Journal of Official Statistics
Vol. 9. No. 4, 1993, pp. 729–745

Error Control of Automated Industry and Occupation Coding[1]

Bor-Chung Chen, Robert H. Creecy, and Martin V. Appel[2]

Abstract: The Automated Industry and Occupation Coding System (AIOCS) was used in the 1990 Decennial Census to classify the natural language responses into one of 243 industry and 504 occupation categories. This paper presents the empirical results from a new error control procedure for estimating the production rates and error rates of the AIOCS. This new procedure consists of the cutoff method (per class threshold) and the weighted approach. It was implemented for the 1990 census to control the production and error rates. One of the basic assumptions of the cutoff method is that there is a positive correlation between the score associated with classifying a response and the probability that the response is correctly classified. For each code category, the magnitude of the score below which selected phrases have an unacceptable probability of error is referred to as the "cutoff score." The use of the weighted approach was to compensate for the procedure used to select validation data used to evaluate the AIOCS. This paper also shows that the potential bias problem from estimating the cutoff score and the production and error rates with the same set of data is very small if the sample size is large.

Key words: Classification; estimations; cutoff.

1. Introduction

As part of its mission to collect, tabulate, and disseminate information about the U.S. economy and population, the U.S. Census Bureau collects Industry and Occupation (I&O) information from individuals in the labor force. These hand-written (natural language) responses are solicited from individuals during Decennial Censuses and demographic surveys. There are six I&O questions. The following is an example of the responses (capitalized) to the I&O questions (Appel and Hellerman 1983):

Industry Questions:

- For whom did this person work: POST
- What kind of business or industry was this: NEWSPAPER PUBLISHER
- Is it mainly (manufacturing, wholesale trade, retail trade, or other): OTHER

Occupation Questions:

- What kind of work was this person doing: SALESPERSON
- What were this person's most important activities or duties: SELLS ADVERTISING
- Was this person an employee of private company, government employee, self-employed, or working without pay: PRIVATE

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

[2] Statistical Research Division, Bureau of the Census, Washington, D.C. 20233-4200, U.S.A.

This example would be classified into industry code 171, *newspaper publisher* and occupation code 256, *salesperson, advertising*.

1.1. *The problems associated with manual coding*

In the past, the I&O classification coding has been completely clerical and extremely expensive and time consuming. In the above example, a coding clerk would look up the industry response *Newspaper Publisher* in a list or dictionary of phrases and retrieve its industry code 171. The clerk would then look up the occupation response *Sells Advertising* and get the occupation code 256. This clerical assignment of codes is subject to human error and subjective judgement which are not consistent when tracked over time. Clerical error rates in the 1980 Decennial Census were higher than ever before. Recruiting, training, and managing large numbers of coders has not been cost-effective. I&O coders require substantial special training and a long learning process. The clerical coding operation also requires strong technical support and is difficult to control. The cost of clerical coding and verification associated with the I&O classification of natural language descriptions is extremely high. It is about $1.2 million per year for all demographic surveys and was approximately $7.7 million for the 1980 Decennial Census.

1.2. *Bureau experience with automated coding*

To overcome the problems with manual coding, the Census Bureau started to experiment with automated I&O coding prior to the 1967 Economic Census. The objective was to have a computer assign classification codes based on the natural language responses. It was expected to provide a better coding quality, to improve consistency, to systematically strengthen the coding structure and procedures, and to reduce the managerial burden of recruiting and training I&O coders.

The first automated I&O coding algorithm was called the O'Reagan algorithm (O'Reagan 1972). It was developed to construct a reference list for industry codes based on empirical response/code sets and then to code freshly received response patterns provided by small business establishments. The O'Reagan algorithm, and the two which followed in sequence, placed great emphasis on the observed frequency of responses and associated codes.

The Corbett algorithm (Corbett 1972) was devised in 1972 but not tested until several years later. The intent of this algorithm was to find the minimal set of word strings that would define the code. Testing in 1975 and 1976, on 1970 Decennial Census records, yielded results inferior to those obtained by the O'Reagan algorithm, which coded 72% of the industry records with 88% accuracy. The Corbett method was also slower (Appel and Hellerman 1983).

Then came the Information Measure Processor (IMP) (O'Reagan 1972) with an attempt to make explicit use of information theory concepts and notations. For each record/code pair in the historical file a weight was computed and stored in a matrix. The mutual information (MI) measure was calculated for every code by a summation over all words in the code. To code an incoming record, each response word was searched in the weight matrix. For words found, the appropriate weights were added. The sum for each code was then divided by the MI of the code, and the code with the highest result was the code assigned. Restrictions on this simplest scheme were tested. The code with the highest score was required to exceed a threshold

and also to exceed its nearest competitor by some specified percentage. Utilizing the same files on which the earlier methods were tested, IMP could code the entire input with 77% accuracy or code nearly half the input with virtually no errors, depending on the restrictions. The best balance seemed to be a 73% coding rate with 87% of those codes correct. The history file was processed at nearly 10,000 records per minute, but coding speed was only 500 records per minute (Appel and Hellerman 1983).

Major and minor variations on all these schemes were tried to the extent that resources permitted. Taking account of word order and syntax produced no benefit. Efforts to use synonyms, truncate, or compress provided ambiguous results. A cascade of approaches, e.g., passing a record on to IMP if Corbett failed to code, was considered but never tested. By late 1976 the Census Bureau decided that no system achievable by 1980 appeared practicable for census processing.

In late 1976, the automated I&O coding research took a significant change in direction by relying on the *coding manual* to provide the initial set of patterns for matching and the numerical weights for scoring. The result of this research effort was the Hellerman I&O Coding System (HIOCS) (Hellerman 1982), which is the earlier version of the Automated I&O Coding System (AIOCS) (Appel and Hellerman 1983, Appel and Scopp 1987). The basic idea of the HIOCS was to simulate in the computer the processes that a production coding clerk employs: such as recognizing "meaningful" words, misspellings, synonyms, and abbreviations.

