SCB STATISTISKA CENTRALBYRÅN

1986-02-03

PROMEMORIOR FRÅN P/STM

NR 21

ON THE USE OF AUTOMATED CODING AT STATISTICS SWEDEN

AV LARS LYBERG

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1986:21. On the use of automated coding at Statistics Sweden / Lars Lyberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1986-21

SCB STATISTISKA CENTRALBYRÅN

1986-02-03

PROMEMORIOR FRÅN P/STM

NR 21

ON THE USE OF AUTOMATED CODING AT STATISTICS SWEDEN

AV LARS LYBERG

On the Use of Automated Coding at Statistics Sweden

Lars Lyberg

0 Abstract

Manual coding is time-consuming and costly, difficult to control, error-prone and boring. To cope with these drawbacks, it appears inevitable to focus on the very basis of manual coding and to consider the possibilities offered by access to a computer of developing a basically new approach.

During the last 15 years Statistics Sweden has conducted a series of experiments in order to investigate the possibilities to automate the coding process. Some of these experiments have been so promising that the method has been used in some of our surveys. This paper reviews these efforts with an emphasis on results and practical problems.

I. Introduction

1 The coding operation and the characteristics of the control problem

Examples of data-processing operations in a survey are editing, coding, key-punching and tabulation. Consider a collective of objects ("elements") of some kind and a set of mutually disjoint categories. Each element belongs to one and only one of these categories. Coding denotes the act of assigning the elements into these categories.

In practice the coding is based upon access to verbal information about the elements of the population or sample under study. This information is usually obtained on schedules in the data collection operation and is entered either by the respondents themselves or by interviewers or enumerators. Unlike certain other kinds of information (numerical data on household expenditures, for instance), verbal information cannot be processed immediately into statistical tables. It must first be coded into different categories where each category is labeled with, for instance, a number. These numbers are called code numbers and the key to these code numbers is called the code. (Naturally, numerical data also may be subject to coding; thus in a census of businesses the objects enumerated can be assigned to categories defined with respect to e.g. total turnover).

The term "coding" is admittedly ambiguous. Attempts have been made to replace it by the term "classification"; this term may be better than coding but it has certain disadvantages. Throughout this paper the term "coding" is used since it is the one most frequently used in the literature. Some other terms in this area are ambiguous as well. For instance, what in this article is called "code number", is in the literature often referred to as "code", and what here is called "code" is often referred to as "code list", "coding standard", "lexicon", or "nomenclature". Still another ambiguity concerns what is to be coded. In the definition above it was postulated that a given element belonged to a certain category. In the literature coding is often described as an operation in which the verbal descriptions or the responses are coded rather than the elements themselves. This common way of describing the situation is easily understood since in many surveys each element is coded with respect to more than one variable.

The coding operation has three components:

(1) Each element in, for instance, a population is to be coded with respect to a specific variable by means of verbal descriptions.

(2) There exists a code for this variable, i.e. a set of code numbers in which each code number denotes a specific category of the variable under study.

(3) There is a coding function relating (1) and (2), i.e. a set of coding instructions relating verbal descriptions with code numbers.

Coding is a major operation in such statistical studies as censuses of population, censuses of business and labor force surveys. Examples of variables are occupation, industry, education, and status.

The problems with coding are of different kinds. As with most other survey operations, coding is susceptible to errors. The errors occur because the coding function is not always properly applied and because either the coding function itself or the code is improper. In fact, in some statistical studies, coding is the most error-prone operation next to data collection. For some variables error frequencies at the 10 % level are not unusual. Another problem is that coding is difficult to control. Accurate coding requires a lot of judgement on the part of the coder, and it can be extremely hard to decide upon the correct code number. Even experienced coders display a great deal of variation in their coding. Thus, there are problems in finding efficient designs for controlling the coding operation. A third problem is that many coding operations are difficult to administer. Coding has a tendency to become time-consuming and costly: for instance, in the 1970 Swedish Census of Population carrying out the coding took more than 300 man-years. In many countries coders in large-scale operations must be hired on a temporary basis, and the consequences for maintaining good quality are obvious. There are even reasons to believe that in the future it might be difficult to obtain even temporary coders for this kind of relatively monotonous work. So there is certainly room for new ideas on the effectiveness of the coding operation.

An overview of the problems with control of coding is given in Lyberg (1981).

2. Coding errors in Sweden

2.1 The 1965 Swedish Census of Population

In 1967 an evaluation study of coding errors in the 1965 Swedish Census of Population was conducted (see Lyberg (n.d.) and Dalenius and Lyberg (n.d.)). From a population of census material comprising about 70 percent of the 1965 population a two-stage sample of verified census schedules was selected. The population was partitioned into four strata and four subsamples were obtained. The evaluation study was confined to the following variables:

(1) Relationship to head of household
(2) Type of employment
(3) Status
(4) Industry

The codes used for variables 1-3 were one-digit-codes; the code used for "industry" was a three-digit-code.

Since we were dealing with four variables and four subsamples we obtained 16 different estimates of error rates. These are given in Table 1 below.

Table 1. Estimates of error rates (%) in production coding in the 1965 Swedish Census of Population.

Subsample	Variable 1	2	3	4
1	1.6	2.7	1.0	14.5
2	1.4	1.6	.6	8.2
3	1.5	3.0	1.3	14.5
4	.7	1.3	1.2	8.7

Subsamples 1 and 3 consisted of totally verified schedules and subsamples 2 and 4 consisted of sample verified schedules. Most of the total verification was done for still inexperienced production coders; this explains the differences in error rates between total and sampling verification.

2.2 The 1970 Swedish Census of Population

In the 1970 Swedish Census of Population the number of variables to be coded increased over that in 1965. For evaluation purposes a sample was drawn from the population of census schedules. A pool of expert coders was used to generate a set of "true" evaluation code numbers for each schedule in this sample. These code numbers were compared with the production code numbers after verification, and this led to estimates of error rates for the different variables on economic activity. These variables were

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Place of work
(7) Type of conveyance to place of work
(8) Number of hours at work

Estimates of error rates for these variables are given in Table 2.

Table 2. Estimated error rates in coding economic activity in the 1970 Swedish Census of Population

Variable	Code	Percent error rate (total population)
(1)	1-digit	4.3
(2)	1-digit	4.7
(3)	3-digit	13.5
(4)	1-digit	3.7
(5)	4-digit	9.9
(6)	1-digit	8.9
(7)	1-digit	11.5
(8)	1-digit	4.4

The error rates for the variables (1), (6), and (7) are probably overestimated, since the code numbers were processed by an optical character recognition machine and we have reason to believe that technical errors in this phase had a minor effect on the error rates for those variables.

The table shows that the multi-digit variables are difficult to code. But also the one-digit variables, a priori considered easily coded, are erroneously coded relatively often. One reason could be that the coding situation is too complex for one coder; i.e. each coder has more variables to manage than he/she can handle.

2.3 The 1975 Swedish Census of Population

The number of variables was smaller in the 1975 Census of Population than in the 1970 Census. Evaluation studies show that the error rates also were smaller in this census than in the 1970 Census. The following variables were studied:

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Type of employment
(7) Type of conveyance to place of work

All of these are one-digit variables except for (3) and (5). In Table 3 estimated error rates are given for these variables.

Table 3. Estimated error rates in coding economic activity in the 1975 Swedish Census of Population.

Variable	Code	Percent error rate (total population)
(1)	1-digit	.6
(2)	1-digit	.6
(3)	3-digit	7.8
(4)	1-digit	.5
(5)	4-digit	3.5
(6)	1-digit	1.0
(7)	1-digit	.5

The results given in this table differ strikingly from those obtained in the 1970 evaluation study. The error rates have dropped for every variable and the fact that the one-digit variables now really seem to be easily coded is most encouraging. The occupation error rate of almost 8 percent is still very serious, but compared to the 13.5 percent rate in 1970 it is a good result. Even better is the estimate for industry.

