AUTOMATED CODING RESEARCH AT THE CENSUS BUREAU[1]

by
Daniel W. Gillman
Martin V. Appel
Statistical Research Division
Bureau of the Census

ABSTRACT

Research on improved methods of performing automated coding is being conducted by the Census Bureau. The Automated Industry and Occupation Coding System (AIOCS) was developed for the 1990 Census Industry and Occupation (I&O) Coding operation. It performed very well. AIOCS has now been certified for use in current surveys I&O coding processes as well. After the 1990 Census, research was begun to improve AIOCS productivity, reduce errors, and understand the algorithm. New methods of performing automated coding are also being analyzed. Some of the new methods can easily be applied to other coding structures. This paper details the research and results which have been done so far.

0. INTRODUCTION

Research on improved methods for performing automated coding is being conducted by the Statistical Research Division (SRD) of the Census Bureau. This research program was begun just after the Industry and Occupation (I&O) coding processing for the 1990 Census started. The Technology Research Staff in SRD designed and developed the system used for the 1990 Census automated I&O coding processing. This system is called the Automated Industry and Occupation Coding System (AIOCS).

There are three arenas where computer coding systems are useful: fully automated such as AIOCS, computer-assisted clerical, and CATI/CAPI[2]. The fully automated systems are used for large batch operations; the computer-assisted clerical systems are used for residual coding of cases which a fully automated system could not decide; and the CATI/CAPI systems are used in the field during interview time. The CATI/CAPI systems will probably be a combination of automated and computer-assisted clerical.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

[2] CATI/CAPI is Computer Assisted Telephone Interviewing and Computer Assisted Personal Interviewing.

The research program has focused on three areas. First, there has been an effort to understand AIOCS algorithm characteristics, improve its productivity, reduce its errors, and apply it to new situations. Second, new methods of performing automated I&O coding are being investigated. This primarily involves giving data to software vendors and analyzing their results. This has also led to investigating methods of general automated coding and clerically-assisted computer coding, especially in the CATI/CAPI environments.

This research will be described in more detail below. A description of each algorithm under consideration will be included with the results of the analysis of that algorithm.

0.1 ALGORITHM ANALYSIS

The analysis of coding algorithms for calculating productivity and errors is performed in a uniform way (Chen, Creecy, and Appel, 1993). This is described as follows:

The industry (occupation) production rate for an automated coding system is the ratio of the number of cases for which it assigns industry (occupation) codes over the total number of cases.   A system assigned code is considered correct when it agrees with a "truth" code assigned by coding experts for the case in question. Which "truth" code to assign to a case is not always apparent, so a disagreement between the computer assigned code and the "truth" code does not always represent an error. Cases for which the computer assigns an incorrect code are called mismatches. The error rate is the number of mismatches over the number of cases with codes assigned. Each automated coding system calculates a score for each code which it produces, a measure of the reliability of that code. The reliability also depends on which code is produced; computer coding systems do better with some code categories than others. In order to control errors, a minimum acceptable score, or cutoff score, is determined for each code category. A computer-assigned code is one where the score is above the cutoff for that code category.

Cutoff scores are determined by examining the cumulative match rate calculated for every case (from some test file) which has been assigned to a particular code category by the computer. Cases within a code category are sorted in decreasing order by score, and the computer code is compared against the "truth" code previously assigned by a coding expert. The cumulative match rate is the ratio of correct cases divided by the number of cases encountered so far in the list (see Attachment 1).

Cumulative match rate distributions for each code category can be graphed by score or case. These graphs can be used to analyze cutoff score calculations or for comparing similarity of data from different sources. Score distributions graphed by case are also used for comparing different data sets. For examples, see figures 1 and 2.

1. AIOCS

In general terms, AIOCS is a natural language translation program implemented as an expert system. It uses an algorithm developed at the Census Bureau by Eli Hellerman (Appel and Hellerman, 1983, and Appel and Scopp, 1987).

There are three main subsystems which comprise AIOCS:

1) The Knowledge Base System (KBS) establishes and maintains a set of coding databases of files and pointer arrays. It consists of a coding database (CDB) of industry and occupation (I&O) descriptor phrases with associated constraints and restrictions, a synonym and abbreviation list, and a lexicon of all descriptor phrase words with measures of their importance for coding.

2) The Coding System (CS) incorporates those functions necessary to enter and classify a respondent's reply.

3) The Quality Measurement System (QMS) measures the reliability of codes assigned by AIOCS. It detects special or new situations requiring improved algorithms or CDB changes.

The CDB in the KBS is actually two separate databases of descriptors: one for industry and the other for occupation. These descriptors were originally extracted from the 1980 Alphabetical Index of Industries and Occupations, the set of descriptions used by clerks for manual coding. Each descriptor has a code associated with it. After the CS selects the best descriptor for describing a response, the associated code is assigned to the case. A more detailed description is given below.

