Survey Methodology, June 1991
Vol. 17, No. 1, pp. 43–56
Statistics Canada

The Blaise System for Integrated Survey Processing

JELKE G. BETHLEHEM and WOUTER J. KELLER[1]

ABSTRACT

Application of recent developments in computer technology allow national statistical offices to produce high quality statistics in an efficient way. At the Netherlands Central Bureau of Statistics (CBS) an increasing use is made of microcomputers in all steps of the statistical production process. This paper discusses the role of software and hardware in data collection, data editing, tabulation, and analysis. To avoid the negative effects of uncontrolled de-centralized data processing, the importance of integration is stressed. This makes the statistical production process easier to manage, and moreover it increases its efficiency. The Blaise System, developed by the CBS, is discussed as a data processing tool that encourages integration. Using a description of the survey questionnaire, this system is able to automatically generate various computer programs for data collection (CAPI or CATI), or data entry and data editing (CADI). The system can also create interfaces to other packages. Particularly, the link between Blaise and the internally developed packages Bascula (for weighting) and Abacus (for tabulation) is described. In this way the Blaise System controls and co-ordinates, and therefore integrates, a large part of the survey process.

KEY WORDS: Integration; Survey processing; CAPI; CATI; Microcomputers; Decentralization; Standardization.

1. INTRODUCTION

The Netherlands Central Bureau of Statistics (CBS) makes an increasing use of microcomputers in survey data processing. The introduction of microcomputers has a considerable impact on the way the work of the statistical office is carried out. Subject matter statisticians become increasingly aware of the potential of the new technology, and consequently use it more and more in their daily work.

This paper discusses the role of the new automation technology in data collection, data editing, tabulation, and analysis. We will stress the importance of standardization and integration. These working policies have three advantages: they enable us to avoid the negative effects of uncontrolled de-centralized data processing, they make the statistical production process easier to manage, and they increase efficiency.

The Blaise System, developed by the CBS, is discussed as the backbone of an integrated survey processing system. On the one hand, the power of this system lies in the consistency it enforces in the various steps of data collection and data processing. On the other hand, it also promotes standardization between different departments. Since all departments use the same software for processing their surveys, everybody speaks the same "language", and so exchange of information between departments is easier and less error prone.

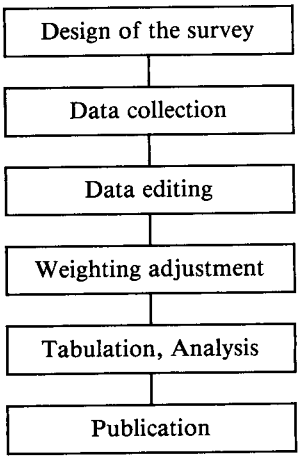
2. THE STATISTICAL PRODUCTION PROCESS

National statistical offices collect data on persons, households and establishments and transform this information into useful statistics. Production of statistical information is often a complex, costly and time-consuming process. This section describes the various steps the

[1] Jelke G. Bethlehem and Wouter J. Keller, Netherlands Central Bureau of Statistics, Automation Department, P.O. Box 959, 2270 AZ Voorburg, The Netherlands.

Table 1

The statistical production process



statistical office has to go through, the problems that it may encounter, and the decisions it has to make. An overview of the process is given in table 1.

The first step is, of course, the design of the survey, in which the statistician specifies the population to be investigated, the data to be collected, and the characteristics to be estimated. Since statistical offices collect most data by means of (sample) surveys, a questionnaire has to be defined, containing the questions to be asked of the respondents. This questionnaire is the first practical description of the data to be collected. Furthermore, in the case of a sample survey, the statistician also has to specify a sampling design, and he must see to it that the sample is selected properly.

The second step in the process is *data collection*. Traditionally, in many surveys the questionnaires are completed in face-to-face interviews: interviewers visit respondents, ask questions, and record the answers on (paper) forms. The quality of the collected data tends to be good. However, since it typically requires a large number of interviewers, who may all have to do much travelling, it can be expensive and time-consuming. Therefore telephone interviewing is sometimes used as an alternative. The interviewers call the respondents from the statistical office, and thus no more travelling is necessary. However, telephone interviewing is not always feasible: only connected people can be contacted, and the questionnaire should not be too long nor too complicated. A mail survey is cheaper still: no interviewers at all are needed. Questionnaires are mailed to potential respondents with the request to return the completed forms. Although reminders can be sent, the persuasive power of the interviewer is lacking, and therefore response tends to be lower in this type of survey, and so does the quality of collected data.