2.4 Some other studies of error rates at Statistics Sweden

Most of the coding studies at Statistics Sweden have been carried out within the censuses. This is natural since the coding is a very extensive operation in a census. During the last decade interest in coding errors has grown and as a result some evaluation studies have been carried out in other surveys as well. Here some estimates of coding errors from such studies are given.

In Olofsson (1976) an industry error rate of 5.7 % is noted in the 1974 Labor Force Survey. Occupation in the same survey had an error rate of 6.2 %. In Harvig (1973b) an 11 % error rate in occupation coding is estimated for coding data for university graduates. In Harvig (1973a) a 3.2 % error rate is estimated when coding underlying causes of death. In Lyberg et al. (1973) an 8 % error rate is estimated when coding teachers' education. In this case the 95 % confidence interval was 5.9 % - 10.3 %.

Extensive reviews of studies of error rates in industry and occupation coding are given in Lyberg (1983).

3. The need for control

It is imperative in most statistical series that coding control is made part of the overall program for producing the statistics. However, knowledge of the error rate is not enough if we want to be far-sighted. We need to know about the error structure, the reliability of the coding process, different types of errors, the seriousness of different errors and the effects of errors, in order to take suitable corrective measures with respect to the code or the coder.

Several control options are available.

Firstly, by means of, say, the U.S. Bureau of the Census' survey model, it is possible to dissect the coding error in a given coding operation. Such a model can also help strike an appropriate balance between various control efforts with respect to all survey operations. (See Bailar and Dalenius (1969)).

Secondly, manual coding is rather well suited for the application of statistical quality control schemes as originally developed for industrial applications. (See Minton (1969)).

Thirdly, there are certain control schemes designed specifically for coding. Two such main schemes are called dependent and independent verification. (See Lyberg (1981)).

Fourthly, evaluation of coding results provides a basis for the allocation of quality control efforts. We have already given examples of results from different evaluation studies. The results of such studies give suggestions concerning the size and emphasis of the necessary quality control program.

Fifthly, it appears inevitable to focus on the very basis of manual coding and to consider the possibilities offered by access to a computer of developing a basically new approach. This idea is, of course, not in principle new: for instance, at the U.S. Bureau of the Census geographic coding has been conducted by means of computer since 1963. What is new is the suggestion in that agency that the computer be used extensively in the coding of such complex variables as occupation and industry. This suggestion may be viewed as a natural extension of earlier uses of computers in the editing operations.

The remaining part of this paper describes the Swedish efforts in this specific field of automation.

II Automated coding - an overview

4. A bird's-eye view of automated coding

In automated coding we distinguish four operations:

i) Construction of a computer-stored dictionary;
ii) Entering element descriptions into the computer;
iii) Matching and coding;
iv) Evaluation.

4.1 Construction of a computer-stored dictionary

In automated coding a dictionary stored in the computer takes the place of the coding instructions and the nomenclature used in manual coding. Obviously the construction of such a dictionary is a very important task. The construction work could be carried out manually but, when dealing with complex multi-digit variables, using the computer seems to be a better alternative. The resulting dictionary should consist of a number of verbal descriptions with associated code numbers. The descriptions could be a sample from the population to be coded or a sample from an earlier survey of the same kind. Of course an important problem is the size of the sample underlying the dictionary construction. Whether the dictionary is constructed manually or by computer the code numbers appearing in it should ideally be those assigned by the best of the available coders and controlled by means of efficient independent verification procedures.

4.2 Entering element descriptions into the computer

Verbal descriptions are to be entered into the computer. One possible method is to punch the descriptions in a more or less free format. However, this method has some serious drawbacks: first, it consumes a lot of "space", and second, the errors involved in large-scale keypunching of alphabetic information are relatively unknown; moreover, such keypunching is rather costly.

A better alternative would be to have the verbal information directly available for optical character recognition. Unfortunately the recognition of hand-written letters is not yet sufficiently developed for this purpose.

There are reasons to believe that at present the entering of verbal descriptions to the computer is the most important practical problem in designing systems for automated coding.

4.3 Matching and coding

Each element description now put into the computer is compared with the list of occupation descriptions in the dictionary. If an element description agrees with an occupation description (is a "match"), it is assigned the corresponding code number; otherwise it is referred to manual coding.

In an automated coding system we will obtain exact matching for a fraction of all elements only. A primary task in developing such a system is to design criteria for the degree of similarity between input words and dictionary words necessary for them to be considered to match.

4.4 Evaluation

The system must include continuing evaluation studies. Such studies aim at

i) controlling the quality of computerized coding;
ii) improving the dictionary and;
iii) controlling the cost.

Whether automated coding is economical or not is a question to be answered by the evaluation. Are the referred cases more difficult to code than those taken care of by the computer? Does the dictionary need improvement? These and other questions are to be resolved by evaluation.

5. The dictionary

There are two general kinds of algorithms for automated coding: weighting algorithms and dictionary algorithms. Weighting algorithms assign weights to each word-code combination using information from a basic file: when a new record is to be coded the program chooses the code number which is assigned the highest weight for the specific record word. Dictionary algorithms look in a dictionary for words or word strings which imply specific code numbers: when a new record is to be coded the program determines whether the word or word string matches any word in the dictionary. If no match occurs the record is rejected and referred to manual coding.

At the U.S. Bureau of the Census a number of different algorithms have been developed and investigated during the last decades. In some straightforward applications like the geographic coding automated coding has been quite successful. Recent efforts deal mainly with the more complex coding of occupation and industry. Four algorithms are described in Lakatos (1977a, b). Two of them, the O'Reagan and the Corbett algorithms, use dictionary methods. The remaining two, the IMP and the INT algorithms, use the weighting method. The INT algorithm is due to Rodger Knaus, and is further described in Knaus (n.d., 1978a, b, 1979, 1983). Current development work at the U.S. Bureau of the Census is described in Appel and Hellerman (1983)and Appel and Scopp (1985). At Statistics Sweden we have worked with the dictionary approach only. Therefore we have nothing to add with respect to other algorithms.

Thus the computer-stored dictionary is a parallel to the dictionary and the coding instructions used in manual coding. In order to create such a computer-stored dictionary a number of operations must be carried out:

i) Choice of basic material;
ii) Sampling a basic file from the basic material;
iii) Expert coding of the basic file;
iv) Establishing inclusion criteria for dictionary records;
v) Construction of a preliminary dictionary;
vi) Testing and completing the preliminary dictionary.

5.1 Choice of a basic material

The most suitable basic material is the set of filled out forms in the survey under study. To use this is rarely possible - time is not on our side. Instead the basic material must often consist of

i) material from an earlier survey of the same kind; or
ii) material from a pilot survey; or
iii) material from another kind of survey in which the same variable was included.

It should be pointed out, though, that basic material of the desirable kind implied above could be efficiently used when revising a dictionary that has been used in production for a while.

It is important that the basic material be up to date. Structural changes occur in the population; e.g. entry and exit of industry and occupation denominations occur frequently. Also it is possible that the respondent reporting pattern changes over a period of time. One example could be the following. In the 1965 Swedish Census of Population, cleaners used to describe their occupation as "cleaner". In the 1970 census a new term, "local keeper", was used by some cleaners. That term had not existed in 1965 and as a consequence was not represented in the basic material. The result was that the dictionary based on the 1965 census material was not able to code 1970 census individuals describing their occupation as "local keeper".

Basic material as in iii) should only be used in exceptional cases, since the reporting pattern for a certain variable could differ substantially between different surveys due to different modes of data collection.