The lexicon consists of all the words which are contained in the descriptors in the CDB. Each word is kept in standardized, compressed form to reduce storage. Also, there are two entropy weights associated with each word, one each for industry and occupation. These weights measure the discriminatory power of each word. The weight goes up if the word is associated with fewer codes and descriptors.

The CS is the software which assigns the codes. For each case, it tries to assign the industry code first then the occupation code. If an industry code cannot be assigned, it tries to assign an occupation code anyway.

The coding procedure can be summarized as follows (Appel, Hellerman, 1983):

1) Each response word is converted to a standardized form. A lexical search is performed and the word's compressed form, weight, and database pointers are located. The words in each response field are sorted by weight.

2) If certain key words appear in the response, a special logical analysis is performed. If this is successful, then I&O codes are assigned without going to the knowledge base and coding is complete. Otherwise, continue to the next step.

3) Retrieve all CDB phrases containing the current key word and calculate a score for each. If any one of the CDB descriptors has a score above the minimum threshold, then sort the phrases by score and assign to the response the codes associated with the phrase having the highest score. If no phrase has a satisfactory score, then select a new key word and repeat this step.

The QMS is the part of AIOCS which makes use of experts. Files of test cases are maintained in databases and these cases are assigned "truth" codes by the experts. As the first step in the measurement process, the files are coded through AIOCS. The experts review the coding results using a benchmark report. Based on these results, they make updates to the CDB and the "truth" codes in the test files. These reviews and updates are designed to improve AIOCS. This process of coding, reviewing, and updating is repeated until the desired error and productivity rates are achieved.

AIOCS was used during 1990 Census I&O Coding processing. Results from the 1990 Census, as measured by the I&O Coding Quality Assurance program, were 57% industry production rate, 37% occupation production rate, 6.2% industry error rate, and 11.8% occupation error rate (Scopp and Tornell, 1991). AIOCS will also be used for I&O Coding processing of current surveys, such as the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). A certification of AIOCS for use in current surveys processing was completed in September 1992 (Jablin, 1992; Gillman, Appel, and Jablin, 1993). The estimates for current surveys, based on results of coding test files for the certification, are 44% industry production rate, 14% occupation production rate, 5.3% industry error rate, and 6.3% occupation error rate. Decennial Census and current survey coding and error rates are different because the decennial automated coder was tuned to the expected error rates of decennial clerical coders and the current survey automated coder was tuned to the error rates of our permanent coding staff.

2. RESEARCH AND RESULTS

The following sections describe the major research efforts which have been applied to AIOCS over the last few years and the results of those efforts.

2.1 DECISION MATRIX ANALYSIS

A decision matrix is a table whose cells represent all the ways AIOCS can make a decision when deciding a case. AIOCS makes decisions by comparing the input fields with pre-coded phrases in a database. The input field which supplies the key word used in the database phrase search and the input field which best matches the database phrase are the parameters which determine the cells in the matrix.

We took the results of coding one of our census test files, the 1980 Large Sample[3], and outputed the necessary data from AIOCS to fill a decision matrix (Appel, 1991).

Four possibilities exist for any assigned code: 1) coded correctly, score above cutoff; 2) coded incorrectly, score above cutoff; 3) coded correctly, score below cutoff; and 4) coded incorrectly, score below cutoff. Cases which fell into category 2 are false positive, and cases which fell into category 3 are false negative. Separate analyses were performed for industry coding and occupation coding (with self-employed occupation tabulated separately), so twelve decision matrices were calculated.

The results for industry coding were very good. Over 76% of all industry codes assigned were decided when the key word came from the industry field. All but 5% of the cases were decided either by logical analysis or when the key word came from a major field. The industry field is the first one used for key words during industry coding. The combination of the key word coming from the industry field and the final score resulting from the comparison between the CDB descriptor and the industry response was responsible for 64.5% of industry production with an error rate of 6.5%. This indicates that the industry response field is the best for deciding an industry code. The industry pseudo-phrase accounted for 22.7% of all codes assigned, however, this resulted in only 16.6% of production. There was an error rate of about 25% for these cases, and they contributed 40% of the total error, i.e. the false positive cases.

For occupation coding, some of the results were similar. Over 66% of all occupation codes were decided when the key word came from the occupation field. Again, all but 6% of the cases were decided either by logical analysis or when the key word came from a major field. The combination of the key word coming from a major field and the final score resulting from the comparison between the CDB descriptor and the occupation response field resulted

[3] 1980 Large Sample - 132,000 cases from the 1980 Decennial Census stratified by code category

in 62.6% of all occupation production. The error rate for these cases was also 6.5%. This indicates the importance of the occupation response field for accurate occupation coding. The occupation pseudo-phrase only contributed 6.5% of production with an error rate of 40.9%. In addition, these errors were over 44% of the total false positives.