If the data are collected by means of paper forms, completed questionnaires have to undergo extensive treatment. In order to produce high quality statistics, it is vital to remove any errors. This step is called *data editing*. Three types of errors can be distinguished: A *range error* occurs if a given answer is outside the valid set of answers, *e.g.* an age of 348 years. A *consistency error* indicates an inconsistency in the answers to a set of questions. An age of 8 years may be valid, and a marital status "married" is not uncommon, but if both answers are given by the same person, at least in the Netherlands, there is something definitely wrong. The third type of error is the *routing error*. This type of error occurs if the interviewer or the respondent fails to follow the specified branch or skip instructions, *i.e.* the route through the questionnaire is incorrect: irrelevant questions are answered, or relevant questions are left unanswered.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterwards, at the office. In many cases, particularly for household surveys, respondents cannot be contacted again, so other ways have to be found to do something about the problem. Sometimes it is possible to determine a reasonable approximation of a correct value by means of an imputation technique, but in other cases an incorrect value is replaced by the special code indicating the value is "unknown".

In addition to data editing, another activity is sometimes carried out during this stage of the production process: *coding of open answers*. A typical example is the question about the occupation of the respondent. Questions are easiest to process if a respondent selects one possibility from a list of pre-coded answers. However, for a question like occupation this set of pre-coded answers would be very long, and thus it would be very hard for the respondent to select the proper answer. This problem is avoided by letting the respondent formulate his own answer, and then literally copying the answer on the form. To enable analysis of this type of information, answers must be classified afterwards. This is a time-consuming and costtly job, which must be carried out by experienced subject-matter specialists.

After data editing, the result is a "clean" file, *i.e.* a file without errors. However, this file is not yet ready for tabulation and analysis. In the first place, the sample is sometimes selected with unequal probabilities, *e.g.* establishments are selected with probabilities proportional to their size. The reason is that a clever choice of selection probabilities makes it possible to produce more accurate estimates of population parameters, but only in combination with an estimation procedure which corrects for this inequality. In the second place, representativity may be affected by nonresponse, *i.e.* for some elements in the sample the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, the results will be biased.

In order to correct for unequal selection probabilities and nonresponse, a *weighting adjustment* procedure is often carried out. Every record is assigned some weight. These weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status and area reflects the known distribution of these characteristics in the population.

In the case of item non-response, *i.e.* answers are missing on some questions, and not all questions, an *imputation procedure* can also be carried out. Using some kind of model, an estimate for a missing value is computed and substituted in the record.

Finally, we have a clean file which is ready for *analysis*. The first step in the analysis phase will nearly always be tabulation of the basic characteristics. Constructing a table is not as simple as it may look at first sight. The composition of rows and columns (often built from a number of variables), the quantities displayed in cells (counts, means, percentages), the way in which percentages are computed, treatment of multiple-response variables, the position of totals and subtotals, and many other things, can make life very difficult.

Many statistical offices also carry out analysis on their data in order to reveal the underlying structures, and thus to gain insight in the data. Information obtained in this way may improve a later survey, and thus improve quality or reduce costs.

The results of the analysis will be *published* in some kind of report. Usually it will contain tables and graphs. It is important to present the statistical information in such a way that the proper "message" is conveyed. Graphs, tables and text should be simple and clear. Particular attention should be paid to graphs, because a visually ambiguous or confusing graph will quite easily lead to wrong interpretation.

3. THE NEED FOR INTEGRATION

The computer has always been important in statistical information processing. In the beginning to computer was only used for activities like sorting, counting and tabulation. In the sixties and seventies, with the emergence of mainframes and statistical packages, it become possible to carry out extensive analysis. The computer was also increasingly used for data editing, weighting adjustment and imputation. The use of computers for data collection is more recent. This first occurred for telephone interviewing (CATI). In the last decade, the advent of the small laptop computers has made it possible for interviewers to take the computer with them to the homes of the respondents. This way of computer assisted face-to-face interviewing is denoted by CAPI.

It will be clear that the computer is used for more and more activities. Hardware and software are available for nearly every step in the production process. Also an increasing number of people are making use of the automation tools. At first, only the computer specialists had access to "their" machines, but now statisticians and subject matter experts have become computer-oriented, and therefore make increasing demands for suitable software and hardware to do their jobs. Simple and straightforward electronic data processing can, and is, now carried out by the subject matter departments themselves, leaving design and maintenance of complex information systems to be carried out by the computer specialists of the automation department. As a consequence of these developments the work of the statisticians and subject-matter experts have changed. They used to be specialists in their own (narrow) field, but now they have acquired more general knowledge and experience in a much broader field containing subject matter aspects, statistical methodology and computer processing. So the specialists have vanished, and a new group with general knowledge of all aspects of survey processing has emerged.