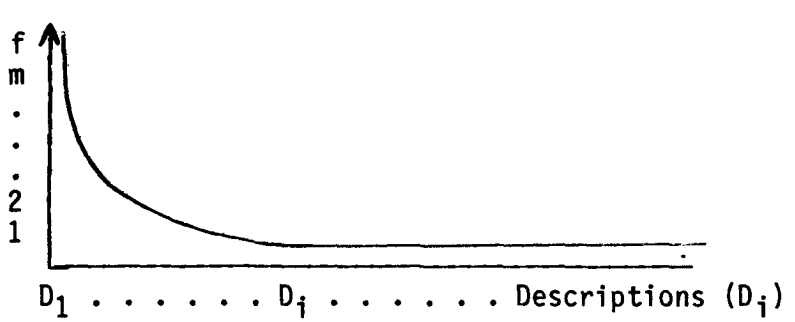
5.2 Sampling a basic file from the basic material

From the basic material we must sample a number of records in order to construct a dictionary. The sampling of records could be carried out in different ways, for instance

- a simple random sample,
- a controlled random sample, or
- a subjective sample.

With the first approach, descriptions with low frequencies have a small probability of being included in the file. This is generally not a negative consequence. Therefore, in almost all of the experiments and applications conducted at our agency we have used simple random sampling. The sample size is a problem, irrespective of the kind of approach we use, since each description should be coded by "experts".

In some of the experiments with automated coding conducted at the U.S. Bureau of the Census, a very large initial random sample of records was chosen: sample sizes of about 100 000 records have been used. In the experiments at Statistics Sweden the basic file has consisted of at most 14 000 records. Despite that, evaluation studies show comparable results. Possible explanations are that a few code numbers and a few dictionary descriptions are, for many variables, sufficient to code a large portion of the records and that the Swedish language is less complex (at least in this context) than English. A typical frequency diagram for unique descriptions is the following:



The typical diagram has a very straggling tail provided that the descriptions are ordered with respect to the frequencies with which they occur. In fact, in some applications many unique descriptions occur only once or twice. In O'Reagan (1972) a closer look revealed that, for one variable, 7 % of the code numbers could handle 50 % of the records. Thus, by means of a rather small initial sample it is usually possible to get a decent dictionary. Our experiences show that vast increases of the basic file (once the "decent" criterion is fulfilled) do not add much with respect to coding degree. An efficient strategy seems to be to concentrate ones' efforts on the most frequently used categories and accept manual coding of most of the remaining part.

5.3 Expert coding of the basic file

In order to construct a good dictionary the basic file has to be coded with high quality, and for this work we have to use the best coders available. Since even "expert" coding is susceptible to errors the expert coding of the basic file must be carried out in conjunction with a control operation.

5.4 Establishing inclusion criteria for dictionary records

The verbal descriptions in an expert coded basic file can be classified into different categories:

a) Descriptions of high frequency which all point at some specific code number;

b) Descriptions of low frequency which all point at some specific code number;

c) Descriptions of high or low frequency with which different code numbers are associated.

In principle, all descriptions pointing at some specific code number should be included in the dictionary. Whether this can be done in practice depends on how large a dictionary we can accept. This in turn is a function of the searching time of the matching program. If the searching time is independent of the size of the dictionary and if having an extensive dictionary does not imply lots of manual administrative work, then all descriptions pointing at specific code numbers should be included. Otherwise we must define what is meant be "high frequency". This decision depends on sample size and number of categories of the code among other things; for instance a small enough basic sample generates no highly frequent descriptions at all. A simple piece of advice is to have a low value of the concept "high frequency" f, say $f \geq 3$, since it is always easier to remove than to add descriptions to the dictionary.

Descriptions belonging to category c) should not be part of the dictionary. There are possible exceptions, though. If, for a vast majority of the cases, a high-frequency description is associated with a specific code number, then an inclusion might be considered. Of course, if such a description is included we end up with deliberately built-in erroneous classifications. Even if such error rates are admittedly small, it is probably better to change the nomenclature so that the coding of this specific description becomes unambiguous in the first place.

5.5 Constructing, testing, and completing the preliminary dictionary

A dictionary can be constructed by man or by computer. Presumably a combination of the two is the most effective. In our first experiments at Statistics Sweden we used manually constructed dictionaries but nowadays we have access to a computer program for dictionary construction.

The manual construction of dictionaries can be characterized as trial and error. At Statistics Sweden we have worked with two lists: list No. 1 is the expert coded file sorted with respect to code number and list No. 2 is the same file sorted alphabetically. These lists form the basis for the construction. List No. 1 is used to get some hints about the structure of the verbal descriptions sorted under a specific code number. We choose a frequency limit f for defining "high frequency" descriptions. All descriptions occuring f or more times are stored in the preliminary version of the primary dictionary which is scanned first in automated coding. We call this dictionary PLEX.

In order to increase the coding degree we must include some variants of the high frequency descriptions already stored. One possibility is to recognize discriminating word strings. In the ideal situation one such string represents many variants of a certain description. Thus, after storing the high frequency descriptions we start looking for discriminating word strings. These strings (or rather, parts of words) are stored in a secondary dictionary. This secondary dictionary, called SLEX, is scanned if PLEX fails to code.

List No. 2 is used as a check. Has a description preliminary stored in PLEX been assigned any other code number except for the specific one under study? It is common that a certain description can be associated with different code numbers depending on the code, the coding instructions, and the auxiliary information used by the coders. The alphabetic list helps us identify such descriptions. When they are identified they can be omitted from the preliminary PLEX. The same goes for the associated word strings in the preliminary SLEX. However, as mentioned above, if we deliberately permit a certain degree of erroneous coding some of these ambiguous descriptions may remain. The probability for such a misclassification should be small, though.

Often a number of highly frequent descriptions are omitted because of their lack of unambiguousness. Then one might reconsider the inclusion of low frequent but unambiguous descriptions in PLEX. Another approach is the possibility to transform some ambiguous descriptions into unambiguous ones by means of auxiliary information.

The word strings in SLEX should be common to several descriptions or be parts of special highly frequent descriptions. We have to be sure that SLEX words do not fit PLEX descriptions for other code numbers. SLEX can never be allowed to expand because of the difficulty to keep up its accuracy. The main problem with SLEX is that we do not know in advance how it behaves when new records are coded.

The manual work described above (or similar manual procedures) can to an important extent be carried out by a computer. Two approaches developed by the U.S. Bureau of the Census are presented in O'Reagan (1972) (O'Reagan's algorithm) and in Corbett (1972) and Owens (1975) (Corbett's algorithm).

The computerized dictionary construction system at Statistics Sweden generates a dictionary with two chapters, PLEX and SLEX. PLEX contains unequivocal descriptions and is scanned first. SLEX contains discriminating word strings that fit several different input descriptions. As a consequence SLEX is not as accurate as PLEX and it is scanned only if PLEX fails to code. Our experience shows that it is rather easy to construct a PLEX manually, but that manual SLEX construction is much harder to manage. Thus we have made a program for computerized construction of SLEX. (As a consequence a computerized PLEX is obtained as a simple special case.)

We have tried a few different versions of the program. The present version, a package called AUTOCOD, is described in Bäcklund (1978). All programs are written in PL1. AUTOCOD contains routines for

- the creation of computer, stored dictionaries (PCLEXK)
- the coding of descriptions (PCAUTOK)
- the updating of dictionaries (PCLEXUP)
- the evaluation of dictionaries (PCLEXT)

PCLEXK creates a PLEX and a SLEX. The procedure involves three steps. The program LEXLADD creates space for a possible SLEX. LEXKONS creates PLEX and SLEX. For each PLEX description, say, a six-character abbreviation starting with the first character is tested for inclusion in SLEX. If that abbreviation fits another PLEX description it is rejected and a new abbreviation is created starting with the second character of the PLEX description. The procedure is repeated at most six times; if no valid abbreviation is obtained the procedure goes on to the next PLEX description. Finally LEXLIST lists the dictionaries by means of EASYLIST. Parameters that can be varied include

- possible use of a list of prefixes which, when making a dictionary of, say, goods, removes such word strings as pounds, roll, and pairs
- minimum frequency f_0 (the dictionary inclusion criterion)
- tolerated degree of equivocalness
- minimum length of words in SLEX.

PCAUTOK codes new records by means of PLEX and SLEX. PCLEXUP is used when we want descriptions to be removed from or added to an existing dictionary. PCLEXT is used to evaluate a dictionary when we have access to a material with manual code numbers assigned.

