A COMPARISON OF TWO METHODS OF AUTOMATED INDUSTRY CODING

John H. Rowland, Mark D. Kinack, Statistics Canada
John H. Rowland, 5A4 Jean Talon Bldg., Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

Key Words: Business Register, matching database, data quality

1. Introduction

The Canadian Labour Force Survey (LFS) is conducted monthly on a representative sample of 58,000 households to collect labour force data for civilian Canadian citizens aged 15 and over. Sampling is performed using a stratified multi-stage probability design, and households remain in the survey for six consecutive months with one sixth of the sample replenished each month. For those persons who are currently employed or have worked in the last five years a set of job description questions are asked, including the name of employer and the kind of business or industry, for their most recent job. Of the 100,000 respondents interviewed each month, approximately 77% have a job description collected. The LFS uses a form of dependent interviewing when collecting the job description information in subsequent months, so that only first-time respondents or those whose job description has changed since the previous month require industry and/or occupation coding.

Automated Standard Industrial Classification (SIC) code assignment was introduced into the Labour Force Survey coding procedures in 1978. The procedure involved two steps: the first (known as the Keyword system) was a screening of descriptions based on the presence of certain key words or part of words to detect descriptions which would not be correctly coded in the second step; and the second step was a simple exact matching of descriptions which required SIC code assignment against description/code combinations on a library of descriptions coded in previous months' surveys. If a description requiring SIC code assignment exactly matched one of the descriptions on the library then the code for the matching description was applied. Any descriptions not coded automatically were output for manual coding.

In 1986 the automated coding system was enhanced with the use of the Hellerman algorithm (Hellerman, 1982). The Hellerman matching routine searches a file of previously coded descriptions for one which best matches the description requiring coding. Both the coded description and the respondent's description are standardized such that trivial words have been removed and suffixes taken off words left in the description. The matching is performed using a formula based on common words and the heuristic weights of the words. The heuristic weight of a word measures how specific a word is to a code, i.e., if a word only appears in descriptions with the same code, this word would have a high heuristic weight; conversely if a word appears in a number of descriptions all with different codes, the word would have a low heuristic weight. The phrase with the highest score will be the one which best matches the respondent's description. In practice, before a phrase is used to assign a final code, it must exceed a threshold score and must be at least a certain percentage higher than the phrase with the next highest score.

The data presently being collected includes two descriptions that can be used in assignment of a single SIC code: the name of the respondent's employer and the kind of business or industry in which the respondent worked. The actual questions that are used in collecting this information are:

1) For whom did you work?
2) What kind of business, industry or service was this?

To date only the kind of business or industry has been used in automated SIC code assignment, with approximately a 76% assignment rate and a 7% error rate at the raw code level. This paper compares the rates and quality of automated SIC code assignment using each of the two descriptions.

2. Methodology

To allow valid comparison of automated SIC code assignment using the two different descriptions, it was necessary to use a single automated coding system. The Automated Coding by Text Recognition (ACTR) system is a generalized system developed by Statistics Canada (General Systems Sub-Division, 1989). It includes both direct match and Hellerman routines, and is similar to the system that is currently used in production for the LFS.

Separate matching databases were created for each description. The database for automated SIC code assignment using the kind of business or industry description was created from the files currently used for that purpose in the monthly LFS. These files contain descriptions that were used at least twice in the previous year for automated assignment of SIC codes.

The database used for automated SIC code assignment using the name of the respondent's employer was created from the Canadian Business Register or Central Frame Database (CFDB) (Colledge & Armstrong, 1989). This database contains data on businesses from across Canada. Among the data available are name of business (both legal and operating names), geographical area, number of persons employed, and SIC code. The file used for the study was the final December 1992 file, which contained all of the updates for the last quarter of 1992. This file contained approximately 864,000 establishment and enterprise names, which were reduced to 240,000 by eliminating all establishments with less than ten employees. The reason for this reduction in the number of descriptions input to the database loading process was that the number of records on the full file was too large for the system to handle. As a result the coding rates using the CFDB files found in the evaluations are lower than would have been the case if the full CFDB had been used, since fewer descriptions on the matching database were available for coding. Also, duplicate descriptions (multi-location establishments) were removed, which reduced the number of records input to the loading process to just over 189,000.