Automation of the statistical production process is nice, but one should be aware of the dangers. Although application of computers promises increased efficiency and quality, an uncontrolled and unco-ordinated use of the new technology may easily lead to chaos, and hence to less productivity. Factors affecting the efficiency of the statistical production process are:

- **Different departments are involved.**
  Many people deal with the information: respondents fill in forms, subject-matter specialists check forms and correct errors, data typists enter the data in the computer, and programmers construct editing programs. Transfer of material from one person/department to another can be a source of error, misunderstanding and delay.

- **Different computer systems are involved.**
  Various data processing activities may be carried out on different computer systems. Transfer of files causes delay, and incorrect specification and documentation may produce errors.

- **Repeated specification of the data.**
  In almost every step of the process, the structure of the data must be specified. The particular system or department has to know about the data: What is the meaning of the variables? Which values are permitted? Are there any constraints on the routing? Which relationships between variables have to be checked? Although essentially the same, the form of specification may be completely different for every step. Every system uses its own "language". The first specification is the questionnaire itself. Another specification may be needed for data-entry, and yet another for the checking program, for tabulation and analysis, *etc.* It is clear that this is not the most efficient way to deal with the information.

The CBS solution to these problems is *integration*. In this context, integration has three different aspects: integration of work, hardware, and software. Let us first have a look at integration of the work.

Traditional data processing consist of what we call *macro cycles*. All survey data as a whole goes through cycles: from one department to another, and from one computer system to another. First the paper forms are cleaned manually by the subject matter department, then data on the forms are entered by the data entry department, next the files are transfered to a mainframe computer system. A program checks the data for consistency, detected errors are printed on lists that are send back to the subject matter department for corrections. This process of data entering and data editing has to be repeated a number of times before the data can considered to be "clean".

The idea behind integration of work is that the macro cycles should be replaced by *micro cycles*. Not the whole data file, but instead only one record at a time should cycle around. Micro cycles means that cycling should take place within one computer system, and that this should be controlled by one department. Going from macro cycles to micro cycles comes down to concentrating all data processing activities in one department, and that is the subject-matter department. Since the subject-matter statisticians are the ones with most knowledge about the area covered by a survey, they are best equipt to deal with the data, to solve problems, and to produce high quality statistics. Of course, they need proper instruments to do their job, *i.e.* powerful and user-friendly software and hardware.

The idea that automation of data processing activities should be carried out exclusively by computer specialists is out of date. More and more the subject-matter statisticians become aware of the possibilities and usefulness of the computer for their own work. So the time has come for subject-matter departments to take simple and straightforward survey data processing into their own hands. Of course, the automation department is responsible for providing the proper automation infrastructure. And this department stays in charge of design and maintenance of complex information systems.

The second aspect of integration is integration of hardware. The idea is to concentrate work on one type of computer as much as possible. Taking into account that a large number of inexperienced statisticians will have to use the computer, the obvious choice is the microcomputer. Microcomputers offer user-friendliness at a relative low price, and moreover, there is an abundance of useful software.

Being aware of the fact that statistical offices process huge quantities of data, one may wonder whether microcomputers have the capacity to carry out all work, and indeed can take over from the large workhorses, the mainframes. To be able to answer this question, it is useful to distinguish between two kind of data processing activities. In the first place there are record oriented activities. These are activities for which only one record at a time is needed. Examples of record oriented activities are data entry and data editing. Record oriented activities are generally very well suited for interactive processing. In the second place, there are file oriented activities. These activities can only be carried out properly if the whole file is available. Examples are the computation of weights and tabulation. Because of their size, file oriented activities are often processed in a batch-wise fashion.

The viewpoint of a few years ago was that record oriented activies could be carried out on microcomputers but file oriented activites had to take place on mainframes. With the increasing power of microcomputers, attention is shifting in the direction of the microcomputer. At this moment, the policy of the CBS is that all record oriented activities have to be carried out on microcomputers and file oriented activities can in many cases (say, with data files of less than 50 megabytes) also be carried on microcomputers. However, for data storage and large batch jobs we still need mainframes.