PLEX and SLEX can be updated simultaneously or separately.

6 The use of auxiliary information

In manual coding we often use not only the verbal descriptions for the variable to be coded but also different kinds of auxiliary information. Typically this information consists of descriptions on some related variable. For instance, information on education or industry is sometimes used as auxiliary information when coding occupation.

Of course, auxiliary information can be used in automated coding as well. The necessary conditions are that the auxiliary information is given together with the record descriptions to be coded and that the auxiliary information is also present in the dictionary. Storing auxiliary information in the dictionary and designing the computer pro-

grams to allow this kind of matching and coding present no serious problem per se. Especially the auxiliary information can be used efficiently if the coding is conducted in two steps; results obtained in the first step coding can be used as auxiliary information in the second step coding. If the first step variables are coded manually the resulting code numbers can be punched together with the verbal descriptions of the variables to be coded in the second step. Since punching of verbal descriptions is a time-consuming operation a faster publication of first step results is made possible. The time saving in an extensive investigation such as a census of population may be considerable; especially this is the case if the second step variables are difficult to code. An example of such a case is the occupation coding in the 1980 Census of Population.

7 Evaluation and control

A final and necessary step in an automated system is evaluation and control. Its primary goal is to maintain the prespecified level of accuracy.

The coding degree, p, and the proportion correctly coded, q, are the main characteristics studied for control and evaluation purposes. If N is the number of elements entered into the computer and n is the number actually coded, then $p = n/N$. If m out of the n coded elements are correctly coded, then $q = m/n$. When evaluating an automated coding procedure, p must be judged together with q. Obviously it is more important to have a large value of q than a large value of p. When comparing different results the product pq might be helpful. Unfortunately, this measure must be used with great care. For instance, the combination $p = .5$, $q = .9$ is much better than $p = .9$, $q = .5$. One should strive primarily for a q-value as high as possible. After that one can concentrate on increasing p. This proportion could be increased until q starts to decrease. It is even possible to increase p to the price of a reduction in q, but then the monetary payoff must clearly outweigh the loss in quality.

The cost for manual coding of the proportion 1-p plays an important role in calculating the costs of the entire coding operation, including both automated and manual steps. The descriptions which the computer is unable to code can be more complex than those it does code. Besides, there is a relatively higher fixed cost associated with the manual coding of the proportion 1-p compared with manual coding of all elements and furthermore all manual code numbers must be keypunched. These costs must be considered when evaluating automated coding. However, recent experiences show that in the census application a good bit of the 1-p may be coded without access to the questionnaire which makes the process faster than conventional manual coding.

A secondary goal of the evaluation and control operation is to gather information that can be used as a basis for changes in the dictionaries and the matching programs. Samples can continuously be drawn from the production and coded by skilled verifiers according to some suitable scheme (for instance independent verification). Thus q can be estimated continuously. If q does not meet quality standards the sample under study must be analyzed. What types of descriptions have been erroneously coded and which have not been coded at all? Is the sample extreme in some sense? Are special sections of the code difficult for the computer? Of course we could try to answer these and other questions even if

the actual q-value meets the prespecified standards. However, adjustments should generally be carried out only when the process is out of control or in danger of becoming so.

III Experiments

Over the years Statistics Sweden has experimented a lot with automated coding. Eventually the experimental results became so convincing that it was decided to implement the technique in ongoing surveys. In this section we will present some general experiences from the early experiments that might be informative to other researchers.

8 Industry

The very first experiments with automated coding at Statistics Sweden concerned the industry variable. These experiments were not very successful. The q-values did not exceed .83 and in one experiment it was as low as .69. During this first phase we thought that a system for automated coding had to be rather complicated and even sophisticated. We were convinced that relying on exact matching only would be highly unsatisfactory. The response pattern, we thought, is so complex that the coding degree obtained by exact matching rules would be much too low to pay off. Therefore, we tried some special matching rules. We experimented with measures of the "distance" between respondent descriptions and dictionary descriptions. We also tried to apply Spearman's rank correlation coefficient as a measure of the similarity between these two kinds of descriptions. More recent experiences have shown that, depending on the level of ambition, this might not be necessary.

After these initial experiments we have not been working with the industry variable very much. In fact, almost all the "trial and error" work, in our opinion the very essence in developing methods for automated coding, is still waiting to be done for this variable. It could even be argued that most of the time, industry descriptions without access to auxiliary information are more or less useless to manual coders as well. Thus, descriptions of industry only are unsuitable for automated coding of that variable. We were anxious to show the sponsors some results and we started to cast our eyes in another direction, towards the occupation variable.

9 Occupation

Most of our experiments have concerned the occupation variable. In Table 4 the main experiments are summerized.

As can be seen from the table the q-values range from .85 to .95 which we considered gratifying. During this experimental period the program for automated dictionary construction was refined. The different sophisticated matching rules could be used on request, dictionary sizes varied between 900 and 11000, and we found that SLEX should be used with great care. The quality never increases by means of SLEX. The coding degree increases of course but always to the price of decreased quality. For instance, in experiment No. 7, the quality increased from .84 to .92 when using PLEX only while the associated coding degree decreased from .84 to .69. It is interesting to note that minor changes in PLEX have no effect on the coding degree and the quality. However, there exists an alphabetic list of approximately 11 000 occupation

descriptions. This list, used by manual coders, is not based on knowledge of the empirical response patterns. We were anxious to know what would happen if that list, already available on tape, was used. We did that in experiments 8 and 9 and obtained the coding degrees 40.2 % and 36.0 % respectively. Thus a desk product such as this list cannot compete with a PLEX based on empirical response patterns, although the latter in this study consisted of 1637 descriptions only. When we merged our computerized PLEX with the alphabetic list and used them together as an extended PLEX the coding degree increased to approximately 75 %. Thus, such a combination could be useful.

Table 4. Experiments with automated coding of occupation.

Experiment	Type of dictionary	Survey	Coding degree (%)	Quality or agreement rate (%)
1	Manual	1965 Census	62	95
2	Manual	1970 Census	66	92
3	Manual	1970 Census	74	84
4	Manual	1970 Census	80	90
5	Manual	Labor Force 1974	81	81
6	Computerized (PLEX + SLEX)	1970 Census	69	87
7	Computerized (PLEX + SLEX)	Labor Force 1976	84	85
8	Computerized (PLEX)	Labor Force 1976	69	93
9	Computerized (PLEX)	Labor Force 1976	69	92
10	Computerized and manual combined	Labor Force 1976	74-76	93-94

10 Goods

Goods or purchases is a main variable in household expenditure surveys. Coding of this variable is relatively simple compared with coding, say, occupation. The normal case is not complicated. Observed coding error frequencies from different household expenditure surveys support this assumption. Experiments with this variable, using PLEX only, resulted in q-values around .995 and coding degrees around .68. The fact that this coding is uncomplicated but costly and time-consuming made it an excellent automation prospect.

IV Applications

Automated coding has been applied in some regular productions at Statistics Sweden. The very first application was the coding of goods in the 1978 Household Expenditure Survey. After that automated coding has been applied in coding occupation and socioeconomic classification (SEI) in the 1980 Census of Population and in coding goods in the 1985 Household Expenditure Survey. The procedure is also used in a continuing survey of book loans where authors and book titles are coded, in coding occupation and SEI in the continuing Survey of Living Conditions and in coding occupation in pupil surveys. Other applications are under way.

11 Coding goods in the 1978 and 1985 Household Expenditure Surveys (HES)

11.1 Introduction

In the 1978 HES approximately 5 900 households were supposed to keep a complete diary (CD) of all goods purchased during a two-week period. The rest of the sample, approximately 7 900 households, should keep a simplified diary (SD) of goods purchased during a four-week period.

In the 1985 HES approximately 6 000 households were supposed to keep a diary of goods purchased during a four-week period. The complexity of this diary corresponds to a case somewhere between the 1978 CD and SD. The survey still strives for a reasonable completeness while all the food-stuffs are assigned the same code number.

