ELECTRONIC PUBLICATION OF STATISTICAL DATA

Alex van Buitenen, Anco Hundepool and Aad van de Wetering
Netherlands Central Bureau of Statistics, Voorburg, The Netherlands

1. Introduction

Statistical offices collect, process and produce large amounts of statistical information. The final step in this process is the dissemination of the collected statistical information. Traditionally this is done by publishing a vast number of books and periodicals. Everything is of course published on paper. As (micro-)computers are now a common tool in almost every office, there is a need to reconsider the way statistical offices should publish information.

As almost all statistical information users have a computer at their disposal, they want to use the statistical information on their computer. Of course they do not like the idea of having to key in the figures from the statistical publications into their machine. They want the information in a machine readable way. The statistical offices have to become aware of new publishing media for their information and as a consequence have to offer the appropriate services.

Besides information traditionally published in books and periodicals, there is a growing demand from researchers for the release of more detailed (if not individual) data. As the packages to perform statistical analysis are widely available also on micro-computers these researchers want to do their own analysis on data files with individual data. This leads to a growing demand for these files with individual data.

Of course, there is a great risk of disclosure in releasing files with individual information, but that is not the issue of this paper. See Bethlehem et al. (1990) for this issue. Once a statistical office is willing to release files with individual information to researchers, they are obliged to supply also the necessary meta-information to enable the researcher to understand the meaning of the different variables.

In the remaining sections of this paper we will discuss how we think that individual data files as well as more aggregated data should be made available to the public and how we should supply the necessary meta-data information to the user.

2. Aggregated data

Most of the data published by statistical offices are aggregated figures. These figures are the result of a long production process; beginning at the collection of the data, editing the data, if necessary weighting the data for the sampling method used and for nonresponse and finally, tabulation. These tables will then form the main part of most statistical publications.

As the need for electronic publications is growing, the Netherlands Central Bureau of Statistics has decided to meet these demands. The most simple and most easy way would have been to copy some data files or tables on a diskette and make them available to the users of statistics. This is however not a very friendly way of publishing data. At least the user needs to know what is on the diskettes. What is the meaning of all the variables, what are the codes used etc. This meta-data is almost as important as the data itself.

Therefore, it was decided to develop a general program (CBSview) which would meet these needs. With CBSview it is possible to publish the information together with the necessary meta information. CBSview is a general shell to publish statistical information and can be used for all kinds of publications. CBSview makes it possible for the users of the statistical information to make their own selections from the data, consult the information and to export the selected information from CBSview to almost any statistical package for further processing.

At all stages of the selection of the variables, the user gets the information he needs to make the selection, such as the description of that variable. After having made the selection, the user can specify the format he wants for the selected information. At this moment we support the following formats:

- Tables in ASCII format. This kind of output is always generated, and can be saved on disk for later use. E.g. these tables can be included in your own reports.

- A Lotus 1-2-3 worksheet. This direct link to Lotus and also Quattro is meant for those applications, where the user wants to make his own calculations or wants to use the large graphical capacities of these programs. The current version of CBSview does not have any graphical possibilities nor can it perform calculations. This is left to the spreadsheet package.

- A dBase file. This file can also be read by other database packages like Paradox.

- An ASCII data file. This general file can be used to read the selected data into various statistical packages like SPSS and SAS.

- Setups to read the ASCII file into other packages. CBSview is able to create setups for SPSS, Stata, SAS, Abacus and Manipula. The latter two packages originate from the Netherlands Central Bureau of Statistics and are meant for tabulation and file manipulation.

- A file containing the descriptions of the selected variables.

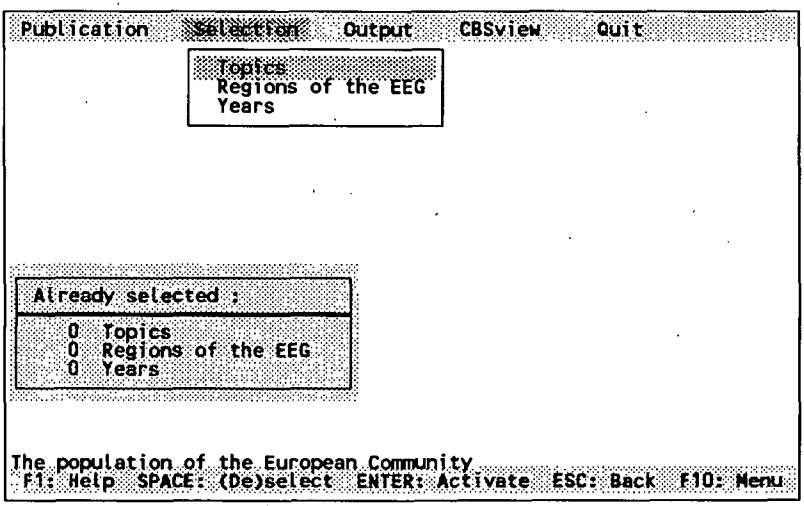
At this moment the structure of the information that can be stored in CBSview is a homogeneous data matrix (up to 10 dimensions). The user can select parts of this data matrix by selecting variables in each dimension.

