AUTOMATED CODING AT STATISTICS SWEDEN

Ronnie Andersson and Lars Lyberg, Statistics Sweden

I INTRODUCTION

1 The drawbacks of manual coding

Coding is a major operation in such statistical studies as censuses of population, censuses of business and labor force surveys.

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 (nomenclature) 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.

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 by the coder 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 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 operation.

2 The challenge

As illustrated 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. This idea is not in principle new: for instance, at the US 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.

During the last decade we have conducted a series of experiments at Statistics Sweden in order to find out whether or not it is possible to automate the coding process. These Swedish endeavors would have been impossible without the access to work carried out at the US Bureau of the Census and the scientific support we have obtained from that agency.

It is quite clear that some of our experiments should have been analyzed in much more detail. However, our aim has primarily been to investigate if coding can be handled by the computer at all. Thus, we have chosen to conduct many experiments at the price of less detailed analysis. Some of these experiments have been so promising that we have dared to tackle some ongoing surveys with this technique. Swedish applications of automated coding are the coding of goods in the 1978 Household Expenditure Survey, occupation in the 1980 Census of Population, the Survey of Living Conditions and the Pupil Surveys, and, finally, book loans for the Swedish Author's Fund bonus disbursements. Automated coding in the Swedish applications implies that the computer manages to code a portion of all elements. The remaining part has to be taken care of manually. Typical automated coding rates have clustered around 70 %. The applications have proven to be rather successful. However, as was the case with the experiments, the results of these applications ought to be analyzed in more detail.

Extensive reviews of coding problems are given in Lyberg (1981) and Andersson and Lyberg (1983).

II THE MAIN COMPONENTS

We distinguish four operations in a system for automated coding:

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

We discuss them, briefly, in turn. In the remaining part of this paper these topics will be discussed more thoroughly.

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 work could be carried out manually but using the computer seems to be a better alternative in most situations. The resulting dictionary should consist of a number of verbal descriptions with associated code numbers. The dictionary descriptions could be a sample of element descriptions from the population to be coded or a sample from an earlier survey of the same kind.

2 Entering element descriptions into the computer

One possible method for entering descriptions into the computer is to punch them in a more or less free format on cards or magnetic tape. However, this method has some drawbacks: the errors involved in large-scale keypunching of alphabetic information are relatively unknown; moreover such keypunching is relatively costly. A better alternative would be to have the verbal information directly available for optical character recognition. Unfortunately the recognition of handwritten letters is not yet sufficiently developed for this purpose. There are reasons to believe that at present the entering of verbal descriptions into the computer is the most important practical problem in designing systems for automated coding.

3 Matching and coding

Each element description put into the computer is compared with the list of descriptions in the dictionary. If an element description agrees with a dictionary description (is a "match"), it is assigned the corresponding code number; otherwise it is referred to manual coding. In a practical situation some element descriptions have no exact counterparts in the dictionary. Therefore, it is necessary to find methods which make it possible to code different variants correctly, including some spelling and punching errors. A primary task in developing an automated coding system is to design criteria for the degree of similarity between element descriptions and dictionary descriptions necessary for them to be considered to match.

4 Evaluation

A system for automated coding must include continuous evaluation studies. Such studies aim at

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

Examples of questions to be resolved by the evaluation are:

Is the fraction coded automatically economical or not?
Are the referred cases more difficult to code than those taken care of by the computer?
Does the dictionary need improvement?

III AUTOMATED CODING PROCEDURES

1 Algorithms for automated coding

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 record 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 US Bureau of the Census different algorithms have been developed and investigated during the last decades. In some straightforward applications like the geographic coding the automation has been quite successful. Recent efforts, though, deal mainly with the more complex coding of occupation and industry. Four algorithms are described in Lakatos (1977). Two of them, the O'Reagan and the Corbett algorithms, are dictionary methods. The remaining two use the weighting method. One of the weighting methods is due to Rodger Knaus and is described in detail in Knaus (1978a, b). Since then further development has taken place at the Census Bureau. One example is the Hellerman algorithm described in Hellerman (1982).

At Statistics Sweden we have worked with the dictionary approach only. We have nothing to add with respect to other algorithms.

2 The dictionary approach

A record in the dictionary consists roughly of a verbal description, possible auxiliary information and the code number.

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

i) Choice of a 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) Testing and completing the preliminary dictionary.

These operations are now briefly discussed in turn.

2.1 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.

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.

2.2 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.

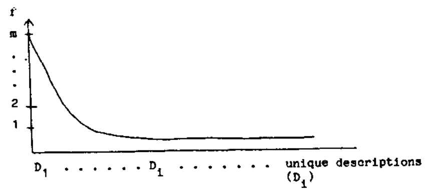
With the first approach, descriptions with low frequencies have a small probability of being included in the file. This is generally not a negative consequence.

