Creating an Automated Industry and Occupation Coding Process for the American Community Survey

Matthew Thompson[1], Michael E. Kornbau[1], Julie Vesely[1]
[1]U.S. Census Bureau, 4600 Silver Hill Rd, Suitland, MD 20746

Abstract
Every year the American Community Survey (ACS) collects data on millions of individuals on a variety of topics, including the industry and occupation in which individuals work. These data are collected in the form of a series of open-ended questions. Clerical coders take these open-ended responses and assign a numeric code for the industry and occupation.

The coding of industry and occupation for the ACS is a massive operation with over 2 million industry and occupation codes assigned every year. To reduce costs, a process was developed to assign industry and occupation codes using the open-ended responses and a logistic regression model. This paper discusses the development of this model and the early results. It is expected that beginning in 2012, 56% of industry codes and 43% of occupation codes will be assigned through this automated coding process for the ACS.

Key Words: Logistic regression, ACS, automated coding, industry coding, occupation coding

1. Introduction

The Industry and Occupation Automated Coding project started as a result of a call by the Director of the Census Bureau for projects that could increase the efficiency of operations at Census. Every year the American Community Survey (ACS) collects data on millions of individuals. Among the data items collected is information on the industry and occupation in which individuals work. These data are collected in the form of open-ended questions. Clerical coders then assign 4-digit industry and occupation codes indicating a specific industry and a specific occupation. In this way, all “chefs” (regardless of how the respondent writes his or her response) will ideally be given the same occupation code, in this case 4000.

Every year over 2 million industry and occupation codes are assigned, and this number continues to grow. Each of these cases is then reviewed by a clerk and assigned a code. A number of ideas have been presented to increase the efficiency and accuracy of this

process: from providing clerks with a list of employer names and possible industry and occupation codes gathered from administrative data, to using administrative data in the ACS coding quality control process.

Among the proposed ideas was the creation of an automated coding process that could have the potential to greatly reduce the amount of clerical coding required to produce industry and occupation estimates. The proposed goal was to create an automated process that could successfully code both industry and occupation for 55% of the incoming ACS data with error rates comparable to those of current clerical process, leaving the remaining 45% to be clerically coded. Ideally this process would also code the majority of the straightforward cases, allowing clerical coders to focus their efforts on coding the more complex responses.

A similar problem was experienced when facing the rising costs of clerically coding industry write-in information on the Internal Revenue Service (IRS) Form SS-4. The Form SS-4 is filled out by all businesses requesting an employer identification number (EIN) from the IRS and the industry information supplied on this form is used by the Census Bureau, IRS, and the Social Security Administration (SSA) as the earliest indication of the industry in which a business operates. In an effort to increase the efficiency of this coding process an automated coding system was developed. In approaching the creation of an automated coding process for the ACS industry and occupation write-ins, the methods developed by Kornbau and Kearney (2005) for coding the Form SS-4 industry write-in responses was used as the starting point.

The remainder of this paper will discuss the specifics of developing and refining this automated coding process for industry and occupation. Section 2 will cover the development of “data dictionaries” – parameter files created from previously clerically coded data. These dictionaries are used to establish relationships between certain words that appear in the open-ended responses and specific industry and occupation codes. The result of applying the relationships established in the data dictionaries to uncoded industry and occupation data is a listing of possible codes. Section 3 will discuss the creation of a logistic regression model designed to determine which of these codes is the “best.” Sections 4 and 5 will cover the adjustments that were made to the model in order to improve its accuracy, as well as, answering the question of how much data to code through this automated process. The remainder of the paper details the use of the Industry and Occupation Autocoder in production and the efforts to maintain and improve upon the initial results.

2. Data Dictionaries

2.1 Dictionary Creation

Data from the 2010 ACS survey year were used in creating the Industry and Occupation (I&O) Autocoder. Of the approximately 2.3 million industry and occupation responses from the 2010 ACS, 1.5 million were randomly selected to be used in creating the

dictionaries and models required for automated coding with the remaining records used to validate the performance of the I&O Autocoder.

