospital', when it was discovered that more businesses with 'Hospital' in the business name are animal hospitals than are hospitals serving humans. The Census Bureau is looking into modifying the coding dictionaries to include entries such as 'Investments', but with an additional exclusion field of 'Real Estate', so that if 'Investments' appears in description and 'Real Estate' does not, the accompanying code would be '523000'. The same could be done with the combination of 'Hospital' and 'Animal' or 'Hospital' and 'Pet'.

The IRS made some changes to the processing of the Form SS-4 and in the description information entered in Item 14 and Item 15 on the Form SS-4 (see Figure 1). These changes were made as an improvement, but required modifications to reading input data for the automated coding program. For example, one change involved the entering of 'None' into an Item 14 description field where IRS previously entered a NAICS sector-level title that corresponded with the checkbox entry for Item 14 (i.e., 'Construction'). IRS wanted to eliminate the unnecessary duplication, so the entry of 'None' made sense. But to be consistent, the Census Bureau adjusted their program to replace the 'None' with the checkbox description to revert to the input best handled by the Autocoder.

7. Coding Rates

The Autocoder will consistently assign a NAICS code to over 80 percent of new EINs, but the quality of some codes will not meet the standards of the Census Bureau and the SSA. The score function gives a good indication of how likely a code assigned by the Autocoder would agree exactly with a code assigned by a coding expert. In general, the higher the score, the more likely the code is correct. At inception, it was determined that an automated coding rate of 60 percent, with clerical coding for the remaining 40 percent, gave overall coding accuracy close to that of 100 percent clerical coding. To reach 60 percent coding, the Census Bureau set a score cut-off of 0.534 for Version 2. Without maintenance it was expected that the coding rate would slowly drop over time due to deterioration, but improvements such as matching part of the full description to the full description dictionary, model improvements, dictionary improvements, and automatic fill-in items enacted by the IRS led to a slight increase in the coding rate when using a score cut-off of 0.534. The following table displays coding rates by quarter:

Table 2. Form SS-4 Record Counts and Coding Rates by Quarter

Quarter	Total SS-4 Records Received	Autocoded SS-4 Records	Coding Rate (%)
2004Q3	653,731	388,133	59.4
2004Q4	652,721	393,581	60.3
2005Q1	800,068	486,612	60.8
2005Q2	793,922	473,233	59.6
2005Q3	666,317	391,918	58.8
2005Q4	752,436	452,397	60.1
2006Q1	871,378	531,204	61.0
2006Q2	845,909	516,408	61.1
2006Q3	725,174	440,829	60.8
2006Q4	709,673	433,304	61.1

Table 2 shows how the move to Version 2 of the Autocoder slightly improved the coding rate starting in 2006Q1.

8. Increasing the Coding Rate

With approval by the Census Bureau and the SSA after several quality control sample reviews at both agencies, the coding rate was raised to 70 percent (by decreasing the score cut-off to 0.414) for most code categories included in the quality control review. Nine code categories were kept at the score cut-off of 0.534 because of consistent error rate failures in quality control. One of the primary reasons for moving to a higher coding rate was to reduce clerical coding costs, but it was still necessary to ensure that quality stayed at acceptable levels. The Census Bureau and the SSA implemented the change with the first SS-4 data received in 2007.

9. Future Research and Improvements

Based on experience, the automated classification of new businesses into a NAICS category requires constant attention. As a result of this attention, the Autocoder process enacted by the Census Bureau and the SSA has maintained its coding ability, and even improved coding since its rollout to production in 2004. The automated coding process has reduced the overall costs of assigning NAICS codes to new businesses by a considerable amount. The large clerical coding costs at the SSA (and shared by the Census Bureau) was reduced by over $1.2 million annually, but with some increase in research and development costs at the Census Bureau to maintain and improve the Autocoder, in addition to the quality

control work. From the start of production, the developers wanted to pursue any significant changes to the program only if it was cost-effective. If substantial work were necessary to increase the coding rate by one or two percent, it would not be a good usage of time and budget. Kearney and Kornbau documents some research into potential improvements based on a sample of cases left un-coded by the Autocoder.

The IRS, SSA and the Census Bureau formed a team in 2006 to develop a new on-line version of the Form SS-4 for filers who use the Internet in applying for an EIN. At least one-half of all SS-4 filers currently use the Internet application. The IRS created a prototype form that includes asking more probing questions based on a filer's response in order to get better information, and a better business description.

For example, if the filer uses a checkbox to indicate that they are involved in real estate, a question pops up to provide the following choices:

1. Rent/lease property you own
2. Use capital to build property
3. Sell property for others
4. Real Estate Management
5. Other (please specify activity)

Based on their response to this question, the filer may get an additional follow-up question. If the filer selects 'Rent/lease property you own' they will get an additional list of choices pertaining to the type of properties:

1. Residential real estate rentals
2. Real estate commercial renting
3. Other real estate rental

In this manner, a series of questions leads to a more specific business description and the assignment of a more detailed NAICS classification.

