Interactive Access to Survey Databases

by Mark Katz and Beverley Rowe
QUANTIME Limited, London, UK

Scenario

As a data-archivist, you have within your computer library tens, perhaps hundreds, of surveys. Many of these are heavily used and you can afford to have them on-line on your small mini-computer. Lack of funds has prevented you from installing CD-ROM or laser disks and you have only a few programmers.

The phone rings: someone looking into the effects of radio-active fallout wants a quick statistic from one of your on-line surveys - one you are not too familiar with. It requires a scan over three years of some 60,000 incompatible questionnaires to select a sub-group of all people who studied physics at university and may have contracted cancer.

Five years ago you would have sent them a tape of data, possibly also some SPSS control commands, and told them to do it themselves, a task that would have taken them days or even weeks to complete. Last year you could have sent them a floppy disk but still a long and daunting task would lie ahead of them.

Now you log into your computer, access the survey and, even though you do not have the questionnaires

handy, find the name of all variables which look at 'physics' and 'cancer'. Within one minute, you dictate the relevant statistics over the phone.

The caller is most impressed with the speed and is quite interested in the results – "How can I find out more?" he asks. "Either dial in to the database or, if you have a PC, I can provide you with the total database on just a few floppies. It's the same interactive system, designed for very fast access by researchers with no previous training" you reply.

An ideal view of the future? No, Quantime offers this type of service today.

The market research industry conducts thousands of surveys each year. Traditionally, the results of surveys are produced as hard copy computer tabulations, but this is changing, and more surveys are ending up as on-line databases.

Quantime is in the forefront of such developments with a user-friendly system called QUANVERT which offers fast access to such databases, some of which exceed two million cases/respondents.

This paper reviews our experiences with offering remote access to large survey databases and working with the data archive at the University of Essex to promote the use of on-line access to the General Household and other surveys.

QUANVERT

Background

Quanvert is the interactive member of a family of software tools for processing survey data, offering non-technical users direct access to data with an interface especially designed for them. It permits very fast exploratory access to data by taking advantage of inverted or transposed file structures.

Quanvert is five years old, written in C and currently runs on DEC-Vac (Unix and VMS), Prime, large IBM mainframes (MVS, VM/CMS), many Unix-based micros and, more recently, the IBM PC/AT.

Quanvert handles numeric, categoric, multi-coded and textual variables with all normal boolean and arithmetic functions. Users may cross-tabulate or interrogate data as well as create new variables, all interactively. Increasingly, users are linking to Quanvert through PC's to download data and use spreadsheets and graphics. An associated package, QUANTUM, sets up the data description.

Some facilities

Quanvert reads transposed files to produce data at either the *aggregate* level (as cross-tabulations) or *disaggreate* level (specific values/responses for selected records).

It interacts with the user to determine the variables to be selected and the types of reports to be produced. The variables (or axes) corresponding to, or derived from, the original fields/questions may be manipulated, tabulated, displayed or used in statistical analysis.

Sub-sets of data

Subsets of the data are extracted by *filter* commands that use logical or arithmetic combinations of existing variables.

The specification of these *filters* processes code text rather than data values. For instance, selecting only women uses the variable *sex* and subset *female* rather than looking at bytes 4-6,

value 2, etc. For frequently used selections a named filter may be created, but Quanvert does not set aside physical subsets.

Types of data

* Categoric (sex, region, etc.)
* Multi-coded (makes of car owned, etc.)
* Numeric (salary, date, etc.)
* Alphabetic (names/addresses, verbatim responses and text)
* Hierarchical (master/trailer)

This last facility allows for analysis at different levels of data for accumulation across levels.

Types of report

* Simple cross-tabulations (up to six dimensions)
* Filtered tabulations using logical combinations of variables
* Means or proportions and table division
* Grossed-up tables (multiple weights if needed)
* Listings of raw data

Operations to look after the database

* Create new variables
* Delete/rename variables
* Create special filters
* Combine similar data for months/areas, etc.
* Join data from different surveys
* Manipulate variables across levels in a hierarchy
* Print a code book, including KWIC index of the database text

Help commands

* Lists of commands and variables
* Detailed explanations
* Marginals (summary statistics) for each variable
* Text search for keywords in the code book

Other features

* Statistical analyses
* Combinatorial analysis
* Sorted/accumulated lists
* Production of sticky address labels
* Data downloading for a micro
* Files for Symphony/Lotus, etc.