The first step in creating and implementing the automated coding process is producing a set of data dictionaries. These dictionaries contain words or phrases commonly found in the ACS industry and occupation write-ins and the specific industry and occupation codes with which they are most commonly associated. To begin preparing a set of data dictionaries, we first must decide which data fields are being used to assign the industry and occupation codes. In our case, employer name (INW2) and industry description (INW3) are, generally speaking, used to assign industry codes, and occupation description (OCW1) and job duties (OCW2) are used to assign occupation codes.

Now we address each of these chosen fields individually; for each of these four fields a set of data dictionaries will be created. First, each field is parsed out into individual "wordbits." These wordbits are single words or strings of consecutive words contained within the larger field. For example,

Example 1:

OCW1 = "WAITER IN RESTAURANT"

Wordbits = "WAITER", "IN", "RESTAURANT", "WAITER IN", "IN RESTAURANT", "WAITER IN RESTAURANT"

Each of these wordbits would be a potential entry in the OCW1 data dictionaries. There will be a separate dictionary for one-word wordbits, another for two-word wordbits, and another for wordbits constituting the full write-in entry. This would result in a total of 12 dictionaries: 3 INW2 dictionaries, 3 INW3 dictionaries, 3 OCW1 dictionaries, and 3 OCW2 dictionaries. It is possible to create additional dictionaries by including wordbits of 3 or more words; however, the number of dictionaries quickly becomes cumbersome.

To create dictionaries, we determine which of the potential dictionary entries are commonly associated with a particular industry or occupation. To determine whether or not "WAITER IN RESTAURANT" should be included in the OCW1 full write-in dictionary, gather all occurrences in the data where OCW1 = "WAITER IN RESTAURANT" and observe the different occupation codes that were assigned.

Example 2:

Wordbit	*OCC Code*	*OCC Count*	*Wordbit Count*	*Frequency Percent*
WAITER IN RESTAURANT	4110[1]	35	45	0.778
WAITER IN RESTAURANT	4120[2]	7	45	0.156
WAITER IN RESTAURANT	4060[3]	3	45	0.067

[1] Waiters and waitresses

[2] Food servers, non-restaurant

[3] Counter attendants, cafeteria, food concession, and coffee shop

In this example, “WAITER IN RESTAURANT” would be associated with the occupation code 4110, but should it be included in the OCW1 full write-in dictionary?

At this point, some criteria need to be set to determine which potential entries should be included in the data dictionaries. After some trial and error and using the criteria of the SS-4 Autocoder as a guide, it was decided that all wordbits that occur at least 30 times in our data and that are associated with a single industry/occupation code at least 50% of the time would be included in the data dictionaries. All other wordbits are excluded with the exception that in the case of the full write-in dictionaries, entries must only appear in our data 15 times. In the case of our example, “WAITER IN RESTAURANT” would be included in the full write-in OCW1 dictionary because it appeared in our data 45 times and was coded to 4110 in 78% of those occurrences. In this way, the final set of dictionaries is created.

2.2 Cross-Code Dictionary Entries

Adding to the complexity of the I&O Autocoding project is the fact that two somewhat related fields are being coded for each record entering the I&O Autocoder process. While, as will be shown later, the coding of industry and occupation are treated essentially as two separate coding processes by the I&O Autocoder, in practice clerks often look at the industry write-ins for information when assigning occupation codes, and vice versa. Additional entries, which will be referred to as cross-code entries, were added to the dictionaries in order to incorporate this practice in some way into the automated coding process.