Interestingly, there were many false negative occupation cases for precisely those decision matrix combinations which resulted in the most production. These false cases accounted for 22% of all occupation codes assigned. Most importantly, the false negatives outnumbered the true negatives almost two to one: 66.3% of the negatives are false. If the false negative rate could be reduced, increases in occupation production would result.

The raw data and a short description for these decision matrices are in Attachment 2.

2.2 SCORING ALGORITHM ENHANCEMENTS

This work was an effort to raise the occupation production rate by developing an algorithm to include the negative cases identified above. We calculated an auxiliary cutoff score for each occupation code category using the negative cases as data. The occupation production rate improved over 4 percentage points. The error rate went down slightly.

While this method was successful, we were concerned that it is ad hoc. We looked for systematic ways of achieving the same result. We decided to test whether modifying the scores of those cases which fell into the above mentioned decision matrix combinations would work. There were three places within the scoring algorithm where the score could be modified: 1) as it is being calculated for each CDB descriptor; 2) after all calculations are finished for each CDB descriptor; or 3) after the final score for a case has been decided. For each of these methods, only the scores generated from the proper decision matrix combinations were modified.

Experiments using the 1980 Large Sample and the 1990 Validation Sample[4] test files were made. In the first experiment, we doubled the score of all appropriate cases. The results of these experiments showed that method 3 was the least successful (see table 1). Methods 1 and 2 produced about the same increase in production and change in errors. The occupation production rate rose over 4% as in the auxiliary cutoff score method. The error rates went up very slightly, less than 0.25%.

[4] 1990 Validation Sample - 69,000 cases from the 1990 Post Enumeration Survey, 150 cases per code category.

Because of the success of doubling the score, further tests were made using method 3. Two tests were run, one where the score was tripled, the other with the score quadrupled. The results showed a gains in productivity with a rising error rate (see table 1), but the best results were obtained using method 2.

TABLE 1

	Production %	Match Rate %
Baseline	37.8	86.11
Method 1	41.7	85.87
Method 2	41.9	85.95
Method 3	41.0	85.72
Triple	41.7	85.63
Quadruple	41.8	85.65

2.3 PSEUDO-PHRASE ANALYSIS

The pseudo-phrase is responsible for 16.6% of industry production and only 6.5% of occupation production. However, the pseudo-phrase is responsible for 40% of the error for each type. An experiment to remove the pseudo-phrase from AIOCS was tried.

We coded the 1980 Large Sample through AIOCS, this time with the pseudo-phrase scoring comparison turned off. The results were not too surprising given the results of section 2.1. The occupation production rate went down from 37.7% to 36.5% with a large drop in the error rate of over 2 percentage points (13.9% to 11.6%). For industry, the production rate went down almost 4 percentage points (58.6% to 54.7%) and the error rate went down from 10.3% to 8.1%.

From these results, we have concluded that the pseudo-phrase should not be changed for industry coding. The reduction in the production rate is unacceptably large when the pseudo-phrase is not used. The slight reduction in the error rate does not justify the drop in production. The reverse is true for occupation. The drop in production is small relative to the drop in errors. Interestingly, the number of occupation cases coded correctly and above the cutoff scores goes up when the pseudo-phrase is shut off. Because occupation production is low, we are considering other methods of calculating the pseudo-phrase.

2.4 CODING DATABASE ANALYSIS

This experiment analyzed what percentage of the coding database is actually used to decide cases. We coded the 1980 Large Sample and 1990 Validation Sample test files through AIOCS and output an identifier and the score for the top three descriptors, called candidates, for each case. Industry and occupation were considered separately.

The results were surprising. Only 50.7% of 20,593 industry descriptors are used, and 43.0% of 30,359 occupation descriptors are used. Fully 85.7% of the industry candidate descriptors and 87.6% of the occupation candidate descriptors are the top choice for some cases.

It appears there are many unused descriptors in the CDB. There are several unanswered questions, however. We need to know: 1) what percent of descriptors decide cases above the cutoff scores; 2) what percent of unused descriptors go to categories with low production; and 3) what percent of used descriptors go to categories with high production. Once these questions have been answered, we can remove some descriptors from the CDB to make it more efficient.

2.5 EMPLOYER NAME LIST LOOKUP

Part of the processing for automated I&O coding includes a comparison of the employer name in the response to a list of employers in the respondent’s geographic region. This list is called the Employer Name List (ENL). It consists of the name of certain employers in the country with the appropriate industry code attached. The employer names are listed separately by geographic region.

For the 1990 Census processing, an exact match algorithm was used. This procedure was able to assign industry codes to about 6% of the total number of cases or just over 10% of the total number of industry codes assigned.

New research is being conducted to improve the exact match algorithm. A fuzzy search algorithm and software from Proximity Technology Inc. is being applied to see if inexact matching is accurate. Preliminary results are not encouraging (see table 2). The system is able to find all the exact matches that the old exact-match system found, but the error rate rises rapidly once inexact matching is needed. Garbage words such as "self", "blank", or "refused" often cause confusion. Filters are being developed and their effectiveness tested.

