# ON ERROR CONTROL OF AUTOMATED INDUSTRY AND OCCUPATION CODING

Bor-Chung Chen, Robert H. Creecy, Martin V. Appel, Bureau of the Census
Bor-Chung Chen, SRD, Rm. 3000-4, Washington, DC 20233

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) data about individuals in the labor force. These hand-written (natural language) responses are solicited from individuals during the Decennial Census and demographic surveys. The 1990 Decennial Census processed approximately 22 million long-form questionnaires with natural language responses to the I&O questions. Based on this set of 6 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: a centralized batch coder called Automated Industry and Occupation Coding System (AIOCS); and a computer assisted clerical coder to aid clerks in coding AIOCS's residuals. The AIOCS is essentially an expert system that builds a lexicon based upon the phrases that appear in the clerical coding manuals and uses pattern matching and a numerical weighting scheme based on an entropy measure ([1], [2]).

A classification problem is to classify an unknown individual $\mathbf{z}$ to one of $k$ populations (or classes) $\omega_1, \omega_2, ..., \omega_k$ on the basis of measurements $z_1, z_2, ..., z_p$ on $p$ characteristics. In this classification problem as in most, there is a tradeoff between classification error rates and production rates. It is obvious that the doubtful cases contribute significantly to the error rates. When a decision on the doubtful cases is not much better than a guess, it might be better not to make a decision at all. Making a decision to reject doubtful cases reduces both the error rates and the production rates. The goal is to reject the fewest cases while maintaining the desired error rate. A number of methods to determine when to reject have been documented. In classification literature, the rejection of classifying a case is known as the *reject option*. Hellman [8] described a classification rule with a reject option using the $(k, k')$ nearest neighbor approach for two classes and provided a bounded value for the error rate of the rule in terms of the Bayes' error rate. Devijver [7] studied a distribution-free lower bound on the Bayes error rate in terms of the asymptotic error rate of the $(k, k')$ nearest neighbor rule with a reject option for two classes described in [8]. The $(k, k')$ nearest neighbor approach is an approach of examining the $k$ nearest neighbors of a test point and making a decision only if they all agree. Both [8] and [7] dealt with only two classes. (Note that the notation $k$ used in the nearest neighbor approach is different from the one used in the number of populations or classes.) Quesenberry and Gessaman [12] described a partial decision discrimination procedure (called the *tolerance region procedure*) using construction of nonparametric tolerance regions with a training sample available from each of $k$ classes. Broffitt, Randles, and Hogg [3] also studied the reject option. They proposed a distribution-free rank procedure in partial decision discrimination problems involving two classes. They provided Monte Carlo investigations on three methods of defining two events so that the occurrence of one event favors classifying an observation to one of the two classes. The three methods are the normal procedure (an approach based on assumptions of normality), the rank procedure, and the tolerance region procedure. The rank procedure was also studied by Randles, Broffitt, and Ramberg [13]. The analysis of the three procedures is performed from the prospect of the respective training samples of the *true* classes. In this paper, we discuss a new reject option called the *cutoff method* which was applied successfully to the 1990 Decennial Census. The analysis of the cutoff method is performed from the prospect of the respective training samples of the *assigned* classes.

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 interviewed for the purpose of validating the 1990 Decennial Census results and those I&O responses were coded by AIOCS as a test. The 1990 Validation Sample was created by randomly selecting no more than 150 cases for each code category from the 1990 PES data set and then having this data set triply coded by clerks. Disputes were 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.

Initially, a method called the *certified method,* was used for controlling error rates. Under this method, a code assigned by the computer was accepted or rejected based on an analysis of AIOCS's coding of the 1980 Large Sample or the 1990 Validation 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 were accepted as final. This computer-expert match rate was called the "certification level." Conversely, code categories with match rates below the certification level were "uncertified" and ALL computer assignments into this category were referred for clerical coding.

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 shows that a significant portion of the sample that was coded 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 automated coder uses for selecting the winning phrase. A description of how to obtain "scores" can be found in [1] and [2].