Cross-code entries are wordbits obtained from the industry write-ins, INW2 and INW3, that are indicative of a particular occupation, or occupation write-ins that indicate a particular industry. The process for adding cross-codes to the dictionaries is exactly the same as that for ordinary dictionary entries with two exceptions. When creating cross-code entries for the INW2 or INW3 fields, wordbits that were commonly associated with a particular occupation were identified. Similarly for the OCW1 and OCW2 cross-codes, wordbits were identified that were indicative of working in particular industries. In addition, the inclusion criteria were changed for determining whether a cross-code should be allowed to enter the dictionaries. Like our more typical dictionary entries, cross-codes were required to appear at least 30 times in our data (15 times for full write-in cross-codes), but for a cross-code to be included in the data dictionaries it was decided that it should be associated with a particular industry/occupation at least 75% of the time.

2.3 Baseline Modeling and Disagreement Rates

The coding procedure starts out very much like dictionary creation. The verification data is taken and each record has INW2, INW3, OCW1, and OCW2 parsed into one-word, two-word, and full write-in wordbits. These wordbits are then matched to the data dictionaries (one-word INW2 wordbits are matched to the one-word INW2 dictionary,

and so on) with any matches to the dictionaries returning the industry and/or occupation code most frequently associated with that wordbit.

In order to make a code assignment from the possibly numerous codes produced by matching to the dictionaries, a model of some sort must be used to select the "best" code from among the list of possible codes. For the purposes of this project, the ultimate goal is to use a logistic regression model to determine which of the possible codes is most likely to match the code assigned clerically. However, a very simple baseline coding model was created to serve two purposes: first, to get some early sense of the quality of our dictionaries and also, to use as a basis for comparisons for the planned logistic regression models.

This baseline model simply sums up the percentages associated with each entry returned from the dictionary match. All wordbits that returned the same code had their percentages summed and the code with the highest sum will be the automated code for the record. This code can then be compared to the clerically assigned code and rates of disagreement between the two codes can be calculated.

In the case of the percentages returned from cross-code entries, it was decided that simply summing the frequencies returned from these entries may not be optimal because this gives industry write-ins just as much weight in assigning occupation as the occupation write-ins. Therefore, the cross-code percentages were weighted down by multiplying by a factor between 0 and 1. The whole range from 0 to 1 was tested using increments of .01. For industry it was determined that a factor of 0.28 resulted in the lowest disagreement rates when the assigned code was compared to the clerically assigned code, and for occupation the optimal factor was determined to be 0.68.

Using this simple model, the records – each with it's associated automated code – were sorted such that the records with the highest "score" were at the top of the file and disagreement rates were calculated. Ultimately only codes with a sufficiently high score will be assigned so we wish to examine the disagreement rates across various levels of coding. Table 1 shows the disagreement rates at various levels of coding. For example, when we only consider those 30% of records with the highest scores, the automated industry code disagrees with the clerically assigned code 2.35% of the time. While if instead we code the 40% of records with the highest scores, the disagreement rate for automated industry coding increases to 3.25%.

Table 1: Baseline Disagreement Rates (DR)

Coding Rate	*Industry DR*	*Occupation DR*
30%	2.35%	7.33%
40%	3.25%	7.52%
50%	4.63%	11.69%
55%	5.22%	14.73%
60%	6.48%	16.39%

In practice, ACS clerical coders are required to maintain an error rate of 5% or lower as determined by a quality assurance process run on all clerical coders. This 5% error rate served as our goal for the I&O Autocoder as well. As can be seen in Table 1, using just this simple model it is possible to code industry for approximately 55% of the cases processed through the autocoder. However, less than 30% of records would be assigned an occupation code.

3. Regression Model

In order to improve upon the results from the baseline model, two logistic regression models were developed – one for assigning industry codes and a second model for assigning occupation codes. When processing records through the I&O Autocoder, industry codes will be assigned first using the industry model, and then occupation will be assigned using the occupation model. In this way industry and occupation assignment are treated essentially as two distinct processes with two exceptions. The first is the use of cross-code entries. In this way occupation write-ins can impact industry assignment and vice versa. In addition, the industry code assigned using the first model will be used as an input into the occupation model.

