RESOURCE DISCOVERY FOR TEACHING DATASETS

Valmira Hoti, Brian Francis and Gillian Lancaster
Department of Mathematics and Statistics, Lancaster University, United Kingdom
V.Hoti@Lancaster.ac.uk

The use of relevant and appropriate datasets is recognised as an important prerequisite in teaching statistics to non-statisticians. Such examples help to provide motivation for the student and can aid both understanding and performance. While impressive resources such as the Data and Story Library and the datasets section of STATLIB exist, there is a need for a more comprehensive index of datasets which are freely available on the web. Datasets exist in a wide variety of locations, however, and it is often a hard task for the lecturer to find an appropriate dataset which both illustrates a particular technique and is relevant to the background of the student. This paper discusses the problem of constructing a resource to allow lecturers to discover appropriate data sources. It reports on a demonstration project which is trawling a wide number of types of data sources for relevant datasets, and describes the successes and pitfalls.

INTRODUCTION

In 1972, Hubert Blacock stated that "One of the most difficult problems encountered in the teaching of applied statistics is the motivation of students". Nearly forty years later, the problem of student motivations still exists despite the availability of modern computer technology, comprehensive statistical software and sophisticated teaching aids. It is generally agreed that the provision of relevant and discipline-specific examples to students who are studying statistics is an essential component of any statistics course to non-statisticians. Singer and Willett (1990) make a strong case for using real datasets in service courses, stating that many datasets used in teaching are artificially created lists of numbers, and do little to help students become competent data analysts. They give three reasons why real datasets provide a more meaningful educational experience.

1. They allow students to acquire data analytic skills in a realistic research environment.
2. Real data allows students to see how statistical analysis can inform topical debate, addressing the "why" of data analysis as well as the "how".
3. Using real data helps integrate statistics into the general education curriculum.

The views of Singer and Willett have been echoed by subject specialists. For example, one of Love and Hildebrand (2002)'s key recommendations for the effective teaching of statistics to business students is to use real data relevant to the topic area. They state that "examples and datasets should be relevant to real business processes and decisions". In social science disciplines, other authors have identified the need to use relevant datasets. Thus in sociology, Paxton (2006) identifies the need to use real datasets which engage the students, and Humphreys and Francis (2009) report on the importance of using real life datasets to motivate criminology students.

We echo these views, and indeed would encourage all statistics courses (whether aimed at statisticians or non-statisticians) to engage with real data, and to focus on interpretation and the substantive meaning of any analysis. For statistics students, real data needs to span a broad range of disciplines, and be sufficiently messy to represent the practical problems which practitioners come across in their day-to-day careers. This is in direct contrast to a beautifully crafted artificial dataset which will illustrate a technique but which will fail to challenge the student with interpretation issues.

The task for any lecturer who is seeking to engage and motivate students on a statistics course is to find suitable datasets which can engage the student's interest. One solution sometimes adopted is to encourage the students to collect their own data but, while this may be valuable in teaching the students the practicalities of design, the formulation of research questions and the difficulties of real data collection, the resulting dataset may not have many of the desired characteristics which are required. In addition, data collection can fail–plants can die and individuals refuse to participate.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

It is better therefore to try to find existing datasets which have been collected and used. But this runs up against another problem–that of trawling for suitable datasets. Singer and Willett identified the problem of finding datasets as one real constraint which stops course developers from using real data. The internet provides a rich source of data, but how can a dataset search be carried out?

Computer scientists recognize this as a major problem to be solved. The inventor of the World Wide Web, Tim Berners-Lee, suggests that in the future, there will be a new web which will focus on information rather than documents. The "semantic web" will consist of linked, openly available datasets, with data having in-built connectivity. He takes the example of government datasets. "Government data is a valuable resource that we have already paid for. This is data that has already been collected and paid for by the taxpayer, and the internet allows it to be distributed much more cheaply than before. Governments can unlock its value by simply letting people use it." (Berners-Lee, 2009).

While such protocols are being developed, the discovery of data for teaching is still a difficult task. In the next section we highlight the available sources of data for teaching. We then focus on the problem of labelling or tagging such data, before describing a potential web-site which would provide a datasets resource for course designers.

DESIRABLE CHARACTERISTICS OF TEACHING DATASETS

We have identified four characteristics of teaching datasets which are in our view desirable.

1. Open Access. There are two views on providing teaching datasets. One view, predominant in Medicine and in the Social Sciences, takes the view that data is private, that investment in collecting data has been made and that data is a private resource. For such researchers, ethics and privacy of information takes precedence over open and freely available data. Thus data usage may require consent from the data subjects as well as the data owner. Sometimes this leads to data not being made available at all, more often, particularly if data collection has been paid for by public funds, the data is made available with charging or other access restrictions. Such information is then protected by what Berners-Lee calls a "walled garden". For teaching datasets, free access to data by both course developers and students is necessary and desirable.
2. Topicality and place of data. These characteristics are more necessary in some datasets than in others. Topicality relates to the year in which the data was collected; place relates to geographical location. In some disciplines such as Engineering, the year in which the dataset is collected and the geographical spread of the data is less relevant; in other disciplinary areas (Social Science, Environmental Science) both characteristics are more important. Topical data will in general be more relevant to students and will help to engage the students more in the research questions being investigated.
3. Documentation of variables and context of the study. Good summary documentation should be available both in providing metadata information (such as the number of cases, a coding book for all variables and what each variable represents) and also the background and context of the study, describing why the data was collected and what research questions it was intended to address.
4. Publications. The dataset should have led to one or more publications which can help both to set the context for the student and to provide background reading.