Total cases = 2515.

TABLE 2

Target Match Rate	Production Rate	Production-Matches
100%	10.22	257-257
95	10.62	267-261
90	11.69	294-266

2.6 GRAPHICAL AND CASE ANALYSIS

The occupation production rate as measured for the current surveys certification was significantly below expectations (Gillman, Appel, and Jablin, 1993). Attempts to find the causes of the difference have begun.

Code categories with little or no productivity as measured through test file data are examined. Usually, these code categories have cases assigned to them incorrectly where the computer generated score is very (relatively) high. Significant increases in productivity can be gained when these few cases are assigned correctly.

Graphs of cumulative match rate distributions for each code category are used to determine the score ranges that have a high percentage of incorrect cases. Database searches using codes and scores as search criteria produce many cases for analysis.

Research of this type is now underway to understand why the occupation production rate is lower than expected for current surveys data. No results are available at this time.

2.7 DATA COMPARISONS

A factor which drove the certification test of AIOCS for use in current surveys I&O coding processing (Gillman, Appel, and Jablin, 1993) was the assumption that current surveys data was different from census data. As a result, the Current Surveys Sample[5] was created for the test.

Due to budget and time constraints during the certification test, it was necessary to determine whether census data were similar enough to current surveys data to allow augmenting the Current Surveys Sample with census data. The measures of similarity were based on results of coding cases from the two sources. Cases coded to the same code category were compared using score and cumulative match rate distributions. It was assumed that if the distributions were similar, then the data were not dissimilar. The score distributions were first chosen for

[5] Current Surveys Sample - 116,000 cases from the Survey of Income and Program Participation, data from the first wave of the 1986 - 1991 panels.

comparison. For these comparisons, we chose the Komolgorov-Smirnov test (Lindgren, 1976), since the score is a continuous variable and the test is non-parametric. The test was conducted at the 10% level for each code category (industry and occupation were tested separately) and the number of categories which failed the test was counted. Using a binomial distribution with 0.1 probability of failure, the probability of finding that number of failed categories was found. If the test statistic was less than 90%, the hypothesis that the data from the two sources were not dissimilar was rejected. The results of this test were dramatic: 2/3 of both industry and occupation categories failed. The probability that this would occur by chance was less than 1 in a billion for each.

We planned to compare the cumulative match rate distributions only if the score distribution comparisons were accepted. This followed from our assumption that the data were similar enough if they passed both tests. The first test failed, so the second was not performed. Results of the K-S tests appear in Attachment 3. From the results of these tests, the assumption that current surveys data are different from census data is justified.

3. SMART SOFTWARE

Smart Software Co. is a software development company in New Jersey. They specialize in neural network software solutions for the PC. They asked for some data from the Census Bureau to help them develop an I&O coding solution for another company. In exchange for supplying data to them, they agreed to code a test file we sent to them. The results and discussion are below.

3.1 RESULTS

Smart Software developed a self-organizing neural network system. They used the 1980 Large Sample as a training set and the 1990 Validation Sample as a test set. The Census Bureau recently sent them the Current Surveys Sample for further training.

There is still much improvement needed before this software can match AIOCS. It is significant that the system assigns occupation codes so well relative to industry. Occupation is always harder to code.

Two sets of analyses were performed. One with the target match rates set as they were for the 1990 Census and the other for the error rates achieved as measured by the QC system in place for the 1990 Census I&O Coding processing. The results follow in table 3.

total cases = 69,301

TABLE 3

	above cutoff	above & correct	total coded	production rate	match rate	target m_rate
Ind	34612	29963	68424	0.4994	0.8657	0.85
Occ	32081	26601	69086	0.4629	0.8292	0.80
Ind	17101	16188	68424	0.2468	0.9466	0.94
Occ	21328	19078	69086	0.3078	0.8945	0.88

At the 1990 Census target match rates, this system has virtually the same overall productivity as AIOCS, about 48%. However, the error rates are much higher. It is interesting that the match rates (1 - error rate) are so close to the target match rates. The match rates for AIOCS based on coding the Large Sample and the Validation Sample were consistently 5 points or more above the targets. As stated above, the 1990 Census QA program measured AIOCS error rates even lower.

The reason that the match rates are so close to the targets is probably due to the way scores are assigned by the neural net software. AIOCS tends to assign the same score to many cases within each code category. This will tend to raise the cutoff score, lower productivity, and lower the error rate (see section 0.1 and the explanation of a similar problem in section 4.1). Based on observation of score data from the neural net, it appears the problem of tie scores is reduced, which reduces the difference between the observed and expected (target) match rates.

The fact that the productivity of the neural net is equal to AIOCS is very encouraging. However, Smart Software has developed a working I&O coding system for another company. There may not be much more room for improvement with this system.