3.1 Independent Variable Selection

When creating a logistic regression model, one of the most important steps is deciding what information to include in the model. In our case, two important decisions had to be made. First, we needed to decide which data fields to include in the modeling process, and once this decision was made, we had to determine how to best categorize those variables for use in the model.

Deciding which data fields to include was fairly straightforward given Census experience with the Form SS-4 Autocoder and the data available to clerical coders. Variables included in the modeling process were gathered from the data dictionaries (e.g. wordbit count, code frequency percent), directly from ACS response data (e.g. sex, age), and calculated from ACS response data (e.g. number of words in INW2). The ACS data used are exactly the same as that seen by clerical coders when assigning industry and occupation codes. This allows the model to most accurately mimic the clerical coding process.

3.2 Independent Variable Categorization

Having determined which data fields to include in the modeling process, the second decision that needed to be made was how to best categorize these data into binary independent variables. Most of the variables need to be converted into multiple binary variables before modeling can take place.

This recode is extremely straightforward in some cases (e.g. “sex” can easily be recoded into two binary variables). However, for many variables, determining the best way to categorize the data into multiple independent variables is a challenge. In addition, the

way in which the variables are categorized may need to be different when modeling industry codes as opposed to occupation codes.

For instance, what is the best way to break age into a reasonably small number of categories? What impact does the age of a respondent being 25 have on our ability to code industry and how might that differ from someone who is 40? Will the categorization differ when considering occupation? To answer these questions without requiring an extremely intimate knowledge of the data, agreement rates for both industry and occupation were studied across the complete range of values for each of the variables to be categorized.

As an example when categorizing AGE for industry modeling, industry agreement rates were calculated for each value of AGE in our data. The total number of wordbit matches to the dictionaries for each age was calculated, as well as the total number of wordbit matches for which the associated industry code agreed with the clerically assigned code. The ratio of matches that agree with the clerical code to the total number of matches is the agreement rate.

Once we have agreement rates for each age value, we can begin to categorize the data into age groups. The question we want to answer is “Are there certain age groupings for which the dictionary matches are more (or less) accurate when compared to other age groups?” By grouping these ages together, our model will be more likely to assign higher parameter values to ages for which dictionary matches often match the codes assigned by clerks and lower parameter values to ages for which dictionary matches are more rarely in agreement with the clerical codes.

Agreement rates were calculated for both industry and occupation for all variables to be included in the models. These agreement rates were reviewed and each variable to be included in the models was broken down into a set of binary variables. The one exception to this was the frequency percent (FREQPCT) used in the baseline model. FREQPCT remained a continuous numeric variable in both the industry and occupation models.

3.3 Regression Modeling and Disagreement Rates

Once the data items to be input into the logistic regression modeling process were decided upon, the logistic procedure in SAS was used with the stepwise variable selection option to add and remove potential independent variables from the model as they meet or fail to meet inclusion criteria.

PROC LOGISTIC was run on the data with the dependent variable, Y, set to 1 if the industry code returned from a dictionary match agreed with the clerically assigned industry code. Out of the 155 possible independent variables supplied, 101 were included in the final industry logistic regression model.

Using the data set aside for verification, the resulting model was used to assign a score, called PHAT, to each of the industry codes returned from the match to the data dictionaries. This PHAT value represents the probability of the industry code returned from the dictionary match agreeing with the code assigned by a clerk. The industry code with the highest value of PHAT was then assigned to each record. Using the codes assigned from the logistic regression model, disagreement rates were calculated at various levels of coding as seen in Table 2.

Table 2: Industry Disagreement Rates Using the Industry Logistic Regression Model

Coding Rate	*Baseline DR*	*Regression DR*
30%	2.35%	0.97%
40%	3.25%	2.24%
50%	4.63%	3.95%
55%	5.22%	4.47%
60%	6.48%	5.65%

PROC LOGISTIC was then run a second time on the data using the occupation variable categorization with the dependent variable, Y, set to 1 if the occupation code returned from a dictionary match agreed with the clerically assigned occupation code. Out of the 153 possible independent variables supplied, 79 were included in the final occupation logistic regression model.