1.3. Current developments

The 1990 Decennial Census processed approximately 22 million questionnaires with natural language responses to the I&O questions. Based on the six responses, each respondent was classified into one of the 243 industry categories and 504 occupation categories. A computerized coding system was developed to classify the 1990 Decennial Census I&O responses. This system had two parts: (1) a centralized batch coder called Automated Industry and Occupation Coding System (AIOCS); and (2) a computer assisted clerical coder to aid clerks in coding AIOCS's residuals. The AIOCS's residuals are the responses that the AIOCS could not code or the code the AIOCS assigned had a high probability of being misclassified.

AIOCS is the latest in a progression of research coding systems. It consists of three major subsystems:

1. The Knowledge Base System (KBS) establishes and maintains a centralized data base of both files and arrays. It consists of a coding data base (DB) of I&O descriptor phrases with associated constraints and restrictions, a synonyms list and a lexicon of all words in the descriptor phrases.
2. The Coding System 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 data base changes.

The general procedure is to select the most meaningful word and to retrieve and score all DB phrases with this word against the respondent's replies. If there is an exact match, the associated code is assigned. If not, the DB phrases are scored on the basis of "closeness-of-fit." For more details on the closeness-of-fit measure and how to

obtain "scores," see Appel and Hellerman (1983) and Appel and Scopp (1987). For each code assigned by the AIOCS, there is a score associated with it. A high score implies a high probability of the response being correctly classified.

The I&O classification problem is to match a response **z** to one phrase and its associated code on the basis of some measurements of closeness-of-fit. Some of the responses classified have a low probability of being assigned to a correct code. These responses are referred to as *doubtful cases*. In this classification problem, as in most, there is a trade-off between classification error rates and production rates, which are the percentage of the total classified responses statistically accepted to their proper code categories. It is obvious that the doubtful cases contribute significantly to the error rates. Making a decision to reject doubtful cases reduces both the error rates and the production rates. The objective is to minimize the proportion of the cases being rejected while maintaining a desired error rate. In this paper, we present a new error control procedure which was applied successfully to the 1990 Decennial Census I&O coding. This procedure consists of the "cutoff method" (per class threshold) and the "weighted approach." They will be discussed in detail in the following sections.

Three data sets were used to evaluate the AIOCS and set cutoff scores (rejection thresholds). They were the 1980 Large Sample, the 1990 PES (Post Enumeration Survey) data set, and the 1990 Validation Sample. The 1980 Large Sample was a sample of more than 132,000 I&O responses from the 1980 Decennial Census. It was triply coded by clerks and reviewed by experts to provide a good data set for evaluating the AIOCS system. The 1990 PES data set contained 361,306 cases which were used for the purpose of validating the 1990 Decennial Census results. Those I&O responses were coded by AIOCS as a test. The 1990 Validation Sample was created by randomly selecting at most 150 cases for each code category from the 1990 PES data set and then having this data set triply coded by clerks and disputes adjudicated by experts.

There are two types of cases that are referred for clerical coding by the AIOCS. The first type consists of the cases which are not coded by the AIOCS and not included in the estimation of the production rate. The second type consists of the cases assigned a code which has a high probability of being misclassified and not included in the estimation of the error rate. The production rate is the percentage of cases classified by the AIOCS.

1.3.1. The certified method

Initially, a method, called the *certified method*, was used for controlling error rates. Under this method, a code assigned by the computer is accepted or rejected based on an analysis of AIOCS's coding of the 1980 Large Sample. If the AIOCS code assignments, for an entire code category, matched those of the experts at or above a target percentage, the computer was "certified" to code this category and *all* of the computer's assignments into this category are accepted as final. This computer-expert match rate was called the "certification level." Conversely, code categories with match rates below the certification level are "uncertified" and *all* computer assignments into this category are referred for clerical coding.

For example, assume that there are only six industry code categories: 930 (Administration of Environmental Quality and Housing Programs), 922 (Administration of Human Resources Programs), 921 (Public

Finance, Taxation, and Monetary Policy), 910 (Justice, Public Order, and Safety), 901 (General Government), and 900 (Executive and Legislative Offices), and we are interested in the AIOCS coding performance of two of them: 901 and 900. Table 1 shows the AIOCS assigned code (under the heading "Code"), the expert assigned code ("Truth"), and the cumulative match rate (CMR). The scores are also shown for the illustration of the cutoff method described below. The certification level is assumed to be 80%. The match rate of the AIOCS code assignments for Code 901 (Example 1) is 66.7% which is below the certification level and the entire code category is uncertified. The match rate for Code 900 (Example 2) is 80.6% which is above the certification level and the computer is "certified" to code this category. The estimated production rate for both code categories is $31/55 = 56.4\%$. The estimated error rate is $1 - (25/31) = 19.4\%$.

1.3.2. The cutoff method

The certified method can be characterized as all or nothing. All responses coded to a certified code are accepted; nothing coded to an uncertified code is accepted. Even exact phrase matches that code to an uncertified code category are referred for clerical coding. A review of the 1980 Large Sample Benchmark Reports showed that a significant portion of the sample that was coded

