



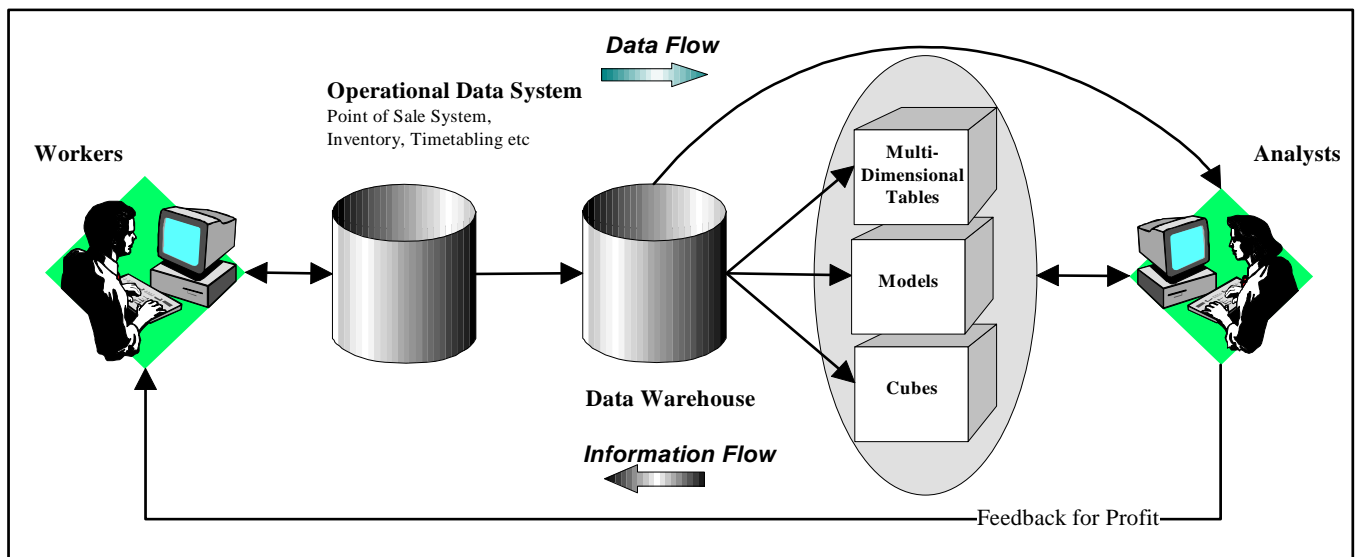
STATISTICS MATHEMATICS INFORMATION  
**NEWSLETTER**

November 1999

*From its foundation, Data Analysis Australia has emphasised the necessity of statistics, mathematics and information technology working together. This issue illustrates some of the benefits for our clients of such a multidisciplinary approach.*

*Dr John Henstridge  
 Managing Director*

## Data Mining - Statistics and Information Technology



Companies are increasingly focusing on the area of data mining to increase profit by searching for the most relevant information. Data Analysis Australia has the cross section of skills essential for effective data mining.

The term data mining was first used by statisticians and was not meant to be complimentary. It referred to unsystematic searching through an often inadequate dataset in order to find *something* that might be significant. This process broke the formal rules that underpinned the statistician's concept of what was and was not significant. This led to claims being made that could not be validated by further data analysis.

More recently computer scientists have used the term 'data mining' to describe methods that scan through large volumes of data to find structures that might not be immediately apparent. The original algorithms were developed for physics applications where the volume of data was enormous and the relationships, once found, were precise. (cont)

## Data Mining - Statistics and Information Technology (cont)

This eliminated statisticians' concerns about significance since the volume of the data and the exactness of a physical relationship meant that random artefacts were unlikely.

Data mining has moved into the commercial world where customer information systems and point of sale equipment collect large volumes of data that can be used to improve marketing.

For example, supermarkets had long known how much of each product they sold but now they can analyse consumer behaviour. Placing certain products together in the supermarket was found to assist in 'impulse buying'.

These new applications of data mining require computer skills and statistical understanding. The relations that are being looked for are no longer exact and consequently the search is for useful relationships rather than exactness.

To do this efficiently requires a number of tools. Modern databases can handle large volumes of data but frequently there is a need to make this more accessible - consequently the development of online analytical processing (OLAP) by many information technology vendors.

OLAP covers the development of data warehouses - repositories of information kept in a form where it can be readily analysed from a variety of angles.

Less well known are computer intensive statistical tools like modern non-parametric regression techniques, which are able to build models with a minimum of assumptions and large volumes of data.