Using the data set aside for verification, the resulting model was used to assign occupation codes in much the same way as industry codes were assigned. The model was used to assign each dictionary match a PHAT and the match (and associated occupation code) with the highest value of PHAT was assigned. Table 3 shows the disagreement rates at various levels of coding.

Table 3: Occupation Disagreement Rates Using the Occupation Logistic Regression Model

Coding Rate	*Baseline DR*	*Regression DR*
30%	7.33%	4.86%
40%	7.52%	6.29%
50%	11.69%	10.00%
55%	14.73%	12.79%
60%	16.39%	15.73%

4. Hardcodes

Once the parameter values for the two models had been determined, we could begin analyzing the results of the ACS Industry and Occupation Autocoder in more detail. In

order to give our investigation some direction, a simple chi-square statistic was calculated for industry assignment as:

$$X_I^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories, O_i is the observed frequency in the i[th] category, and E_i is the expected frequency in the i[th] category. Here the categories are the industry codes assigned. The expected frequency for each industry code is the number of times the clerks assigned that code and the observed frequency is the number of times the autocoder assigned that code.

Similarly, a chi-squared statistic was calculated for occupation assignment as:

$$X_O^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j},$$

where m is the number of occupation codes assigned, E_j is the number of times the clerks assigned the j[th] occupation code, and O_j is the number of times the autocoder assigned that code.

Using this information, the industries and occupations that contributed the most to this statistic were identified. In other words, the "i" and "j" values that produced the largest values of $\frac{(O_i - E_i)^2}{E_i}$ and $\frac{(O_j - E_j)^2}{E_j}$ were flagged for more detailed review.

Having identified industry and occupation assignments that often are in disagreement with the clerical codes assigned, we went further and identified the clerical codes most often assigned when there was a disagreement. For example, Example 3 shows the clerical codes most often in disagreement with an automated industry assignment of 7860 – elementary and secondary education.

Example 3:

Automated Industry	*Clerical Industry*	*Industry Description*	*Number of Occurrences*[4]
7860	7890	Other schools and instruction	153
	7870	Colleges and universities	149
	8470	Child day care services	86
	9480	Administration of human resources programs	54

[4] Counts are from verification file of more than 800K ACS records from 2010.

At this point, a thorough examination of the ACS microdata within each of these targeted industries and occupations took place. Subject area experts were brought in to comment on hard-to-code cases and a series of direct assignments – referred to within the autocoder as "hardcodes" – were established to correct the common and straightforward errors. In this way, significant improvements were made to the quality of the output from the Autocoder, as seen in the Tables 4 and 5.

Table 4: Industry Disagreement Rates Using the Final Industry Model with Hardcodes

Coding Rate	*Baseline DR*	*Original Regression DR*	*Final Regression DR*
30%	2.35%	0.97%	1.01%
40%	3.25%	2.24%	2.12%
50%	4.63%	3.95%	3.96%
60%	6.48%	5.65%	5.39%

Table 5: Occupation Disagreement Rates Using the Final Occupation Model with Hardcodes

Coding Rate	*Baseline DR*	*Original Regression DR*	*Final Regression DR*
30%	7.33%	4.86%	4.49%
40%	7.52%	6.29%	5.30%
50%	11.69%	10.00%	8.35%
60%	16.39%	15.73%	13.43%

5. Determining Production Coding Rates

Using the initial goal of no more than 5% disagreement between automated codes and clerical codes resulted in coding approximately 56% of the industry cases processed and 40% of the occupation cases processed – short of our original goal but still a dramatic reduction in the clerical workload. But does coding 56% of industry and 40% of occupation cases return codes of equal quality to that currently being produced by the clerical coders? While the 5% error rate is the minimum requirement, many clerks may maintain a lower error rate.

