USAR – PUTTING THE 1991 SAMPLE OF ANONYMISED RECORDS ON YOUR WORKSTATION

Ian Turton and Stan Openshaw

WORKING PAPER 94/14

SCHOOL OF GEOGRAPHY • UNIVERSITY OF LEEDS

Abstract

The paper describes the the development of a customised computer software package for easing the analysis of the 1991 Sample of Anonymised Records. The resulting USAR package is designed to be portable within the unix environment. It offers a number of features such as interactive table design, intelligent data interpretation and fuzzy query. Finally, an example of SAR analysis is provided.

1 Background

A major UK census milestone was reached in 1993 with the release of Britain's first ever, official, sample of anonymised census microdata. In fact this amounts to a 3% sample of the 1991 census being released in anonymous form, without any names and addresses and only coded to large geographical areas. This followed almost a decade of discussion and debate about the content, need and confidentiality aspects of possible microdata from the census.

The key needs case is contained in Marsh *et al* (1991) and a broad description of the SAR itself is provided in Marsh and Teague (1992), Marsh (1993) and Middleton (1994).

The two SARs have been provide: a 2 % sample of individuals and a 1 % hierarchical sample of households and individuals within those households. These datasets are extremely important because they offer census analysts an opportunity to design their own tabulations free of many of the restrictions that normally apply to census data. As Marsh (1993) puts it "Users can explore relationships on the sample data, interacting until they reach the table that they feel gives the best information. They may still decide to commission a special tabulation, perhaps to get a level of geographical specificity not available in the SAR or because small numbers demand a 100% run, but they will be much more confident than in the past that they will be getting what they want" (p296). However, the SAR is much more than a source of enhanced tabulations, it also allows users to explore and analyze census data in ways they have previously restricted to other kinds of sample survey; for example, analysis of variance and regression at the individual level free of any of the danger ecological fallacy and spatial aggregation effects (Openshaw 1984). The SAR sample is sufficiently large to investigate subpopulations in society and permits the hierarchical aspects of household structure to be studied. Equally importantly, standard variables can be redefined to match user needs; for example, measures of deprivation, head of household, household composition, definitions are no longer fixed but can now reflect what users need. Again we agree with Marsh (1993) when she summarised the advantages of the SAR as follows "In the final analysis, however, the value of the SAR lies in the fact that, in contrast to tabulations produced by the Census Offices, we do not have to plan in advance all the useful ways in which the data can be presented. ... Almost every census user will have experienced the frustration of finding that the table they required was slightly different from

the one they in the end had to use." (p297). This is a slight exaggeration. Nevertheless it is the historically relatively small number of essentially non–spatial census users who will probably benefit most from the SARs, indeed this group may well be instrumental in opening up the census data resource to a broader social science constituency (Openshaw 1994).

The problem is essentially how to develop a software environment that will allow typical social scientist census data users to make the most of the SAR. This is a major challenge because it is not sufficient to merely provide users with the means of generating tables of their choice. This only works well if the users know precisely what they want. The size, complexity and novelty of the SAR makes this assumption highly questionable. Left to their own devices, there is a danger that instead of practicing their crosstabulations on a small sample survey of 100 or so households, the availability of the SAR will merely result in many social scientists practising their crosstabulations on the SAR. Apart from these concerns, there are a number of more basic issues concerned with how best to diffuse the SAR into the research community. In particular, how

1. to make the SAR easily available and usable; and
2. to trivialise the tabulation task in a highly interactive computing environment; and
3. to provide the table design tools needed if most users are going to get maximum benefits from the flexibility offered by the SARs.

This paper describes one way of attempting to meet these three objectives by putting the SAR into a unix workstation environment so that it can sit on a researcher's desk. Section 2 is concerned with the basic design of a user friendly system for accessing the SAR. Section 3 examines more closely the technical aspects whilst section 4 presents an example of the use of the system. Finally section 5 outlines some plans for further development.

2 Designing a SAR Specific Software System for Unix: USAR

2.1 Coping with the 1991 SAR

The SAR is a large dataset, though not excessively by modern standards. The original SAR data files amount to 126 Megabytes (79Mb for the 2% sample of individuals and 47Mb for the 1% sample of households). If the data are loaded for SPSS then the system files amount to 93Mb and for SAS then the system files go up to 176Mb.

The original intent in 1991 was to load the SAR data files on to a large mainframe at Manchester Computing Center (MCC) to provide a table output service for users using the same relational database package (model 204) as used by the UK census agencies in processing

the 1991 census. Indeed, the Census Microdata Unit at the University of Manchester was established in 1992 to provide three sorts of dissemination service relating to the SARs:

1. an online service, available over a network using database software such as SIR and Quanvert and statistical packages such as SPSS and SAS;
2. distribution of the raw SAR data; and
3. a customised tabulation service.

This paper describes a fourth diffusion and dissemination path that is now also supported by the CMU. The idea is to develop portable data access software designed for use with the SAR and which can be ftp'ed over JANET to any unix workstation with sufficient disc space. The aim is to provide an alternative standard means of accessing and using the SAR that is likely to become increasingly relevant during the 1990s. Indeed the continued fall in cost of workstation hardware emphasised the importance of storing the SAR data in a form suitable for a unix workstation and even a PC environment. This provides academics (and others who buy the data) a very different access path, one based on the distribution of the SAR data with associated specialist software designed to allow both expert and non–expert users to easily and quickly obtain maximum benefit from the SAR datasets. Indeed, it is thought likely that both experienced and enthusiastic SAR users will increasingly want the SAR data available on their local unix systems and that they will find specially developed SAR relevant software of considerable practical assistance, to complement the widely available general purpose statistical packages.

The SAR data was originally conceived as a means by which users could develop their own customised tabulations. The underlying presumption is that the user knew precisely, in advance, the content of this tabulation. To this extent, the SAR provides a substitute for the OPCS traditional census tabulation service which might well be regarded as clumsy, haphazard, expensive and slow. However it soon became apparent that users will increasingly by interested in the SAR as an analyzable data source in its own right, and with out any strong interest in using it only as a source of special tabulations; to supplement the tables contained in the small area statistics or local base statistics set of up to 20,000 crosstabulated counts. This reflects the attractions of public use microcensus data, but it also raises many interesting research related analysis questions and increasingly draws attention to whether or not social scientists can actually cope with the SAR. It is noted that traditional sample survey data sets are fairly small is size. There has really not been a very large sample survey equivalent to the SAR available for general and unrestricted analysis. This raises the need for appropriate analysis technology suitable for handling such a dataset.

The main problems raised by the SAR include the following:

1. Its size may result in extended tabulation times if inefficient software is used. There are

1,116,181 individual records and 215,789 household records containing records for 541,894 individuals.

2. Its complexity. The data contain hierarchical relationships that are important but have never been easily and efficiently handled by many statistical packages.

3. Its volume. There are 48 individual variables and 44 household variables with varying numbers of categories; which increases uncertainty in table design and seems likely to make recoding a major activity.

4. Many of the SAR variables relate to research themes for which pre–existing theory and knowledge are probably inadequate, putting emphasis on data exploration within both strongly focused and weakly focused areas of interest.