Table 1. Examples

Example 1: Industry Code 901					Example 2: Industry Code 900 (Continued)				
Case #	Score	Code	"Truth"	CMR	Case #	Score	Code	"Truth"	CMR
1	5600	901	901	1.000	4	7000	900	900	0.750
2	4300	901	901	1.000	5	7000	900	900	0.800
3	4300	901	901	1.000	6	6700	900	901	0.667
4	4300	901	901	1.000	7	6700	900	900	0.714
5	4000	901	922	0.800	8	6700	900	900	0.750
6	4000	901	901	0.833	9	6600	900	900	0.778
7	3800	901	901	0.857	10	6600	900	900	0.800
8	3700	901	930	0.750	11	6600	900	900	0.818
9	3700	901	921	0.667	12	6600	900	900	0.833
10	3700	901	921	0.600	13	6500	900	922	0.769
11	3700	901	921	0.545	14	6500	900	921	0.714
12	3700	901	910	0.500	15	6500	900	921	0.667
13	3700	901	910	0.462	16	6500	900	901	0.625
14	3700	901	901	0.500	17	6500	900	900	0.647
15	3700	901	901	0.533	18	6500	900	900	0.667
16	3700	901	901	0.563	19	6300	900	900	0.684
17	3700	901	901	0.588	20	6300	900	900	0.700
18	3700	901	901	0.611	21	6300	900	900	0.714
19	3700	901	901	0.632	22	6300	900	900	0.727
20	3700	901	901	0.650	23	6300	900	900	0.739
21	3700	901	901	0.667	24	6300	900	900	0.750
22	3700	901	901	0.682	25	6300	900	900	0.760
23	3700	901	901	0.696	26	6300	900	900	0.769
24	3700	901	900	0.667	27	6300	900	900	0.778
Example 2: Industry Code 900					28	6300	900	900	0.786
1	7200	900	900	1.000	29	6300	900	900	0.793
2	7000	900	901	0.500	30	6300	900	900	0.800
3	7000	900	900	0.667	31	6300	900	900	0.806

to uncertified codes does in fact agree with the expert's code (38% Ind, 36% Occ). The problem is to identify, within uncertified categories, the coded cases that have a high probability of being correct. In order to do this, a discriminator with a predetermined level of accuracy is needed to identify individual responses that are coded correctly. A new method, called the *cutoff method*, described in this paper, uses as this discriminator, the "score" or closeness-of-fit measure that the AIOCS uses for code assignments.

To use the cutoff method, a target match rate (equivalent to the certification level in the certified method) must be specified. The responses (or cases) the AIOCS has assigned to a code category are arranged in descending order of the scores and the cumulative match rates (CMR) are calculated. The cutoff score is the score of the last case that has a CMR greater than or equal to the target match rate. If the AIOCS code assignments, for each code category, have a score greater than or equal to the cutoff score, the codes are accepted as the final code. For example, if the target match rate is 80%, then, in code 901 of Table 1, the last case which has a CMR greater than or equal to 80% is case 7. The cutoff score is, therefore 3,800. In code 900, the last case which has a CMR greater than or equal to 80% is case 31, and the cutoff score is 6,300. The estimated production rate for both code categories is $38/55 = 69.1\%$. The estimated error rate is $1 - (31/38) = 18.4\%$.

1.4. Main results

In what follows, we present the empirical results of the estimates of the production rates and error rates of the AIOCS running on a Hewlett-Packard 9000/825 workstation with HP-UX 8.0 and HP FORTRAN 8.14. The cutoff method was implemented for the 1990 census to control the production and error rates. Compared with the certified method, the cutoff method provided a productivity increase of 6.7% to 27% for industry coding (5.6% to 12.6% for occupation coding) at the price of error rate increase of 0.7% to 2.6% (0.1% to 3.2% for occupation coding). The use of this method reduced the clerical effort for industry and occupation coding by about 10% with an estimated saving of hundreds of thousands of dollars over the certified method. A separate cutoff score for each industry category and each occupation category was determined from the coding of the 1980 Large Sample and the 1990 Validation Sample. The target match rate was set separately for industry and occupation. Other studies of automated industry and occupation coding include those of Creecy, Causey, and Appel (1990); Chen, Appel and Creecy (1990); Masand, Smith, and Waltz (1990); and Creecy, Masand, Smith, and Waltz (1992).

The remainder of this paper is organized as follows: Section 2 describes the estimators of the production rate and error rate, an initial experiment of the 1980 Large Sample using the estimators described in Section 2 is given in Section 3, a description of some empirical results is the content of Section 4, Section 5 describes an approach with the weighted method, Section 6 is about the computer quality assurance sample for the 1990 census industry and occupation coding operation, and the conclusions are given in Section 7.

2. The Estimators of Production Rate and Error Rate

Assume that there are k populations: $\omega_i, i = 1, 2, \ldots, k$ (where k was 243 for industry and 504 for occupation in the 1990 Census I&O Coding); and the classification rule is D, where $D = \langle D_1, D_2, \ldots, D_k \rangle$ and D assigns an individual x to ω_i if and only if $x \in D_i$.

Also, $\forall\, x \in D_i, g_i(x)$ is the highest score among the candidate code categories and is called the discriminant score, such that D assigns x to ω_i; and $p(g_i(x))$ is the probability of x being correctly classified; i.e., $\forall\, x \in D_i$, the distribution of x being correctly classified is a Bernoulli distribution with parameter $p(g_i(x))$. Then, we assume that $p(g_i(x))$ and $g_i(x)$ are positively correlated. Reviewing the 1980 Large Sample Benchmark Reports indicates that the positive correlation assumption is likely to be correct.

Let $g_i(x_1) \geq g_i(x_2) \geq \cdots \geq g_i(x_n), x_j \in D_i$, $j = 1, 2, \ldots, n$, and C_j is the cumulative match rate; i.e., $C_j = k_j/j$, where k_j is the number of matches in the first j cases. Then, given a target match rate $t, 0 \leq t \leq 1$, if $\exists\, m(t) \ni$

$$m(t) = \max\{j \mid C_j \geq t, 1 \leq j \leq n\}$$

$g_i(x_{m(t)})$ is defined as the cutoff score. If $m(t)$ does not exist, the cutoff score is infinity. Figure 1 shows three examples, two of them are from Table 1, to illustrate how to obtain a cutoff score with the cutoff method, where $t = 0.80$ for Examples 1 and 2 and 0.85 for Example 3. Examples 1 and 2 were made by the authors and Example 3 came from the 1990 census data. The solid lines in Figure 1 are the graphs of the cumulative match rates, the horizontal lines at t are the levels of the target match rates, and the dot lines are the graphs of the scores. Point A is the last case which has a CMR greater than or equal to t. The correspondent score of point A can be found by drawing a vertical line from Point A to meet the dot line where the cutoff score is defined. Theoretically, point A is the intersection of the CMR curve and the horizontal line at t. In Figure 1(b), point A is not located at the intersection because in the figure it clearly shows that case 31 is the last case that has a CMR greater than or equal to t.