Due to the number of non-duplicate descriptions, it was impossible to load the file into a single database. It had to be split into five files by region, each of which was loaded into a separate database. In each of the five cases, approximately 96% of the descriptions were successfully loaded; those records not loaded were rejected as duplicates after standardization.

The default ACTR parsing data were used in the phrase file parsing (or description standardization) routine for loading both databases. This involved two stages: breaking the text into words, and reducing the words to a standard form and order. The standardization of words included replacement of short forms or contractions with complete words, deletion of double characters and suffixes in words, and removal of trivial words and hyphens from the descriptions.

The default ACTR parameters for the Hellerman coding routine were deliberately used in the phrase matching routine. These parameters are used to distinguish "definite" phrase matches from "other" phrase matches through the use of the scores that are assigned to the phrases on the matching database during the automated coding process. If the score of the highest scoring phrase is above the parameter upper threshold and its score exceeds that of the next highest score by at least the parameter percent difference specified, then that phrase is declared a definite winner. If the percent difference in scores between the two highest scoring phrases is less than the percent difference parameter or no score exceeds the upper threshold parameter, then all phrases with scores above the lower threshold parameter are declared to be other winners.

Descriptions requiring SIC code assignment were extracted from the January 1993 final LFS file. All records that were assigned industry and occupation codes in that month were retrieved and any records that were coded by the Keyword system were eliminated on the assumption that such records would continue to be coded in this manner. ACTR was then run in parallel using the name of the employer and the description of the kind of business or industry, which allowed comparisons to be made of both rates of automated SIC code assignment and the codes themselves. Note that while use of a single month's data simplified the coding procedures, it could somewhat limit the interpretability of the results since no attempt was made to control for factors such as seasonality or the state of the labour market.

Codes assigned automatically using the kind of business or industry description were compared against the codes assigned automatically using the name of employer description. If both the name of the employer and the kind or business or industry were uniquely coded to the same SIC, that code was considered correct. If either the name of the employer or the kind or business or industry were assigned multiples codes, the record was output for manual review to determine the correct code. Also output for manual examination were cases where different codes were assigned, or where a code could not be assigned to one or both of the descriptions. Comparisons were then made between the codes assigned automatically using either the kind of business or industry or name of employer descriptions or manually, and data on the quality of the automated coding were derived. The provision of the indeterminate automatically-assigned codes to the manual coders could have had an affect on the manual coding process. This approach was followed in the study since in practise such results would be made available to manual coders to facilitate their task.

3. Results

The data presented below are divided into three parts: definite winners are records to which only a single code was assigned, either by direct match or in the Hellerman routine due to a sufficiently high score; other winners are records to which one or more codes was assigned but the scores for the codes were too close for definite assignment or none had a score sufficiently high for definite assignment; and not coded are records to which a code was not assigned. Note that other winners cannot be considered to be completely coded, since a decision by a manual coder among the alternate codes must still be made.

Table 1

Summary of Coding Rates		Kind of Business/Industry SIC Coding			
		Definite Winners	Other Winners	Not Coded	Total
Name of Employer SIC Coding	Definite Winners	2,184 27.3%	157 2.0%	163 2.0%	2,504 31.3%
	Other Winners	1,917 24.0%	286 3.6%	235 2.9%	2,438 30.5%
	Not Coded	1,787 22.4%	398 5.0%	866 10.8%	3,051 38.2%
	Total	5,888 73.7%	841 10.5%	1,264 15.8%	7,993 100%

The comparison of the rates of automated SIC assignment using the two descriptions clearly demonstrates that the rate of automated SIC code assignment using the kind of business or industry description is substantially higher than the rate of automated SIC code assignment using the name of employer description (see Table 1). Overall, while the rate of automated SIC code assignment using the kind of business or industry description was 84.2%, the rate using the name of employer was 61.8%. In terms of definite winners, the rate using the kind of business or industry description is 73.7%, more than double the rate of automated SIC code assignment using the name of employer description.

Three sets of quality comparisons were performed. The first used raw SIC codes at the three digit level, the second comparison used the first two digits of the SIC code, and the third used only the first digit. This collapsing of codes by digit is appropriate because it results in meaningful grouping of the raw codes into higher level industry categories. These analyses provided a general indication of the relative quality of the codes assigned. Naturally, one would expect the quality of automated SIC code assignment using either the name of employer or the kind of business or industry to improve as the degree of precision decreases.