5. The SAR is expected to become an extremely popular data resource due to its novelty, the potential new census insights it offers, and because the census is already a major data resource for a wide range of researchers who may well discover that the SAR provides an additional dimension to their work. As a result it cannot be assumed that the SAR end–user will possess any particularly expert level of skill in computing and, or, social survey analysis. There could also be a very large number of them. This combined with the other characteristics of the SAR makes it important that the data are: easily accessed with a minimum of alien computer control language, analysis is fast, the data is largely self–documenting, and a wide range of different social scientist skill levels can be catered for. Access has to be easy, straight forward and intuitively obvious, but also table creation is not sufficient by itself, so new tools are needed to help users cope with th SARs.

It is assumed that these problems can be best addressed by developing specialist SAR specific software. It is also believed that such a system would have to offer the intending SAR user sufficient added functionality to make it worthwhile acquiring and then discovering how to use it.

2.2 Basic Design of a Unix Based System

A case has been made that the nature of the SAR data and the problems it will present many users suggests that specialist software might be the best way of meeting their needs. In principal the task is fairly straightforward and can be specified as in figure 1. The process is essentially that of providing tabulation functions but in a context of broader system design. Tabulation of fixed pre–specified variables is not sufficient by itself because many users will need help with the table design process. It is noted also that statistical packages whilst good at tabulation and analysis, provide surprisingly few tests or methods designed to assist the user specify "good" tables. Additionally there is not even any statistical concept of what might distinguish a "good"

table from a "bad" table. Yet this is central to the use of the SAR. The immense freedom that a large sample data set provides the user with, puts considerable emphasis on the table design process. In many ways, the most relevant table design tools still need to be invented. Tools are also needed to help the user explore the SARs by both providing recoding and other table design optimisation functions as well as cater for more broad brush, tentative, highly explorative searches and queries. Maybe the expert analyst of multi–dimensional contingency tables will not need to use such technology; however, our view is that most of the potential users will not be experts. Nor is the attainment of an expert level of statistical skills and advanced training in user hostile statistical packages such as GLIM ever likely to be a reasonable pre-requisite for making use of the SAR.

It is useful then to view the development of the proposed unix SAR diffusion system (USAR) to be based around the table design process. The principal technical problem is not specifying the process but of providing viable methods for the function boxes and of making user access both as simple and as safe as possible. A key element to the design of USAR is the design of a data structure that allows a table to be passed from function to function as the user adds features and variables to the table they ultimately wish to see displayed on the screen. This is handled by the use of a global structure which contains references to the variables used for the rows and columns of the table, the recodes applied to these variables and to the names of these recoded groups. The structure also keeps track of the data file to be used and the filters to be applied when that file is read from disk. At each stage in the trip from the top of figure 1 to the bottom the user can display details from this structure and modify or add to them. This is complicated by the numerous opportunities for the user to change their mind as to the form of the table and to return to an earlier part of the table design process to achieve this.

2.3 User Interface and Portability

In operationalizing figure 1 it is critically important that the user interface is as simple and intuitively obvious as possible whilst still being portable within the unix world.

The usual way of providing a simple interface for the user is a graphical user interface (GUI). However the need for portability across different unix systems (which whilst superficially similar are very different to the programmer) and an ability to function on very different hardware, restricts what is feasible. To achieve this end USAR uses the unix curses library (a standard UNIX software library) a which allows the use of intelligent terminals found on local and wide area networks. It is proposed to develop a second version of USAR which makes use of X-windows (USARx) to allow researchers with access to a suitable terminal to make use of the advantages of this system. The requirement that USAR should be portable and as hardware independent as possible is a major and inviolable constraint on the implementation of USAR. It certainly limits the level of GUI sophistication; for example, no use can be made of the function

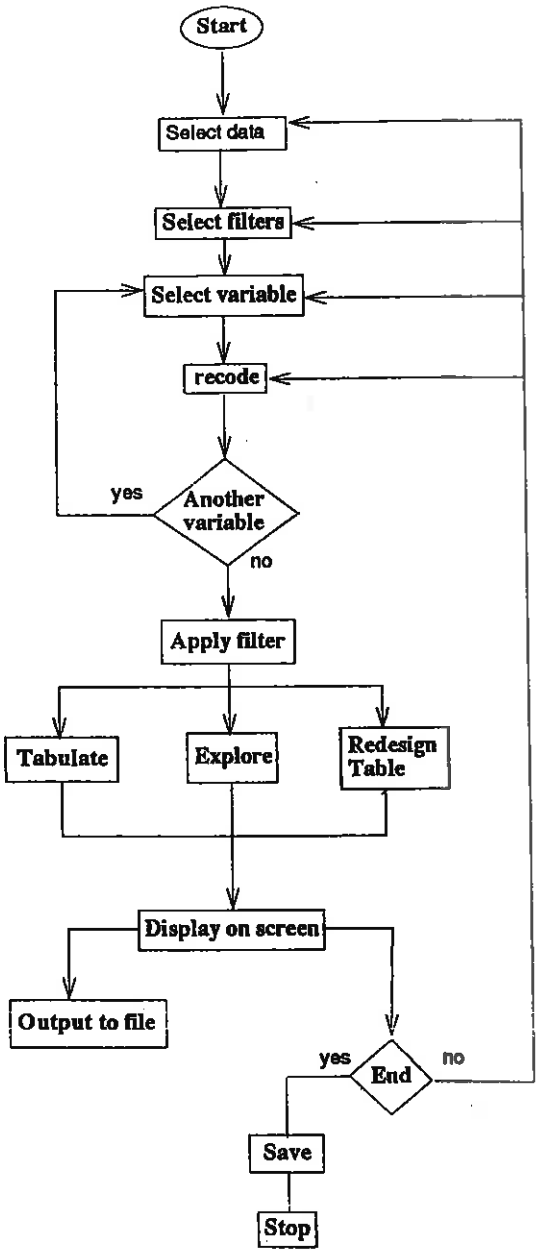

Figure 1: Graphic representation of USAR

keys or many other non–standard keys on the keyboard; but it also, perhaps even perversely, imposes a degree of simplicity that users might actually discover to be helpful.

Another design requirement is that USAR should be able to produce output which can be easily imported as flat ascii text files into standard graphics, spreadsheet and wordprocessing packages running on both UNIX machines and PC's with the minimum of user effort.

2.4 Ease of Use

It is extremely desirable that the system should be easy to use and not require either special training courses or lengthy manuals containing hundreds of pages of documentation. The ideal system would be intuitively obvious and contain all the necessary help information so that both the complete novice and computer illiterate social scientist could learn about USAR from within it. This goal is assisted by a number of aspects of the system design context:

1. the restricted nature of the functions that the system provides avoids the need for general purpose commands in the context of totally flexible open–ended systems;

2. the simplicity of the interface precludes elaborate windowing techniques that might otherwise have tempted the authors to develop a much more elaborate, perhaps over elaborate, design;

3. the target end–user is assumed to be a social scientist non–expert with computers which rules out the more terse unix style of modes of input and indeed any keyboard operation that requires the user to use more than one finger at a time.