The users of the computer environment should be confronted as little as possible with the mainframe. Therefore, the CBS is moving in the direction of front end/back end systems. The front end consists of microcomputers, and that is what the statisticians use to specify their problems. The back end is a mainframe or mini-computer, and is used bulk work, maybe even without the user knowing it. Particularly for database applications the client/server approach looks very promising. In this approach, the real database activities take place on a dedicated minicomputer, whereas the activities are specified, initiated and controlled by the microcomputers at the desks of the users.

4. STANDARDIZATION

The CBS makes an increasing use of microcomputers (running under MS-DOS) in many steps of the statistical production process. On the one hand, this opens new ways towards efficient information processing, but on the other hand, it creates new problems that have to be dealt with. If every department is free to select and purchase its own type of computer and software, the automation infrastructure may easily get out of control, and turn into chaos. Departments will not talk the same "language" anymore, because they use different data formats and different software. It is clear that this calls for a strong policy on standardization. The CBS has adopted such a policy, and in practice it means that there are only one or two software packages available for a particular task.

Another advantage of standardization is that it limits the amount of training that has to be provided for the users. In order to cope with the problem of training a large number of new microcomputer users, the CBS runs an average of 50 one-day courses per month (occupying three fully equipped lecture rooms every working day).

Attention should also be paid to the way in which the microcomputers are used in the organization. Distribution of a lot of stand-alone microcomputers may seem a simple solution, but there are also problems that have to be solved. In the first place, it is very easy to copy (confidential) data files on local hard disks, so we have a data security problem. Furthermore, activities like making back-ups and archiving are often neglected by the users in the subject-matter departments. Also communication between departments (*e.g.* sharing data files) is only possible by exchanging floppy disks. Finally, distribution of new releases of software packages, including their documentation, is often cumbersome in large organizations with a lot of stand-alone microcomputers.

To avoid the above mentioned problems, the CBS has installed approximately 60 local area networks (LANs). Every department has its own LAN. Ten to sixty microcomputers are connected to a high-end 386-based fileserver with a storage capacity of up to 600 Megabytes. In this environment there are in total nearly 2,300 microcomputers, half of them based on the Intel 386SX micro-processor. Security is guaranteed by means of password protection in a login-procedure, by encryption, and by using floppy-less workstations (of the 2,300 microcomputers only 60 have a floppy or hard disk drive). Archiving and backing-up the LANs is carried out in a centralized way by the automation department. A full backup of more than 15 Gigabytes is carried out every night. It is clear that version control and updating software can more easily be realized in such an environment. Distribution and installation of new software releases on a LAN is easy, since, with one command one can upload the new version to all fileservers. All software licenses are based on concurrent usage, which is checked by home-made software.

The role of microcomputers in the statistical production process is growing, but for the time being, there are still applications (like the use of large databases) that need mainframe or minicomputer systems. In this environment, the CBS has adopted Oracle as the standard database system. Development of a database application is preferably carried out on a microcomputer, whereas actually running it takes place on a mini computer. Recently, the CBS realized a client/server architecture based on a distributed database system. Microcomputers in the network serve as front ends and the minicomputers as back ends.

So, as the use of the data processing instruments is brought closer to the subject-matter specialists at the departments (de-centralization), standardization and coordination of the work environment of the subject-matter users demands strong centralization. More details about the automation infrastructure can be found in Keller, Metz and Bethlehem (1990).

5. INTEGRATION OF THE SURVEY PROCESS

The previous section discussed the need for integration in the survey process. Particular attention was paid to concentrating the work in subject-matter departments, and standardization of the hardware and software instruments. But standardization of software is not enough. The efficiency of the production process can be increased even more by integrating the required standard software into one system. This section describes how such an integrated system for survey processing is implemented at the CBS.

An integrated system for survey processing should be based on a powerful language for the specification of questionnaires. This specification is the "knowledge base", containing all knowledge about the questionnaire and the data. The system should be able to exploit this knowledge, *i.e.* it must be able to automatically generate all required data processing applications. On the one hand it means the automatic generation of software for data collection, data entry and data editing, and on the other hand the automatic generation interfaces for other data processing software, *e.g.* for tabulation and analysis. In this way repeated data specification is no longer necessary, and consistency is enforced in all data processing steps.

The backbone of the integrated survey processing system developed by the CBS is the Blaise System. In the design phase of the survey, the questionnaire is specified in the Blaise language. And it is this specification that is used throughout the whole survey process to extract the information necessary to carry out the various data processing steps. Table 2 summarizes the integrated system for survey processing.

