Editorial

Concern with the quality of official statistics is widespread, not least among statisticians themselves. The report of the Working Party of the Royal Statistical Society, "Official Statistics: Counting with Confidence" was the result of nine months deliberation, preceded by one of the best attended general meetings in recent history. This report, proposing a number of changes to the structuring of the GSS and to the production and dissemination of official statistics, has apparently been rejected by this government.

Radical Statistics has had considerable input into this process, in submission of evidence to the working party, in a Press Release on the report, in contacts with individual journalists, and in an article in the press by Ludi Simpson (reproduced here). A critical report on out involvement in this process is given by Alison Macfarlane.

Where we go from here? The Government response to the RSS report, from what we know so far, welcoming the reassurance as to the integrity of the government statisticians themselves (see article in Sunday independent by Rosie Waterhouse, reproduced in this newsletter) has (deliberately?) missed the point; the absence of any explicit rebuttal implies tacit agreement that the other explanations given in the report, of structural deficiencies in the service (particularly the lack of autonomy) and a consistent reduction of resources (even preceding the Rayner implementations) are nearer the mark (note that the November issue of RSS News & Notes will refer to the Government response in more detail).

In this newsletter, Andrew Philpott Morgan gives a critique of the report and Ray Thomas argues for a radical reconceptualisation of the service. Also highly relevant is Alison Macfarlane's article on Official Statistics, written originally for Critical Public Health, and before the RSS report, reproduced here with a postscript. Finally Andrew Philpott-Morgan proposes a book on Official Statistics which will clearly be an important contribution to this debate.

Another enduring current issue is the Poll Tax. William Low's article compares the Poll Tax to the replaced Rating system, examining differences in the charges levied under both systems in comparison to differences in the take up of services between socio-economic and income groups within a Local Authority. He finds more equity in the rates. This is an important complement to Charlie Owen's national analysis (in Issue No. 44) of differences between individuals by virtue of their residing in different Local Authorities. For anyone interested in this area your attention is also drawn to Ray Thomas's computer conferencing initiative

Conall Boyle spells out the implications of considering the unemployed in relation to those of working age, in contrast to the Government's (varying) definitions. David Hutton examines the quality and consistency of statistics on cooperatives.

Note that book reviews are now reappearing in this journal, thanks to Brendan Burchell. He welcomes any requests to review.

1991 AGM and Conference

Note that the AGM and Annual Conference will be held on 23rd February 1991 (provisionally) at the Department of Public Health and Medicine at Leeds University, kindly hosted by Waqar Ahmed, Trevor Sheldon & Colin Thunhurst.
Anyone with ideas as to the programme should get in touch with Waqar on 0274-733466 Ext 6262 up to 1st December and either Trevor (0532-344854) or Colin (0532-459034) after.

Book Review:

Mapping and Measuring the Information Economy

by Ian Miles and others at the Science Policy Research Unit. Library and Information Research Report 77, the British Library.

(review by Ray Thomas)

This is an unusual book. Its aim is to give a critical guide to data and data sources relevant to the development of activities using new information technology. It is not about the information industries themselves (the subject of a parallel study at the Polytechnic of Central London), and in spite of the use of the term 'mapping' in the title, it is not about the spatial impact of IT developments (the subject of a parallel study at the Centre for Urban and Regional Studies at the University of Newcastle). Nor does it aim to be a definitive guide to IT on the lines of the joint ESRC/RSS Reviews of UK Statistical Sources.

Rather, the book tentatively presents a conceptual framework for collecting information on IT and reviews currently available surveys and statistics in the light of that framework.

Miles identifies the convergence of computer and telecommunications technologies, underpinned by the binary representation of data, as the crux of IT. This approach leads to a focus on the 'heartland' industries of chip production, computers, software, telecomunications equipment, and telecommunications services. Other parts of the book are concerned with research and development, with the diffusion of IT, on major application areas, on IT and people (as students and trainees as well as employees), and on some of social implications of IT developments. Miles discusses the idea of an information economy and concludes, pragamatically, that the concept is useful for distinguishing between economic affairs before and after the development and diffusion of IT.