The second approach could be realized by means of stratifications. If it is known that the pattern of descriptions varies a great deal within certain parts of the basic material and varies very little within other parts, it may be desirable to use a disproportionate sample by sampling proportionally more records from the parts which vary more.

The third approach can be used when we must consider special cost situations or when we have access to different kinds of subjective prior information.

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 US 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' effort on the most frequently used categories and accept manual coding of most of the remaining part.

2.3 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. Different schemes for independent verification can be possible solutions. Such schemes for independent verification can be possible solutions. Such schemes are discussed in Lyberg (1981). There are reasons to believe that some scheme that is less attractive from a production point of view, can be efficient when coding a basic file, a procedure that is basically a one-time effort.

2.4 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) Variants of descriptions of type a);
c) Descriptions of low frequency which all point at some specific code number;
d) Descriptions of high or low frequency with which different code numbers are associated
e) Variants of descriptions of type d).

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, then all descriptions pointing at specific code numbers should be included. Otherwise we must define what is meant by "high frequency". This decision depends on sample size and number of categories of the code among other things; for instance a small enough sample generates no highly frequent descriptions at all. A simple piece of advice is to have a low value of the concept "high frequency" say ≥ 3, since it is always easier to remove than to add descriptions to the dictionary.

A dictionary can contain descriptions in categories a), b) and c) above. Including descriptions in categories d) and e) results in bad coding quality. We should aim primarily at including descriptions in categories a) and c). After that we consider how many variants should be included. If we want a reasonable automated coding degree, though, a number of variants must be included in the dictionary one way or the other.

2.5 Construction, 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 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 coding of "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 get a coding degree of some magnitude we must include some variants of the high frequency descriptions stored. A possible solution 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 part 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 what supplementary information the coders use. The alphabetic list helps us identify such discriptions. 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, if we deliberately permit a certain degree of erroneous coding some of these ambiguous descriptions may remain. The probability for such misclassification should be small, though.

Often a number of highly frequent descriptions are lost 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 procedures) can to an important extent be carried out by a computer. One approach is given in O'Reagan (1972) (O'Reagans algorithm) and another is given in Corbett (1972) and Owens (1975) (Corbett's algoritm). The computer program created at Statistics Sweden is described in Gustavsson and Karlsson (1978) and Bäcklund (1978).

3 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 auxiliary information is also present in the dictionary. Storing auxiliary information in the dictionary and designing the computer programs to allow this kind of matching and coding present no serious problem. The auxiliary information can be used efficiently if the coding is conducted in two steps; variables coded in the first step can be used as auxiliary information in the second step coding. Such two step coding can be advantageous in a system for automated 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. Punching of verbal descriptions is a time-consuming operation and 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 (see Section V).

4 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. 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 payoff must clearly outweight 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 did 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 some applications 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.

IV EXPERIMENTS

1 Industry

Over the years only three automated industry coding experiments have been conducted and two of them were minor. In the main experiment we tested a computerized dictionary for census data based on 6 000 descriptions in the basic file. The descriptions came from the 1970 census, the computer coded 61 %, and 83 % of these were correctly coded.

The experiments are summarized in Table 1.

The bad results as to quality for the labor force survey experiment is explained by the fact that in that survey the interviewers collecting the data strive for detailed descriptions, and as a consequence the descriptions are sometimes composed of whole sentences. In the censuses the respondents themselves fill in the answers and this usually results in short descriptions. Anyway, in this experiment the SLEX dictionary played a too important role. Many SLEX words fitted the long descriptions provided by the interviewers resulting in bad quality.

The experiments with the industry variable show that the coding degree is relatively low, which perhaps can be accepted, but the errors are too frequent. However, we have not been working with this 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.

2 Occupation

When we first started to deal with the occupation variable we were convinced that the size of the dictionary had to be quite large. In our first experiment the basic file consisted of 14 000 descriptions from census material which was coded using an independent verification scheme. The final dictionary, constructed manually, consisted of 900 descriptions. This first dictionary was a type of PLEX dictionary, but different sophisticated matching rules could be used on request. The first experiment, which was carried out with an independent set of 3 800 occupation descriptions from the 1965 census, gave some encouraging results: the coding degree was 62 % and 95 % of these were correctly coded (having verified manual coding as evaluation standard). This was considered very satisfactory.

Later on several trials were carried out and we introduced both PLEX and SLEX. With both PLEX and SLEX together we sometimes obtained coding degrees around 80 % and qualities around 90 %.

We also tested our program for computerized dictionary construction on 1970 census descriptions. This dictionary has been refined and used in the coding of 1980 census descriptions (see Section V).