The curses interface restricts the user input to a small number of keys. Here the keys are restricted to the arrow keys for selection, return as an action on selection, and occasionally alphanumeric input. This is operated with a range of fixed screens and a context dependent dialogues and help box at the bottom of each screen.

The system is designed so that the novice user can easily and quickly create a table and then be lead via help information to explore the more advanced facilities. A number of cookbook examples (Turton and Openshaw 1994) that lead the nervous through the process have also been provided. There is no need to have a different systems interface for the advanced user since the basic commands are terse, and the help information that appears in the dialogue area can simply be ignored. It must be noted that this simplified design prevents the advanced user producing tables of extreme complexity with several tables concatenated together, however nothing prevents them from carrying out this sort of table editing with other unix software tools.

An important design feature that affects ease of use and also user psychological attitudes towards USAR and their confidence in the system is the inability of the system to crash, barring

external problems with the machine running USAR. No matter what the user does the system should not crash but return a fairly understandable message that identifies the problem and suggests a solution. This is only possible because of the fairly restricted functionality that the system provides. A further important design feature is that when a table is generated perhaps as a result of a long and complex series of recodes, maybe split over more than one session, the recode details and any filters imposed on the data should be printed as footnotes to the table.

2.5 Data Security in an Open Systems Environment

USAR is intended to be run in an open systems environment. It is designed to be copied and transferred between sites; at the same time it is recognised that the SAR data is a valuable commodity.

There is a basic design requirement to ensure that the raw SAR data cannot be easily reconstituted and that the system even when stolen will not readily work. The aim is not to provide comprehensive guarantees of system security. This is unnecessary because the Manchester CMU distributes the raw SAR database only to registered users. From a census confidentiality viewpoint, the SAR contains no confidential data; it is an open database and was designed to be so. Instead the objective is to offer users an incentive to acquire access rights via 'official' channels.

3 Technical Aspects

3.1 Easy and Flexible Table Generation

The task of storing multi–dimensional tables which can contain variables with up to 278 classes is not easy. It cannot be done simply; i.e. by using each variable as a coordinate in an N–dimensional array, where N is the number of variables to be tabulated. This process only works well for small values of N. It rapidly becomes memory hungry and extremely sparse. The first task was to devise a highly efficient but flexible table storage structure.

The basic algorithm is as follows:

Step 1: The user specifies which datafile is to be used and which variables are rows and which are columns. USAR stores this information in a global table structure:

Step 2: A tree structure is constructed for each side of the table. Each node of the tree is a value of variable. The tree is constructed for the innermost variable first, by using a recursive binary method (Sedgewick 1983). If a second variable is required this tree is then constructed and the first tree added to each node of the new tree. This process is repeated for all the variables (see figure 2).

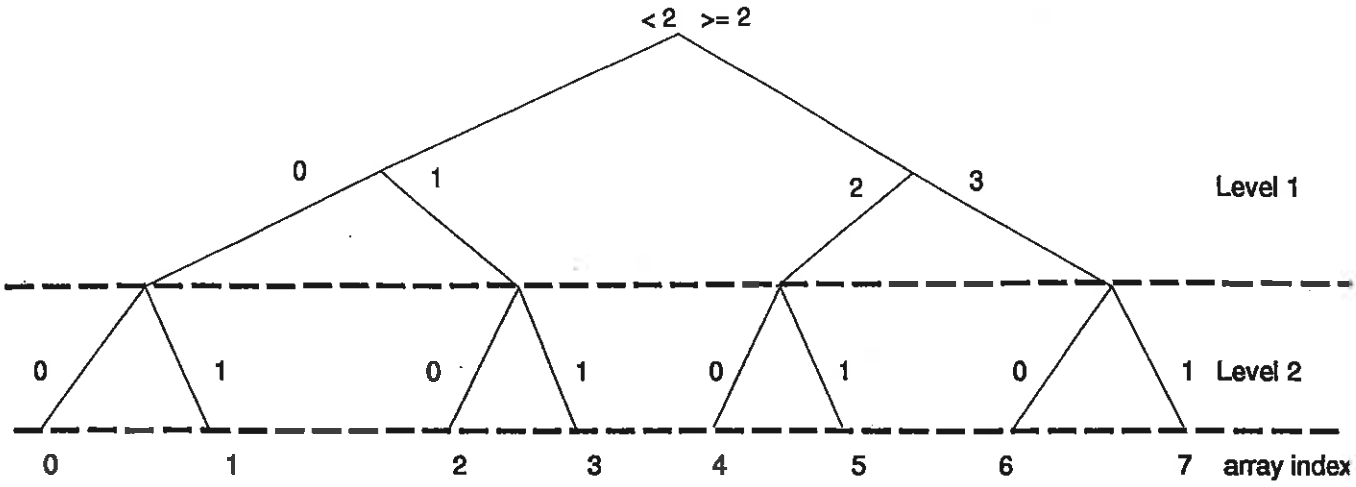

Figure 2: An example of the USAR multidimensional table tree structure

Step 3: The final row of nodes in the tree are numbered to provide indexes to the array used to store the table.

Step 4: Each record is read from the data file and the value of each variable of interest is extracted from the record. These variables are then used to move from the root of the tree downwards to the base. When the node that matches the case for a variable is found, the program moves down a level into the subtree attached to the tree at that node. When all the variables have been processed the number of the final node is returned to the main program.

Step 5: The two numbers returned in the previous step are used as the indexes of the dynamically sized 2–dimensional array that holds the table.

This algorithm allows the USAR system is designed to handle up to 20 dimensions of table in a highly efficient manner. For example a 4 dimensional table of 8 by 9 by 6 by 12 is stored in a 72 by 72 array, and the time taken to find the array indexes for and given record is proportional to $\log_2(8) + \log_2(9)$ or generally $\log_2(N)$ where N is the range of the variable. This level of performance is quite adequate for current purposes.

3.2 Responsive Table Generation

It is axiomatic that the system should provide a fast response to table requests. In designing tables, particularly complex ones, it is essential to obtain a "feel" about the sparseness or range of values of the likely range of values of the results as quickly as possible *if not immediately.*

USAR's speed of response is a function of four factors:

1. platform speed which might also relate to network speed if remote file mounting is used.
2. data compression technology employed to code the SAR;

3. input blocking size; and

4. whether all the SAR or just a sample is read.

The USAR system reads the data in a specially created binary unformatted structure that uses a minimum space for the time taken to read the records. This reduces the size of the SAR to 39Mb for the individual data (from 76Mb) and 22Mb for the household data (from 40Mb). This is achieved by packing the integers of the SAR two to a byte, which halves the space required. Further savings are made by recoding the data values which in some cases use two or three integers to code nine classes (i.e. lowest floor). USAR is highly efficient compared with the equivalent disk space requirements for SPSS and SAS. However, although it compares poorly with the unix compress algorithm (24Mb) it is many times faster to read. Furthermore for users with slow disks or a shortage of disk space a compilation option allows USAR to work directly from unix compressed versions of the USAR files (26Mb). This option approximately doubles the time taken to read the data files, but it halves the required disk space needed to store the data.