The survey designs allow continuous delivery of diaries from the respondents during the survey year. Thus the material can be processed in cycles, which might be advantageous in a system with automated coding. In 1978 we were not at all convinced that our system should work in a production environment, so it was decided that, to start with, the coding should be carried out by two parallel systems, one manual and one automated. After two months production the systems were evaluated in order to decide a preferred system to be used during the remaining ten months of production, and automated coding passed the test. In the 1985 survey, automated coding was used from the start.

11.2 The automated system

In 1978 the dictionary construction was step-wise. Extensive efforts were made in creating an initial dictionary. After that continuous revisions were made prior to many cycles. The initial dictionary was based on the dictionary used in the experiments mentioned in Section III together with a list of all descriptions in the experimental material. Each unique description was coded by HES experts. The construction involved a lot of manual work, since the pattern of descriptions had changed during the nine years that had passed since the last 1969 HES from which we had gathered the experimental material. Only a PLEX was constructed, with a 100 % unequivocal rate. This initial dictionary consisted of 1 459 descriptions. In the automated coding procedure, uncoded descriptions were listed alphabetically on an optical character recognition form and code numbers were assigned directly on it. Some of the uncoded descriptions were added to the dictionary in the updating process.

In 1985 the 1978 dictionary (the last version, 4230 descriptions) was used as a starting point. This manual procedure resulted in a first dictionary consisting of 1985 PLEX descriptions. When this is written the 1985 HES is still going on.

11.3 Results

In the 1978 survey 33 cycles were run. During this period 17 different versions of PLEX were used; thus, only a few cycles were coded with identical dictionaries. In Table 5 below the dictionary sizes and coding degree for the cycles are given.

Table 5 Dictionary size and coding degree for the 33 cycles in the 1978 HES

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1	1459	56
2	2	1554	63
3	3	1760	67
4	4	2228	66
5	5,6,7	2464	68,68,63
6	8	1632	64
7	9	1990	53
8	10,11	2451	69,66
9	12	2866	61
10	13	3065	68
11	14	3613	58
12	15,16	3752	72,73
13	17,18	3832	39,70
14	19,20	4011	65,73
15	21,22	4229	51,72
16	23,24,25,26 27,28,29,30	4230	64,67,62,67, 72,65,67,50
17	31,32,22	4230	65,39,67

The coding degree over all cycles was 65 %. As can be seen from the table, the coding degree decreases sharply now and then. This is explained by the fact that CD's are easier to code automatically compared with the SD's and that the proportion of CD's varies between the cycles. As a matter of fact, some cycles contain only one type of diary. An estimate of the coding degree for CD is 70 % and for SD 38 %.

The dictionary was modified prior to most of the cycle runs, at least for the major part of the production. As shown in Table 5, the additions have generally outnumbered the removals. These modifications did not change the coding degree very much, though. A closer look reveals that many dictionary words are used very seldom or not at all and that relatively few dictionary words can take care of most of the input descriptions. All new uncoded descriptions were gathered in a special file. As soon as a new description occurred in at least three households it was included in the dictionary, provided it could be unequivocally coded to a specific category.

Essentially the same procedure is currently used in the 1985 HES. In Table 6 the dictionary size and coding degree for the first 17 cycles are given.

Table 6 Dictionary size and coding degree for the first 17 cycles (until September 1, 1985) in the 1985 HES

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1,2	1985	81,78
2	3	2029	82
3	4,5	2063	82,81
4	6	2126	82
5	7	2156	83
6	8	2176	82
7	9	2207	81
8	10	2228	83
9	11,12,13	2272	83,83,82
10	14,15,16	2370	83,80,81
11	17	2446	83

In coding these HES's it was decided to use PLEX only because of the inefficiency of the SLEX file. We assume that the coding degree goes down 10-15 percentage points when SLEX is dropped. However, for the

65 % coded, the coding quality is high with an error rate less than 1 %. Special evaluation studies showed that the quality of the coding of the remaining part was very good too: the error rate was around 1 %. This rate is far better than the one a SLEX would give. In all, the coding of the 1978 HES was a smooth operation. The key operators found it less boring to punch verbal information for a change. The cost calculations point to the fact that automated coding was 2-5 % cheaper than a conventional manual system. Besides, the system provided some further advantages. Since all descriptions are key-punched the primary material is better documented than when merely the code number is keyed. Thus, it is possible to give more detailed descriptions of the goods contained in the groups for which estimates are provided. Furthermore, since the dictionary manages to code most straight-forward descriptions the remaining manual coding becomes more interesting to the coder. The coding of the 1985 HES is a smooth operation, too. We have learned from the experiences gained in 1978 and cut down on administrative routines.

As can be seen from the tables above, it does not seem worth the effort to make extensive dictionary revisions after a specific point. Quite soon a rather stable coding degree is obtained, which cannot be substantially altered without changing the dictionary construction principle. For the 1978 survey we note that with the third version already we have obtained a coding degree of 67 %. Despite much work and repetitive modifications after that point, we have at best obtained 73 %.

The 1985 survey has a similar pattern so far. The dictionary has not grown as much as it did in the 1978 survey but the growth that actually has taken place (from 1985 to 2446 descriptions) has not affected the coding degree, which is very stable in the interval 81-83 %.

The computer coding is very inexpensive. The coding of tens of thousands of descriptions costs less than 300 SEK. Computer costs for updating dictionaries are about the same. The expensive part is the manual preparation and administration. As mentioned, this part is conducted more efficiently in the 1985 survey.

12 Automated coding of occupation and socio-economic classification in the 1980 Census of Population

12.1 Introduction

In the 1980 Census of Population the coding of occupation and socio-economic classification (SEI) is automated. In short, this automated coding means that personal identifications and the occupation descriptions are punched and matched against a computer-stored dictionary. The dictionary contains occupation descriptions with associated occupation and SEI code numbers.

The occupation code used in the census is built upon the "Northern Standard for Classifications of Occupations" (NYK) which in turn is built upon the "International Standard for Classification of Occupations" (ISCO). The code contains roughly 280 different three-digit categories. The code for SEI is a two-digit code with 14 different categories.

Here we shall concentrate upon the coding of occupation, since the system was originally constructed for this coding. The coding of SEI was added later on and the system is not "perfect" for coding that variable.

12.2 The coding system: an overview

In Figure 1 below, a chart of the coding system is shown.

First, the occupation descriptions and the personal identifications on the census questionnaires are keypunched. The punched information from a questionnaire is called a questionnaire record. A questionnaire record may contain one or two individual records. After the keypunching, the questionnaire records are split into individual records and at the same time punched occupation descriptions are edited.

In the editing process special signs (points, lines etc) and prefixes (1st, vice, etc) are removed and the remaining parts of the occupation description are brought into one sequence.

The punched file is matched against a file containing the economically active population in the census. In this matching we get some unlinked punched records, for example due to the fact that occupation is punched for an individual who is not economically active. These unlinked punched records are not used henceforth. Punched occupation descriptions will be missing for some economically active individuals. This may be due to the fact that some occupation description on a specific questionnaire is missing. Sometimes the description may be present on the questionnaire but it has been omitted in the keypunching process.

All economically active individuals must of course be coded, at least into some of the "trash" categories designed for situations where the occupation is unknown. In connection with the matching, code numbers for type of activity, industry, institutional classification and so on, are obtained from the file of the economically active.