Note that the CMR curves in Figure 1 do not steadily decrease, but oscillate. Although there is a positive correlation between $p(g_i(x))$ and $g_i(x)$, it may occur that a case with a lower score is correctly classified and another case with a higher score is not. Another reason contributing to the oscillations of the CMR curves is that several cases with tie score may be arranged in any order and only part of them are correctly classified. In Figure 1(b) (also see Example 2 of Table 1), cases 13 to 18 have tie score and the CMR curve decreases between cases 13 and 16 and increases after case 16. In Figure 1(c), there is a quasi-constancy of the CMR curve between cases 29 and 116 and it has oscillation with steadily decreasing after case 116. After examining the data, the scores from case 1 to case 28 are between 43,200 and 8,800 with a match rate of 89.3%, the scores from case 29 to case 116 are between 7,600 and 3,000 with a match rate of 84.1%, and the scores after case 116 are between 2,800 and 622 with a match rate of 25.7%. This is an example which shows that the positive correlation assumption is likely to be correct.

Let

t = a target match rate,
P = the true production rate,
R = the true error rate,
$g_i(x_{m(t)})$ = cutoff score for D_i,
M_i^c = number of x being correctly classified to D_i and with discriminant score $\geq g_i(x_{m(t)})$,
N_i^c = number of x being classified to D_i and with discriminant score $\geq g_i(x_{m(t)})$, and
N_i = number of x being classified to D_i,
$K_i^c = N_i^c - M_i^c$,

then the estimated error rate is

$$\hat{R} = \frac{\Sigma_i K_i^c}{\Sigma_i N_i^c} \tag{1}$$

and the estimated production rate is

$$\hat{P} = \frac{\Sigma_i N_i^c}{\Sigma_i N_i}. \quad (2)$$

Since every coded case is classified to a class, $\Sigma_i N_i$ is a constant and equal to the total sample size.

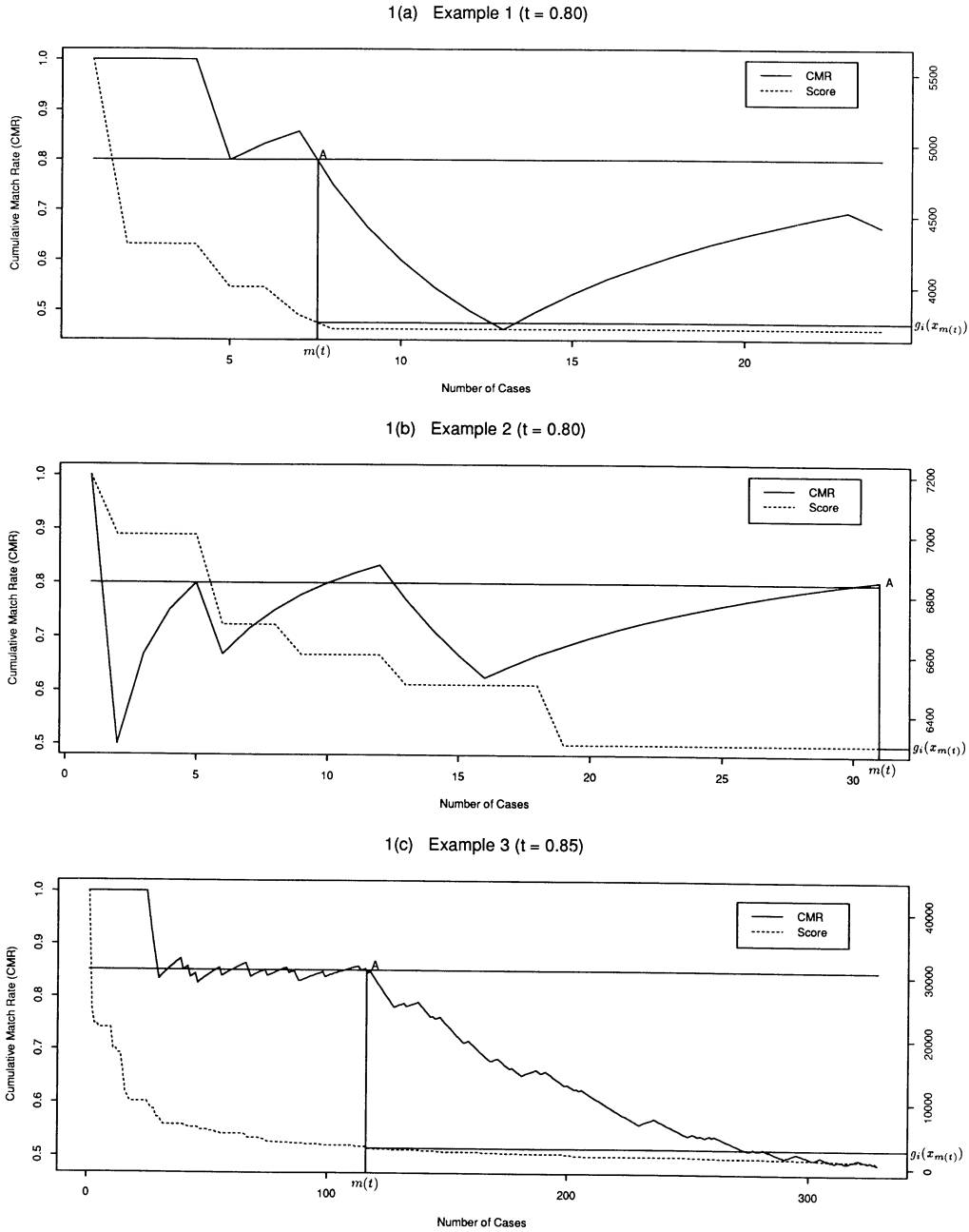

Fig. 1. Illustration of the cutoff method

3. Initial Experiment on the 1980 Large Sample

In this experiment, assuming that the target match rate is 83% for industry and 80% for occupation, the cutoff score, $g_i(x_{m(t)})$, is first estimated for each code category i. Then, the production and error rates are estimated from the estimated cutoff scores.