The data is also read in blocks to maximise input speeds, the block size is a compile time variable to allow users to experiment as to the most efficient size of input block size for their particular hardware. Though on a Sun workstation it appears that the caching carried out by the operating system makes this largely unnecessary, however it cannot be assumed that this is so for all types of hardware especially at the lower end of the market. The result is that a complete read of the data on a Sun Sparc10 workstation takes 31 cpu–seconds for the individual data and 15 cpu–seconds for the household data. On the Manchester Computing Centre (MCC) cray-6400 server (MIDAS) the individual file takes 19 cpu–seconds and the household file takes 2 cpu–seconds, this can be compared to times for SPSS on the same system of 2 minutes 42 seconds for the individual file and 1 minute 29 seconds for the household file. Elapsed time for the user of course varies depending on other activity on the workstation but is still much less than a minute.

A more dramatic improvement is obtained by using only a sample of the SARs instead of all of it. A size adaptive sampling approach has been adopted to meet this objective, with the sampling fraction being modified to yield real–time results on different speed platforms. This might be most useful when trying out different recodes and table layouts. The USAR data files are stored in random order so that sequentially reading the first 10% gives the equivalent of a 10% simple random sample. This allows the level of interactivity to be adjusted according to the computing environment in which USAR is being run; although obviously all the data should be used when creating final results. Nevertheless samples for the SAR tend to produce very good estimates, this reflects the efficiency of simple random sampling. Table 1 show the effects of taking a 1% and 10% samples from the SAR and factoring up. It is noted that the 10% sample is 10 times faster than the full data set and the 1% sample 100 times faster; so

Table 1: Sampling from the SAR

	Full Sample		10% Sample		1% Sample	
SEX	Male	Female	Male	Female	Male	Female
REGIONP						
North	29991	31967	29560	32360	28600	36100
Yorks and Humb	47603	50701	47290	50740	47000	52700
East Midlands	39275	41128	40260	40440	40500	42700
East Anglia	20530	21030	20710	21670	22100	20400
Inner London	23537	25357	23880	26030	28800	25100
Outer London	40192	43498	39260	44150	38500	46500
Rest of S.East	104882	110423	104950	108390	103500	100900
South West	46150	49100	44970	49150	44200	47300
West Midlands	50929	53338	51040	52810	47900	50900
North West	60948	65303	62220	66010	64400	68400
Wales	28070	29909	28650	29640	28300	27200
Scotland	48860	53460	48150	53850	49200	54900

these results appear almost instantaneously. Tests on a heavily loaded server (with more than fifty users) over a heavily loaded network showed that tables using the full data set required approximately two minutes to complete whereas using a subset of the data allowed tables to be created in a much more reasonable 30 seconds. As hardware speeds continue to improve so the need for sampling will disappear.

Finally to help the user see what the table will look like USAR provides a preview function which shows the layout of the table with no values in it, but gives the user an insight into the shape and size of the table.

3.3 Fast and Easy Table Design

There is a comprehensive table design capability able to handle up to 10 by 10 tables (i.e. 10 variables on each axis of the table) with variables being grouped or recoded in any arbitrary fashion. This is important because it is expected that many users will use the SAR to disaggregate standard tables in various ways to obtain the benefits of micro–census data.

The ability to easily recode and regroup the SAR variables is another key need. There are a number of functions that need to be handled:

1. an ability to load and use standard recode dictionaries either previously prepared in USAR by the user or supplied as part of the standard recode library.

2. to easily recode SAR variables in whatever way is convenient, relevant and easy; and

3. to save recodes for subsequent recall; there is no point in having to retype them each time.

Facilities are provided in USAR for these functions. The recoding task is made particularly easy with the user being provided with a list of labels and text that describes the various codes. The use of standard recodes is also very important. The SAR allows the user to create new definitions of variables and of households. Indeed, there are already various new derived variables that can be added to those provided by OPCS, for example the number of children under 16 in the household. With time the numbers of derived variables may greatly exceed these real ones in the SAR. USAR provides a natural means of loading from standard libraries new definitions so that they may be used without risk of typing errors causing problems.

USAR always works in terms of a two dimensional output table, although each output dimension can have up to 10 different variables. This is unavoidable in presenting multidimensional tables on a flat screen or piece of paper, and has long been a feature in the printing of census tabulations. It is also manifest in the SAS and LBS field definitions. This explains the importance of designating row and column variables. The table can be transposed subsequently.

3.4 Table Output

A further aspect concerns table output, and especially printing. This is a tedious area that many statistical packages tend to ignore. Yet most users of the SAR are interested in table production. It follows then there should be a means of producing generating high quality printable tables. USAR addresses this problem in a unix fashion by allowing the user to output the table in a variety of formats. Other more specialised tools can then be used to produce the printed output. For printed tables the preferred method is to use LaTeX a high quality, free typesetting package[1] which can be used to convert the table and any associated text in to high quality output with very little editing. The tables are complete entities and include notes on filters used, recodes and a copyright permission of the approved format. It is therefore a simple matter to include the USAR output into a document and supply it with a caption and print it out.

Tables can also be output in ascii format with columns separated by tabs which allows import in to most spreadsheet and graphics programs available in the unix and PC worlds. This should encourage users to produce graphic representations of complex data rather than clogging papers with impenetrable tables.

3.5 Handling Data Uncertainty

Sparsity is another problem. With the SAR it is expected that even seemingly mild levels of data disaggregation will produce massively empty (i.e. sparse) tables or tables populated by large quantities of small numbers. This may well be intended but the resulting unreliable

[1] LaTeX can be obtained by ftp from ftp.latex.ac.uk.

nature of the tables needs to be detected and the user at least warned about it. The resulting feedback is in our view an important part of the table design process. Indeed, a balance needs to be struck between the degree of crosstabulation requested and the level of grouping of individual variables needed to produce meaningful results. In this context "meaningful results" is interpreted as values that can be shown to be different from zero when sampling uncertainty is taken into account, at a reasonably modest level of significance (i.e. a type I error level of 5%). It is imagined that such values should populate most of the cells in a tables, if the user wishes to use the results to demonstrate something of substance. To aid the user in this task USAR highlights the cells which fall within two standard errors of zero as calculated by :

$$c - \left[2 \times \sqrt{(c \times (N-c)/N)}\right] < 0 \qquad (1)$$

where c is the count in the cell of the table
and N is the total number of items in the table (Cochran 1963).
At the user's request USAR will also display this value.

Once a table has been displayed the user has the opportunity to recode some or all of the variables used in the table, USAR will instantaneously redisplay the table and the uncertain cells. This should help users to create meaningful tables most of the time, with a little care.

3.6 Automated Table Design

There are probably three principle types of SAR user.

1. Those who know precisely what they want in a table and can provide a detailed, unchanging, *a priori* specification;

2. Those who know more or less what they want but a little recoding will still be needed as they to–and–fro from results back to table via recode changes in an iterative manner; and

3. Those who have no firm idea of what tables they want although they have a good idea of what variables are of interest and a much larger set of variables that are of possible interest.

