USE OF MODERN COMPUTING TECHNOLOGY IN LARGE SCALE SURVEYS

Svein Nordbotten
P. O. Box 309 Paradis
5856 Bergen
Norway
svein@nordbotten.com
http://nordbotten.com

1. Purpose

The purpose of this presentation is to review and discuss the potentials of present and anticipated computing technology for the production of official statistics and large-scale surveys, i.e. the work in the national statistical organizations. The computer technology will probably be the single factor having most impact on the development of statistical production in the first part of the next century. The computing technology is also expected to contribute more than anything else to fill the gaps between statistical systems in industrial and developing countries.

2. Technology

We will distinguish between soft and hard computer technologies. In the last 50 years, the development of hard technology, i.e. the physical equipment has been far ahead of the development of soft technology, i.e. the systems and programs, needed to make full use of the hardware.

Hard technology

The development of the PC from the early 1980's to the current powerful **workstations/ servers** is the first technical precondition for the present use of computer technology in statistical processing. A modern workstation of 1999 has both processing speed and storage capacity far above that of the big mainframes first introduced in production of official statistics [Daly and Eckler 1960].

During the last 10-15 years powerful **client/server** technology has replaced traditional main frame computers. In addition to general efficiency, the advantages included more flexible management and a less fault sensitive environment. The combination of powerful, often specialized, servers with clusters of workstations permit optimal distributed processing.

The second most important event on the hard technology side has been computer access to the **communication networks** including fiber optical cables, radio and satellite links, which permitted transfer of information. From a communication point of view, existing technology makes it possible to establish connection between any two computers worldwide to exchange statistical information.

We can expect that all the above technologies will be further refined and developed in the future permitting applications, which so far have been considered out of reach. Already well developed, **parallel-computing technology** may be introduced for more general use. The new hard technology, which probably will have the greatest impact on statistical production in the next, is the fusion of mobile wireless telephone and computer communication technologies. In few years, the use of computers and access to servers anywhere at the globe will be technically feasible for the producers of official statistics. In parallel with technical development, unit prices for as well speed as capacity have been continuously falling making the technologies feasible in applications for which they previously were prohibitive.

Soft technology

The soft part of computer technology, **systems and programs,** permitting users to take advantage of the hard technology, has in most fields lagged years behind hardware development. In the infancy of computers, it was a usual opinion in statistical organizations that statistical processing was so particular that the best road to computational success was to design and develop dedicated programs and application systems for statistical production.

This opinion has changed gradually. Today, many statistical organizations rather **buy** standard software tested and developed for general use, and the resources are now focused on how to **apply** the soft technology as effectively as possible on statistical tasks.

The basic soft technology, **operating systems**, has developed in small steps during the last decades, and will mainly be adjusted as new hard technologies are introduced. The statistical producers use commercial soft technologies in 3 main areas, **statistical analysis**, **database management**, **communication**. In specific fields such as **survey design**, **editing** and **imputation**, statistical **estimation** and **confidential data protection**, the statistical producers continue to develop their own products, mainly because commercial software developers have not yet found the market attractive.

In the future, the communication systems will most likely be the greatest interest for the statistical producer. The number of computer clients is increasing fast and it has been estimated that in about 5-7 years, as many as one fifth of the world population may have access to the global network. The communication is expected to become **multimedia** in the sense that the message sender will be able to submit her message in alternative forms. On the receiver side, the user will have the option to select the media he prefers. The travelling message may be in some generalized form.

3. Applications

Computer technology will continue to revolutionize the production of statistics. So far, software for **optimal survey design** is only in its infancy. We should expect to get **CAD** (computer-assisted design) software, which will allow the statistician to compute the optimal design for his survey in much the same way as the engineer computes the specifications for a ship or an airplane.

One of the fields in which we have seen great development during the last 10-12 years is the application of computer technology in collecting observations. The introduction of **computer-assisted interviewing** etc. has made the collection of survey data faster and more accurate in permitting the reliability of the data to be controlled in presence of the respondents. **OCR** (optical reading) has also contributed significantly to efficient data transfer. However, as the uses of computers expand, more facts will be recorded at their sources. **EDI** (electronic data interchange) is therefore expected to take care of an increasing proportion of the statistical data input in the future.

The ideas of statistical data centers, statistical registers and archives, etc. have been discussed for several decades as means to improve the basis for official statistics [Nordbotten 1967]. These ideas are now being realized as **register statistics** and statistical **warehouses** for microdata and in **statistical databases** for macrodata all described by metadata in **meta databases.**

Data collections may originate from many different sources and need transformations as well as control and adjustments to be useful for the different applications. Coding from one representation to another has developed from primitive transformation to computer based **smart coding** taking advantage of contextual knowledge from background data.

Statistical producers have always been observant of the relationship between product value for the users and **product quality.** Statistical **data editing** was introduced to detect suspicious records. New technology has both permitted more efficient editing by computerization of old methods as well as investigating new editing approaches previously infeasible, for example the use of neural networks, which can effectively make use of knowledge already known about the statistical units [Granquist 1998].