The affect of this new method of collecting information about the type of business will not preclude the use of the Autocoder, but it will in many cases lead to the automatic assignment of a NAICS code outside of the Autocoder. For other businesses, the applicant will need to key in additional information that the Autocoder will use in making an assignment. It is intended that industry coding for the new Internet SS-4 will combine the Autocoder and automatic classifications with an overall improvement in classification. The new Form SS-4 is scheduled to become available in July 2007.

Several other improvements are under investigation at the Census Bureau. One is to build coding dictionaries using the SS-4 business name and description combined with NAICS codes from sources other than the SSA clerical codes. The advantage of this approach is that we don't need to rely on clerical coding to update dictionaries or create new dictionaries, which is important to eliminate clerical coding costs when the classification system goes through a considerable change. One drawback in using other sources is that the coding dictionaries built from those sources carry along any confidentiality requirements, such as those specified in U.S.C. Title 13. If the Census Bureau used industry codes from the Economic Census or other surveys to build the dictionaries they could not be used by SSA, which cannot have access to Title 13 data. However, the dictionaries could be used at the Census Bureau. Efforts are also slated to use a string comparator to allow for inexact matches to the coding dictionaries so that misspellings – such as 'Restaurant' compared to 'Resturant' – and other minor variations can still result in a match.

The coding dictionaries will need to be updated with 2007 NAICS codes by January 2008 to convert to the new classification system. It is not expected that the conversion will cause significant changes overall, but it can have a high impact for certain code categories including the information and real estate sectors where the 2002 to 2007 NAICS revision is most significant. It's possible that the coding rate will slightly decrease to allow for sufficient clerical coding to re-build affected dictionary entries with the proper NAICS codes, counts and frequency percentages that are integral to the Autocoder.

10. Acknowledgments

The authors would like to acknowledge the contributions of Franklin Winters, Paul Hanczaryk, Carol Comisarow, Anne Kearney, Leann Weyl, Jeff Pearson, Scott Handmaker and James Restivo at the Census Bureau for their contributions to ensuring that the rollout of the Autocoder was successful at the Census Bureau and the SSA. The authors would also like to acknowledge Larry Katz, Penny Metz, Margie Aviles and Fred Galeas at the SSA for their contributions and advice in implementing the Autocoder at the SSA and in pushing for quality improvements.

11. References

[1] 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)

[2] 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 April 2, 2007 (to be released)

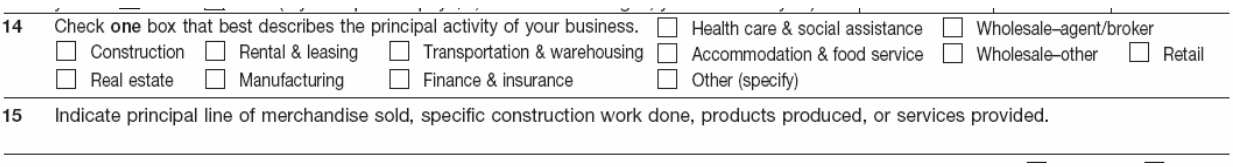
[3] Kearney, Anne T., Kornbau, Michael E., (2005), "An Automated Industry Coding Application for New U.S. Business Establishments." Proceedings of the American Statistical Association, Business and Economic Statistics Section, 2005.

[4] 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

Table 1. Logistic Regression Estimated Match Probability (Score) by Agreement Level Between the Autocoder NAICS and the Clerically Coded NAICS

Score	Number of Records	Agreement Rate Between Autocoder and Clerical Code (in percent)		
		Exact Agreement	Partial Agreement	Some Level of Disagreement
0.95 to 1.00	103,675	96.8	1.6	1.4
0.90 to 0.95	95,051	92.0	4.1	3.4
0.85 to 0.90	63,580	87.5	6.3	5.4
0.80 to 0.85	59,596	82.8	9.1	7.1
0.75 to 0.80	51,866	75.4	13.4	10.0
0.70 to 0.75	37,114	72.4	13.0	13.0
0.65 to 0.70	27,823	68.4	12.9	16.8
0.60 to 0.65	23,824	62.6	15.3	19.7
0.55 to 0.60	25,055	59.2	15.7	22.4
0.50 to 0.55	30,024	53.9	19.9	23.3
0.45 to 0.50	35,047	46.6	26.2	24.0
0.40 to 0.45	36,546	41.6	25.2	29.0
0.35 to 0.40	32,005	37.4	27.1	32.4
0.30 to 0.35	27,265	32.7	28.1	37.0
0.25 to 0.30	18,801	29.2	26.1	43.3
0.20 to 0.25	14,803	24.7	29.0	45.5
0.15 to 0.20	19,905	18.3	33.6	47.3
0.10 to 0.15	40,842	12.7	39.0	47.3
0.05 to 0.10	26,572	10.5	31.2	56.9
0.00 to 0.05	1,185	9.5	20.2	69.8
Total	770,579			

Figure 1. Part of IRS Form SS-4: Items 14 and 15 (see http://www.irs.gov/pub/irs-pdf/fss4.pdf for entire form)


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.