Researchers, and others, often complain that advances in knowledge and understanding are hindered by lack of information. It is noteworthy that Miles does not often make this complaint. In the concluding chapter of the book Miles claims (with full justification) that he has demonstrated that "a wide range of data are available on different aspects of IT use and the information economy". Miles points to the diversity of IT applications and the difficulty of assembling a coherent picture from many different sources. He

suggests that though newer services are poorly covered, there is a wealth of data on the IT-producing sectors and on the diffusion of IT. He calls for more analysis of existing studies of professional IT employment. On R&D he says "little effort has been made to pull together the many statistics".

Miles shows how the broad pattern of IT developments can established on the basis of existing statistical series and surveys. The association often made between the information economy and the growth of service industries, for example, is indicated by the standard input-output tables which show that service industries account for eighty percent of IT investment. But, since no breakdown of services industries is given, there is no way to go beyond the broad pattern.

The diffusion of IT in manufacturing is traced by surveys conducted by Northcott at PSI in 1981, 1983, 1985 and 1987. A number of surveys indicate the use of IT is positively associated with size of organization and that sectors with the highest use of mainframes also tend to be the highest users of mini and microcomputers.

Miles includes a table which shows that the Ministry of Defence accounts for 32% of central government IT expenditure, the Department of Inland Revenue 16%, and the Department of Health and Social Security 11%. No other central government department accounts for more than 4% of the total. This Table makes the reader wonder how the share of the Department of Trade and Industy of government IT expenditure in this country compares with that of MITI in Japan.

The nature of the problems encountered in getting meaningful detail rather than a picture of the broad pattern are well identified. Problems arise, for example, from existing industrial classifications which include electronic equipment with electrical equipment. Problems arise from the rapid rate of technological progress (equivalent to a rise of about 20% per annum in the performance to cost ratios of most kinds of computer equipment) which makes it difficult to maintain any kind of comparability over time.

As the costs of hardware fall, the contribution made by software, by computer services and by value added services can all be expected to increase. But measurement problems in the non-hardware areas are greater. Miles reports a difference between private and governemnt estimates of the size of the computer services industry, for example, of nearly 100%.

One of the questions raised by this book is the relevance of the education and training of statistician. This is a book about statistics, but its author is not

ntified as a professional statistician, nor is he listed as a Fellow of the Royal tistical Society. Does this mean that dealing with statistics relevant to IT velopments gets into areas which are not covered by the traditional training d expertise of the statistician?

T developments are having a considerable influence on the way statistics are used and on the number of people using them. The spreadsheet is a significant example which has contributed to a growth in both the number of users and in the range and variety of statistical calculations made. But it is not clear that this impact is reciprocated in the sense that statisticians are making a contribution to understanding IT developments.

Should the expertise of the statistician be limited to traditional areas such as the evaluation of experimental data and the analysis of well established and reliable time series? Or should the job of the statistician be defined more widely in terms of the identification of facts about society, and the way society is changing. If the latter, there is no doubt that Ian Miles, through the publication of this book, would have established himself as one of Britain's leading statisticians.

Ray Thomas
Open University

Software Review

PC/BEAGLE:

Software to do the thinking for you?

Price £60+VAT (£50 Education Rate) from Pathway Research Ltd, 8 GrovenorAvenue, Maperley Park, Nottingham, NG3 5DX (0602 621676)

(review by Brendan Burchell)

It's not very often that a new data analysis software package comes along which claims to be able to revolutionise the way in which you will analyse data. It was those claims that made me request a copy to review for the Radical Statistics Newsletter. PC/BEAGLE seemed to be doing a lot of things that a good exploratory data analyst should be doing, and claimed to have "beaten the experts" already in several different domains.