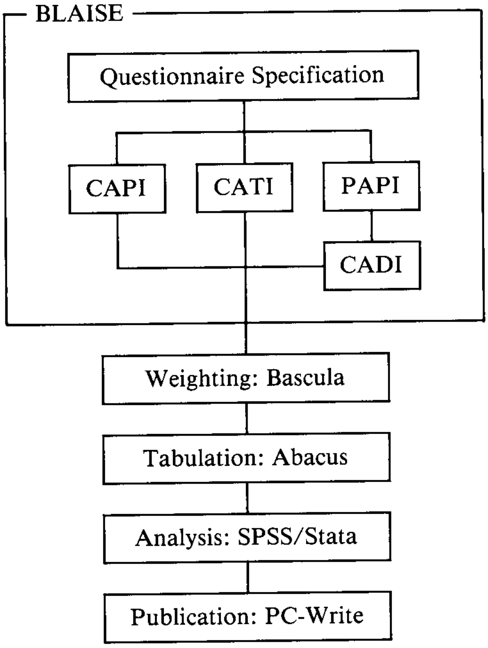
The Blaise System can produce three kinds of programs: CADI, CAPI and CATI programs. CADI stands for Computer Assisted Data Input. It integrates data entry and data editing by offering an interactive environment for processing paper questionnaire forms. The Blaise System can also produce the software required to carry out CAPI or CATI interviewing. The Blaise System is discussed in more detail in section 6.

Whatever form of data collection is used, the result will be a "clean" data file, *i.e.* a file in which no more errors can be detected. The next step in the process will often be the computation of adjustment weights. The program Bascula will take care of this. It is able to read the Blaise data files directly, and extract the information about the variables, *i.e.* the meta-information, from the Blaise specification. Running Bascula will cause an extra variable to be added to the data file containing the adjustment weight for each case. More about Bascula can be found in section 7.

Now the file is ready for tabulation, and for that, the integrated system offers the program Abacus. This program is also able to read and understand the data files created in the previous step of the process. See section 8 for details. Tabulation may be followed by a more extensive analysis of the data. For that purpose the Blaise System can generate interfaces for the statistical packages SPSS and Stata. More about this in section 9.

Table 2

Integrated survey processing



Finally, a publication will be prepared using the standard wordprocessor PC-Write. Since this wordprocessor runs on the same computer system as the other software, it is easy to import generated tables and results of statistical analysis into the text.

6. THE BLAISE SYSTEM

The Blaise System was developed by the CBS, and it derives its name from the famous French theologian and mathematician Blaise Pascal (1623-1662). The basis of the Blaise System is the Blaise language, which is used to create a formal specification of the structure and contents of the questionnaire. The Blaise language has its roots, in large part, in the programming language Pascal.

The Blaise System runs on microcomputers (or networks of microcomputers) under MS-DOS. It is the backbone of the integrated survey processing system, and as it is intended to be used by the people of the subject-matter departments, one need not be a computer expert to use the Blaise System. The design goal of the system was to provide subject-matter experts with a powerful but user-friendly tool that enables them to input their knowledge about a survey into the system, and to take care of all subsequent data processing steps.

In the Blaise philosophy, the first step in carrying out a survey is to design a questionnaire in the Blaise language. Such a specification of the questionnaire contains more information than a traditional paper questionnaire. It not only describes questions, possible answers, and conditions on the route through the questionnaire, but also relationships between answers that have to be checked.

The Blaise System can produce programs for CADI, CAPI or CATI. A CADI program is an intelligent and interactive system for data entry and data editing of data collected by means of paper forms. The subject-matter specialist works through a number of forms with a microcomputer, processing them one-by-one. He enters answers to questions at the proper places and, after completion of the form, he activates the check option to test routing and consistency. Detected errors are reported and explained on the screen. Errors can be corrected by consulting the form or calling the supplier of the information. After elimination of all errors, a clean record is written to a file.

The CADI program can also be used for a different way of data processing not mentioned thus far. Sometimes statistical offices do not carry out their own data collection, but they have to create statistics using data files that were generated elsewhere, outside the statistical office. In these cases the data still has to be checked. The Blaise System has a facility to import this kind of data files. With a CADI program, an integral check can be carried out on all records in a batch-wise version. Thus the records are assigned either the status "clean" or "dirty". And the dirty records can be corrected interactively, again with the CADI program.

A CAPI/CATI program can be used for computer assisted interviewing. The paper questionnaire form is replaced by a computer program containing the questions to be asked. This computer program is in control of the interview. It determines the proper next question to be asked, and checks the answers as soon as they have been entered. In the case of CAPI, the interviewing program is loaded into a laptop computer, and the interviewer takes this computer to the homes of the respondents. In the case of CATI, the program is in a desktop computer. The interviewer calls the respondents from a central unit, and carries out the interview by telephone.