Another problem with using neural net software is training the system. If one were to apply the neural net to another type of coding (not I&O), substantial time and resources would be needed to train the software for that type of coding and data. Even switching from census data to current surveys data might require retraining. The training step requires an analyst who is thoroughly familiar with neural nets, Smart Software's system, and the data. This requirement might make the neural net impractical.

4. LAWRENCE TECHNOLOGIES

Lawrence Technologies is a small company that until recently was based in San Antonio, TX. Their business is derived from contracts for the development of specialized text search and retrieval software. The principal in the company has developed a technique emulating a holographic data storage model. They boast very fast database search and retrieval of full text documents. They develop their software mainly for the PC.

This company agreed to develop a prototype industry coding system without charge. They claimed almost no effort was needed in the development process and that many improvements can be made to the system. The results and discussion are below.

4.1 RESULTS

The holographic data storage technique requires the use of a program to store the file in the proper formats. The stored file is called the training set in what follows. The company claims very fast processing times for this transformation. In practice, this took much longer than we expected, but not long if the computer were being set up to process many thousands of records in production. Typically, the training step took between 2 and 3 hours for a file with 132,000 records.

The algorithm uses the holograph model to find the cases in the training set which are candidates for best match. The search technique through the holograph employs "correlithms" (Lawrence, 1992a and 1992b). Nearest neighbor and fuzzy search techniques are used to distinguish among the candidates. The system is optimized by compiling and loading some natural language and dictionary information before the training set is stored.

The Large Sample and the Current Surveys Sample were used as training sets. They and the Validation Sample were used as test sets. The Large Sample and Current Surveys Sample were split in half, and six automated coders were created by training on the two full files and the four half files separately. We processed the Validation Sample through every automated coder we created, and we processed the Large Sample and subsets through the Current Surveys Sample trained coders and vice-versa. We also processed the Current Surveys Sample and subsets through coders trained on themselves.

The results were surprising. The details are in the tables which follow. Several important points are worth noting. First, the production rates are good for a first attempt. There is still much room for improvement to match the results of AIOCS during the 1990 Census. Of course, a true comparison requires the ability to code occupation as well.

The production rates suffer somewhat from an anomaly caused by the way the cutoff scores are calculated. There are many tie scores among the cases within each code category. Often, the cumulative match rate will rise above and fall below the target match rate within a group of cases with the same score. If this happens and the cumulative match rate of the last case is below the target, that score cannot be the cutoff. This problem was observed in each output set studied for this system. The problem can also be seen indirectly, by noting that the overall match rates are significantly above the target match rate for every run. The target match rate is a sharp bound if the cases in each code category have unique scores.

Most surprising are the results from using the Current Surveys Sample. When coding the Current Surveys Sample through software trained with the Large Sample (or subsets), the production obtained was almost nothing: less than 0.1%. However, when the Current Surveys

Sample (or subsets) are used as training sets, the coding results were the best! The results of coding the Validation Sample through all six automated coding systems revealed this.

There are two important implications of this. First, it is apparent that interchanging pairs of files between training set and test set is not a commutative operation using this software. In other words, coding the Large Sample with software trained on the Current Surveys Sample produces good results whereas coding the Current Surveys Sample with software trained on the Large Sample produces bad results.

The other implication is that current surveys data are different from census data. This confirms the finding (see section 2.7) from the certification of AIOCS for current surveys: AIOCS processed current surveys data differently than census data with respect to distributions of scores by code category. One reason for the difference may be due to how the data are collected. The census relies on self-reporting which is often sloppy and is surely inconsistent. Current surveys use a permanent, trained interviewing staff to collect the data. The interviewers know how to get respondents to give complete information. This fact helps explain why the Current Surveys Sample was the best training set for Lawrence Technologies software.

The test results displayed in table 4 were created by coding the Validation Sample data using a coder that was trained using all of the CPS data file.

TABLE 4

Training Set	Entire CPS File		
Test Data	1990 Validation Sample		
Target Match Rate	85.0	83.0	80.0
Total Cases	69279	69279	69279
Coded Above Cutoff	29319	30977	33930
Coded Above Cutoff + Correct	26172	27233	29254
Match Rate	89.3	87.9	86.2
Production Rate	42.3	44.7	49.0

The test results displayed in table 5a and 5b were created by coding half of the CPS file using a coder that was trained using the other half of the CPS data file. The two parts are called Half A and Half B.