Only the type 1 user avoids the dangers of *post hoc* analysis, and can reasonably test hypotheses, and thus justify the title "scientific". However it is believed that the potential SAR user base greatly exceeds the limited supply of census analysis experts, so it might be expected that an increasing number of users will be interested in some kind of table exploration. Collectively they may be termed table dabblers, although as knowledge builds up over time they will tend to become experts and the uncertainty in their table design work may well tend towards zero. Likewise, it is likely that those who start by thinking they are experts, will in fact

turn into table dabblers, albeit in a more controlled fashion, as they respond to the immense variable richness provided by the SAR. Users interested in avoiding *post hoc* analysis problems, may well do so by table dabbling on data for one region and then having fixed their table design applying it to data for another.

It should be recognised that the results do depend on the table, on the recodes and filters used. There is a kind of modifiable table problem here that needs to be addressed. This is important because by a judicious juggling of the table design parameters a wide range of different results, and even perhaps contradictory results, might be generated. This is not a new development but has traditionally been rendered inactive by the difficulties of table design in semi–interactive environments or by software that make it difficult or laborious to change tables, or by small data sets that render it pointless. The SAR accessed via USAR escapes these constraints and places, therefore, a much greater burden on the user to be responsible, careful and aware of the dangers of unsafe tabulation.

Unsafe tabulation can be the result of accidental as well as deliberate and careless table design. A well planned table can be ruined by the injudicious recoding of a single variable. One way out is to use only pre–defined category groupings. Science can escape *post hoc* problems only by a rigorous *a priori* definition of recodes. For example the use of standard 5 year age cohorts: or the use of a standard higher order grouping of occupations. Uncertainty is removed by using these devices. However, this rigidity conflicts with the benefits of flexibility provided by the SAR. The user no longer *has to use* standard recodes but, if not, the question then arises as to which other non–standard recodes would be better and how can their usage be justified if they are *post hoc*. It is here where the user needs help and current statistical packages offer only very limited assistance, if any at all.

One way of avoiding some of the problems is to draw the users attention to tables that are unduly sparse and to allow the user to recode the variables used in the table and display the new results on screen for the user before committing the table to print. Accordingly, some basic tabled data visualisation tools are needed to assist the table design process, mainly by drawing the users attention to those parts of the table where the reported counts are likely to be so unreliable as to be useless. To this end, the median polishing method linked to simple table colouring displays is provided as aids to the user in designing "useful" tables that contain "meaningful" information. The flexibility of the SAR emphasises this need for a basic table design tool–kit that goes beyond the table generation task. Previously this seems to have been a greatly neglected topic. In this process the median polish method (Tukey 1977) is applied to the table and then residuals are displayed on the screen as highlighted areas layed over the original table. This draws the users attention to the areas of the table that are not explained by a simple plane fitted to the table. A further example of this technique is given in Marsh (1988).

USAR also addresses these problems by providing several functions for the semi–automated design of tables. A number of objective functions are provided in USAR, and it is hoped that users will submit more over time. At present there is a choice between minimisation of entropy (Shannon 1948), maximisation of χ^2 and the minimisation of small values:

The entropy of a table is defined as:

$$E = \sum_{i,j} \frac{c_{ij}}{N} \times \log\left(\frac{c_{ij}}{N}\right) \tag{2}$$

where c_{ij} is the value of cell row i, column j, and N is the total of the table, which is the inverse of the Shannon information content of the table. So the process of minimising entropy can be seen as maximising "information". For further discussion see Webber (1979).

The χ^2 function is simply the standard form:

$$\chi^2 = \sum_{i,j} \left(c_{ij} - \frac{c_j r_i}{N}\right)^2 / \frac{c_j r_i}{N} \tag{3}$$

where c_{ij} is the value of cell row i, column j, N is the total of the table, c_j is the total of column j and r_i is the total of row i.

The user creates a table and then selects one or more variables to be regrouped. The table optimiser does not change the number of categories present so the user must provide an initial recode for the selected variable with the desired number of classes.

Step 1: The user specifies a table in the usual way;

Step 2: An objective function is selected from minimisation of entropy (equation 2), maximisation of χ^2 (equation 3) and minimisation of small values in the table.

Step 3: USAR passes the table and grouping information to the optimiser.

Step 4: Taking each selected variable in turn a group is randomly selected and then the function attempts to add classes to this group (respecting the order of ordered variables).

Step 5: The new table is evaluated and if there is an improvement this table is stored as the best so far.

Step 6: Repeat the above step until no further improvement is found;

Step 7: Select a new group and repeat the last 2 steps until the maximum number of iterations is exceeded.

Step 8: Repeat until there are no more variables to be used.

This process is fast because the table is not recreated from the raw data file each time a change is to be evaluated. The use of grouping matrices provides a highly efficient framework for handling these function optimisation problems. Experience suggests that this simple Monte Carlo optimisation method works well and that there is no need to use more sophisticated methods such as simulated annealing or tabu procedures (Dowsland 1993, Glover and Laguna 1993).

3.7 Query Facilities, Data Explorers and Fuzzy Searches

Database query is another need. The user may wish to know how many people or households possess a certain combination of characteristics. This can be regarded as a special case of cross tabulation with recoded presence or absence values, or as a series of 'select if' statements linked by boolean operators such as AND, OR and NOT and shrouded in parenthesis. Nevertheless, it is a valid SAR query and these needs have also been addressed by USAR.

An extension of this facility is provided to allow a user to "explore" the SAR database by specifying a variable and a value of interest and then progressively finding others that are strongly associated or strongly not associated with it.

Step 1: The user selects variables and their coded values that are of interest

Step 2: USAR then constructs a linked list of AND filters for each variable selected, multiple selections from the same variable are assumed to be OR choices.

Step 3: Once this chain of filters is constructed and it can be as short as one variable or as long as the researcher requires, the data file is read and each record that passes the list of filters is examined one variable at a time and the counter for the class of the variable is incremented.

Step 4: The user is shown a sorted list of variable classes that occur in records of the type selected.

Step 5: The user then has the choice of selecting more variable classes to extend the list of filters, deleting filters from the chain or constructing a table using the filters already selected. The process can be repeated until too few records match the increasingly large set of filters.

The user may then wish to use the results to create one or more tables to describe the data structure that has been found. This is a very simple but effective search tool. It provides one way of answering what other variables are closely associated with a given variable and its value.

Extensions to handle "fuzzy" queries also needs to be considered even if they might initially be considered somewhat fanciful. A fuzzy query on the SAR could be considered as a probability

that a particular record met the stated 'select if' criteria even when it would fail the criteria if applied deterministically. The fuzziness could be a reflection of measurement error in the SAR or in the precise specification of the 'select if' statement itself. In the later case the user might be a little uncertain about precisely what is needed or would wish to count cases that nearly met the stated criteria. This would not be uncommon as census analysis is an art not a science.

A variation of the fuzzy query is the fuzzy search where an attempt is made to define record profiles that occur most often with out insisting that all the records are identical.