This paper presents the empirical results of the point estimates of the production rates and error rates of the AIOCS. The cutoff method was implemented for the 1990 census to control the production and error rates. 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. The basic idea is to use a score that is positively correlated with 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." A separate cutoff score for each Industry category and each Occupation category is determined from the coding of the 1980 Large Sample, the 1990 Validation Sample, or the combination of both, such that the match rate is above a specified target match rate. The target match rate can be set separately for Industry and Occupation, or, if desired, separately for each I&O category. Other studies of automated industry and occupation coding include those of [4], [5], [6], and [9].

This paper is organized as follows. Section 2 describes the estimations of the production rate and error rate. The conclusions are given in Section 3.

## 2. Production Rate and Error Rate Estimations

Assume that there are $k$ populations (where $k$ was 243 for industry and 504 for occupation in the 1990 Census I&O Coding): $\omega_i, i = 1, 2, ..., k$; and the classification rule is $D$, where $D = \langle D_1, D_2, ..., D_k \rangle$ and $D$ assigns an individual to $\omega_i$ 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 $\omega_i$; 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 $g_i(x_1) \geq g_i(x_2) \geq ... \geq g_i(x_n)$, $x_j \in D_i$, $j = 1, 2, ..., n$, and $C_j$ is the cumulative match rate; i.e., $C_j = \frac{k_j}{j}$, 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 is an example to illustrate how to obtain a cutoff score with the cutoff method, where $t = 0.85$. Let
$t =$ a target match rate,
$P =$ the true production rate,

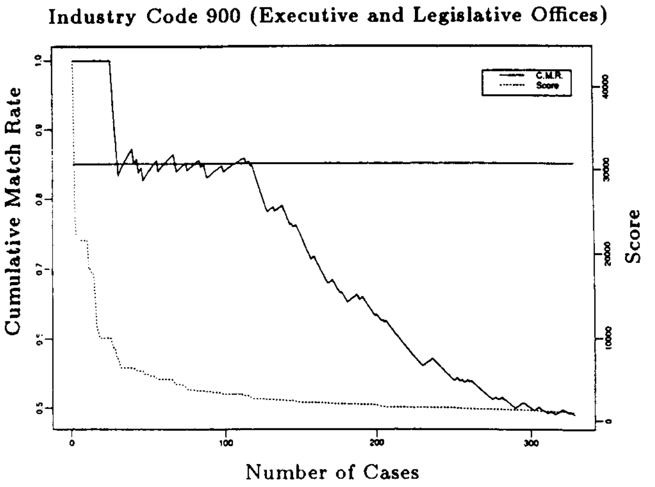

Figure 1. Illustration of the Cutoff Method

$R =$ 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 $\geq g_i(x_{m(t)})$,
$N_i =$ number of $x$ being classified to $D_i$, and
$K_i^c = N_i^c - M_i^c$,
then the estimated error rate is

$$\widehat{R} = \frac{\sum_i K_i^c}{\sum_i N_i^c}, \qquad (1)$$

the estimated production rate is

$$\widehat{P} = \frac{\sum_i N_i^c}{\sum_i N_i}. \qquad (2)$$

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

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, the jackknife [10] may be used for each $D_i$. To simplify the computing process of the jackknife, a typical discriminant analysis method [11] is 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. If (1) and (2) are estimated from the test sample and the independent cutoff scores, which are estimated from the cutoff sample, then the bias in the estimations can be reduced.

### 2.1. Initial Experiment on 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.

To reduce the potential bias of the estimates, each code category in the 1980 Large Sample is randomly divided into two subgroups. One subgroup is the cutoff sample; the other is the test sample. The results indicate that there is a small sample size problem. For those codes with large sample sizes, the estimates of cutoff scores, production rates, and 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.

Since there is already 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 size, 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

$$\Pr(\text{M.R.} > p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}) = 0.90 \qquad (3)$$

with $\alpha = 10\%$, where "M.R." denotes "Match Rate." The sample size, $n = 26$, was selected so that the estimated match rate is at least $p - 0.10$ with 90% confidence. Because of the lack of information on the cutoff 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 99999; i.e. all cases were referred to clerical coding.

### 2.2. Some Empirical Results

In this subsection, some empirical results are presented and shown in Table 1. 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 without using independent cutoff scores are also listed in Table 1.