Micro data imputation is frequently considered part of editing and used to fill out the blank fields of partial non-response and, in some cases also change the values of suspicious fields. Techniques as **hot deck,** etc. have been used for more than 40 years by means of computers and are still used in improved versions for imputing non-response.

An application of increasing importance is statistics based on survey data combined with background data from previous censuses and administrative data sources. Previously, this was a typical application of ratio or regression estimation. With new technology, a **virtual** census may be administered. The strategy for such a census is to compute the relations between the required variables from the sample and the background data for the corresponding units. Then use these relationships for imputing values for each of the units *not* in the sample and finally develop the desired statistics by simple aggregation of the observed and imputed micro data. The tools typically used for computing and using the relationships are **neural networks** and **regression analysis** [Nordbotten 1998]. The application just described can also serve as an example of **data mining**, a new and still quite controversial approach to knowledge acquisition.

The ultimate aim of official statistics is to be useful and informative. Modern computer technology also provides the statisticians with tools to customize statistics to the needs of the individual user or user group. From a modest beginning 30-40 years ago, when some producers offered statistics printed in customized forms to the last decade in which dissemination of statistics has propagated to **multimedia forms**, **CD** and **web statistics** offered optionally as **numeric**, **textual** and **graphical** presentations. The flexibility by which statistics can be customized increases the risk for disclosure of confidential data. Significant progress has been made in developing methods for **protecting the confidentiality** of statistical data needing computer technology to be implemented [Fienberg and Willenborg 1998].

Computer technology has already had important impact on the **accessibility** of statistics. The communication technologies of the mobile phone and the laptop computer are expected to merge, permitting the user to bring the computer technology and access to statistics with her wherever she moves. Access will not be limited to her national statistical service, but to an increasing number of statistical services.

Statistical systems based on new computer technology create new challenges to management. New long-term **management strategies** are currently discussed [Cook 1999]. The users of the future and their needs will be oriented internationally. They will demand statistics that require combinations of data from different national producers. To serve the users, a **worldwide distributed and networked statistical system** must be developed which, from the users point of view, can be imagined as a single virtual provider of statistics. The hard computer and communication technologies for such a system already exist, but more soft technologies are still needed for extracting statistics from national producers and harmonizing the statistics as demanded by the users.

4. Implications for developing countries

There have been 3 factors, which have limited developing countries to take full advantage of computer technology: need for computer maintenance, computer staff, and available technology. Representatives of the manufacturer usually took care of the computer maintenance. Staff was trained abroad and advisors from cooperating countries and organizations gave assistance, while technology was made available according to local funds and grants from development funds. The solutions frequently became copies of solutions designed for work under different conditions [Sadowsky 1988].

The current trends give improved outlook for developing countries. First, the prices of computer hardware performance and capacity have decreased and the hardware becomes more affordable for each year. Second, the hard technologies have reached a level of reliability, which has reduced the dependence of expert maintenance significantly. Third, flexible soft technologies, which can be adjusted to local requirements, are being developed. Fourth, the software is becoming much more user friendly and requires much less training. Fifth, the industrialized countries cannot afford to be without the statistics of developing countries and a global networked statistical system will become a mutual responsibility.

5. Conclusions

Computing technology has during the last 40-60 years had a major influence on statistical production and systems. Development dominated by computer technology can be expected to continue in the next century. Two factors will be particularly important. Technology prices will continue to decrease and make the technology available for more producers while the development of communication technology will increase cooperation among statistical producers and make the statistical products globally more available for users. The demands for international statistics will be a strong incentive for developing a global networked statistical system.

REFERENCES

Cook, L. (1999): Managing in a networked statistical system. Invited paper to the 52nd Session of the International Statistical Institute. Helsinki.

Daly, J. and Eckler, A.R.. (1960): Application of Electronic Equipment to Statistical Data-Processing in the US Bureau of the Census. Proceedings from the 33rd Session of the International Statistical Institute. Paris,

Fienberg, S.E., and Willenborg, C.R.J. (Guest Editors 1998): Special Issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. Journal of Official Statistics. Vol. 14, No.5.

Granquist, L. (1997): The New View on Editing. International Statistical Review. Vol. 65, No.3. pp. 381-387.

Nordbotten, S. (1967): Purposes, Problems and Ideas Related to Statistical File Systems. Proceedings from the 36th Session of the International Statistical Institute. Sydney.

Nordbotten, S. (1998): New Methods for Editing and Imputation. Proceedings from Agriculture Statistics 2000 in Washington D.C. International Statistical Institute, The Haag. pp. 220-208.

Sadowsky, G. (1988): Statistical Processing in Developing Countries: Problems and Prospects. Interregional Workshop on Statistical Data Processing and Data Bases. UNDP, UNSO and UNECE. Geneva.

(FRENCH RESUME

The present state of and trends in computer technology development for the future of large-scale statistical production is discussed, and a worldwide statistical network based on computer technologies is anticipated.)