Perhaps the most powerful facility is that separate surveys can be stored as individual data sets and then 'joined' together. Thus, data for different years may be aggregated to compare results over time. Quanvert automatically introduces a new variable (*years* or whatever the appropriate unit) which may be used as a

breakdown. Quanvert looks after minor changes between questionnaires from different years. Providing the code text and options remain constant, Quanvert transparently combines data even though the position on the questionnaire has altered.

Data reading time depends on the degree of filtering. Unfiltered requests are processed at speeds of 500–1000 cases per second, irrespective of the size of database or number of variables. On the Vax 750, it is not uncommon to reach speeds of up to 30,000 cases per second on heavily filtered tables. Even on the Compaq (IBM/AT compatible PC) Quanvert processes up to 15,000 cases per second. Speeds of up to 200,000 respondents per second have been recorded on an IBM mainframe.

Defining the Data

Introduction to QUANTUM

Quanvert is closely linked to the package Quantum. This batch-oriented program edits, recodes, and tabulates survey data. The user sets up a specification file which describes the data together with the recoding and analyses required. This file is compiled by Quantum and run on the original raw data file, a multi-stage process involving data reading, data accumulation and report printing.

A typical Quantum specification file has two sections:

The *EDIT* section uses a special language that combines many features of Fortran with special facilities for handling survey and structured data. It includes powerful data checking commands and an online data correction facility.

The *TABULATION* section contains non-procedural statements that:

a. Define the axes. For instance, the following statement specifies a variable *Sex* which may be found on position 6 of the data file, where 1 denotes *Male* and 2 denotes *Female*

l sex
col6;hd=Sex of Respondent;Base=Total Sample;Male;Female

b. Define the tabulations, specified as a series of *TAB* statements. These use the predefined axes as rows, columns and filters. This section offers a large selection of options for format control for mathematical computations and percentage calculation, as well as detailed layout of headings, row/column text, figures and labelling.

Setting up the data description for Quanvert

Quantum is used for this. The user specifies the variables plus associated headings, text and location using the tabulation section. Any recording or derivation of new variables would be included in the edit section.

Quantime has developed a semi-automated SPSS-Quantum conversion package.

Preparing the transposed file

The *flip* program is now invoked to read the Quantum specification, extract and recode data from the original data file and prepare the transposed file. Since the transposed file contains the original data *and* the data description, it is not necessary to retain the original data files. The time taken to invert the General Household Survey (12,000 households,

23,000 people and some 120 variables) was only two hours on the Compaq. The Appendix contains details of this transposed file.

Extending the Data

Even though Quanvert works with transposed files, it is possible to add new cases or variables. It is rarely necessary to go back to Quantum to create variables.

Shorthand methods are provided to copy a variable with the addition of an existing filter or set of filters. The new variable becomes part of the database.

More generally, the user sets up in a separate directory a mini database containing the new records and this is added to the database. It is not necessary to reprocess the entire database. Secondary databases are simply appended to the main database, i.e. each file in turn is appended to the relevant variable–file.

However, in many situations it is better to keep additional sets of records separate. For instance, data may arrive in monthly batches or from different areas. In this case all the secondary databases are *joined* together in a *MULTI-FLIP* structure, such that there is one master directory and multiple subdirectories. Quanvert creates a new variable which contains as elements each of the subdirectories, e.g., month or country. This allows the user to tabulate any variable by, for example, month or select any number of sub–directories for an analysis.

Multi–flip looks after changes to the variables between batches of data. If the number of elements and the code text are unchanged (even if they come from different parts of the questionnaire), those variables are assumed to be accessible to all sub–directories.

If a new *variable* needs to be added to the database (or an existing one replaced), it is not necessary to set the database up again. The user may create a new variable directly within Quanvert, using logical combinations of existing variables. Alternately, if new data have been provided or additional external variables are required, these may be prepared separately and *merged* into the main database. This will add or replace those with identical names.

Post–processors

Quanvert has facilities to select variables from specified respondents and to write this out to a file. This file may then be downloaded to another system for statistical or graphic operations. An option provided will convert the values of variables into numeric fields, rather than the text of the value (e.g. value 1 for *male* and 2 for *female*) and thus simplify the interface to statistical systems. This option also provides the SPSS variable and value labels. A useful facility on Unix–based systems is to pipe this output to user defined post–processors directly rather than to an external file.