PC/BEAGLE is a "Rule Finding Programme" which claims to use artificial intelligence technology. It reads normal rectangular datafiles and can produce rules for the way in which those variables inter-relate. A "target rule" is put forward and its state (ie True/False) is predicted by the others in the dataset, either in simple bivariate rules or more complex interactions involving several variables.

An Example – investigating the North-South divide

Perhaps the best way to get a feel of what BEAGLE (which stands for Biological Evolutionary Algorithms Generating Logical Expressions) does isto start with an example of a set of analyses where it produced some useful results, before going on to expose some possible shortfalls in the software and the logic behind it. The dataset that I used was the one produced by CURDS (Centre for Urban and Rural Studies, University of Newcastle), and made available from the Essex Archive in the form of ateaching dataset to accompany Marsh's book "Exploratory Data Analysis". It consists of data on 280 towns (or "local labour market Areas" to be precise, geographical functionally defined units.) I set out to explore the way in which the towns in the South East, South West and East Anglia(taken together, hereafter called "the South") could be differentiated from the rest of the UK ("the North"), using the following variables: Population in 1971, Population in 1981, Change in employment from 1971-78, Change in employment from 1978-81, Rate of

unemployment in May, 1985 and Proportion of dual car ownership. In order to test the software's "intelligence" to the full, I did not give it any prior hints about the way in which those variables might be related to North-South differences.

The best results I obtained were on the third attempt to run the programme (Some might find it rather disconcerting to use a statistical tool likely to give very different results on consecutive runs, but I find it a welcome reminder of the arbitrary nature of a lot of what we do. More of this later.) On that run the two rules which it produced to predict that a town was in one of the three Southern regions were:-

1. The unemployment rate should be below 9.3 (no surprises here).
2. If and only if a town has a population greater than 56485.5 it should have a dual car ownership rate of greater than 17.855.

These two rules give a crude success rate of 78.2% in classifying all 280 towns (or, as a more rigorous test, when these rules were derived from a randomly selected subset of 182 towns and applied to the other 98 towns, a crude success rate of 76.5%) Where both of the rules were true, 39 out of 42 (93%) towns were in the South - The exceptions being Macclesfield, Harrogate and Kendal. Of the 192 towns where both rules were false, 39 153 (79%) were correctly classified.

Rule number 1 would have been easy enough to find by some simple exploratory analyses. It is Rule number 2 that is more intriguing. A human geographer would have told you that dual car ownership was related to urban/rural dimension, but it is exactly that sort of interaction that is very easily missed by the analyst who does not have a priori grounds to hunt for it. Yet BEAGLE revealed a sizable interaction that is both interesting and meaningful. In the South, there is no difference between the rate of dual car owner households between small towns (67% of them having rates greater than 17.9, the cutoff point identified by BEAGLE) and large towns (64%). However, in the north, the rate of dual car ownership drops markedly from 41% of small towns being classified as having high ownership rates compared to only 9% of the large towns.

Clearly that is not the end of the story as far as an inquisitive analyst is concerned, but it's an excellent start. Before going on to describe situations where BEAGLE was unhelpful, and my worries about those situations where it might be positively dangerous when applied to social statistics, I'll give an overview of the process that the user and software goes through in order to get from the raw data to the sorts of rules described above.

Running the Software

Three input files are needed, a data file containing the dependent and independent variables (a simple comma or space delimited rectangular file), a data-definition file simply listing the names of the variables and their type (numerical, string or simply labels - in this case the names of the towns were also read in to make the output more readily interpretable) and a rule file - in this case simply containing the rule "REGION < 4" as the South East, East Anglia and South West were numbered 1, 2 and 3 respectively).
There are six main parts to the programme that one goes through sequentially, checking the output at each stage, and sometimes acting on that output. This semi-automated nature of the programme is nice in as much as it keeps you informed, allows you to spot when you or it make silly mistakes, and gives you a feel for what is going on, but may become annoying if you were to use it on a regular basis.
The first module is called SEED (Selectively Extracts Sample Data), which reads in the data and splits it randomly into two groups – one to derive the rules, and the other to test them. This helps counter the problem of spurious inflation of statistics, whereby parameter estimates are more successful on a sample that they are derived from than when applied to another sample drawn from the same population. What BEAGLE does to overcome this is to take out about a third of the sample and ignore it until the rules are derived from the rest of the batch, and then, at the very last stage, see how well the rules work on that "virgin data".