As a result of the matching we get a file which contains among other things:

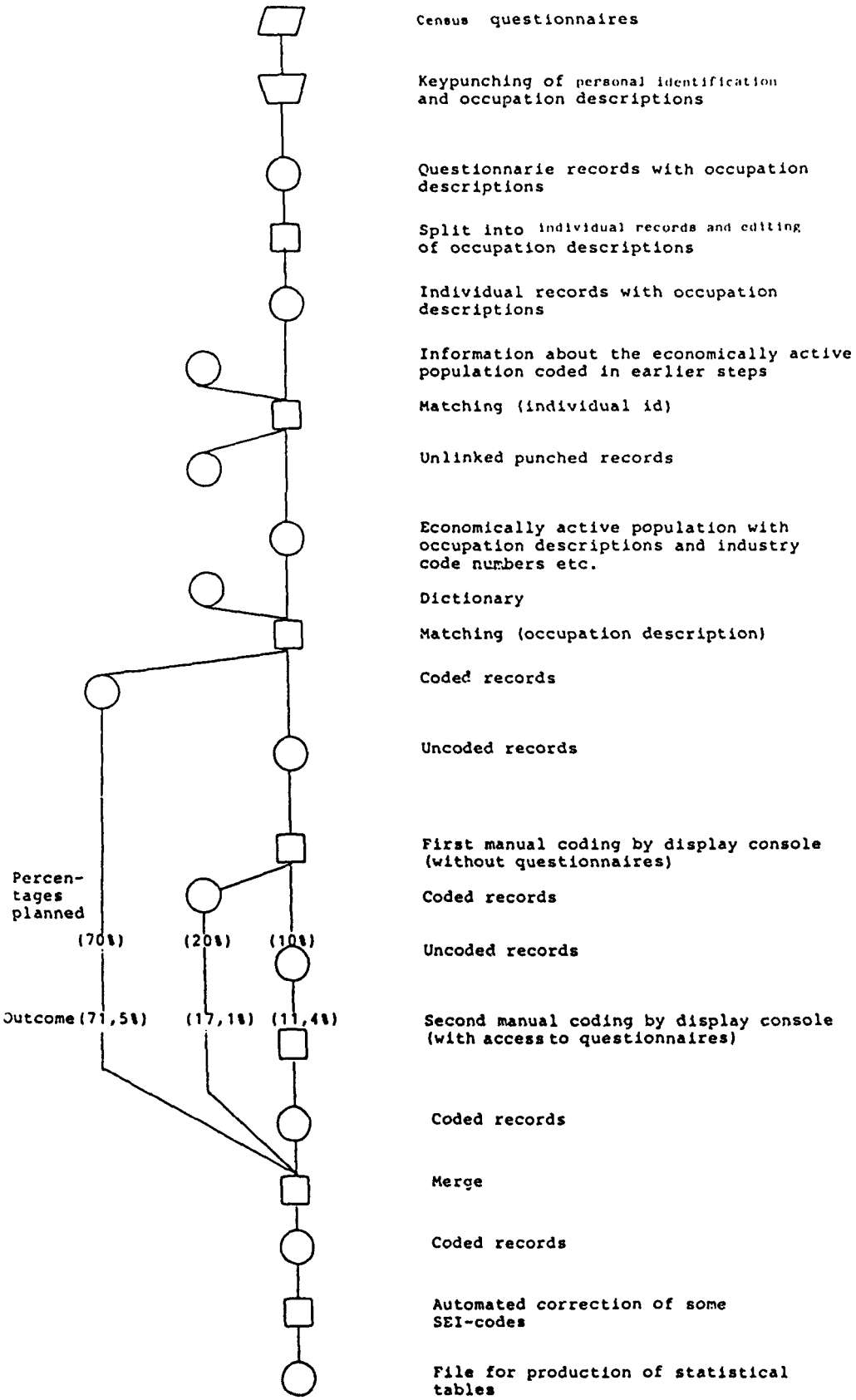
- personal identification
- punched and edited occupation description (with the exception mentioned above)
- industry code number
- institutional classification code number
- size of establishment.

This file is sorted according to edited occupation descriptions and industry code numbers and matched against the computer-stored dictionary. If an edited occupation description is found in the dictionary, then occupation and SEI are coded.

The census occupation dictionary contains both PLEX and SLEX and it is described in more detail below.

The manual coding is carried out on display consoles and in two steps. The first manual coding is carried out without access to the questionnaires. The records which cannot be coded are left "empty" and are coded later on in the second manual coding. In the second step the

Figure 1.
Coding of occupation and SEI in the 1980 Census of Population
An overview



questionnaires are used. Then, the automatically coded records and the records coded in the first and second steps are merged into one file.

At last some SEI-code numbers are automatically corrected. This correction is made by means of a specific question on the questionnaire, where the variable associated with that question has been coded earlier. This question provides information whether the respondent is an employer or an employee, which is an important aspect of the SEIcode.

12.3 The dictionary

The dictionary consists of a PLEX and a SLEX. An excerpt from PLEX is given below.

ALFABETISKT PRIMÄRLEXIKON

YKOD	SEI		NÄRG	SEKT	S
793	11	TAKARBETARE			
793	11	TAKLÄGGARE			
793	11	TAKMONTÖR			
032	60	TANDL			
032	60	TANDLÄKARE			
044	36	TANDSKÖT			
044	36	TANDSKÖTERSKA			
044	36	TANDSKÖTERSKEELEV			
044	36	TANDSKÖTERSKEPRAKTIKANT			
744	46	TANDTEKNIKER			
032	60	TANDVLÄKARE			
044	12	TANDVÅRDSBITRÄDE			
633	12	TANKBILSCHAFFÖR			
801	11	TAPETFABRIKSARBETARE			
781	21	TAPETSERARE	50		
714	21	TAPETSERARE	*		
781	21	TAPETSÖR	50		
714	21	TAPETSÖR	*		
801	21	TAPETTRYCKARE			
504	11	TAPPARE	2		
821	11	TAPPARE	31		
772	11	TAPPARE	33		
736	11	TAPPARE	37204		
731	11	TAPPARE	37		

There must be an exact agreement between an input occupation description including any auxiliary information and a PLEX dictionary description to be considered a "match". PLEX is using industry (NÄRG), institutional classification (SEKT) and size of establishment (S) as auxiliary information.

Since the coding operation was carried out in two steps, we have a most favorable situation for automated coding. First, type of activity, industry and some other variables were manually coded. Then the automated coding of occupation and SEI was carried out. As already pointed out, this results in certain time-savings when it comes to publishing results concerning the variables coded in the first step. Besides,

two-step coding makes it possible to use the auxiliary information in the automated coding process. We believe that the good result of the automated coding in the 1980 Census (presented below) is, to a large extent, due to the fact that auxiliary information was used.

When studying the excerpt above it is, for instance, seen that the description "TAKLÄGGARE" (Roof builder) always gets the code number 793 and 11 for YKOD and SEI, respectively. The description "TAPETSERARE" (Upholsterer) gets the code numbers 781 and 21, respectively, provided the industry code number is 50 (construction). For all other industry code numbers (=*) the description "TAPETSERARE" gets the code numbers 714 and 21, respectively.

SLEX contains parts of words of the type "ADJUNK" (part of the word ADJUNKT which means something like "assistant master at secondary school"). The purpose is that such a part shall fit many variants of a specific occupation description. "ADJUNK", for example, fits all variants in the following example:

ADJUNK	052	56	ADJUNKTBIMAKEHÖGSTADIET 93101
ADJUNK	052	56	ADJUNKTBIOLOGIKEMI 93101
ADJUNK	052	56	ADJUNKTBIOLOGIKEMI 93101
ADJUNK	052	56	ADJUNKTBIOMATEMATIK 93101
ADJUNK	052	56	ADJUNKTEKOÄMNEN 93102
ADJUNK	052	56	ADJUNKTENG 93102
ADJUNK	052	56	ADJUNKTENGELSKAFRANSKA 93102
ADJUNK	052	56	ADJUNKTENGELSKAOCHTYSKA 93101
ADJUNK	052	56	ADJUNKTENGELSKATYSKA 93101
ADJUNK	052	56	ADJUNKTENGELSKATYSKA 93102
ADJUNK	052	56	ADJUNKTENGELSKATYSKAMATEMAT 93102
ADJUNK	052	56	ADJUNKTFILOSOFIOCHMATEMAT 93102
ADJUNK	052	56	ADJUNKTFYSIKOMATEMATIK 93102
ADJUNK	052	56	ADJUNKTFÖRETAGSEKONOMI 93102
ADJUNK	052	56	ADJUNKTGRUNDSKOLANSHÖGSTADIU 93101
ADJUNK	052	56	ADJUNKTGYMNASIE 93102
ADJUNK	052	56	ADJUNKTGYMNASIESKOLANKOMVUX 93102
ADJUNK	052	56	ADJUNKTGYMNASIET 93101
ADJUNK	052	56	ADJUNKTHISTORIASVENSKARELIO 93102
ADJUNK	052	56	ADJUNKTHÖGST 93101

Of course, it happens easily that a certain SLEX-word fits the "wrong" occupation description. It is almost impossible to avoid such mistakes when building SLEX. One way to reduce the coding errors due to a "course" SLEX is to use auxiliary information, for example industry code numbers. In the example above "ADJUNK" in industry 931 (education) is coded with the code numbers 052 and 56, respectively.