TABLE 5a

Training Set	Half A CPS File		
Test Data	Half B CPS File		
Target Match Rate	85.0	83.0	80.0
Total Cases	58106	58106	58106
Coded Above Cutoff	30382	32046	32969
Coded Above Cutoff + Correct	27588	28864	29497
Match Rate	90.8	90.1	89.5
Production Rate	52.3	55.2	56.7

TABLE 5b

Training Set	Half B CPS File		
Test Data	Half A CPS File		
Target Match Rate	85.0	83.0	80.0
Total Cases	58108	58108	58108
Coded Above Cutoff	30734	31802	33545
Coded Above Cutoff + Correct	27794	28603	29585
Match Rate	90.4	89.9	89.0
Production Rate	52.9	54.7	57.7

The results in table 5a and 5b above display a bias that is evident when a data set is split in two parts and one part is used to train the software and the other is used as a test set. For this test, a random number generator was used to split the file. Similar results were obtained when the Large Sample was split into two files. Those data have not been reproduced here. The production rates for the Current Surveys Sample split files are significantly higher and the error rates somewhat lower than other results using the Current Surveys Sample as a training set (see table 4, for instance).

The test results displayed in table 6 were created by coding the Validation Sample data using a coder that was trained using all of the Large Sample data file.

TABLE 6

Training Set	Entire Large Sample File		
Test Data	1990 Validation Sample		
Target Match Rate	85.0	83.0	80.0
Total Cases	69279	69279	69279
Coded Above Cutoff	27696	29038	31226
Coded Above Cutoff + Correct	24512	26849	25432
Match Rate	88.5	86.0	87.6
Production Rate	40.0	41.9	45.1

The test results displayed in table 7 were created by coding the CPS file using a coder that was trained using all of the Large Sample data file.

TABLE 7

Training Set	Entire Large Sample File		
Test Data	Entire CPS File		
Target Match Rate	85.0	83.0	80.0
Total Cases	116658	116658	116658
Coded Above Cutoff	32	470	1107
Coded Above Cutoff + Correct	28	393	895
Match Rate	87.5	83.9	80.9
Production Rate	0.0	0.0	0.0

5. CONNECTION MACHINE

The most successful research to date has been the use of the Connection Machine (model CM-2) from Thinking Machines Corp (Creecy et al, 1992). The CM-2 is a massively parallel computer based on the SIMD (single instruction - multiple data) model using a hypercube network and containing 8192 single bit processors. Each processor has 128K of local memory.

The work was a research project funded by the Census Bureau to Thinking Machines. Their researchers used a technique called Memory Based Reasoning (MBR). A description of this and the results are below.

5.1 RESULTS

MBR is similar to the technique used by Lawrence Technologies. It takes a file of cases and uses it as a training set. Each case is assigned to a separate processor and a test case is compared to each case in the training set simultaneously. Software creates virtual processors for data sets which are larger than the number of physical processors on the CM-2. Nearest neighbor algorithms are used to decide which case in the training set matches the test case best. Features, which are combinations of fields or parts of fields, and the probabilities that a feature is classified to a particular category are used to enhance the matching procedure.

The tests were conducted using the 1980 Large Sample. Half the Large Sample was used as training set and the other half was used as a test set. The results are in table 8.

TABLE 8

	Production Rate	Error Rate
Industry	63%	10%
Occupation	57%	14%

As compared with all the other systems, these numbers are the best, although we were never able to verify these results independently. The drawback with this system is the expense of the computer: over $1M. Also, the systems which fit on the PC might be useful in CATI/CAPI applications; this system clearly cannot be used there.

There is one important point to note about these results. The production and error rates are based on using the Large Sample only. As we discussed above in section 4.1, the results are biased in favor of the algorithm if the training set and test set come from the same data sample file. The work with the CM-2 was done before the Validation Sample and Current Surveys Sample were created. Better estimates of the production and error rates will be obtained if the system can be benchmarked using separate sample files for the training set and test set.

6. OTHER SYSTEMS

There are several other systems which have been reviewed, but currently, there are no results for them. These are discussed below.

6.1 INFERENCE GROUP

The Inference Group of Australia is a private company whose sole business is developing automated and computer-assisted coding software for statistical agencies. They claimed they had an all purpose coding system which could be used for any coding structure. This system is based on natural language analysis and the use of an extensive thesaurus. Inference Group claims very high production rates for installed systems (about 80%) with error rates of less than 5%. Agencies with installed systems from Inference Group are very happy with the results.

The Census Bureau sent Inference Group several test files, coding structures, and classification indexes with the promise of fantastic results. They have informed us that they have to redesign their system and have been unable to produce any results for us so far.

6.2 FRENCH SYSTEM

The French have demonstrated a prototype all purpose coding system which is also based on natural language analysis and an extensive thesaurus. The unique aspect of this new system is that they took all of the country's coding structures and installed them into the system and created a conformance between each pair of structures.

The French claim that their system will code a case to whatever coding structure is most appropriate based on the input, and through the conformance, output the appropriate codes. Unfortunately, no reliable data are available yet to test their claims. The Canadians may have some results soon.

6.3 HNC

HNC Corporation, based in San Diego, has a unique data search technique based on linear algebra and vector arithmetic. Their system works much the same as Lawrence Technologies, Smart Software, and Thinking Machines. They take a training set and install it in a database. Test cases are coded by matching to the most appropriate training set case. This case is found through the vector analysis and nearest neighbor algorithms.