Although this format is suitable for most publications, it has proved to be too restrictive for other publications. For these other publications the basic information consists of a (large) set of small tables. Many statistical publications are a compilation of a set of tables. In version 2 of CBSview it will be possible to put these kind of publications on diskette. This will make it possible to inspect the basic tables on the screen, export them to a file to include them in reports and convert the tables to a Lotus 1-2-3 worksheet.

This version of CBSview will also include a thesaurus to help the users to locate the information they are interested in.

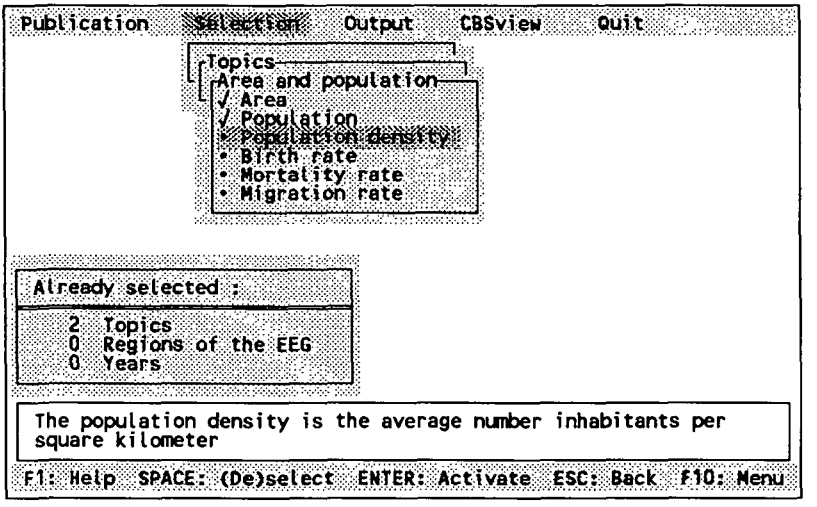
3. The use of CBSview

The use of CBSview can be divided into two stages. In the first stage, after the publication has been chosen, the user makes a selection of the data available in the publication. A menu driven program helps the user to select the information he is interested in. As the information in a CBSview publication is a homogeneous (more dimensional) data matrix the user must select items from each dimension. In the following screen the user chooses one of the (in this example) three dimensions, where he wants to make a selection. The first dimension is always called topics (variables), while the names of the other dimensions can vary depending of the publication. For example regions for a regional publication or years for time series.

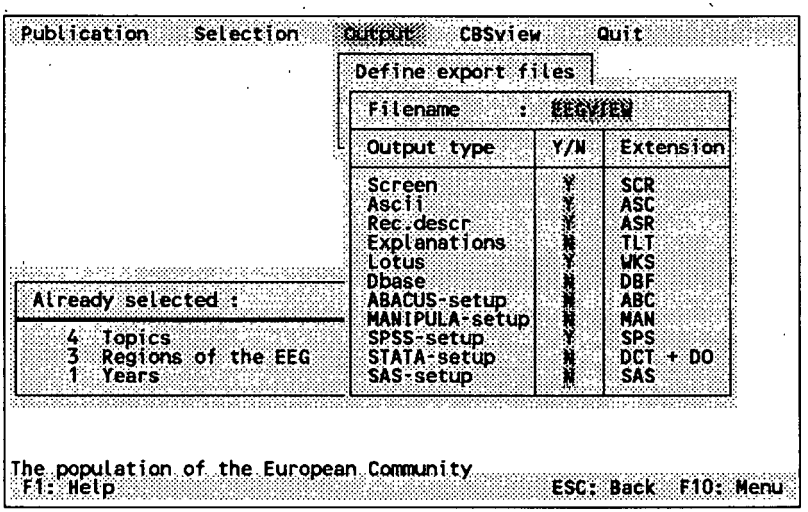


After selecting the item 'Topics' the user will access a hierarchical tree structure in which all the items are structured. This tree structure can be