The estimations of (1) and (2) have a potential problem of bias if the same sample set is used to estimate both the cutoff score, and the production and error rates. To reduce the bias of the estimates, a typical discriminant analysis method (Panel on Discriminant Analysis, Classification, and Clustering 1989) was used. The method involves two analysis stages. The first stage is concerned solely with a training or cutoff sample, and the second stage is concerned with a test sample. A set of independent cutoff scores is first estimated from the cutoff sample, and (1) and (2) are estimated from the test sample and the independent cutoff scores, then the bias in the estimations can be reduced. Therefore, each code category in the 1980 Large Sample is randomly divided into a cutoff sample and a test sample. The results indicate that there is a small sample size problem. For those codes with large sample sizes, the estimates of the cutoff scores, the production rates, and the error rates, are consistent over several runs. Those estimates for the codes with smaller sample sizes have greater variations. However, there is a slight difference between the estimates with and without using independent cutoff scores. The summary for all codes using (1) and (2) is shown in Table 2. The replications in Table 2 were obtained by dividing the 1980 Large Sample into two subgroups with a random number generator such that a number between 0 and 32,767 was generated for every case in the sample. If it was an even number, the case was assigned to the first subgroup; if it was an odd number, the case was assigned to the second subgroup. Replication I was obtained with seed 1 given to the random number generator while Replication II with seed 13.

From Table 2, there is a slight difference between the estimations with and without using independent cutoff scores. The table also indicates that there is a possibility of overestimations of the production rate and underestimations of the match rate (except production rate estimation of the first subgroup for industry of both replications and for occupation of Replication I, and of the second subgroup for occupation of Replication II, which are at most 0.2% difference) without using independent cutoff scores. These slight overestimations/underestimations are acceptable with the available samples.

Since there is a small sample size problem, if a group is divided into subgroups, the problem for some codes with small sample size will be even worse. To minimize the effect of the problem, we excluded some codes with small sample sizes, less than 50, in the estimations of the overall production rate and error rate. The number 50 was chosen for several reasons. First, it was based on a binomial distribution with parameter $p = 0.80$, and its normal approximation. The rule of thumb indicates that the approximation is "good" if $np(1-p) > 9$. The value of n is obtained from

$$\text{Prob}\left(\text{Match Rate} > \text{p} - \text{z}_{1-\alpha}\sqrt{\frac{\text{p}(1-\text{p})}{\text{n}}}\right) = 0.90 \qquad (3)$$

with $\alpha = 10\%$. The sample size, $n = 26$, was selected so that the estimated match rate is at least $p - 0.10$ with 90% confidence. Due to the lack of information on the cutoff

Table 2. Summary of the 1980 large sample

Seed	Replication	Total Cases	1st Subgroup	2nd Subgroup
1	I	132744	66350	66394
Industry with Target Match Rate 83%				
Note	Total Cases(A)	Number Coded(B)	Production Rate (B/A)	Error Rate
(1)	66350	36821	0.555	0.103
(2)	66350	36976	0.557	0.106
(3)	66394	36946	0.556	0.106
(4)	66394	36599	0.551	0.108
Occupation with Target Match Rate 80%				
Note	Total Cases(A)	Number Coded(B)	Production Rate (B/A)	Error Rate
(1)	66350	19392	0.292	0.133
(2)	66350	19437	0.293	0.147
(3)	66394	19525	0.294	0.136
(4)	66394	18912	0.285	0.144
Seed	Replication	Total Cases	1st Subgroup	2nd Subgroup
13	II	132744	66363	66381
Industry with Target Match Rate 83%				
Note	Total Cases(A)	Number Coded(B)	Production Rate(B/A)	Error Rate
(1)	66363	36626	0.552	0.102
(2)	66363	36727	0.553	0.106
(3)	66381	36932	0.556	0.105
(4)	66381	36766	0.554	0.107
Occupation with Target Match Rate 80%				
Note	Total Cases(A)	Number Coded(B)	Production Rate(B/A)	Error Rate
(1)	66363	20150	0.304	0.142
(2)	66363	19717	0.297	0.150
(3)	66381	19656	0.296	0.134
(4)	66381	19729	0.297	0.146

(1) First subgroup is both the cutoff sample and the test sample
(2) First subgroup is the test sample; Second subgroup is the cutoff sample
(3) Second subgroup is both the cutoff sample and the test sample
(4) First subgroup is the cutoff sample; Second subgroup is the test sample

score variance, we doubled the number and picked 50 as a discriminator in the estimations. Second, there are over 165,000 cases used in the estimations and 50 cases is a small proportion over the combined 1980 Large Sample and 1990 Validation Sample. Third, by examining the benchmark cutoff score reports, most of the codes with cases below 50 have a cutoff score of 99,999; i.e., all cases were referred to clerical coding.

4. Some Empirical Results

In this section, some empirical results are presented and shown in Table 2. The target match rates used were 85% for industry and 80% for occupation. The cutoff sample and test sample were created with the combined 1980 Large Sample and 1990 Validation Sample. The code categories with sample size less than 50 were excluded from the estimations. For comparison purposes, the estimations from the available samples

Table 3. Production rate and error rate estimations

Industry with Target Match Rate 85%				
Note	Total Cases(A)	Number Coded(B)	Production Rate(B/A)	Error Rate
(1)	83215	43588	0.524	0.105
(2)	83068	43323	0.522	0.103
(3)	83215	43579	0.524	0.100
(4)	83068	43326	0.522	0.109
(5)	132744	72282	0.545	0.096
(6)	33539	14636	0.436	0.106
(7)	166283	87007	0.523	0.100
(8)	361306	209088	0.579	
(9)	361306	170488	0.472	
(10)	361306	205510	0.569	
Occupation with Target Match Rate 80%				
Note	Total Cases(A)	Number Coded(B)	Production Rate(B/A)	Error Rate
(1)	91210	34731	0.381	0.150
(2)	91374	34427	0.377	0.148
(3)	91210	34389	0.377	0.141
(4)	91374	35102	0.384	0.142
(5)	132744	47334	0.357	0.137
(6)	49840	23173	0.465	0.138
(7)	182584	71949	0.394	0.142
(8)	361306	135185	0.374	
(9)	361306	136295	0.377	
(10)	361306	136217	0.377	