This fuzzy query facility is implemented in USAR by having the user either specify a template or ideal set of values, or by selecting variables of interest and then request the identification of cases that have a certain fraction of the variables in common (i.e. 4 out of 8) but not necessarily all of them.

Step 1: The user specifies the variable class combinations required as a pattern, and the number of required matches.

Step 2: USAR converts this list of combinations into a binary pattern, with one for a match and zero for no match.

Step 3: This list is then checked for patterns with too few matches which are removed. The remaining binary numbers are then used as array indexes for an array of counters.

Step 4: Each record is examined in turn to see if it passes the filters in place, and then to see which of the patterns if any it matches. The counter associated with this pattern is then incremented.

Step 5: The user is shown a list of patterns and the number of records that matched each pattern, together with a listing of the variables and classes it represents.

This fuzzy query or search function might also be considered as providing another form of categorical data exploration. The identified records might be regarded as providing clues about the multivariate structure present in the SAR. Any interesting results might then be used to specify tables.

4 A Case Study of Using USAR with the 1991 SAR

4.1 Analysis of Lone Parent Families

To demonstrate the usefulness of both the SAR and USAR a brief analysis is made of lone parent families. An examination of the local base statistics' (LBS) index, via MetaC91 (Williamson, Rees and Birkin 1994), shows that four census tables deal with lone parents in some form: Table

Table 2: Comparision of LG87 and factored SAR results

	Lone Parent Dep Child		Lone Parent No Dep Chilren	
TENURE	LBS	USAR	LBS	USAR
Owned Outright	7081	9110	24404	25010
Buying	32040	35950	22306	22870
Private Rent	8272	10650	3516	5000
HA Rent	6995	7650	2107	2130
LA/NT Rent	56068	60290	25191	25060
SH Rent	650	650	375	350

Notes

HA rent	Housing Association rent (England and Wales)
Rent LA/NT	Local Authority or New Town rent
SH rent	Scottish Homes

L37 gives the age sex distribution of lone parents between 16 and 24 years of age for Great Britain, table L40 shows the primary economic position of lone parents by the age of children in the family, table L80 contains the number of hours worked by working lone parents, and table LG87 possibly contains the most relevant information showing tenure and car ownership for lone parents.

Table 2 shows the LBS and SAR equivalent table LG87. The differences are due to sampling error, although with the 100% tables other differences may be caused because the SAR contains no imputed households.

However these tables tell us relatively little about the status of lone parents as we are unable to cross reference the groups from one table to another, for instance it is impossible to determine how number of hours worked affects tenure for the group, or the marital status of lone parents or their ethnic characteristics.

USAR allows the results of Table 2 to be disaggregated by many other variables; There are a large number of other disaggregations that can be performed and this is one of the principal attractions of the SAR. However care must be taken to avoid excessively small numbers in which case some regrouping would be desirable.

4.2 Exploring the SAR

Data exploration is a simple task when using USAR, since the user need only have a single variable in mind when starting and the program will lead them through to discover other variables of interest. Initially the individual SAR data file was selected for the search and the family type variable was fixed as lone parent with dependent children. Table 3 shows some of the results of this search. This table indicates that the vast majority of single parents are female (87%) and 55% do not have access to a car, which is widely seen as a deprivation indicator. Nearly half (49%) are single and a further quarter (29%) are divorced. Employment status

Table 3: Some related variables for lone parents with dependent children

Variable	Class	Percentage
Sex	Female	87
Cars	0	55
Tenure	LA Rent	39
Marital Status	Single	49
Marital Status	Divorced	29
Employment Status	Economically Inactive	49
Employment Status	In employment	44
Ethnic Group	White	90

Table 4: Some related variables for lone parents with dependent children, with no employment

Variable	Class	Percentage
Sex	Female	93
Ethnic Group	White	90
Cars	0	76
Tenure	LA Rent	54
Tenure	Buying	17
Marital Status	Single	47
Marital Status	Divorced	28

is split nearly equally between in and out of employment. However only 39 % of lone parent families are in Local Authority rented housing and that the only ethnic group that matters is white.

The search was continued to look at lone parents who are not in employment since they can be considered to be a more marginal group of society, the results of this search are summarised in table 4. For this more disadvantaged group, the sex of the parent increases to 93% female, the proportion without access to a car also rises to 76%. Local Authority renting also increases to 54%. The remainder of variables stay at much the same levels as for table 3. These results are interesting in that they effectively provide a counter to the view held by the media that young women jump council housing waiting lists by becoming unmarried mothers.

4.3 Designing your own SAR tables

Following the search in the previous section it is now possible to start converting these results into tables from the SAR. The first table of interest (table 5) shows the distribution of the sex of the head of family by family type, which allows us to put the question of lone parent families into perspective. As can be seen lone parent families account for only 16% of the total, with lone parent families with dependent children making up only 10% of the total.

Table 6 shows the sex distribution of economic activity for the heads of lone parent families with dependent children. This clearly shows the inequalities of society's expectations that

Table 5: Percentage of family type by sex of family head

	SEX	Male	Female	Total
FAMTYPE				
Mar No Ch		40.17	10.19	18.70
Mar Dep Chil		36.64	11.70	39.92
Mar No dep Ch		14.47	2.60	15.30
Cohab No Ch		4.16	7.73	5.99
Cohab dep chil		1.88	4.83	3.01
Cohab No dep chil		0.19	0.83	0.37
lpt Dep chil		0.91	40.42	10.13
lpt n/dep chil		1.59	21.70	6.55

Notes

Mar No Ch	Married no children
Mar Dep Chil	Married with dependent children
Mar No dep Chil	Married with non–dependent children
Cohab No Ch	Cohabiting no children
Cohab Dep Chil	Cohabiting with dependent children
Cohab No dep Chil	Cohabiting with non–dependent children
lpt Dep chil	Lone parent with dependent children
lpt n/dep chil	Lone parent with non–dependent children

women should stay at home to care for their children with 51% of female parents describing themselves as inactive, whilst 47% of male lone parents are full time employees.

It is also possible for USAR to investigate further the status of the categories shown in table 6 by considering the commonly used indicators of deprivation, access to a car and tenure (see for example Hirschfield (1993) or Davies, Joshi and Clarke (1993)). These results are shown in table 7. It can seen that even for women who have full time employment that a much higher proportion of them fall in the left most column of no car and rented accommodation than for men in the same circumstances, with less than half occurring in the least deprived (right most) group where as for men this group comprises 61%.

A further area of interest is highlighted in table 8 which shows the relationship between marital status and deprivation. As would be expected a majority of single lone parents (66%) fall into the most deprived group. However this also occurs for separated (i.e. still married and remarried) and divorced lone parents, though both these groups have a higher proportion of members in the least deprived group. The only group to achieve a majority of members in the well–off group are the widowed, possibly as a result of life insurance and pension provision.