Our experience is that a SLEX of occupation descriptions without auxiliary information produces too many coding errors. On the other hand, we believe that it is possible to build a powerful SLEX if one can use words of different length and other auxiliary information besides industry.

Sweden is divided into 24 counties and the census coding is carried out one county at a time. When a county has been matched, two lists are generated.

The first list is the frequency list, from which an excerpt is given below.

STATISTISKA CENTRALBYRÅN 1982-08-19

FOLK-OCH BOSTADSRÄKNINGEN 1980

YRKEN SOM EJ MATCHAT MOT PLEX MED FREKVENS STÖRRE ÄN 1

KOMPRIMERAD YRKESBESKRIVNING	ANTAL
BYGGTRÄ	002
BYRÅDIREKTÖRNATURVÅRD	002
BYRÅINTENDENT	002
BYRÅSEKRFÖRSÄLJNING	002
BÅTTRAFIKÄGARE	002
BÄDDBITRÄDE	002
CEMENTKVARNSOPERATÖR	002
CHARGERARE	002
CHARKARB	002
CHARKARBETARE	003
CHARKFÖRESTÅNDARE	003
CHARKUTERIBITR	003
CHARKUTERISTSTYCKARE	002
CHARKUTERISTTILLVERKNING	002

This list contains those occupation descriptions that the dictionary has failed to code and which occur at least twice in the input file.

In the example above it is seen that, for instance, three records with the occupation description "CHARKFÖRESTÅNDARE" (Butchershop manager) have not been coded.

When the coding of a county is terminated the frequency list is scanned and new occupation descriptions are entered into PLEX. Furthermore, the control lists provide supplementary information for corrections in PLEX. The size of PLEX increased from about 4 000 records to more than 11 000 during the production.

The other list, the SLEX-list, shows the occupation descriptions which have been coded by SLEX. An excerpt is given below.

FOLK-OCH BOSTADSRÄKNINGEN 1980 DATUM 1982-10-06 SIDA 76

POSTER SOM MATCHAT MOT SLEX

ORDDEL	YRK	SEI	REDYRKE	NÄRG
ÖVERLÄ	031	57	ÖVERLÄKAREKIRKLIN	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKIRURGI	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKIRURGKLINIKEN	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKLINIKCHEFLÅNGVÅRD	93310
ÖVERLÄ	031	57	ÖVERLÄKAREKLINIKERCHEF	93310
ÖVERLÄ	031	57	ÖVERLÄKAREMEDKLIN	93310
ÖVERLÄ	031	57	ÖVERLÄKARERÖNTGEN	93310
ÖVERLÄ	031	57	ÖVERLÄKARERÖNTGEN	93310
ÖVERLÄ	031	57	ÖVERLÄKGYNEKOLOGI	93310

In the SLEX-list, coarse coding errors are easily discovered.

SLEX has not increased as much as PLEX, because we have not had enough time to find and try new SLEX words. It contains slightly more than 500 words. As pointed out before, we believe that it would be possible to create a much more powerful SLEX, provided we could use auxiliary information.

The coding degree for the entire production was 71.5 %, roughly 68 % by PLEX and 3 % by SLEX. The coding degree varied between the counties from 67.2 % to 76.6 %. Our overall goal was 70 % so everything went slightly better than planned.

The cost for running the matching program is negligible. Look at the following example. The descriptions for one county with 341 529 economically active individuals were matched against a PLEX containing 10 291 descriptions and a SLEX containing 513 words. The result was:

	Number of coded records	Coding degree (%)
PLEX	246 652	72.2
SLEX	8 339	2.2
Total	254 991	74.7

The cost for this matching and automated coding was 303 SEK.

It should also be mentioned that, according to our census experiences, the keypunching personnel shall be instructed to punch exactly what is written on the questionnaires (up to a pre-specified number of characters, in this case 30). We believe this gives the best combination of punching rate and quality.

12.4 First manual coding

After the matching against the dictionary, almost 30 % of the economically active population remains uncoded. This part must be coded manually. In the census this is carried out in two steps. The first manual coding is carried out on display consoles without access to the questionnaires. Twenty records are shown at the same time on the display console. For each individual the screen shows the occupation description, the code number for industry, institutional classification, size of establishment, and type of activity.

The coders use an alphabethic occupation list containing more than 12 000 official occupation descriptions with associated occupation and SEI code numbers. The principle rule is that the coder must find "exactly" the same occupation in the occupation list as the one shown on the screen. When the occupation is found in the list, the associate code numbers are keyed on the display. Occupation descriptions which cannot be coded are left uncoded and these records are coded in the second manual coding.

We had predicted that 20 % of the total number of records should be coded in the first manual coding. The outcome was 17.1 %. The coders involved in the first manual coding managed to code an average of 217 records per hour (including those that were left uncoded).

12.5 Second manual coding

In the second manual coding the last portion of the records are coded. This coding is carried out on display consoles with access to the questionnaires. We predicted that about 10 % of the records would remain at this last stage. The outcome was 11.4 %.

The second manual coding is very time-consuming. In fact, this step is very similar to conventional coding of the roughly 10 % most difficult descriptions. The coders involved in the second manual coding managed to code an average of 27 records per hour.

As can be seen, this coding in two steps makes the coding life a lot easier. The coding speed of the first step can be kept very high since the coding is carried out without access to the questionnaires.

12.6 Some general remarks

As already mentioned, the resulting coding degree of occupation and SEI in the 1980 Census of Population was 71.5 %. Calculations made prior to the decision to use automated coding showed that a coding degree of 60 % would be profitable.

It should also be pointed out that the high coding speed obtained in the first manual coding is to a large extent due to the fact that the occupation descriptions are entered into the computer, which means that the coding can be carried out without consulting the questionnaires. This saves a lot of time.

Of course, we do not know the exact cost for an imagined system of conventional manual coding of occupation and SEI in this census, but we have reasons to believe that the automated coding saved us at least one million SEK, i.e., about 10 % of the total coding cost for occupation and SEI.

Money, however, was not the only reason for using automated coding. It would have been impossible to get enough coding personnel at Statistics Sweden to do the coding within reasonable time. Automated coding reduced the number of records to be manually coded regarding occupation and SEI from about 4 000 000 to 1 200 000 and made it possible to use two manual coding systems.

We also believe that there is a great value having the occupation descriptions entered into the computer per se. As was the case with purchase descriptions, an occupation description contains more information than does a single code number. This "extra" information might be useful to, for instance, medical researchers.

Unfortunately, no evaluation of the coding quality has been undertaken. However, we know that PLEX results in practically error-free coding of almost 70 % of the economically active individuals. The SLEX-lists were carefully checked throughout the entire production process and bearing in mind that SLEX was used for a few percent of the cases only, we can assume that its relative inaccuracy has no serious impact on the total error rate. We have also, as soon as the coding of one county was terminated, scanned the control lists, i.e., we have listed a sample of records and checked the code numbers in each county. This procedure has lead to continuous improvements of the dictionary and the coding instructions. The lists have given us coarse estimates of the error rates and we strongly believe that the occupation error rate is lower in this census compared with the estimated 8 % rate obtained in the evaluation of the 1975 Census.