(1) with the test sample and the cutoff sample cutoff scores
(2) with the cutoff sample and the test sample cutoff scores
(3) with the test sample and its own cutoff scores
(4) with the cutoff sample and its own cutoff scores
(5) with the 1980 large sample and its own cutoff scores
(6) with the 1990 validation sample and its own cutoff scores
(7) with the combined 1980 large & 1990 validation sample and its own cutoff scores
(8) with the PES data set and the 1980 large cutoff scores
(9) with the PES data set and the 1990 validation cutoff scores
(10) with the PES data set and the combined 1980 large & 1990 validation cutoff scores

without using independent cutoff scores are also listed in Table 3.

Comparing the results of the first four lines in Table 3 indicates there are no significant changes (within 0.6% for industry and 0.9% for occupation) in the estimations with and without independent cutoff scores. Possible reasons are explained below. Consider using the jackknife method (Miller 1974) to estimate the cutoff score for a code category. First, assume that $k_{m(t)}/m(t)$ is the match rate above the cutoff score estimated from the whole sample of the code category. By removing one case from the sample, it has the following possibilities in estimating a new cutoff score:

1. The case removed has a score $\geq$ the cutoff score: There are two possibilities:
   a. The case matches: The new match rate above or equal to the cutoff score is $(k_{m(t)} - 1)/(m(t) - 1)$. The difference between the old match rate and the new match rate is

$$\frac{k_{m(t)}}{m(t)} - \frac{k_{m(t)} - 1}{m(t) - 1} = \frac{m(t) - k_{m(t)}}{m(t)} \times \frac{1}{m(t) - 1}. \qquad (4)$$

The first term of the right hand side in equation (4) is the estimated error rate which is controlled by the target match rate t. The second term is very small when the sample size is large.

Table 4. Cutoff score frequencies with jackknifing

Industry code	total	cutoff	frequency	percent	Occupation code	total	cutoff	frequency	percent
932	88	99999	88	100.0	867	15	99999	15	100.0
931	159	1518	158	99.4	866	165	2132	139	84.2
		1575	1	0.6			2100	26	15.8
930	128	5000	123	96.1	865	74	2475	63	85.1
		3900	5	3.9			2400	11	14.9
922	197	3000	156	79.2	864	24	99999	24	100.0
		3600	34	17.3	859	129	99999	129	100.0
		2850	7	3.6	856	285	784	284	99.6
921	123	962	123	100.0			832	1	0.4
910	782	962	782	100.0	855	135	666	134	99.3
901	834	3800	833	99.9			832	1	0.7
		4300	1	0.1	853	120	1200	119	99.2
900	156	2800	155	99.4			1600	1	0.8
		3000	1	0.6	849	139	800	138	99.3
893	202	99999	202	100.0			832	1	0.7
892	214	99999	214	100.0	848	40	99999	40	100.0
891	350	99999	349	99.7	844	187	99999	160	85.6
		12800	1	0.3			3400	27	14.4
890	264	2200	242	91.7	843	40	99999	40	100.0
		2000	22	8.3	834	29	99999	29	100.0
Occupation					833	7	99999	7	100.0
889	619	99999	617	99.7	829	29	99999	29	100.0
		8000	2	0.3	828	41	99999	41	100.0
888	234	2800	233	99.6	826	15	99999	15	100.0
		3200	1	0.4	825	105	870	104	99.0
887	114	1400	113	99.1			980	1	1.0
		1600	1	0.9	824	111	784	110	99.1
885	170	1000	136	80.0			1012	1	0.9
		980	34	20.0	823	81	622	80	98.8
883	249	3000	248	99.6			784	1	1.2
		3200	1	0.4	814	23	99999	23	100.0
878	104	8400	103	99.0	813	68	4000	57	83.8
		9600	1	1.0			1600	11	16.2
877	342	784	341	99.7	809	174	3600	173	99.4
		800	1	0.3			3800	1	0.6
876	20	99999	20	100.0	808	267	800	266	99.6
875	45	99999	45	100.0			1000	1	0.4
874	186	99999	184	98.9	806	155	9200	154	99.4
		8000	2	1.1			9600	1	0.6
869	375	5000	375	100.0	804	703	622	703	100.0
868	2	99999	2	100.0	total	9118	same	8918	97.8

Note: The first cutoff score in each code category is the estimated cutoff score from the whole set of sample. "percent" = "frequency"/"total".

Therefore, the value in equation (4) is insignificant and the probability of getting a new cutoff score will be very small. If a new cutoff score exists, it has a larger value.

b. The case does not match: The new match rate above or equal to the cutoff score is $k_{m(t)}/(m(t)-1)$. The difference between the new match rate and the old match rate is

$$\frac{k_{m(t)}}{m(t)-1} - \frac{k_{m(t)}}{m(t)} = \frac{k_{m(t)}}{m(t)} \times \frac{1}{m(t)-1}. \quad (5)$$

The first term of the right hand side in equation (5) is the estimated match rate. The value in equation (5) is greater than that in equation (4). If the sample size is large enough, the probability of getting a new cutoff score is also very small. If a new cutoff score exists, it has a smaller value.

2. The case removed has a score < the cutoff score: The new match rate above or equal to the cutoff score will not change. If a new cutoff score exists, it will have a smaller value. However, the probability of getting a new cutoff score is very small.

Also, the score of each case is assumed to be continuous from 0 to infinity. In AIOCS, the assigned scores are integers, and there are many tie scores in each code category. That also contributes to not getting a new cutoff score when the jackknife method is used.