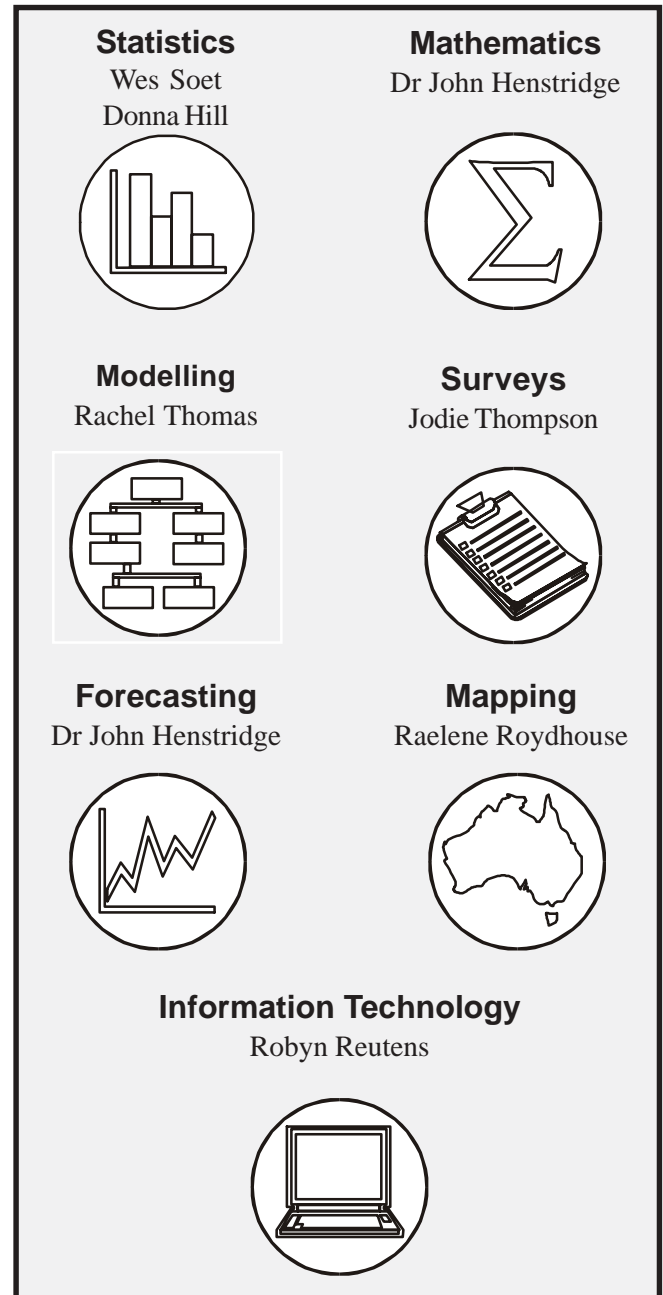
Data Analysis Australia frequently works at the intersection of these two approaches to data mining. The professional staff are a combination of statisticians and computer scientists, but all work at the interface where the aim is to understand data and make it productive for the client's needs.

Data Analysis Australia frequently works with large corporate databases measured in thousands of megabytes, from both a data mining and an information management perspective.

For further information contact **Dr John Henstridge** at Data Analysis Australia.

## Data Analysis Australia - How to describe our company?

A number of Data Analysis Australia's clients have asked the question: 'What do you actually do?'. In response to this, Data Analysis Australia has initiated points of contact for major areas of expertise, to make it easier for clients to refer to their specific areas of interest.



More about these areas is given in Data Analysis Australia's new brochure 'Areas of Expertise', enclosed with the newsletter, which gives a full description of the company's areas of strength and how these areas are applied to diverse organisations.

For further information, please contact staff for each area or reception.

## Making Cyanide Work Faster

Most gold extraction is carried out by a leaching process using cyanide solution. Costs can be reduced by making the leaching process work faster. It is also critical to leach as fully as possible, leaving little gold in the tailings.

When approached by Multi Mix Systems, a supplier of oxygenation equipment aimed at improving the leaching process, Data Analysis Australia developed an automated curve fitting program that provided understanding of leach rates and the grade of gold in the tailings (see figure below for an example of a curve fitting a set of data).

More importantly, the program provided rigorous statistical evaluation of the effects of the equipment.

The program was implemented as a menu driven function in the statistical package S-Plus, giving access to advanced statistical features and graphics. The result is a system that Multi Mix Systems can use to truly demonstrate the effectiveness of their improvements.

The project illustrated the continuing development and application of Data Analysis Australia's software development and statistical sections.

The project was managed in Data Analysis Australia by **Donna Hill** and **Wesley Soet**.