Post–processors are provided to reformat cross–tabulations into a form acceptable to other packages. This uses the SYLK file–format or shortened character format for input to Symphony/Lotus with the *FILE-IMPORT* option.

To summarise, then, Quanvert offers very fast analysis of survey data. It combines simplicity of use with a wide range of facilities. The interface to other packages, its linkage facilities to download to micro–computers and its availability on a broad range of computers, from large IBM, most major mini–computers, to the IBM–PC, makes it a leading package for the analysis of large survey databases.

What is the GHS?

The General Household Survey (GHS) is one of many surveys conducted each year by OPCS (Office of Population Census and Surveys) in London, a government department. It is considered a cornerstone of social research in the UK.

The GHS is carried out each year with some 12,000 households/23,000 people as a hierarchical data set. It is normally available within 6 – 9 months of the end of fieldwork.

The survey covers a broad range of topics: health, education, car ownership, use of energy, employment, income, family size, age, and so on. There are some 800 basic variables.

OPCS carry out the data collection, cleaning and preliminary analysis. They have switched to Sir for data management but still use a fairly old system on their ICL computer for the main reporting.

A Monitor appears each year as the first indicator of social change but the OPCS is unable to provide much or fast response for further reporting. They use the Data Archive at the University of Essex as a distribution point and invite bonda fide researchers to conduct any further work themselves on their own computers, using the raw data.

The ESRC Data Archive

The Data Archive at the University of at Essex is funded by the ESRC (Economic and Social Research Council) and is one of the largest collections of machine–readable survey data files in Europe. They have thousands of surveys, many with an associated SPSS control files. They publish a quarterly newsletter and hold an important position in the social survey world. Many surveys from the private sector are held by the Archive, and it is a condition of all ESRC grants that resulting data are deposited there.

The GHS Experiment

By mid–1984, Quantime had considerable experience with remote databases and a well–established UK–based service for the private sector. Quantime felt that this concept needed to be introduced to the public and academic sectors and initiated discussions with the Archive to take an important and well used dataset for implementation as part of the Quanline service. After lengthy discussion and approval from OPCS, it was decided to use three years of GHS data, and work began in mid–1985.

Quantime set up this important public database and made it available at no charge to bona fide researchers in the academic sector through the Quanline time–sharing service. Agreement was reached whereby Quanline joined with Essex in low–level marketing to the academic and public sectors. It was hoped that the experience would help the Data Archive in any plans to make datasets available interactively, rather than by mailing tapes or floppy disks.

After unsuccessful attempts to obtain external funding, Quantime allocated over 50,000 dollars to the project from internal funds. This included recruiting a consultant to develop the database and market the concept as well as an allowance of computing resources to store the data, set up the database and provide free access time to users.

Raw data and a list of variables were supplied in July 1985. At the time, we were unable to obtain an SPSS file for this dataset and it had to be set–up 'manually' in Quantime, a daunting task. The variables were prepared and the Quanvert database was available in September 1985.

A subset of the data for 1980, 1981 and 1982 was put up, including most of the household level data and important parts of the individual level data; in all, some 120 variables of a total of 600. However, the choice of variables turned out to be a poor one and insufficient key topics were available. The database is available on the Vax under Unix as a time-sharing service through Quanline. The user has access to all years and may produce comparative reports between years. The 1980 data are also available on the IBM PC/AT (and Compaq). On all machines, the average time to scan one year's data is under 10 seconds for either household-level or individual-level reports.

GHS Data On-Line

Progress in marketing has been steady. Contact, sometimes to considerable depth, has been made with over fifty academic, public sector and quasi-public consultancy organisations. Marketing has focused on mailshots, telephone calls, direct mail and press releases.

This has lead to very positive interest in the public sector and given an interesting insight into the (largely unsatisfied) demand for GHS data, into the research and thinking of users attempting to obtain statistical information from large survey databases, and the constraints under which they operate.

University and polytechnic users are being offered Quanvert for GHS at no charge. There are now eleven committed users at six sites.

The main problems with promoting the current version of GHS have been:

- the non-coverage of important areas of interest
- rather old data.

Whether interested in research or reference, the user must be able to find everything that was collected. The lack of income variables in the earlier releases of the service was particularly disabling, but other areas have proven important to particular users.