In order to verify the use of these coding percentages in production, a sample of 2,000 cases for which the automated code disagreed with the clerical code were selected from the verification dataset – 1,000 industry assignments and 1,000 occupation assignments. These 2,000 cases then went through a second round of clerical coding. It is important to note that each sample was selected to span all score values output by the logistic regression models. The most experienced clerks available were selected to code these cases and they were instructed to take as much time to research each case as necessary to assign the most accurate industry and occupation codes possible. Once this second round

of clerical coding was complete, each of the 1,000 industry assignments had been assigned 3 industry codes and each of the 1,000 sampled occupation assignments had been assigned 3 occupation codes: a clerical code, an automated code, and a second clerical code (which will be referred to as the expert code).

An "agree flag" was created for both the automated codes and the original clerical codes. If the automated code agreed with the expert code then the agreement flag for the autocode was set to 1. If the automated code did not agree with the expert code then agreement flag for the autocode was set to -1. Similarly, an agreement flag for the clerical code was set to 1 or -1 depending on whether or not it agreed with the expert code. Then a variable called DIFF was created as the difference between the two agreement flags. For example,

Example 4:

If
Automated Occupation: 2300
Clerical Occupation: 2310
Expert Occupation: 2310

Then
Automated Agreement Flag = -1
Clerical Agreement Flag = 1
DIFF = (-1) – 1 = -2

The 1,000 sampled industry records were then sorted by PHAT (highest to lowest) and a cumulative total for DIFF was calculated. Automated codes with high PHATs were more often in agreement with the expert code and thus the cumulative DIFF was positive. However, as the PHAT values dropped clerical codes agreed more often with the expert code and the cumulative DIFF dropped. The point at which the cumulative value of DIFF becomes consistently negative indicates the level where automated coding is starting to give lower quality results than the clerical coding. The PHAT value where this shift from positive to negative occurs in the cumulative value of DIFF becomes the PHAT cutoff for automated coding. In this way, the coding rate was set so that automated coding with that rate gave a similar level of quality to that of 100 percent clerical coding.

After analyzing the results of the expert coding, the score cutoffs were set such that 56% of cases input into the industry model would receive an industry code and 43% of cases input into the occupation model would receive an occupation code. Note that the PHAT cutoffs are held constant here, not necessarily the coding rates. While the PHAT cutoffs currently correspond to coding 56% of industry and 43% of occupation, these values may fluctuate somewhat over time. This also allows for the coding rates to increase as improvements are made to the autocoder process.

Also, note that the occupation coding rate of 43% will result in a disagreement rate of greater than 5% on our verification dataset. This is because though the switch from positive to negative in the cumulative DIFF occurred at approximately 40% for occupation coding, the cumulative DIFF was only slightly negative until a large drop off at 43%. While coding records using the autocoder between the 40% and 43% cutoffs will result in a slight decline in quality, this decline will be small and a significant amount of additional data will be coded.

Tables 6 and 7 below, show the disagreement rates at these new cutoffs. When coding 56% of industry records input into the I&O Autocoder, we see a disagreement rate of 4.53% on the verification dataset when comparing the clerically assigned codes to the autocodes. In the case of occupation coding, a disagreement rate of 5.86% is observed.

Table 6: Industry Disagreement Rate at Final Coding Rate

Coding Rate	*Baseline DR*	*Original Regression DR*	*Final Regression DR*
30%	2.35%	0.97%	1.01%
40%	3.25%	2.24%	2.12%
50%	4.63%	3.95%	3.96%
56%	--	--	**4.53%**
60%	6.48%	5.65%	5.39%

Table 7: Occupation Disagreement Rate at Final Coding Rate

Coding Rate	*Baseline DR*	*Original Regression DR*	*Final Regression DR*
30%	7.33%	4.86%	4.49%
40%	7.52%	6.29%	5.30%
43%	--	--	**5.86%**
50%	11.69%	10.00%	8.35%
60%	16.39%	15.73%	13.43%

6. Implementation

In March 2012, the ACS Industry and Occupation Autocoder was implemented into production processing of the 2012 American Community Survey. Through the first several months of production the Autocoder has been coding approximately 56% of industry records and 43% of occupation records, as expected. This results in approximately 30% of cases processed receiving both an industry and occupation code with another 40% receiving only one code – industry but not occupation, or vice versa. Clerical coding continues on both those records that received neither an industry nor an occupation code and those records that received only a single code.