Data was sent to HNC for their use. They developed a very rudimentary prototype which could not be analyzed. They declined to continue with our work unless the Census Bureau could provide funds.

6.4 TEXT RETRIEVAL

Various companies with general text retrieval engines have been contacted. They have been told the Census Bureau is interested in developing a CATI/CAPI automated/computer-assisted coding system for use by interviewers. These companies include ConQuest of Columbia, MD.

ConQuest was given the I&O coding test samples, the coding structure, and the classification index. They have not yet produced a working prototype.

6.5 CANADIAN SIC CASES CODER

Statistics Canada has produced a clerical computer-assisted coding system for their SIC structure using CASES. This system is used to classify the companies who are applying for a new payroll deduction number from Revenue Canada. The U.S. has a similar process involving the IRS.

The coding system, which is tied in with the entire questionnaire instrument, is not automated at all. The index is a computerized version of the classification system. The software makes use of a tree of menus four levels deep. Each level corresponds to a digit in the four-digit

SIC. The menus have statements which correspond to each separate digit used at that point in the classification.

The interviewer traverses through the menu system picking the appropriate description at each step. The description chosen determines the digit at that level and the menu to appear on the screen at the next lower level. After four levels, the system prints the code and description that were found.

Observing the system in use showed that the trained interviewers preferred to code with the printed manual rather than the computer. More work needs to be done to speed the system. Work must be started to determine if it can be automated at all.

6.6 ACTR

The Automated Coding by Text Retrieval (ACTR) system is the generalized automated coding system developed and used by Statistics Canada. It is a based on the Hellerman Algorithm (Rowe and Wong, 1994), similar to AIOCS (see section 1). The Census Bureau evaluated a beta version.

Unlike AIOCS, ACTR is a generalized coding system. It is designed to be useful for a wide range of coding applications. The input text parsing strategy can be customized to optimize the system for each coding application. Also, the project database may be dynamically updated, i.e. if a case must be manually coded, its input can be added to the coding database so similar cases will be automatically coded in the future.

The function of ACTR is to assign a code to a single input text string. This makes ACTR unsuitable, currently, for Census Bureau I&O coding. Another limitation is, unlike the parsing strategy, the weighting scheme cannot be altered. Lastly, there is no spell checking in ACTR. This can greatly affect coding results since slightly misspelled words will not be identified correctly. Statistics Canada is working to overcome these limitations and others.

In consideration of using ACTR for the Canadian 1991 Census, a research project to test ACTR was established. The results from that study are shown in Table 9 below:

TABLE 9

VARIABLE	MATCH %	ERROR %
Mother Tongue	92.1	3.4
Place of Birth	91.6	2.0
Ethnic Origin	93.8	1.3
Major Field of Study	78.0	4.4
Industry - Company Name	31.5	8.2
Industry - Kind of Business	38.0	25.5
Industry - Linked Files	22.3	2.4
Occupation - Line 1	42.7	31.1
Occupation - Line 2	19.2	37.0

It is apparent that ACTR performs much better with single word answer questions than ones that require more detailed responses. Place-of-Birth and Ethnic Origin are easier for the respondent to describe than Industry or Occupation.

6.7 NCHS

NCHS works with I&O codes for birth and death certificates. They are developing an automated I&O coding system called Classifier for Industries and Occupations (CLIO). They claim that this system performs very well but they have no statistics yet.

CLIO is based on a dictionary of word pairs with codes attached. The development process includes adding new phrases to the dictionary to increase production. NCHS is working with several states to help improve the system. They have requested that the Census Bureau send them some additional data. We will also send them a file with no codes for a test.

7. CONCLUSION

The Census Bureau is conducting research on a wide range of different automated coding and computer-assisted coding systems. At this time, many of the research projects are at a preliminary stage. The work analyzing AIOCS has produced many results.

There are three arenas where computer coding systems are useful: fully automated, computer-assisted clerical, and CATI/CAPI. The fully automated systems are used for large batch operations; the computer-assisted clerical systems are used for residual coding of cases which a fully automated system could not decide; and the CATI/CAPI systems are used in the field during interview time. The CATI/CAPI systems will probably be a combination of automated and computer-assisted clerical.

An aim of the research is to find a system which could be tailored to perform in all three arenas. Such a system has to be small to fit in a CAPI laptop and must be fast and accurate to handle large batch operations. The main focus of the research so far is to find a system which has productivity and error rates that are better than those for AIOCS. Speed is also very important, about five cases per cpu second is the lowest acceptable limit. The speed of all the systems tested so far meets or exceeds the criterion.

The best system tested so far is the Memory Based Reasoning (MBR) approach used by Thinking Machines on their CM-2. Those results may be high due to the same sample bias which was seen using the Lawrence Technologies system (see section 4.1), because the Large Sample was split into two files for the CM-2 work. At the time this study was performed, the Validation Sample and the Current Surveys Sample did not exist. More extensive work needs to be done to get an accurate measure of the capabilities of MBR on the CM-2 (or a more advanced machine).