With regard to the quality of the codes assigned at the three digit level, 90.8% of the records coded using the kind of business or industry description had a correct code assigned, while only 68.4% of the records coded using the name of the employer had a correct code assigned. The difference is even more apparent when only the definite winners are considered. Of the definitely winning codes assigned using the kind of business or industry description, 96.2% were correct while only 71.9% of the definite winners assigned using the name of the employer description were correct (see Table 2).

Table 2

Summary of 3-digit Codes Assigned	Definite Winners	
	Kind of Business SIC Coding	Name of Employer SIC Coding
Correct	5,666 96.2%	1,800 71.9%
Incorrect	222 3.8%	704 28.1%
Total	5,888 100%	2,504 100%

At the two digit level the increase in the percentage of the number of records with correct codes assigned (as compared with the percentage correct at the three digit level) is higher for assignment using the name of employer description than for assignment using the kind of business or industry description. Specifically, of the records coded with the name of employer description, 75.8% had a correct code assigned when examined at the two digit level, an increase of 7.4% over the level of correctness at the three digit level, while of the records coded with the kind of business or industry description, 92.8% had a correct code assigned, an increase of only 2% over the three digit level. Similar changes are found when only the definite winners are examined.

At the single digit level, again the increase in the percentage of the number of records with correct codes assigned is higher for assignment using the name of employer description than for assignment using the kind of business or industry description. Specifically, of the

records coded with the name of employer description, 84% had a correct code assigned when examined at the single digit level, an increase of 15.6% over the level of correctness at the three digit level, while of the records coded with the kind of business or industry description, 94.3% had a correct code assigned, an increase of only 3.5% over the three digit level.

4. Conclusions

In general, it appears that more and better SIC codes are assigned when the kind of business or industry description is used than when the name of employer description is used. There could be many reasons for this. One possibility is that respondent provided information may be better for the kind of business or industry than for the name of employer. Since LFS information is accepted from proxy respondents (on average, slightly more than half of the data is provided by proxy reporters), the exact name of a respondent's employer may not be known to the person providing the information and therefore will be less accurate and result in the description not being coded or inaccurately coded, while the kind of business or industry description requires less precision in reporting and therefore would be more frequently automatically coded and with more accuracy.

It is clear that the quality of the automated coding based on the kind of business or industry is higher than that based on the name of the employer, and that holds regardless of the level at which the assigned codes are examined. However, the quality gap narrows moving from the three digit level to the single digit level, and this supports the idea that it is more difficult to collect name of employer descriptions that can be automatically coded in a household survey with proxy respondents than it is to collect kind of business or industry information that can be automatically coded.

It is also possible that the removal of the small employers from the CFDB would result in a difference in the rates of automated SIC code assignment. The fact that these employers were not included in the matching database, plus the inherent lack of precision in proxy reporting of employer names, could explain a majority of the differences between the rates and quality of automated SIC code assignment. Further research into this question would be required for a definitive answer.

The results of this study clearly demonstrate that a combined system could yield higher automated coding rates than either alone. From Table 1, automated coding using the name of employer description results in assignment of definite winners to 2.0% of records not coded using the kind of business or industry description. This represents 13% of the records currently requiring manual coding.

LFS will be exploring the possibilities for improvements to the existing automated system in the future. The scenario that has been proposed involves attempting automated SIC code assignment using each of the name of employer and kind of business or industry descriptions. More codes would be assigned than are currently, simply because coding using the name of employer description is not presently used. If the potential benefits of adding name of employer coding are cost effective it may be introduced as an enhancement - but that will be only to complement automated coding using the kind of business or industry description.

5. Acknowledgements

We would like to thank the members of the LFS Redesign Automated Coding Working Group for their contributions towards this research, and Doug Drew for helpful comments on an earlier version of this paper.

6. References

Colledge, M., and Armstrong, G. (1989). Statistical Units, Births and Deaths at Statistics Canada after the Business Survey Redesign. Internal paper, February 1989 Revision, Statistics Canada, Ottawa, Canada.

General Systems Sub-Division (1989). Automated Coding by Text Recognition Version 1.06, User's Guide. System Development Division, Statistics Canada, Ottawa, Canada.

Hale, A. (1988). Computer-Assisted Industry and Occupation Coding in the Canadian Labour Force Survey. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, 387 - 395.

Hellerman, Eli (1982). Overview of the Hellerman I&O Coding System. Draft memo. Bureau of the Census, Washington D.C.