An experiment was performed on the cutoff sample described in Section 3 by using the jackknife method to estimate new cutoff scores. The results are consistent with what we discussed above; some of them are shown in Table 4 in which the first cutoff score in each code category is the estimated cutoff score from the whole cutoff sample. The "frequency" column in the table gives the number of times that the jackknife method described above provided the estimated cutoff scores in the "cutoff" column. The total estimated probability of having the same cutoff score is 0.978 from Table 4. If the combined 1980 Large Sample and 1990 Validation Sample was used in this analysis, the probability of having the same cutoff score would be even higher. Therefore, the results from Tables 2 and 3 indicate that the estimates are not seriously affected by using the two-stage analysis. In Section 5, a weighted method is proposed without using the two-stage analysis.

5. Approach with Weighted Method

As mentioned in Section 1.3, the 1990 PES data set contained 361,306 cases which were used for validating the 1990 Decennial Census results and those I&O responses were coded by AIOCS as a test. Of the 361,306 cases, at most 150 cases for each code category were randomly selected and triply coded by clerks and experts without regard to the population size (i.e., the number of cases in each code category of the 1990 PES data set). In order to compensate for this sample selection procedure for the validation data, a weighted approach can be used to estimate the error rate and production rate for 1990 production coding. This weighted approach provides a justification of large code category in the 1990 PES data set getting higher weights toward the estimations of the production rate and match rate. It also deals with correcting the earlier elimination of small categories since the elimination of small categories overestimates the production rate and match rate. In this section, the results were obtained without using the

independent cutoff scores. The weighted approach is described below. Let

N_i^p = number of cases in the PES data set for code i,
T_p = number of cases in the PES data set,
C_i^p = number of cases coded in the PES data set for code i,
K_i^p = number of cases matched in the coded PES data set for code i,
N_i^v = number of cases in the 1990 Validation, 1980 Large, or combined sample for code i,
C_i^v = number of cases coded in the 1990 Validation, 1980 Large, or combined sample for code i,
K_i^v = number of cases matched in the coded 1990 Validation, 1980 Large, or combined sample for code i,
P_i^p = proportion of the sample size in the PES data set for code i, i.e.,

$$P_i^p = \frac{N_i^p}{T_p} \tag{6}$$

where

$$T_p = \sum_j N_j^p.$$

The underlying assumptions of this approach are that the estimated production rates and match rates for each code are equal for the PES data set and the 1990 Validation Sample; i.e., for each code i, the production rate is

$$\frac{C_i^p}{N_i^p} = \frac{C_i^v}{N_i^v} \tag{7}$$

and the match rate is

$$\frac{K_i^p}{C_i^p} = \frac{K_i^v}{C_i^v}. \tag{8}$$

After algebraic manipulation, the estimated production rate is

$$\hat{P}_r = \sum_i \frac{C_i^v}{N_i^v} P_i^p \tag{9}$$

and the estimated match rate is

$$\hat{M}_r = \frac{\sum_i \frac{K_i^v}{N_i^v} P_i^p}{\hat{P}_r}. \tag{10}$$

The estimated error rate is

$$\hat{R}_r = 1 - \hat{M}_r. \tag{11}$$

Equations (9), (10), and (11) were used to estimate the production rate and error rate for industry and occupation with target match rates between 65% and 95% (or target error rates between 5% and 35%). Figure 2 shows the graphs of the results. For purposes of comparison, the results from the certified method are also shown in Figure 2. The graphs from Figure 2 indicate that the cutoff method is superior to the certified method. The superiority came from the improvement of efficiency and reliability of the uncertified categories in the certified method if the same target match rate was used. Figure 2 also shows that there is a trade-off between production rate and error rate. Although the basic estimates using the cutoff method may be biased, we believe that the comparisons between the cutoff and certified methods are still valid as described in Section 4. Note that the estimations of the production rate when applying independent cutoff scores to the PES data set (see Table 3) are consistent with the results using the weighted approach. The results from Figure 2 were used to decide which target match rates to select by specifying a desired error rate. A cutoff score for each industry and occupation code category was produced on the basis of the selected target match rates. This cutoff method was successfully implemented in the 1990 Decennial Census I&O coding production.

6. Quality Assurance Sample

A quality assurance (QA) program was conducted for analyzing the computer QA sample in order to determine whether the AIOCS was performing as expected and to determine the error rates by code category. The computer QA sample was selected from the set of responses that the AIOCS assigned both industry and occupation codes, i.e., above the respective cutoff