The Census Bureau will continue researching new methods of automated coding. As more data become available, decisions will be made as to which technologies to pursue.

8. REFERENCES

Appel, M. V. (1991). "Field Weights for the Automated Coder", Memorandum to John Priebe, Census Bureau, Washington, DC, July 25.

Appel, M. V. and Hellerman, E. (1983). "Census Bureau Experiments with Automated Industry and Occupation Coding," Proceedings of the American Statistical Association, 32-40.

Appel, M. V. and Scopp, T. (1987). "Automated Industry and Occupation Coding", presented at Development of Statistical Expert Systems

(DOSES), December 1987, Luxembourg.

Bureau of the Census (1990). Official 1990 U.S. Census Form, Form D-2 (Long Form Questionnaire). U.S. Department of Commerce.

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, 729-745.

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.

Gillman, D. W., Appel, M. V., and Jablin, C. (1993). "Certification of Decennial Automated I&O Coder for Current Surveys," Proceedings of Census Bureau's Annual Research Conference (ARC), March 21-24, 766-782

Gillman, D. W., Appel, M. V. (1993). "Analysis of the Census Bureau's Automated Industry and Occupation Coding System Algorithm", 1993 Proceedings of the Government Statistics Section of the

American Statistical Association, Aug. 8-12, 234-239

Jablin, C. (1992). "Evaluation of Automated I&O Coding for Current Surveys", Memorandum to Sherry L. Courtland, Census Bureau, Washington, DC, October 15.

Lawrence, P. N. (1992a). "Correlithms", unpublished, copyrighted, available from the author at email address nick@lt.com.

Lawrence, P. N. (1992b). "Introduction to Dataware Engineering", unpublished, copyrighted, available from the author at email address nick@lt.com.

Lindgren, B. W. (1976). "Statistical Theory", 3rd ed., Macmillan Publishing Co., New York.

Rowe, E., Wong, C. (1994). "An Introduction to the ACTR Coding System", Bureau of the Census Statistical Research Report Series No. RR94/02.

Scopp, T. S., Tornell, S. W. (1991). "The 1990 Census Experience with Industry and Occupation Coding," presented at the Southern Demographic Association Annual

Meeting, Jacksonville, FL, October 11, 1991.

We coded the entire 1980 Large Sample, for both industry and occupation, and obtained the computer generated code, the truth code, the computer score, the cutoff score, the field from which the key word in the winning phrase came, and the field which best matched the winning phrase. Three sets of tables were generated from these data: industry, occupation, and self-employed occupation. They are described below.

The coding results were organized into four categories (icat or ocat = 1,2,3,4) depending on how the automated coder decided each case. These categories are defined as follows:

_cat = 1: computer code matched truth and computer score above the cutoff score
_cat = 2: computer code not equal to truth but computer score above the cutoff score
_cat = 3: computer code matched truth but computer score below the cutoff score
_cat = 4: computer code not equal to truth and computer score below the cutoff score.

The other variables used in the tables were key word field (ifno or ofno) and match field (ifdrw or ofdrw). The _fno variables took on one of five values each:

0: case coded by logical analysis
1: key-word obtained from employer field
2: key-word obtained from industry field
4: key-word obtained from occupation field
5: key-word obtained from duties field.

The ifdrw variable took on one of four values:

0: case coded by logical analysis
1: winning phrase matched with employer field
2: winning phrase matched with industry field
3: winning phrase matched with industry pseudo-phrase.

The ofdrw variable also took on one of four values:

0: case coded by logical analysis
4: winning phrase matched with occupation field
5: winning phrase matched with duties field
6: winning phrase matched with occupation pseudo-phrase.

The cells in each table contain four data items:

frequency: the number of cases in this cell
percent: the percentage of the total in this cell to the total in the table
row pct: the percentage of the total in this cell to the total in this row
col pct: the percentage of the total in this cell to the total in this column.

The tables themselves are organized as follows:

Table Number	Table Variables
Industry	
1.0	icat by ifno
1.1	ifno by ifdrw (for icat = 1)
1.2	ifno by ifdrw (for icat = 2)
1.3	ifno by ifdrw (for icat = 3)
1.4	ifno by ifdrw (for icat = 4)
Occupation	
2.0	ocat by ofno
2.1	ofno by ofdrw (for ocat = 1)
2.2	ofno by ofdrw (for ocat = 2)
2.3	ofno by ofdrw (for ocat = 3)
2.4	ofno by ofdrw (for ocat = 4)
Self-Employed	
3.0	ocat by ofno
3.1	ofno by ofdrw (for ocat = 1)
3.2	ofno by ofdrw (for ocat = 2)
3.3	ofno by ofdrw (for ocat = 3)
3.4	ofno by ofdrw (for ocat = 4).