The generation of a Blaise CADI/CAPI/CATI proceeds in a number of steps. First, a text editor is used to enter the Blaise specification of the questionnaire, after which it is checked for syntax errors. Detected errors must be corrected, and to do that the system returns to the text editor and places the cursor on the approximate location of the error. After correction, the specification is checked again. If no errors are detected, the specification is transformed into Pascal source code, which in turn is compiled into an executable program.

The Blaise language must serve two somewhat conflicting purposes. On the one hand it must be powerful enough to be able to deal with all kinds of large and complex surveys, and on the other, Blaise questionnaire specifications must be readable enough, for use by subject matter specialists. In fact, a Blaise questionnaire must be self-documenting, i.e. it is the basic description of the survey which can be used by all people involved. Table 3 gives an example of a simple questionnaire in Blaise.

The first part of the questionnaire specification is the QUEST section, containing the definition of all questions that can be asked. A question consists of an identifying name (for internal use in the questionnaire), the text of the question as presented to the respondents, and a specification of valid answers. The next part of this sample Blaise questionnaire is the ROUTE section. It describes under which conditions, and in which order the questions have to be asked. Consistency checks are specified in the CHECK section.

The description above does not exhaust the power of the Blaise language. An overview of the Blaise language can be found in Bethlehem *et al.* (1989b), and more details in Bethlehem *et al.* (1989c).

The Blaise System contains a module for *interactive coding*, thereby providing the possibility of integrating coding either in the data collection phase or in the data entry and data editing phase. The module contains two different tools. The first tool implements a hierarchical approach to coding. Coding of an answer starts by entering the first digit of the code by selecting

Table 3

A simple Blaise questionnaire

```
QUESTIONNAIRE Work "The Work Survey";

QUEST
  SeqNum    "Sequence number of the interview?": 1..1000 (KEY);
  Age       "What is your age?": 0..99;
  Sex       "Are you male or female?": (Male, Female);
  MarStat   "What is your marital status?":
              (Married   "Married",
               NotMar    "Not married")
  Job       "Do you have a job?": (Yes, No);
  JobDes    "What kind of job do you have?": STRING[20];
  Income    "What is your yearly income?":
              (Less20    "Less than 20,000",
               Upto40    "Between 20,000 and 40,000",
               More40    "More than 40,000");
  Travel    "How do you usually travel to your work?":
              SET [3] OF
              (Walking   "Walking",
               Bicycle   "By bicycle",
               Car       "By car or motorcycle",
               PubTrans  "By bus, tram, train or metro",
               Other     "Other means of transport");
  OthTrans  "What other means of transport?": STRING[20];

ROUTE
  SeqNum; Age; Sex; MarStat; Job;
  IF Job = Yes THEN
    JobDes; Income; Travel;
    IF Other in Travel THEN OthTrans ENDIF
  ENDIF

CHECK
  IF Age < 15 "Respondent is younger than 15" THEN
    MarStat = NotMar "he/she is too young to be married!"
  ENDIF

ENDQUESTIONNAIRE.
```

the proper category from a menu. After the user enters a digit, the program presents a subsequent menu containing a refinement of the previously selected category. So the description becomes more and more detailed until the final digit is reached. The second tool consists of a dictionary approach to coding. It tries to locate an entered description in an alphabetically ordered list. If the description is not found, the list is displayed, starting at a point as close as possible to the entered description. The list can be made so that almost any description, including permutations, is present. The advantage of this method is that it is simple, fast and controllable. Both coding tools can be used simultaneously.

7. BASCULA

The clean file with sample survey data produced by the Blaise System is usually not ready yet for making inference about the population from which the sample has been drawn. The problem is that the data do not constitute a representative sample, and so some adjustment procedure has to be carried out.

In order to account for unequal selection probabilities and nonresponse, one often has to compute adjustment weights. Post-stratification is a well-known technique. Every record is assigned some weight, and these weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population. Two major problems can make application of post-stratification difficult: empty strata and lack of adequate population information. Research has been carried out at the CBS in order to improve weighting techniques. The result was a new general method for weighting, in which weights are obtained from a linear model which relates the target variables of a survey to auxiliary variables. Post-stratification is a special case of this method. Because of the generality of the method, different weighting schemes can be applied that take advantage of the available population information as much as possible, and at the same time avoid the above mentioned problems. See Bethlehem and Keller (1987) for more details.