We can expect broadly two uses of the GHS or other large survey databases: active research and casual reference. We would expect academics to fall in the first category, non-academic researchrs (public or private sector) in the second. Because of the computing and staff resources required to obtain information quickly from archived survey data, most people terminate their research prematurely, or turn to other sources (often at great cost) to find information that is duplicated in GHS.

Discussions are now taking place to open up GHS data to the private sector and to charge for this service.

Other datasets were selected to supplement the GHS, namely the WFS (World Fertility Survey) Fiji survey of 4,900 respondents and 300 variables, and NCDS (National Child Development Survey) of some 18,000 children and 350 variables. We hope to have the three-year British Social Attitudes Survey on-line before the end of 1986. These surveys run alongside other private datasets resident on the Quanline computers and include:

British Telecomm's Telecare project (data from 3 million respondents);
Manpower Services Commission (400,000 Youth Training Scheme trainees);
British Gas's NDES project (55,000 establishements).

All of these are accessed by regional marketing and research staff at hundreds of offices around the UK.

Although we have dedicated this section of the paper to our work with the Data Archive, it represents less than 5% of the work of Quanline

UK, measured in terms of computing resources and usage by researchers.

A Summary of our experiences

Some Observations

We are able to assess the impact of interactive survey analysis based on our experience of some five years of Quanvert, many hundreds of users and some 2–3,000 connect hours each month on databases ranging in size from a few hundred to a few million cases.

a. Users cannot grasp the concept

Since most researchers are unable to obtain really fast or simple access to large databases, ad-hoc or reference interrogation has been largely overlooked. It cannot be understood without a demonstration or trial evaluation and an element of retraining. It is so alien to most people that there is a barrier to its introduction. They say: *What extra benefit is there to me if the results come back in two minutes instead of two hours?*

b. Software development is misdirected

A lot of human resources are spent on developing easy analysis systems, but not enough on good data organisation or easy data access. User-friendliness only comes with many users and long sessions by people other than the primary user or programmer. The use of 'laser disks' demands a new type of storage mechanism. If they are to be used speedily and effectively, we cannot simply replace the old floppy or Winchester disk but use the same old software.

Many people are using tailor-made programs and re-inventing wheels in software development. There are too many government departments using (and even re-writing) Cobol programs for survey analysis.

Archived data require a *read-only* strategy. Conventional DBMS programs place too much emphasis on updating rather than (fast) reporting. They are also greedy for computing and storage requirements.

c. The need for Statistics is exaggerated

The importance of statistical reporting is over-emphasised; in fact it represents less than 20% of access. The 80% can be achieved by (complex) cross tabulations. Despite this, the availability of good statistics seems to be a more important criterion than speed, flexibility or user-friendliness. In practice a 'database server' is required as a front end to the statistical software.

What Changes are Necessary?

There are now some 2,500 bibliograhic and textual databases available. Users spend millions of dollars each year; an entire industry has been built up, with conferences, training, newsletters, books and software investment. But this information is mostly textual and is difficult to manipulate arithmetically. It is also at a very high level of aggregation.

As the provision of fast, interactive tabulations from large databases gains momentum, a number of key issues are evolving:

a. Data consistency is important. Data under intensive scrutiny will lose credibility if badly formed.

b. Help information is needed, down to the

variable level.

c. A support team must be available to deal with queries on computing, telecommunication, and data problems.

d. A database of databases is required to indicate the best source for information.

e. There must be a common format for a data description language so that users can provide translators to/from a common language. Unfortunately this probably has to be SPSS, but a more comprehensive dictionary approach would be better.

f. Users must stop developing their own special analytical tools and rely on those already in use. The public sector must be prepared to go to the private sector and to scour the world for the best system.

g. The main software developers must turn to inverted files as a basis for fast access.

h. Specialised software on micros (spreadsheet, graphics, modelling) is being developed far more quickly on mainframes. Our emphasis should be on interface techniques.

i. There is money to be made by selling data interactively to the private sector. This money will help to cover the cost of data storage and computing and contribute to future development costs. This is particularly important when government is reducing the support given to academic and research institutions.

j. Software must be portable across computers. Unix has established itself as market leader and the language C may be even more important. Software should not be constrained to operate within the current limitation of memory/disk of today's micros.

k. In view of the data compression techniques now available, it is possible to store and analyse very large databases on micros and distribute the data on floppy disks. It should not be assumed that these large surveys can only be handled on large mini or mainframe computers.

l. More research needs to be put into Expert systems that ask users what they want. The software then does the searching and decision making jointly with the user.