Table 1. Production Rate and Error Rate Estimations

| | Industry ($t = 0.85$) | | Occupation ($t = 0.80$) | |
|---|---|---|---|---|
| Note | Prod. Rate | Error Rate | Prod. Rate | Error Rate |
| (1) | 0.524 | 0.105 | 0.381 | 0.150 |
| (2) | 0.522 | 0.103 | 0.377 | 0.148 |
| (3) | 0.524 | 0.100 | 0.377 | 0.141 |
| (4) | 0.522 | 0.109 | 0.384 | 0.142 |
| (5) | 0.545 | 0.096 | 0.357 | 0.137 |
| (6) | 0.436 | 0.106 | 0.465 | 0.138 |
| (7) | 0.523 | 0.100 | 0.394 | 0.142 |
| (8) | 0.579 | | 0.374 | |
| (9) | 0.472 | | 0.377 | |
| (10) | 0.569 | | 0.377 | |

(1): test sample & cutoff sample cutoffs
(2): cutoff sample & test sample cutoffs
(3): test sample & its own cutoffs
(4): cutoff sample & its own cutoffs
(5): large sample & its own cutoffs
(6): validation sample & its own cutoffs
(7): combined sample & its own cutoffs
(8): PES data & large sample cutoffs
(9): PES data & validation sample cutoffs
(10): PES data & combined sample cutoffs

Comparing the results of the first four lines in Table 1 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 to estimate the cutoff score for a code category. First, assume that $\frac{k_{m(t)}}{m(t)}$ is the match rate above the cutoff score estimated from the whole sample of the code category. If a case is removed from the sample, one of the following possibilities occurs in estimating a new cutoff score:

- the case removed has a score $\geq$ the cutoff score: There are two possibilities:

  1. the case matches: The new match rate above or equal to the cutoff score is $\frac{k_{m(t)}-1}{m(t)-1}$. 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. 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.

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

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

  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.

- 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 2.1 by using the jackknife method to estimate new cutoff scores. Twelve Industry and thirty-eight Occupation code categories with a total of 9118 cases were used for the experiment. The results are consistent with what we discussed above. The total estimated probability of having the same cutoff score was 0.978. 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, those results indicate that the estimates are not seriously effected by using the two-stage analysis. In Section 2.3, a weighted method is proposed without using the two-stage analysis.

### 2.3. Approach with Weighted Method

In order to compensate for the 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. 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

$$\widehat{P_r} = \sum_i \frac{C_i^v}{N_i^v} P_i^p, \tag{9}$$

and the estimated match rate is

$$\widehat{M_r} = \frac{\sum_i \frac{K_i^v}{N_i^v} P_i^p}{\widehat{P_r}}. \tag{10}$$

The estimated error rate is

$$\widehat{R_r} = 1 - \widehat{M_r}. \tag{11}$$

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%). Figures 2 and 3 are the graphs of the results. For purposes of comparison, the results from the certified method with and without weighting are also shown in the figures. The graphs from Figures 2 and 3 indicate that the cutoff method is superior to the certified method. They also show that there is a tradeoff 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 2.2. Note that the estimations of the production rate when applying independent cutoff scores to the PES data set (see Table 1) are consistent with the results using the weighted approach. The results from Figures 2 and 3 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.