The Second stage, ROOT (Rule-Oriented Optimization Tester), reads in the target rule that you want to use to classify the data into two groups (North-South in the example above) and if you have any hunches about rules that might turn out to be effective they could be put in at this stage. Otherwise BEAGLE randomly creates a number of rules (about 10) to be tested in the next module. Usefully, it also provides a file containing minimum, mean and maximum values for each variable at this stage too.

The Third stage, HERB (Heuristic Evolutionary Rule Breeder), is the real guts of the thing. It works on a survival of the fittest evolutionary model. Rules are tested in turn, the ones with little or no predictive power (as measured by the Phi Coefficient) being discarded, the moderately good ones being mutated or "mated" together and the best rule being carried forward unchanged. One chooses the number of generations and cycles to let this process run for; For the TOWNS dataset this took about 10 minutes on a fast 286 machine, but on other, larger datasets I've had to let HERB run all night. The output from this stage is a number of much better rules, typically about three to six.

These rules still have to be combined to see how they work in combination, which is done in STEM (Signature Table Evaluation Module) which looks at the proportion of "True" cases (eg. proportion of towns in the South) for each combination of the rules from HERB. It also tests the chi-square values of all the rules in combination on the target expression, and the model leaving out one rule at a time. Often at this stage a redundant rule is identified and can be dropped.

Finally, this combination of rules is tested on the data using the LEAF (Logical Evaluator And Forecaster) module, usually using that set of data that was put by at the first stage by SEED. It lists the cases, giving the predicted and actual value of the target value for each one. It also gives some summaries, most importantly the success rate, but also success rate omitting the cases where their estimates where based on too few cases to be reliable, and the success rate only for the cases which were positive on all rules or negative on all rules.

For most purposes one would probably stop here, but there are also facilities for creating files of the output suitable for reading back into statistical packages, or for creating code in PASCAL that would divide up the sample according to the rules produced. There are some other variations on the theme too, for instance one can use a continuous variable as a target rather than a dichotomous one, or analyze time- series data by selecting variables to be lagged.

So that's an example of a relatively successful application and a description of how BEAGLE works. So what are the drawbacks? First some practical considerations, then some conceptual worries that I have.

The first couple of examples that I tried proved unsuccessful, and a list of situations where BEAGLE works best is provided at the back of the manual which tends to suggest that it would not be suitable for a lot of social data sets. For instance, it performs poorly on "noisy" data, which rules out much psychological and sociological datasets, particularly ones where the individual is the unit of analysis. It also has problems with large and small datasets – the author suggests hundreds rather than thousands of cases. Similarly, the greater the number of variables, the less likely it is to come up with the best rules. Furthermore, it works best when there are approximately the same number of "true" and "false" cases according to the target rule – ie in my example, towns in the North and towns in the South. I gave it an example of some survey data with approximately a 94%-6% true-false split, and it predicted that all cases were "true", thus achieving a crude success rate of 94%! It thus would not be able to handle a lot of epidemiological datasets concerned where the incidence of risk was appreciably less than 50%. You

also need to think quite carefully about how you arrange the variables that you feed it – for instance which ones to turn into dummy variables, which the manual gives no advice on. Of course, you also need an interesting problem to solve where you have identified the dependent variable of interest. "Ask no sensible questions, get no sensible answers".