Ignoring change will not make it go away. Quanline has shown that the trend is to interactive tabulations, a reduction in printed reports resulting in greater freedom and wider distribution of survey data. Archivists are in a unique position to beat the rest of the world.

We believe that in the long term, the concept of transposed files will become part of conventional DBMS technology, providing the benefits of data compression and fast access for ad hoc interrogations while preserving fast retrieval and update.

The concept of interactive access to survey data will become a small but vital part of the technique of converting survey data into useful information.

Background Information

QUANTIME is a major software and systems house serving the market research industry, with over 70 people worldwide, 50 user sites and some hundreds of clients using Quantum/Quanvert to analyse survey data. Quantime's headquarters are in central London, with offices in New York and Cincinatti and major agencies in Europe. Most of Quantime's work is for the private sector, but increasingly

the public sector is taking advantage of these services.

Almost all of Quantime's development and services are based on DEC/VAX's running under Unix – in fact there are six VAX/750's spread around the world linked with a sophisticated network of telecommunications hardware and software. Quantime is both a developer and user of software, offering a tabulation bureau service, time–sharing and the sale of software and hardware. Software includes highly specialised tools for Computer Assisted Telephone Interviewing, Automatic Questionnaire printing and direct data entry – all closely integrated with data editing and analysis packages.

In 1984, Quantime opened a new division QUANLINE dedicated to the needs of users wishing to load and access remote survey databases. This is based on two of the computers and currently hosts some forty databases, requiring 1,600 Mbytes of disk storage.

Quanvert is available in two ways:

- As a software package in its own right for use on IBM mainframes and PC's, DEC/VAX, PRIME and many other Unix–based minis. Normally, one would take QUANTUM (and FLIP) in order to be able to set up the Quanvert databases. However, where a user wishes to distribute databases, Quantime also supplies a 'read only' version of Quanvert.

- Through the Quanline time–sharing service. The UK operation operates from London and a new US service will be launched in the summer from Quantime's New York office.

Appendix

The Transposed File Concept

- **What is it?**

  A conventional data file may be considered as a matrix, with records as the rows and variables as columns/fields. Any analysis of this file involves scanning sequentially through the matrix but this is wasteful since it is unlikely that any tabulation needs to read **ALL** records and **ALL** variables.

  There are a number of techniques to minimise the time to isolate and select pre–specified records – these include index sequential or random access, heaps, lists and overflows, but they all demand a choice by the user of key variables – a choice that may be difficult to make.

  The concept of a transposed file is the conversion of the data file into a set of smaller files, one for each variable. These files are (unlike a pure relational database) simply sequential files holding the response from each case as a single record. It is not linked in any way to other files – the relationship between them is purely positional, i.e. the 412th record occupies the same logical position in each file.

  To prepare the transposed file, a special program is run, which reads through the data file sequentially and write out a series of subfiles. This is a once only process which requires both the data file and a copy of the variable description file. When this transposed file has been created, there is no further use for the original data and description files.

- **The benefits**

Any program wishing to read the data, need only pull off the relevant variable files, normally a very small subset of the full data file.

The benefit of this approach is that any analysis of the data file can be very fast. It is a function of the number of records, NOT the size of datafiles. Furthermore, all variables have equal importance – there is no need for the user to nominate key variables when setting up the database. But there is another major benefit – data compression. Since each subfile contains values of only one variable, there is significant scope for data compression.

- **Data compression**

There are four main opportunities for such compression:

- Where there are frequently repeated values, e.g. if the data is grouped in geographic order and the first 1,000 records relate to people in Scotland, the next 3,000 in Wales, etc.
- Where data are 'missing', e.g. the variable *salary* only has values for employed people.
- Where only a few records have a specific value, e.g. if on the file only 10% of all people are women.
- Where the data are hierarchical, there is no need to repeat variables at a higher level for variables at a lower level.

There are more advanced methods for data compression using HUFFMAN coding and pattern searching, which is fairly easy to achieve on such files. The use of MONTE CARLO simulations has revealed some useful statistics about the repeatability of bit strings on these types of transposed files.

The result of this is that data storage requirements can be reduced dramatically and data reading time reduced accordingly. Compression ratios of over 50% are often achieved and it is not unusual to see figures in excess of 95% for specific variables. The size of the file, in many cases is less than the size of the raw questionnaire–based data and further research is being carried out to improve these ratios. ¤