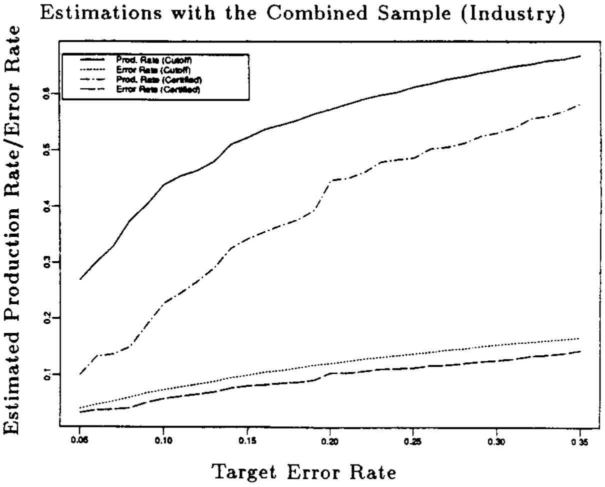

Figure 2. Industry Production and Error Rate Estimations

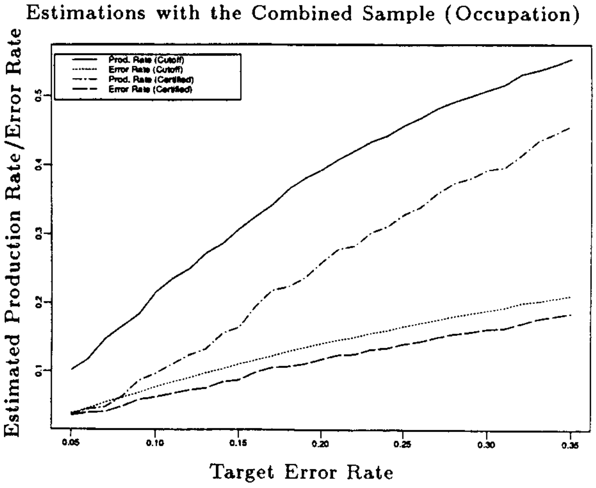

Figure 3. Occupation Production and Error Rate Estimations

## 2.4. Quality Assurance Sample

A quality assurance (QA) program was conducted for analyzing the computer QA sample in order to determine whether the AIOCS is 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 certified both Industry and Occupation, i.e., above the respective cutoff 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, this 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 60611 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.

## 3. Conclusions

In this paper, we presented a new error control method, called the cutoff method, that can be used in classification problems with scores assigned to each classified case, such as the AIOCS. The key feature of the cutoff method is that it uses a multiple-threshold decision rather than a single-threshold decision that most of the other methods do. 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 $1-t$. Therefore, the use of the cutoff method reduced the clerical effort for industry and occupation coding with an estimated savings of hundreds of thousands of dollars.

There are still some open questions which need further research or experiments. Some of them are comparisons of the cutoff method with the normal procedure, the tolerance region procedure, or the rank procedure mentioned in Section 1; more experiments on the variance estimations; estimations of score and cutoff score distributions; and the bias issue of the estimations of the production rate and error rate. The results presented in this paper are very encouraging to continue investigating these issues.

## 4. References

[1] Appel, M. V. and E. Hellerman [1983] "Census Bureau Experiments with Automated Industry and Occupation Coding," *Proceedings of the American Statistical Association,* 32–40.

[2] Appel, M. V. and T. Scopp [1987] "Automated Industry and Occupation Coding," presented at Development of Statistical Expert Systems (DOSES), December 1987, Luxembourg.

[3] Broffitt, J. D., R. H. Randles, and R. V. Hogg [1976] "Distribution-Free Partial Discriminant Analysis," *Journal of the American Statistical Association: Theory and Methods Section.* **71**, 934–939.

[4] Chen, B., M. V. Appel, and R. H. Creecy [1990] "Production Rate and Match Rate Estimation for the Automated Industry and Occupation Coding System," draft report, Bureau of the Census, Washington, DC 20233.

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

[6] Creecy, R. H., B. M. Masand, S. J. Smith, and D. L. Waltz [1991] "Trading MIPS and Memory for Knowledge Engineering: Automatic Classification of Census Returns on A Massively Parallel Supercomputer," Technical Report, Thinking Machines Corporation.

[7] Devijver, P. A. [1979] "New Error Bounds with the Nearest Neighbor Rule," *IEEE Transactions on Information Theory,* **IT-25**, 749–753.

[8] Hellman, M. E. [1970] "The Nearest Neighbor Classification Rule with a Reject Option," *IEEE Transactions on Systems Science and Cybernetics.* **SSC-6**, 179–185.

[9] Masand, B., S. Smith, and D. Waltz [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 20233.

[10] Miller, R. G. [1974] "The jackknife—a review," *Biometrika,* **61**, 1–15.

[11] Panel on Discriminant Analysis, Classification, and Clustering [1989] "Discriminant Analysis and Clustering," *Statistical Science,* **4**, 34–39.

[12] Quesenberry, C. P. and M. P. Gessaman [1968] "Nonparametric Discrimination Using Tolerance Regions," *The Annals of Mathematical Statistics.* **39**, 664–673.

[13] Randles, R. H., J. D. Broffitt, and J. S. Ramberg [1978] "Discriminant Analysis Based on Ranks," *Journal of the American Statistical Association: Theory and Methods Section.* **73**, 379–384.