13 Other applications

13.1 Coding of occupation, SEI, and union membership in the Survey of Living Conditions (SLC)

In the continuing SLC all numeric and some of the verbal information are keypunched in order to make the editing more efficient. As a by-product, punched verbal information can be used for automated coding. That is the case for the occupation and SEI variables. The punched file is edited and the occupation descriptions are matched against a PLEX dictionary. In case of a match, the code numbers for occupation and SEI are listed together with all the other information punched from the questionnaires. After that, the code numbers automatically assigned are checked by the coding personnel and changed if necessary. Furthermore, uncoded descriptions are coded on the list. The PLEX used in the SLC is part of the PLEX used in the 1980 Census, namely the part that contains no auxiliary information.

This "semi-automated" system works well and recently it has been extended to include the variable "union membership" (60 % coding degree) and there are plans to extend it even further to include the variable "education" as well.

13.2 Coding of occupation in pupil surveys

Statistics Sweden carries out continuing surveys of different pupil groups. The surveys are conducted a certain time after the pupils have finished their education. The purpose is to get information on their present work and plans for the future.

Almost all the information obtained on the questionnaires in these surveys has always been punched for different purposes. In two recent surveys this punched information has been used for automated coding of occupation.

The PLEX used in the pupil surveys is part of the PLEX used in the 1980 Census, namely the part that contains no auxiliary information.

The coding degree obtained is around 50 %. A large portion of the remaining punched occupation descriptions could be manually coded without access to the questionnaires. This coding without consulting the questionnaires is done much faster than conventional coding. This is the same experience that we had with the 1980 Census material; the difference in coding speed between the first and second manual coding is substantial. This is, we believe, an often forgotten advantage of automated coding. That is, the coding speed of the manual part of an automated coding system may be substantially higher than the speed of a conventional, entirely manual coding system.

The experiences of automated coding in the pupil surveys were favorable and the system for automated coding of occupation is now used in those surveys.

13.3 Book loans

The Swedish Author's Fund makes disbursements to authors in proportion to the popularity of their books among borrowers at public libraries in Sweden. This bonus is based on sample data from different libraries and it is distributed once a year. The survey is carried out by Statistics Sweden on a commission basis.

The general data processing situation, where a list of alphabetic keypunched names of authors and book titles is produced, is quite favorable to automated coding. In such a situation it is easy for an automated system to compete with a manual. Even a rather modest coding degree makes the automated system profitable, since the punching is "free of charge". The only requirement is that the computer cost should be less than the manual coding cost on a record-by-record basis.

Only a PLEX dictionary with a 100 % unequivocal rate is considered, since each error could have a substantial effect on the bonuses distributed.

The system has been used since 1978. During that period the dictionary has increased from 6 900 authors and book titles to 65 000. During the same time the coding degree has increased from 33 % to 80 %. Evaluation studies carried out a few years ago revealed that the dictionary containing 65 000 descriptions was not 100 % accurate. Therefore, a revised dictionary containing 35 000 descriptions was created. With this dictionary the coding degree has dropped from 80 % to approximately 68 %. Since the system payed off from the start already, it is profitable with a broad margin.

V The future of automated coding

Obviously, automated coding is a possible option when designing a coding operation, at least for some variables. We believe that its success is a function of language complexity, though. It seems as if the Swedish language is more forgiving than English in this respect.

In most of our experiments and applications we have used methods that are clearly unsophisticated. Efforts with sophisticated methods have not been especially successful but not especially extensive either. The methodological development has probably suffered from the fact that relatively modest coding degrees around 65-70 % have payed off. We should strive for even more profitable systems; we should like the coding degree to jump 10 or 15 percentage points in, for instance, the coding of goods or occupation. This could be done by more sophisticated methods but also by changing the code in some respects. Merging of different categories are sometimes prohibited due to obligations towards the data users. Perhaps it is not too preposterous to make changes in the codes in order to obtain a less costly coding. That option should certainly be considered more often in times of scarce financial resources.

The coding degree can also be improved by storing auxiliary information in the dictionaries and by using more efficient SLEX dictionaries.

Automated coding is here to stay. The Swedish labor market legislation makes it difficult to temporarily hire coding personnel for occasional efforts such as the coding in a census. We have to rely on our permanent staff and automated coding has emerged as the rescue when it comes to cutting work load peaks.

So far, our strategy has been to put the easier variables to a test first. Now we have to proceed to the more difficult ones and make the dictionaries and the supporting routines more efficient.

14 References

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

Appel, M.V. and Scopp, T. (1985): Automated Industry and Occupation Coding. Paper presented to the Census Advisory Committee of the American Statistical Association and on Population Statistics at the Joint Advisory Committee Meeting, April 25, 1985 in Rosslyn, Virginia.

Bailar, B.A. and Dalenius, T. (1969): Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model. Sankhyā, Series B, Vol 31, Parts 3 & 4, pp. 341-360.

Bäcklund, S. (1978): Automatisk kodning. Beskrivning av programvara och programvaruhantering. Memo, Statistics Sweden (In Swedish).

Corbett, J.P. (1972): Encoding from Free Word Descriptions. Memo, U.S. Bureau of the Census.

Dalenius, T. and Lyberg, L. (n.d.): An Experimental Comparison of Dependent and Independent Verification of Coding. Memo from Tore Dalenius to Leon Pritzker.

Harvig, H. (1973a): Kontrollkodning av dödsbevis. Memo, Statistics Sweden (In Swedish).

Harvig, H. (1973b): Kontrollkodningsexperiment på blanketter för inskrivningsuppgifter till högre studier. Memo, Statistics Sweden (In Swedish).

Knaus,R. (n.d.): Syntactically Based Classification from Natural Language Responses. Memo, U.S. Bureau of the Census.

Knaus, R. (1978a): Inference by Semantic Pattern Matching in Industry Classification. Memo, U.S. Bureau of the Census.

Knaus, R. (1978b): Automated Industry Coding - An Artificial Intelligence Approach. Memo, U.S. Bureau of the Census.

Knaus, R. (1979): A Similarity Measure on Semantic Network Nodes. Paper presented at the Classification Society Annual Meeting, Gainesville, Florida, 1979.

Knaus, R. (1983): Methods and Problems in Coding Natural Language Survey Data. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 51-60.

Lakatos, E. (1977a): Automated I & O Coding. Memo, U.S. Bureau of the Census.

Lakatos, E. (1977b):Computerized Coding of Free Verbal Responses. Memo, U.S. Bureau of the Census.

Lyberg, L. (n.d.): Beroende och oberoende kontroll av kodning. Rapport nr 4, Forskningsprojektet FEL I UNDERSÖKNINGAR, Stockholms universitet, Stockholm (In Swedish).

Lyberg, L. (1981): Control of the Coding Operation in Statistical Investigations - Some Contributions. Ph.D. Thesis, Urval No. 13, Statistics Sweden.

Lyberg, L. (1983): The Development of Procedures for Industry and Occupation Coding at Statistics Sweden. Statistical Review, pp. 139-156.

Lyberg, L., Nordling, P., and Elmdahl, J. (1973): Kodningskvaliteten i lärarregistret. Memo, Statistics Sweden (In Swedish).

Minton, G. (1969): Inspection and Correction Error in Data Processing. Journal of the American Statistical Association, Vol 64, pp. 1256-1275.

Olofsson, A. (1976): Kvalitetskontroll av näringsgrenskodningen i AKU hösten -74. Memo, Statistics Sweden (In Swedish).

O'Reagan, R.T. (1972): Computer-Assigned Codes from Verbal Responses. Communications from the ACM, Vol 15, No 6, pp 455-459.

Owens, B. (1975): The Corbett Algorithm for Coding from Free Word Descriptions. Memo, U.S. Bureau of the Census.

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

19 Ersättningens och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)

20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

★Garnisonstryckeriet★Stockholm★1986★