It is important to note here, that while the original goal of 55% coding of both industry and occupation was not met, the quality of the code assignment is much higher than that original goal would allow. The original goal was based upon results from a similar autocoding process used for the 2000 Census. And while coding rates of 55% were achieved for the 2000 Census, upon further review it was discovered that disagreement rates were approximately 10% for industry and 13% for occupation as detailed by Gillman and Appel (1999). Similar results could have been achieved for the ACS I&O Autocoder, but it was decided that the additional coding was not worth the loss in quality and so the approximate 5% disagreement rate was maintained.

7. Quality Control

As a means of maintaining and possibly improving the quality of the I&O Autocoder, a quality control (QC) process has also been developed. On a monthly basis, samples are drawn and coded by clerks in a manner similar to the "expert coding" procedures discussed in Section V. The samples are selected so as to cover all industries and occupations assigned and focuses specifically on codes for which we know the disagreement rate with clerical coding is high. This QC design is based in large part on the QC process developed by Kornbau et al. (2007) for the Form SS-4 Autocoder.

Once coding of the QC sample is completed disagreement rates can be calculated and any codes (or code groupings) that have a consistently high disagreement rate – as compared to the baseline rates from the verification dataset – will be subject to a detailed review. In this way erroneous or missing entries in the data dictionaries can be identified and corrected or added as needed.

8. Future Research and Improvements

While the early results of the ACS I&O Autocoder are promising, a number of opportunities for improvements – both to the autocoder and the coding process as a whole – have been identified by the development team. Based upon similar experience with the Form SS-4 coding, the introduction of a drill-down type internet version of the ACS survey form could be extremely beneficial to the hardcoding process. In this scenario, respondents would be presented with a list of very broad industry sectors, and upon choosing an industry sector would be further prompted to select from among a number of industries within that sector. In many instances, this would allow the autocoder to take the auto-filled response and immediately assign an industry code. Occupation codes could be assigned with the help of a similar drill-down process.

In addition, the current version of the autocoder uses raw, unedited data both for dictionary creation and to match new data to the dictionaries. This results in a number of misspelled words being included in the dictionaries as well as a number of words with very similar meanings being entered individually into the dictionaries. Some research could be done on what impact "cleaning" the data before processing would have on

results. This could mean correcting spelling errors, removing prefixes and suffixes from words, and/or combining synonyms into single dictionary entries.

Finally, there is an opportunity to utilize autocoder results in clerical coding. There are many instances in which an industry/occupation code is selected through the automated process, however, the score associated with the code is not high enough to warrant assignment. These cases then continue on to clerical coding. It would be possible to use these unassigned codes to sort the clerical coding files such that all cases "suspected" – by virtue of the unassigned code – of being in a given industry/occupation are grouped together. This could allow clerical coders to work on groupings of very similar cases, focusing their attention on the slight details that might distinguish one from another - details that could be missed if handling the cases individually. This could also have obvious benefits in terms of the speed of the clerical coding operation.

References

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.

Kearney, Anne T., Kornbau, M.E., (2005). "An Automated Industry Coding Application for New U.S. Business Establishments." Proceedings of the American Statistical Association, 2005.

Kornbau, Michael E, Bouffard, J., Vile, M., (2007). "Making Quality Improvements to an Automated Industry Coding Application for U.S. Business Establishments." Proceedings of the Third International Conference on Establishment Surveys (ICES-III), 2007.