Combining the results of tables 7 and 8 allows us to see the effects of marital status and economic position in table 9. This table shows that single lone parents tend to be proportionally worse off than the other groups regardless of economic position. This can be seen as a result of a majority of the other groups receiving some support from the child's absent father either

Table 6: Percentage of economic activity of lone parents with dependent children

	SEX	Male	Female
ECONPRIM			
Employee FT		47.97	18.25
Employee PT		2.03	16.91
Self-emp with		4.50	0.69
Self-emp without		8.11	1.52
Govt scheme		1.15	0.84
Unemployed		13.32	6.31
Student		1.59	1.48
Perm Sick		3.35	1.91
Retired		2.38	0.38
Other Inactive		15.61	51.71

Notes

Employee FT	Employee Full Time
Employee PT	Employee Part Time
Self-emp with	Self employed with employees
Self-emp without	Self employed without employees
Perm Sick	Permanently Sick

directly or in the form of housing.

This simple example of SAR analysis demonstrates both the potential of USAR and the richness of the SAR data resource.

4.4 Table Re–Design

Many of the tables in the previous section have a large number of low values that are not strongly different from zero. The question now arises as to what else can be squeezed out of them. One tool is the median polish which can be used to uncover simple anomalous values for further interest.

Another is the use of the objective functions described in section 3.6 to design more parsimonious or compact tables. For example table 9 can be regrouped using minimisation of entropy to give table 10. Here the regrouping of economic activity (ECONPRIM) from 8 to 5 classes sensibly combined several of the smaller groups. The usefulness of this regrouping depends on the context of the situation. The user also has to select the variables to be used in the regrouping.

Table 7: Percentage economic activity and deprivation of lone parents with dependent children

	SEX	Male				Female			
	CARS	No		Yes		No		Yes	
	TENURE	Renting	Owning	Renting	Owning	Renting	Owning	Renting	Owning
ECONPRIM									
Employee FT		22.34	46.67	47.98	61.60	7.81	22.04	22.65	36.22
Employee PT		2.84	3.33	3.59	0.74	12.38	23.98	18.90	22.42
Self-emp with		0.35	1.11	3.14	7.79	0.07	0.53	0.81	2.00
Self-emp without		0.71	4.44	6.73	13.17	0.26	1.95	1.43	4.08
Govt scheme		2.84	1.11	1.35	0.19	0.96	0.71	0.87	0.62
Unemployed		25.89	18.89	11.21	6.68	7.34	6.11	5.36	4.77
Student		1.77	4.44	0.90	1.30	1.24	1.42	2.00	1.71
Perm Sick		6.03	4.44	4.48	1.30	2.06	2.65	2.00	1.24
Retired		2.48	2.22	1.79	2.60	0.46	0.53	0.37	0.15
Other Inactive		34.75	13.33	18.83	4.64	67.42	40.09	45.60	26.79

Notes

Employee FT	Employee Full Time
Employee PT	Employee Part Time
Self-emp with	Self employed with employees
Self-emp without	Self employed without employees
Perm Sick	Permanently Sick

Table 8: Percentage marital status and deprivation of lone parents with dependent children

	CARS	No		Yes	
	TENURE	Renting	Owning	Renting	Owning
MARSTAT					
Single		66.31	5.36	14.40	13.93
Married		41.39	11.24	13.72	33.65
Remarried		37.79	7.17	16.29	38.76
Divorced		42.06	12.41	16.08	29.44
Widowed		24.38	15.30	10.32	50.00

4.5 A Fuzzy Search

Now supose interest focuses on identifing the characteristics of the households that are associated with lone parents with dependant children. Five variables are used: sex, car access, tenure, primary economic activity and marital status. There are two approches to fuzzy analysis:
(1) A fuzzy search. The five variables are specified with some of the categories selected for each of the five variables. USAR will find the household profiles that occur most often, with no more than the specified number of missmatches. Note that the fuzzyness here only relates to the unselected variable class combinations.
(2) A fuzzy query. An ideal household profile is specified; for example female, single, no car, local authority renting and ecomonically inactive. All households that match, for example, 3 or more of these variable class combinations are identified.

The results are shown in tables 11 and 12. In the fuzzy search (table 11) a large number of profiles are found, but only the five most common are shown. The most common household type is female, single, no car, local authority renting and ecomonically inactive, which accounts for 11.5% of all lone parent households. This was used as the example template for the fuzzy query (table 12) which identifies a number of related profiles. The most common group are those that differ from the idea profile in marital status.

Table 9: Percentage marital status, economic position and deprivation of lone parents with dependent children

		CARS	No		Yes	
		TENURE	Renting	Owning	Renting	Owning
ECONPRIM	MARSTAT					
Employee FT	Single		34.80	10.19	22.24	32.77
	Married		15.37	11.57	17.27	55.79
	Remarried		15.66	8.43	15.66	60.24
	Divorced		16.49	11.32	17.80	54.39
	Widowed		9.09	13.37	9.63	67.91
Employee PT	Single		58.53	5.99	18.66	16.82
	Married		27.05	16.43	12.80	43.72
	Remarried		41.86	6.98	18.60	32.56
	Divorced		35.58	17.06	17.40	29.97
	Widowed		17.88	16.56	9.27	56.29
Self-emp with	Single		7.14	7.14	21.43	64.29
	Married		6.45	6.45	12.90	74.19
	Remarried		0.00	0.00	20.00	80.00
	Divorced		3.33	5.00	15.00	76.67
	Widowed		0.00	5.26	15.79	78.95
Govt scheme	Single		74.47	0.00	14.89	10.64
	Married		33.33	8.33	16.67	41.67
	Remarried		0.00	0.00	0.00	0.00
	Divorced		50.00	17.39	17.39	15.22
	Widowed		66.67	0.00	0.00	33.33
Unemployed	Single		67.91	7.44	10.93	13.72
	Married		52.05	10.27	15.07	22.60
	Remarried		41.18	0.00	17.65	41.18
	Divorced		46.84	13.92	14.77	24.47
	Widowed		41.18	17.65	11.76	29.41
Student	Single		46.99	10.84	16.87	25.30
	Married		43.48	13.04	13.04	30.43
	Remarried		50.00	0.00	0.00	50.00
	Divorced		33.82	10.29	25.00	30.88
	Widowed		42.86	14.29	0.00	42.86
Perm Sick	Single		64.29	10.71	8.93	16.07
	Married		54.29	5.71	22.86	17.14
	Remarried		40.00	0.00	20.00	40.00
	Divorced		56.00	12.80	16.00	15.20
	Widowed		30.30	30.30	24.24	15.15
Retired	Single		100.00	0.00	0.00	0.00
	Married		71.43	0.00	14.29	14.29
	Remarried		100.00	0.00	0.00	0.00
	Divorced		50.00	6.25	31.25	12.50
	Widowed		35.00	17.50	10.00	37.50
Other Inactive	Single		75.60	3.60	12.37	8.43
	Married		60.74	9.59	12.07	17.60
	Remarried		52.99	8.21	15.67	23.13
	Divorced		63.67	10.98	14.15	11.19
	Widowed		37.07	16.33	10.20	36.39