Some experiments have been conducted on labor force survey material. In these experiments the computerized construction program has been used on a basic file consisting of 6 000 descriptions. From these descriptions the program created a PLEX containing 1 637 descriptions and a SLEX containing 1 230 discriminating words.

The main experiments are summarized in Table 2.

3 Goods

A main variable in household expenditure surveys is "goods". Coding of goods demands none of the various kinds of judgements which are the big problem in coding, say, occupation. Blood-pudding is almost always blood-pudding and the normal case is not complicated. Observed coding error frequencies from different household expenditure surveys support this assumption. But, as in many other surveys, the coding is costly and time-consuming. Our objective when experimenting with automated coding of goods was to find out if it could be used in the 1978 Household Expenditure Survey (HES).

A sample of 26 000 records (=goods descriptions) was drawn from the 1969 HES. This sample was divided in four different parts. The largest one consisting of 14 000 records, was used as experimental materials.

The first dictionary construction created 686 PLEX words and 600 SLEX words when using a frequency limit = 3, i.e. only those descriptions which occurred three or more times in the basic material were considered for PLEX. The first experiment using the experimental materials gave a 69 % coding degree and the quality was 93 %. Most of the errors were due to SLEX.

In order to increase the coding degree a new dictionary was constructed based on the same basic material but with the frequency limit decreased to 2. This dictionary created 1140 PLEX and 961 SLEX words. The same experimental material was used and the rerun resulted in 75 % coding degree and 90 % quality. Of course, the error rate was unacceptable, but dropping SLEX would increase the quality to 99 % even if the price we had to pay was a decreased coding degree (down to 63 %). After revising PLEX we eventually obtained a coding degree around 68 % with a quality around 99,5 % when coding the remaining experimental materials.

V 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 socio-economic classification (SEI) in the 1980 Census of Population. These are the two major efforts so far. Besides automated coding is used in a minor continuous survey of book loans where authors and book titles are coded. It is also used in coding occupation and SEI in the continuous survey of living conditions and in coding occupation in pupil surveys.

1 Coding goods in the 1978 Household Expenditure Survey (HES)

1.1 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, was supposed to keep a simplified diary (SD) of goods purchased during a four-week period.

The survey design allowed continuous delivery of diaries from the respondents. Thus the material could be processed in cycles, which might be advantageous in a system with automated coding.

1.2 The automated system

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 described in Section IV 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 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.

1.3 Results

During the period from March 15, 1978, to April 26, 1979, 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 3 the dictionary sizes and coding degree for the cycles are given.

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.

The dictionary was modified prior to most of the cycle runs, at least for the major part of the production. As shown in Table 3, the additions have generally outnumbered the removals. These modifications did not change the coding degree very much. 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.

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

2.1 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 a number of occupation descriptions with associated occupation and SEI code numbers.

The coding system is "tailormade" for the census (see Andersson and Flodström (1982) and Andersson (1983)) but of course we have used the experiences made at Statistics Sweden during the last decade (see Lyberg (1981)).

Here we shall concentrate mainly 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.

2.2 The coding system: an overview

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.

As a result of the matching we get a file which contains among other things: personal identification, punched and edited occupation description, industry code number, institutional classification code number and size of establishment.

This file is sorted according to edited occupation description 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 dictionary contains the usual two chapters, PLEX and SLEX and it is described in more detail in Section V.2.3.

The manual coding is carried out on display consoles in two steps. The first manual coding (see Section V.2.4), 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 (see Section V.2.5). In the second step the questionnaires are used. Then, the automatically coded records and the records coded in the first and second steps are merged into one file.

2.3 The dictionary

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, institutional classification and size of establishment as auxiliary information.

Since the coding operation was conducted 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 the variables coded in the first step. Besides, it 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 is, to a large extent, due to the fact that we could use auxiliary information.

SLEX contains word strings of the type "ADJUNK" (part of the word ADJUNKT which means something like "assistant master at secondary school"). The purpose is that one word string shall fit many variants of an occupation description. The word string "ADJUNK", for example, fits many such variants.

Of course, it happens easily that a certain word string in SLEX fits the "wrong" occupation description. It is difficult to avoid such mistakes when building SLEX. One way to reduce the coding errors due to SLEX is to use auxiliary information, for example industry code numbers.

Our experience is that a SLEX for occupations 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 word strings of different length and other auxiliary information besides industry.

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

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

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 mentioned in Section VI give supplementary information for corrections in PLEX. PLEX has grown from about 4 000 records to more than 11 000 during the production.

The second list, the "SLEX-list", shows the occupation descriptions which have been coded by SLEX and coarse coding errors are easily discovered by means of that list.