I also uncovered a few annoying and potentially misleading "features" in the program. For instance, the first time that I attempted to read the towns data into BEAGLE I had left spaces in the middle of some of the town's names (eg. Milton Keynes, St Andrew's) which meant that they got read as two variables, throwing the whole of the dataset out of synchronisation. It crashed when I tried, in error, to read in Word Perfect file instead of an ASCII one. It is also rather unfortunate that several of the acronyms used to describe the modules have completely different meanings in other exploratory data analysis contexts (eg STEM, LEAF, ROOT). But, for £50 or £60 for a suite of specialist software, one does not expect perfection. All in all it's not got a bad feel about it. And the manual is excellent, often teaching by example (several sample datasets are provided on the programme disks)

The more important questions concern the implications of this completely new approach to data analysis. Is it the extreme example of atheoretical data-dredging? Those statisticians who were brought up on the philosophy that hypotheses should be carefully formulated and analyses specified before the data is even touched would clearly have no truck with such a method. But even exploratory data analysts would probably have reservations about such a thorough sifting of the data.

There are two conceptually distinct problems as I see it. Firstly, and most simply, there is the "type 1 error" problem; if you test enough different rules, you are bound to get some solutions that seem to have some predictive power, even on randomly generated data. This problem is dealt with in BEAGLE by dividing the data into two sets, one for data generation, the other for testing the rules once formulated. This may be very wasteful or impractical with small samples, or situations (such as the towns example above) where the entire population (ie local labour market areas in the UK) is small and finite, but significance testing or model fitting has always been difficult in those situations anyway. The more complex, second problem with BEAGLE is it's totally empirical approach to data analysis. In the early days of the discipline, hypotheses would be specified exactly in advance, and only those calculations to test exactly those hypotheses were performed. This would have been exceedingly convenient in the days before computers. Then, with the advent of the sort of computing power that we have all been getting used to over the past decade or so, experimenting with models until the one

that has the largest impact on the hypotheses being tested became the norm. Now, with the addition of a new genre of software which will reveal all of the potentially interesting relationships in the dataset, perhaps we are entering an era where we will be doing everything the other way around completely: the first stage of an analysis will be to set a rule-finder to work on the data, and when it's done its job, we then stop and think about what to make of those results. The dangers of this approach depend on the epistemological process which the data analysis is embedded in. In some disciplines the results of an analysis stand or fall principally on the predictive power of the relationships found. For instance, forensic scientists might be interested in predicting where a microscopic fragment of glass has come from (a car headlight, a bottle, a window pane, etc), and if BEAGLE can come up with rules that are better than those previously used (as the author claims it has done) then that is an end in itself.

However, in the social sciences, usually we are more interested in finding relationships that further our understanding of rather complex situations, rather than finding facts from the dataset. Finding relationships between variables is usually only a useful exercise in so much as it advances our comprehension of the underlying processes that give rise to those relationships. Thus, I am sceptical of the usefulness of BEAGLE to increase our understanding. BEAGLE is a rule or fact producer, and facts are not what make for intelligence in the social sciences.

But, and this is perhaps my greatest concern, as we have all become increasingly aware over the last decade, statistics concerning the state have been used not so much to inform about the effects of government but to misinform. It is nothing new to suggest that two statisticians can, from the very same set of data, construct two very different stories, both supportive of the values and expectations that they started with. Yet, most of us still believe that data analysis is not an entirely subjective exercise; if good practice are followed, there should be some consensus at the end of the day. And one of the worst practices that data analysts can perform is to continually dredge through their data until they find some isolated morsels of evidence that support their position. Then, by quoting those few bits of supportive evidence, while disregarding the evidence contrary to their position, they can present a persuasive argument based on the facts. If there is a danger of BEAGLE, then it is that it could be used to arm the unscrupulous statistician or politician with so many "facts" that there are bound to be some that can be used to argue from any position. It's something that goes on anyway, but the widespread use of programmes like BEAGLE could make it a lot more prevalent.

But don't let me detract from the quality or value of BEAGLE itself. If you've read my example and think that it would be useful for you to do things like that, it is very good value for money indeed.

Brendan Burchell
Social and Political Sciences
University of Cambridge. University of Cambridge.