Table 10: A possible regrouping of Table 9, using minimisation of entropy

		CARS TENURE	No		Yes	
			Renting	Owning	Renting	Owning
ECONPRIM	MARSTAT					
Employee FT	Single		34.80	10.19	22.24	32.77
	Married		15.37	11.57	17.27	55.79
	Remarried		15.66	8.43	15.66	60.24
	Divorced		16.49	11.32	17.80	54.39
	Widowed		9.09	13.37	9.63	67.91
Employee PT	Single		58.53	5.99	18.66	16.82
	Married		27.05	16.43	12.80	43.72
	Remarried		41.86	6.98	18.60	32.56
	Divorced		35.58	17.06	17.40	29.97
	Widowed		17.88	16.56	9.27	56.29
Self-emp,Govt Scheme, Unemployed	Single		64.79	6.81	12.45	15.95
	Married		33.06	10.08	13.71	43.15
	Remarried		22.22	2.78	19.44	55.56
	Divorced		30.95	12.63	15.16	41.26
	Widowed		17.39	7.61	9.78	65.22
Student, Perm Sick Retired	Single		55.56	10.42	13.19	20.83
	Married		52.31	7.69	18.46	21.54
	Remarried		54.55	0.00	9.09	36.36
	Divorced		48.33	11.48	20.10	20.10
	Widowed		33.75	22.50	15.00	28.75
Other Inactive	Single		75.60	3.60	12.37	8.43
	Married		60.74	9.59	12.07	17.60
	Remarried		52.99	8.21	15.67	23.13
	Divorced		63.67	10.98	14.15	11.19
	Widowed		37.07	16.33	10.20	36.39

Table 11: Results of a Fuzzy Search about Lone Parents

Sex	Cars	Tenure	Economic Activity	Marital Status	Count	% of Total
Female	0	LA/NT Rent	Other Inactive	Single	2529	11.5
Female	0	LA/NT Rent	Other Inactive	Divorced	1604	7.3
Female	1	OO Buying	Employee FT	Divorced	746	3.4
Female	0	LA/NT Rent	Other Inactive	Married	742	3.4
Female	0	LA/NT Rent	Employee PT	Divorced	479	2.2

Table 12: Results of a Fuzzy Query about Lone Parents

Sex Female	Cars 0	Tenure LA/NT Rent	Economic Activity Other Inactive	Marital Status Single	Count	% of Total
•	•	•	•	o	2604	20.8
•	•	•	•	•	2529	20.2
•	•	o	•	o	1911	15.3
•	•	•	o	o	1402	11.2
•	•	o	•	o	1188	9.5
•	•	•	o	•	857	6.8
•	•	o	o	•	581	4.6
•	o	•	•	o	492	3.9
•	o	•	•	•	272	2.2
•	o	o	•	•	221	1.8
o	•	•	•	o	113	0.9
o	•	•	o	•	41	0.3
o	•	•	•	•	34	0.3
o	•	o	•	•	34	0.3

where • indicates a match and o a mismatch

It is apparent that lone parent families are a diverse group which can not be simply categorised. There are large variations between the status and employment conditions of the group which appear to be explained by marital status and the sex of the parent.

5 Future Developments

The SAR provides an extremely interesting data resource. The USAR system described here is an attempt to simplify user access and to provide tools for extracting maximum value from the SAR data resource. There are plans to extend the table design process by incorporating other objective functions and algorithms but without losing sight of the well defined objectives involved in USAR. It is also obvious that USAR could be easily modified to provide a similar system for some other large survey databases and thus ease problems of access and possibly also dramatically improve usefulness by also making their analysis much easier.

Finally it is observed that the SAR can be used to generate about $10^{2,000}$ different tables if you consider recodes as distinct tables, or a mere 4.86×10^{62} if the available variables in the SAR are permutated. There is currently no machine based technology in existence, any where in the (academic) world, that is able to explore this universe of tables in a focused search for the most "interesting results". Yet computer systems are rapidly becoming large enough to store all the data in memory and fast enough to engage in all manner of previously infeasible computational tasks. The technology widely used to analyze survey data, and also the SAR, mainly dates back to the 1960s or before. Survey technology is now lagging behind given the immense changes in computer performance and the explosion of survey–like microdata. There is an urgent need

and also a formidable challenge to look again, in the newly emerging computational era of parallel supercomputing, how best to explore large databases such as the SAR for new theory and knowledge.

Acknowledgments

The research described here was supported by the ESRC under grant number H507255100. The assistance of various people in developing and testing USAR is gratefully acknowledged; in particular Angela Dale, Keith Cole, Phil Rees, John Stillwell, Martin Charlton, Ed Fieldhouse and many others.

The USAR system can be acquired by registered SAR users for free from either CMU or by anonymous ftp from gam.leeds.ac.uk (/pub/usar/usar-nn.tar.Z). The data files used by USAR require proof of SAR registration.

References

Cochran, W. (1963). *Sampling Techniques*, J. Wiley, London.

Davies, H., Joshi, H. and Clarke, L. (1993). Is it cash the deprived are short of ?, Paper presented at Research on the 1991 Census Conference, Newcastle upon Tyne.

Dowsland, K. (1993). Simulated annealing, *in* C. Reeves (ed.), *Modern Heuristic Techniques for Combinatorial Problems*, Blackwell Scientific Publications, pp. 20–69.

Glover, F. and Laguna, M. (1993). Tabu search, *in* C. Reeves (ed.), *Modern Heuristic Techniques for Combinatorial Problems*, Blackwell Scientific Publications, pp. 70–150.

Hirschfield, A. (1993). Using the population census to study deprivation, Paper presented at Research on the 1991 Census Conference, Newcastle upon Tyne.

Marsh, C. (1988). *Exploring Data - An Introduction To Data-Analysis For Social-Scientists*, Polity Press, Cambridge.

Marsh, C. (1993). The sample of anonymised records, *in* A. Dale and C. Marsh (eds), *The 1991 Census User's Handbook*, HMSO, London, pp. 295–311.

Marsh, C. and Teague, A. (1992). Samples of annoymised records from the 1991 Census, *Population Trends* **69**: 17–26.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991). The case for samples of anonymised records from the 1991 Census, *J. of the Royal Statistical Society (A)* **154**(2): 305–340.

Middleton, E. (1994). Samples of anonymised records, *in* S. Openshaw (ed.), *Census Users' Handbook*, Longmans, London.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data, *Environment and Planning A* **16**: 17–31.

Openshaw, S. (1994). *Census Users' Manual*, Longmans, London.

Sedgewick, R. (1983). *Algorithms*, Computer Science, Addison–Wesley, Reading, Massachusetts.

Shannon, C. (1948). A mathematical theory of communication, *Bell System Technical Journal* **27**: 379–423.

Tukey, J. (1977). *Exploratory data anlysis*, Addison–Wesley.

Turton, I. and Openshaw, S. (1994). *A Step–by–Step Guide to Acessing the 1991 SAR via USAR*, Working Paper - 94/6, School of Geography, University of Leeds.

Webber, M. (1979). *Information Theory and Urban Structure*, Croom Helm Ltd., London.

Williamson, P., Rees, P. and Birkin, M. (1994). A meta–database of census variables and tables, *Evironment and Planning A – This issue.*

Produced By
School of Geography
University of Leeds
Leeds LS2 9JT
From Whom Copies May Be Ordered