SLEX has not grown as much as PLEX, because we have not had enough time to find and try new word strings. It contains slightly more than 500 word strings. As pointed out before, we belive 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 goal was 70 % so everything went a little better than planned.

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

	Number of coded records	Coding degree %
PLEX	246 652	72.2
SLEX	8 339	2.4
Total	254 991	74.7

The cost for this matching was 303 Swedish crowns (about 40 US-dollars).

Finally it should 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.

2.4 First manual coding

After the matching against the dictionary almost 30 % of the economically active population remains uncoded. The first manual coding is carried out on display consoles without access to the questionnaires. It is carried out by means of an alphabethical 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 on the display console. When the correct occupation is found in the list, the associate code numbers are keyed on the display. Occupation descriptions which cannot be coded are left empty and these records are coded in the second coding.

We had forecasted that 20 % of the records should be coded in the first manual coding. The outcome was 17.1 %. The rate of coding in the first manual coding was 217 records per hour.

2.5 Second manual coding

In the second manual coding the remaining records are coded. The coding is carried out on display consoles with access to the questionnaires. We forecasted 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 rate of coding in the second manual coding was 27 records per hour.

3 Other applications

3.1 Coding of occupation and SEI in the Survey of Living Conditions (SLC)

In the continuous 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. Then, the code numbers automatically assigned are checked by the coding personnel and altered if necessary. Furthermore, uncoded descriptions are coded on the list.

3.2 Coding of occupation in pupil surveys

Statistics Sweden continuously carries out surveys of different pupil groups. The surveys are taken 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.

3.3 Book loans

The Swedish Author's Fund makes disbursements to authors in proportion to the popularity of their books among borrowers at public libraries. This bonus is based on sample data from different libraries in Sweden and it is distributed once a year. The survey is carried out by Statistics Sweden on a commission basis. The general data processing situation is quite favorable to automated coding: we get a list containing alphabetic keypunched names of authors and book titles; in such a situation it is easy for an automated system to compete with the 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 %. Since the system payed off from the start already, the system is now profitable with a broad margin.

VI EVALUATION OF THE AUTOMATED CODING PROCEDURES

1 Goods

In coding the 1978 HES it was decided to use PLEX only because of the inefficiency of the SLEX file. The price paid is the lower coding degree: we assume that it goes down 10-15 per cent when SLEX is dropped. At the same time, though, coding quality is high with an error rate, for the 65 % coded, of 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 can by no means compete with 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.

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. 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 %.

2 Occupation and SEI

The resulting coding degree of occupation and SEI in the 1980 Census of Population was 71.5 %. Calculations (see Hall (1980)) made prior to the decision to use automated coding showed that a coding degree of 60 % would be profitable.

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 are convinced that the automated coding saved at least one million Swedish crowns (approximately 133 000 US dollars), 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 in time. Automated coding reduced the number of records to be coded from about 4 000 000 to 1 200 000 and made it possible to use two coding systems, first and second manual coding.

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

VII THE FUTURE OF AUTOMATED CODING

Obviously automated coding might be a possible option when designing a coding operation. Its success is a function of language complexity, though. It seems that the Swedish language is more forgiving than English in this respect.

In most of our experiments and applications we have used methods that are rather unsophisticated. Early efforts with sophisticated methods have not been especially successful but not especially extensive either. The methodological development has probably suffered from the fact that rather modest coding degrees around 65-70 % have payed off. We ought to strive for 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 sections. 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. Our labor market legislation makes it difficult to 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.

VIII REFERENCES

Andersson, B. and Flodström, U. (1982): Systemdokumentation delsystem Y, FoB 80, Memo,Statistics Sweden (In Swedish).

Andersson, R. (1983): Automatisk kodning av yrke och socioekonomisk indelning (SEI) i folk- och bostadsräkningen 1980. Metodinformation 1983:3, Statistics Sweden (In Swedish).

Andersson, R. and Lyberg, L. (1983): Automated Coding at Statistics Sweden. Memo, Statistics Sweden.

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.

Gustavsson, R. and Karlsson, J-E. (1978): Automatisk kodning. Memo, Statistics Sweden (In Swedish).

Hall, J. (1980): Kodning av yrke i FoB 80. Memo, Statistics Sweden (In Swedish).

Hellerman, E. (1982): Overview of the Hellerman I&O Coding System. Memo, U.S. Bureau of the Census.

Knaus, R. (1978a): Inference by Semantic Patterns 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.

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

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

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.

Table 1. Experiments with automated coding of industry

Experiment	Type of dictionary	Survey	Coding degree (%)	Quality (% agreement)
1	Manual	1965 Census	50	80
2	Manual	Labor Force 1974	65	69
3	Computerized	1970 Census	61	83

Table 2. 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

Table 3. 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	31,32,33	4230	65,39,67