Bascula is a general weighting program, running on microcomputers under MS-DOS. It combines several weighting techniques. In the first place, traditional post-stratification can be carried out. And if the number of empty strata is small, one can instruct the program to collapse (*i.e.* combine) these strata with neighbouring strata. In the case of many empty strata, or lack of sufficient population information, Bascula can carry out the linear weighting technique described above or apply iterative proportional fitting (also called multiplicative weighting, or raking ratio estimation). The resulting weights can either be added to the data file, or be stored in a separate file.

Bascula is able to read the Blaise data file directly, and also extracts the required information about the variables from the Blaise specification. The information about the population has to be provided by the user. The program is menu-driven, making it user-friendly. It will carry out a complete post-stratification if possible. If not, the user has to decide either to carry out linear or multiplicative weighting.

Presently, Bascula can only be used on a microcomputer. In the future, a back end will be developed that will run on our mainframe environment. Bascula was particularly developed for use in social and demographic surveys, where post-stratification is combined with relatively simple estimation procedures. For use in economic surveys, different software will be developed. This software will concentrate on stratified sampling designs in combination with more complex estimators (ratio and regression estimation).

8. ABACUS

Tabulation is one of the basic activities in the statistical production process, and it was one of the first to be automated. Many tabulation packages have already been developed in the world, but many are not very user-friendly. This is partly caused by the fact that proper generation of a complex table needs a lot of parameters to be specified: the variables to be used in the various dimensions (rows, columns, layers), whether to concatenate variables (display all values of a variable, followed by all values of another variable) or to nest variables (display for every value of one variable all possible values of another variable) within a dimension, the displayed cell quantity (counts, percentages, totals, averages), whether to display totals and subtotals, and many layout features. To be able to cope with all these parameters, traditional packages have control languages to specify tables, and these languages are often not very easy to learn and to use.

Abacus is a tabulation package, running on microcomputers under MS-DOS. While Abacus may be seen as yet another tabulation package, it was developed with very specific design goals. In the first place, no control language is used to specify a table. The program is menu-driven instead. The user designs his table in an interactive, simple, and intuitive way, without having to know about any control language. In the second place, Abacus can directly read the data file created by the Blaise System, as well as Ascii files. The meta-information, *i.e.* the information about the variables in the file, can be generated by the Blaise System, or can (in the case of separate Ascii files) be entered interactively by the user. Thirdly, the program can produce camera-ready tables.

Another striking property of Abacus is its speed. A table produced by SPSS-Tables in 3 minutes was generated by Abacus in about 6 seconds (all timings based on the same 386SX based microcomputer). The reason for this is that the Abacus program is rather small, so it can use a large part of the memory as a working area which allows for a table of up to 90,000 cells.

Tables produced by Abacus can have up to three dimensions (layers, rows and columns). Every dimension can hold up to 10 variables, which may be nested or concatenated. In the example in table 4, the column variables ''Employment'' and ''Sex'' are concatenated while in the row dimension the variables ''Region'' and ''Town'' are nested. In this example no variable has been placed in the layer dimension.

This table contains simple counts, but Abacus can also calculate totals of quantitative variables, make percentages tables, and averages tables. It is also possible to have more than one (up to 10) items in the cells of the table. In that case, the user has to decide to put each item in a separate row, column or layer. If the data has been collected by means of a sample survey, Abacus can accommodate weighted data, using the weights that are, for example, computed by Bascula. The only thing the user has to do is to specify the variable containing the weights.

Table 4

An example of a two-dimensional table

Number of Records	The population of Samplonia				
	Total	Employment		Sex	
		Job	No Job	Male	Female
Total	1,000	341	659	511	489
Agria	293	121	172	145	148
Wheaton	144	60	84	70	74
Greenham	94	38	56	44	50
Newbay	55	23	32	31	24
Induston	707	220	487	366	341
Oakdale	61	26	35	36	25
Crowdon	244	73	171	128	116
Smokeley	147	49	98	80	67
Mudwater	255	72	183	122	133

Source: Samplonian Statistical Office.

Much attention has been paid to the layout of the table, because the tables produced should be camera-ready. Therefore there are many options in Abacus to control the layout. It is possible to specify up to 10 lines of text for the header and for the footer of the table, and one can select both horizontal and vertical rules (as in the example), only horizontal rules or no rules at all. The layout of the text in column headers and the width of the columns can also be influenced.