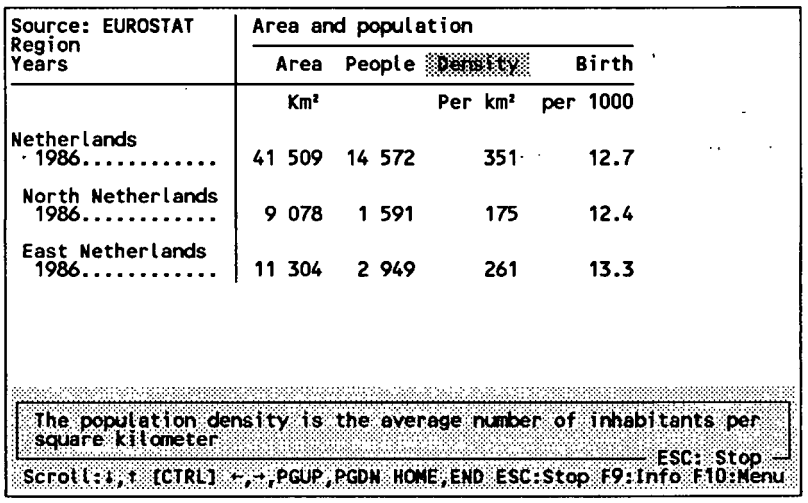
several levels deep. The menu program leads the user through the hierarchical tree structure. At any level the information about the current item is displayed in a box at the bottom of the screen. Also some information about the number of selections made is displayed on the screen.



In a similar way the user makes his choices from the other dimensions. If the selection of the information is completed the user moves to the next step, i.e. the actual retrieval of the selected information. In the following screen the user can choose the kind of output he wants to be generated. More than one choice is possible.



Finally the user instructs the program to do the retrieval and the selected information will be presented on the screen as the following table. If the table is too big to fit on a screen the user can scroll through the table.



4. Individual data

A different approach is followed for individual data. In this situation the user in general does not want to see the individual records of the data file, but either he wants to generate his own tables from the data or he wants to make his own analysis of the data. But here too the availability of good meta information is very important. When publishing individual data there is great risk of disclosure of individual information. At the CBS there is a research project running to study these problems. As a result of the study a prototype of a program ARGUS (De Jong et al., 1992) has been developed, which helps to identify the possible disclosure risks. The problems of disclosure however are not the subject of this paper.

When we publish a data file with individual records, we assume that the data file has been protected against disclosure and therefore can be made available to the public. Here again we do not want to make a simple ASCII file and leave it to the users to work it out. The users need to know at least what the meaning of the variables is, what the coding of the categories stands for, etc.

We have chosen to publish these files in an ASCII format. Not because we think that everyone wants to use this ASCII file, but it is a very general format that can be read by various programs. Besides this ASCII file with the data the user needs the meta information. Therefore, we supply a system which is capable to converting the meta information into a format useful to the user. This format can be a plain record description but also a setup for transforming the data into the format of the various analysis packages like SPSS, SAS and Stata.

5. Publication media

Up to now we have only discussed the publication of information on diskette. Other computer media are available, of which the CD-ROM is going to play an important role in publishing statistical information . The CD-ROM has a very large capacity (about 500 Megabytes) and is also very reliable. Once the data have been written on CD-ROM they are very

secure. Although CD-ROM is a slow medium compared with a hard disk, it will be (and is) used for publishing statistical information. With respect to the software (CBSview) using a CD-ROM to publish information is not a source of great problems. However in the Netherlands we still wait to start the use of CD-ROM because of the relative small availability of CD-ROM players in the Netherlands. Also the capacity of a CD-ROM is a point. We do not have a census any more and therefore we do not have those very large data files to justify the use of a CD-ROM.

An other possibility is the use of a publication computer (server) located at the statistical office. In this situation the data files to be published are on this computer which can be accessed by the use of a modem and a micro computer. The same software (CBSview) can be used to make the information available. In this situation the user runs on his own computer the specification part of CBSview (the 'front end') and only accesses the central computer (the 'back end') to read the selected information from the CBSview databases. This way of working implies that the specification part of the software is made available to the users to run it on their own computer. They download the meta information of a publication only once and use it (off line) to specify a selection.

6. Conclusion

It is very important for the statistical offices to be aware of the changing demand for information from the public. The computer will play (and is already playing) an important role in the dissemination of statistical information. The use of adequate software to disseminate the statistical information facilitates the use of this information. It should be the goal of a statistical office to achieve that the information it has gathered is used in as many places as possible. The use of the statistics produced is the main reason why a statistical offices exists.

The Netherlands Central Bureau of Statistics has developed software (CBSview) to disseminate statistical information on electronic media. The first reactions from the public are very encouraging. Although the use of

CBSview is restricted to CBS publications, the same software is available (as STATview) for other agencies to make their own publications.

References

Bethlehem, J.G., Keller W.J., Pannekoek J.(1990): "Disclosure Control of Microdata", *Journal of the American Statistical Association* Vol.85, No. 409, p 38-45.

Jong W.A.M. de, Willenborg L.C.R.J., 1992, "Argus: an integrated system for data protection", *Proceedings of the International Seminar on Statistical Confidentiality*, Eurostat/ISI.