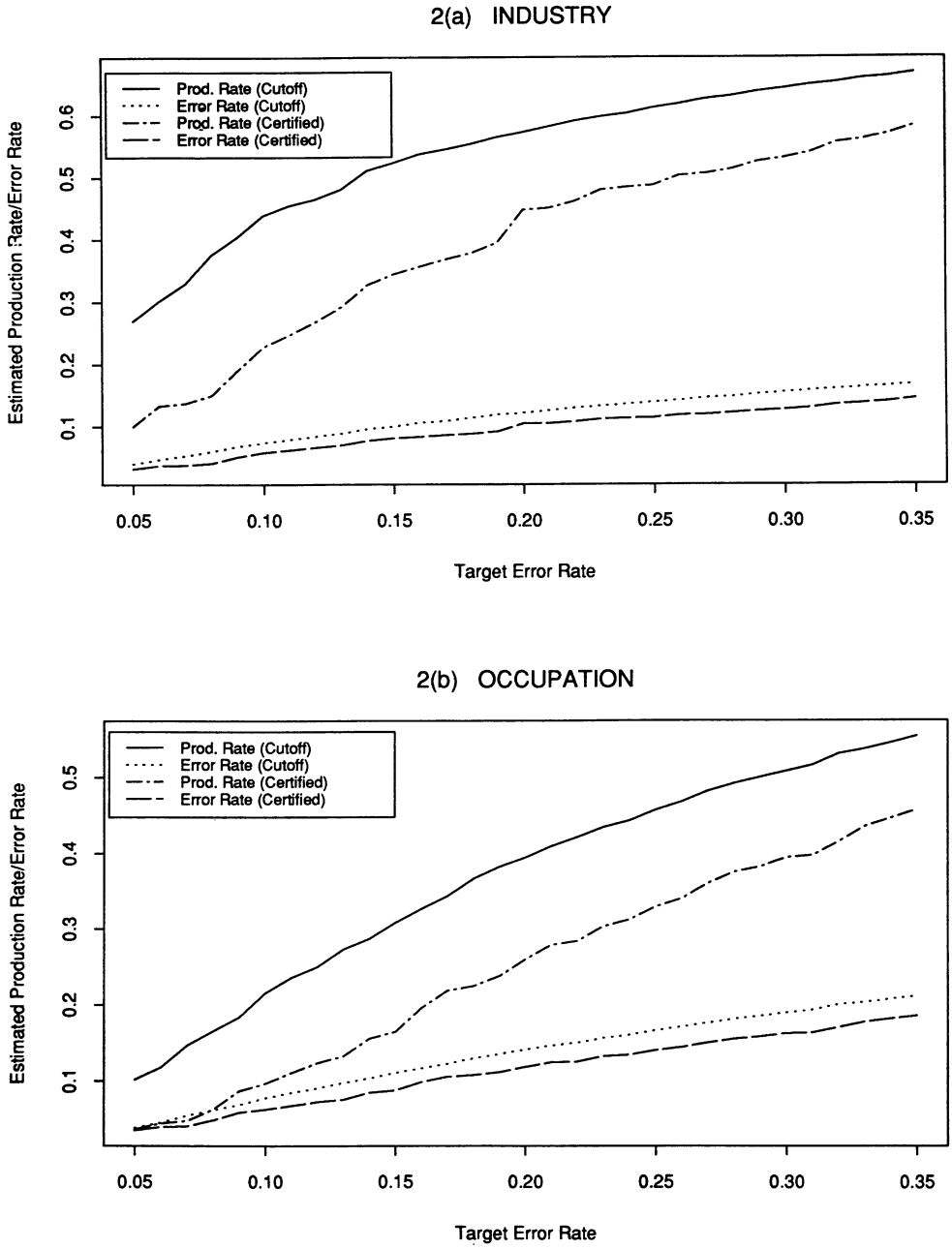

Fig. 2. Estimations with the combined sample

scores. Each sample case was replicated twice and distributed to three different clerks for manual coding. If at least two of the clerks assigned the same code, the majority code was considered the correct code. An error was charged to the clerk who had assigned the minority code. An AIOCS error occurs if the certified code did not agree with the majority code. If a majority code did not exist, the certified code was considered the correct code. The estimated AIOCS error rate is the ratio of the number of computer QA codes in error to the total number of computer QA codes assigned. The estimated error rate from the computer QA sample with total cases of 60,611 was 6.2% for industry and 11.8% for occupation. The actual production rate of the AIOCS was 57.8% for industry and 37.0% for occupation. The estimated error rate from the combined sample that the AIOCS certified both industry and occupation was 7.3% for industry and 12.8% for occupation. Although the error rate estimates from the combined sample are about 1% higher, we think that the estimates are still very close to the computer QA sample error rates.

7. Conclusions

This paper presented the error control procedure that was successfully implemented in the 1990 Decennial Census I&O coding operation. The use of this procedure significantly reduced the clerical effort for industry and occupation coding and saved hundreds of thousands of dollars. The results of this research provide empirical evidence of superiority of the cutoff method over the certified method. For a given target match rate t, the estimated production rate of the cutoff method is higher than that of the certified method and has smaller deviation between the estimated error rate and $1 - t$. We also showed that the effect of using the jackknife method to estimate the production rate and error rate for dealing with the bias problem is very small when the sample size is large and independent estimated cutoff scores are not used.This simplified the computing process of estimating the cutoff score for each code category and significantly reduced the computing time. The key feature of the cutoff method is that it uses a multiple–threshold decision rather than a single–threshold decision. Another key feature of the cutoff method is that the error rate can be controlled to at most $1 - t$. This feature provides users a choice between productivity (high production rate) and quality (low error rate).

Although the only major application of the cutoff method was for the error control of the 1990 Decennial Census Industry and Occupation coding, it can be applied to any classification problem where a decision is made based on a score (or any other measurement). In order to apply this procedure, the score and the probability of the case being correctly classified must be positively correlated.

8. Acknowledgement

We appreciate the comments made by the Chief Editor and three anonymous referees that helped this become a better paper.

9. References

Appel, M.V. and Hellerman, E. (1983). Census Bureau Experiments with Automated Industry and Occupation Coding. Proceedings of the American Statistical Association, Section on Survey Research Methods, 32–40.

Appel, M.V. and Scopp, T. (1987). Automated Industry and Occupation Coding. Paper presented at Development of

Statistical Expert Systems (DOSES), December 1987, Luxembourg.

Chen, B., Appel, M.V., and Creecy, R.H. (1990). Production Rate and Match Rate Estimation for the Automated Industry and Occupation Coding System. Draft report, Bureau of the Census, Washington, DC.

Corbett, J.P. (1972). Encoding from Free Word Descriptions. Draft memo, Bureau of the Census, Washington, DC.

Creecy, R.H., Causey, B.D., and Appel, M.V. (1990). A Bayesian Classification Approach to Automated Industry and Occupation Coding. Proceedings of the American Statistical Association, Statistical Computing Section, Anaheim, August, 1990.

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.

Hellerman, E. (1982). Overview of the Hellerman I&O Coding System. Draft memo, U.S. Bureau of the Census, Washington, DC.

Masand, B.M., Smith, S.J., and Waltz, D.L. (1990). Automated Industry and Occupation Coding on the Connection Machine System. Project report on research at Thinking Machines Corp., Sponsored by Bureau of the Census, Washington, DC.

Miller, R.G. (1974). The Jackknife – A Review. Biometrika, 61, 1–15.

O'Reagan, R.T. (1972). Computer Assigned Codes from Verbal Responses. Communications of the ACM, 15, 455–459.

Panel on Discriminant Analysis, Classification, and Clustering (1989). Discriminant Analysis and Clustering. Statistical Science, 4, 34–39.

Received January 1992
Revised April 1993