A rounding procedure can be carried out to protect confidential data in the table. Cell totals, but also marginal totals are rounded to a multiple of some specified constant, *e.g.* 5. Abacus can provide both normal rounding and random rounding. If the user is not satisfied with the resulting table, he can import the output of Abacus into the spreadsheet program Lotus 123, and carry out further processing there. A final feature to be mentioned here is the possibility of creating new variables by recoding existing variables (*e.g.* from age to age classes). More details about Abacus can be found in Bethlehem *et al.* (1989a).

9. ANALYSIS

The CBS has not developed any software for statistical data analysis, the main reason being that there are already enough good statistical packages available. The CBS itself uses the packages SPSS (both on mainframe and micro) and Stata (on micro). To make these packages part of the integrated system for survey processing, tools have to be available to export the data from Blaise to them. This is realized in two steps. First, the data file is converted from the Blaise format to Ascii format, and second, the information about the variables, as available in the Blaise questionnaire, is translated in such a way that it can be understood by the particular package. Thus, a setup file is created. By running this setup from within the statistical package, a system file is created. And by loading the system file, the user can start straight away with his analysis, without having to bother about specifying the variables, labels, *etc.*

The procedure above only works for SPSS and Stata, and not for other packages. Of course, this approach could be implemented for every known statistical package, but that would require a large programming effort. Instead, a different road was taken. The Blaise System has a special setup generator utility. The user ''paints'' the structure of the setup file in a word processor, and by running the setup generator with this general setup description and the Blaise questionnaire as input, a real setup file is created. So, with the setup generator the user can generate setup files for his own favourite package.

10. CONCLUSION

The advent of the microcomputer has had a considerable impact on the work of the national statistical office. The subject matter statistician is making use of it more and more, and for his work, he needs an integrated survey processing system like the one based on the Blaise System. The power of this system lies on the one hand in the consistency it enforces in the various steps of data collection and data processing. This makes the whole process easier to manage and to control. On the other hand it also encourages standardization between different departments. Since all departments use the same software for their surveys, exchange of information between departments is easier and less error prone.

The integrated approach to survey processing was developed with in mind a highly centralized organization, like that of the Netherlands Central Bureau of Statistics. In such an organization, this approach can lead to a substantial increase in efficiency. However, not all statistical offices have a centralized structure. Particularly in larger countries, data processing is often decentralized. Regional offices take care of data processing in their own regions, and the resulting data files are sent to the central office. In the central office, the regional files are combined into one national file. The integrated approach can also be applied successfully in such an environment: the central office develops the Blaise questionnaire, and copies of the generated data entry program are sent to each regional office. This ensures consistency of data collection and data editing at the regional level. The regional data files will all have the same Blaise format, so combining them into one national file will be a simple job using the tools of the Blaise System. Since all regional files will be "clean", no further editing will be necessary at the national level. The only job for the national office will be to tabulate and analyse, and to publish the results. Furthermore, if either the regional offices or the central office can make regional publications.

The Blaise System has been tested and used since the middle of 1986 for a substantial number of surveys. The system is developed in close cooperation with the users. Every new version contains enhanced features. Abacus has been in use for over a year, and is very popular among its users. The Bascula program is still in development. The first prototype has just been released.

We did not mention the possibility of exporting data and meta-information from the Blaise System to the Paradox database package. In the future, a link will also be established between Blaise and the Oracle database system. In this way, a client/server architecture can be realized for Blaise users.

At the end of the statistical production chain, there are some aspects of publication that still have to be dealt with. In the first place, software will be developed to asses the risk of disclosure of confidential (private) information in statistical information to be published. Tools will also be offered to protect tables or data files against these risks. Finally, statistical offices engage more and more in electronic publication of statistical information, *i.e.* statistical information on floppy disks, CD-ROM, *etc.* To help the users of this type of information in selecting the subset of information they need, user-friendly software must be made available to them. This software is now being developed.

REFERENCES

BETHLEHEM, J.G., VAN BUITENEN, A.A.A., HUNDEPOOL, A.J., ROESINGH, M.J., and VAN DE WETERING, A. (1989a). Abacus 1.0, A Tabulation Package, Compact Guide. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN, L.F.M. (1989b). Blaise 2.0/An Introduction. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN,L.F.M. (1989c). Blaise 2.0/Language Reference Manual. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Offical Statistics* 3, 141-154.

KELLER, W.J., BETHLEHEM, J.G., and METZ, K.J. (1990). The impact of micromputers on survey processing at the Netherlands Central Bureau of Statistics. *Proceedings of 1990 Annual Research Conference, U.S. Bureau of the Census*, 637-645.