Apart from these four characteristics, there are other issues to consider. What form can the datasets be downloaded in? Is there any teaching material associated with the data? What forms of analysis does the dataset lend itself to?

SOURCES OF DATASETS

Statistical packages

There are a wide variety of datasets available on the websites of statistical packages. We take two examples. For **R**, a wide variety of datasets are available in the `datasets` library which is provided in the core download; however each extra user contributed package usually has additional datasets associated with it. There is no central indexing of all datasets. For SPSS, a wide variety of datasets is provided, but many are hypothetical and documentation, while available, is hard to find.

Statistical journal webpages

An increasing number of journals provide a datasets archive which can supply datasets which were used in their journal papers. Such datasets are usually open access. The quality of documentation is variable from dataset to dataset. Some datasets are also artificial. As archive examples, the *Journal of the Royal Statistical Society* has provided datasets from 1998 for many of the more applied papers, particularly for papers in Series A and C; the *Journal of the American Statistical Association (JASA)* provides a dataset archive through the `statlib` website, with most datasets in the areas of economics, social health science, engineering, methods in economics and statistical education. The *Journal of Statistical Education* provides a large number of teaching-related datasets. All of the datasets are well documented and they can be downloaded straightforwardly.

Statistics books

Most modern applied statistics textbooks provide datasets as part of the resources of the book. Links to datasets relating to books can either be found on the publisher websites or on sites such as `statlib`. These datasets naturally have good teaching materials associated with them and provide a primary source for teaching. There are also specialist books, such as the "Handbook of Small Datasets" (Hand, 1994).

Dataset websites

There are many websites specialising in providing links to data resources. Some sites also specialize in delivering the datasets themselves. For example, the `statlib` (http://lib.stat.cmu.edu) website provides datasets for a wide number of statistics textbooks as well as a large list of classical datasets. It also hosts the datasets for JASA. Other sites such as http://statsci.org provide links to other sites and resources. In addition, some sites specialise in particular form of data–thus the UCI repository (http://archive.ics.uci.edu/ml/) is a collection of databases that are used for empirical analysis of machine learning algorithms and is useful for data mining.

Data archives

Most countries have one or more national data archives, which act as a repository for government surveys and for other research datasets. For examples, the data archive in the UK is based at Essex University (www.data-archive.ac.uk), and hosts over 5000 datasets collected since the 1970s. The datasets are not open access, and registration at the site is essential. While documentation is excellent, the datasets are often large, and need considerable investment of time. Services such as ESDS (www.esds.ac.uk) are providing teaching versions of large datasets together with teaching materials to go with them. However, such datasets are unfortunately also not open access.

Data and story library websites

One of the most interesting resources available on the web is that of the data and story library. The original concept was constructed by Cornell in the 1990s, with the aim of providing teaching datasets with a story attached (http://lib.stat.cmu.edu/DASL/). While documentation is good, the dataset is embedded in the documentation and may need additional work before use. Specific country versions also exist–for example OzDASL (http://www.statsci.org/data/).

CONCLUSION: THE NEED FOR A DATASETS DISCOVERY RESOURCE

There are now a vast number of datasets freely available on the web. However, it is equally clear that the problem of dataset discovery still exists. Most websites provide restricted lists of datasets, but there is no central repository which can allow datasets to be searched by topic, disciplinary area, etc. While the semantic web may in the future provide a solution, there is a need for a dataset resource which could provide dataset searching to allow suitable datasets to be discovered.

We see such a web resource as providing and delivering both datasets and links to datasets, teaching materials and references to publications and code. Datasets would ideally be delivered in a number of formats. The Data Documentation Initiative (DDI)–an international effort to establish a standard for technical documentation for datasets–could provide a suitable metadata description. Tagging and labelling of the data is crucial, and in particular the issue of the classification of statistical methods needs to be addressed.

While classification schemes of mathematical methods exist, such as the AMS classification (http://www.ams.org/mathscinet/msc/msc2010.html), the statistics classification scheme within it is old-fashioned. There is in our view no suitable existing ontology of statistical methods, and an additional task is to develop an appropriate classification scheme so that datasets can be labelled.

Finally, we would see the resource as collaborative, allowing users to deposit their own datasets or links to the datasets into the database, and to comment and to add additional information to existing datasets. While this might introduce problems of spam, this seems to offer the best way forward to allow such a resource to grow and become useful to a wide number of course developers.

REFERENCES

Berners-Lee, T., & Shadbolt, N. (2009). Put in your postcode, out comes the data. *The Times*, November 18, 2009.

Blalock, H. M. (1972). *Social Statistics.* 2nd ed. New York: McGraw Hill.

Hand, D. (1994). *A handbook of small datasets*. London: Chapman and Hall.

Humphreys, L., & Francis, B. (2009). Developing numeracy in criminology students through crime data. In D. Green (Ed.), *CETL-MSOR Proceedings 2008.* Birmingham: Maths, Stats and OR Network. Online: www.ltsn.gla.ac.uk/repository/CETLMSOR2008_Proceedings.pdf.

Love, T. E., & Hildebrand, D. K. (2002). Statistics Education and the "Making Statistics More Effective in Schools of Business" Conferences. *The American Statistician, 56*(2), 107-112.

Paxton, P. (2006). Dollars and Sense: Convincing Students That They Can Learn and Want to Learn Statistics. *Teaching Sociology*, *34*(1), 65-70.

Singer, J. D., & Willett, J. B. (1990). Improving the Teaching of Applied Statistics. Putting the Data back into Data Analysis. *The American Statistician, 44*(3), 223-230.