Single Model Fit

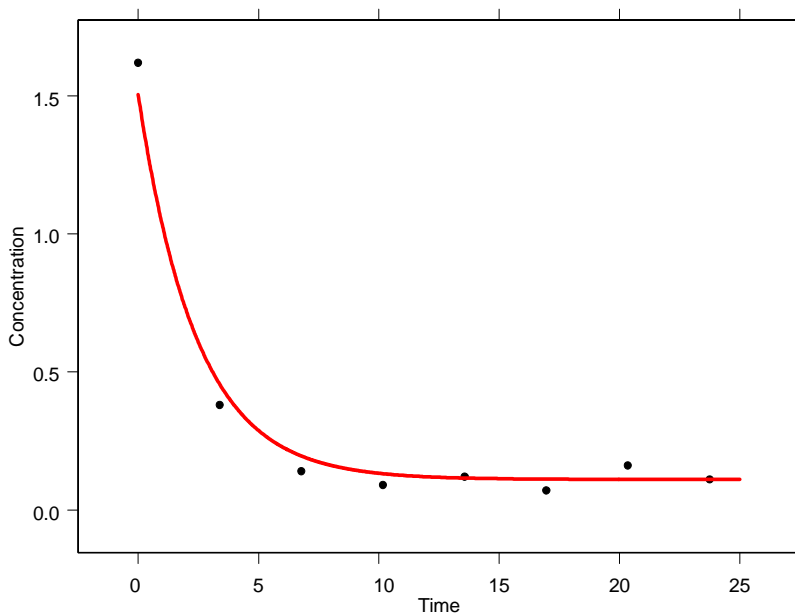


Figure 1: Graph illustrating the 'fit' of a model to a set of results.

## How to?... Understand the Y2K Date Problem

With the Year 2000 (Y2K) almost upon us it is appropriate to consider how dates are best managed in statistical data.



Dates are difficult for a number of reasons. Often the day of the week is just as important as the date itself but this is not apparent when a date is presented.

Unequal length months create the problem, which is compounded by different holidays and numbers of working days and the fact that some areas measure months differently, ie accounting months are four week periods.

Experiences of Data Analysis Australia in this area have led us to frequently work with daily or weekly data rather than aggregating to monthly or quarterly, retaining the integrity of the data.

Computer programs usually store dates either as a text-like file or as a number. For example, Microsoft Excel stores dates by counting the days since 1<sup>st</sup> January 1900 (almost), whilst many database systems descended from dBase use a "yyyymmdd" format. Both of these are Y2K compliant **if used properly**.

Astronomers have a need for accurate dates and have long used the system of Julian dates – the Julian date is the number of days that have passed since noon on January 1st, 4713 B.C. and days begin and end at noon, instead of midnight. This is not related to the Julian calendar (named after Julius Caesar) but rather was named after the Frenchman Julian Scaliger.

It has the advantage that it is well defined independently of any particular software and provides a means of transferring data information between packages.

Data Analysis Australia has updated all hardware and software to ensure the greatest possible Y2K compliance. For Data Analysis Australia's Y2K Statement, please contact **Carole Lefort**.

## Internal Technology and Products

Important resources used by Data Analysis Australia in many projects are the proprietary models and methodologies that have been developed internally. These include:

Detailed population forecasting model for Western Australia. This complements the models of the ABS and the Ministry for Planning, giving, for example, more information on the indigenous population, and allowing consultants to consider complex 'what if' scenarios.

SETI (SocioEconomic Trends Indicators). SETI measures changes in socioeconomic status for each geographical area in Australia.

For further information on these products, please contact **Dr John Henstridge**.

## Classic Quote

The Met Office give a 40% chance of clear skies for the eclipse, but they admit to a 30% chance that they might be wrong.

*Daily Telegraph*  
5 August 1999

## Staff News

**Wesley Soet** joins Data Analysis Australia as a consultant statistician. Wes is currently completing his PhD in mathematics at Curtin University specialising in the modelling of queues. Wesley's initial projects with Data Analysis Australia include the modelling of chemical reaction rates.

Congratulations to **Rachel Thomas** on the completion of her Masters degree in Mathematics from the University of Western Australia. Anyone who sees Rachel's thesis with 130 pages of closely argued mathematics, would recognise the effort in this major achievement.

**Dr John Henstridge** participated in a conference regarding the future of statistics in Australia. The conference was sponsored by The Statistical Society of Australia and was held in Wollongong in New South Wales. In September, John presented a paper at the Australian Transport Research Forum.

Data Analysis Australia would also like to congratulate **Prof. Cheryl Henstridge AM**, who recently became a Member of the Order of Australia. Cheryl, a director of Data Analysis Australia, was honoured for service to mathematics, particularly in the areas of research and education and through professional organisations.



Data Analysis Australia Pty Ltd provides a complete service in the fields of statistics and mathematics - consulting, research, software development and the provision of data. We serve clients from all sectors of industry and government.

**97 Broadway**  
**(P.O. Box 3258, Broadway, Nedlands, 6009)**  
**Nedlands, WA, 6009**  
**Australia**

**Telephone:** (08) 9386 3304  
**Fax:** (08) 9386 3202  
**Web Site:** <http://www.daa.com.au>  
**Email:** [daa@daa.com.au](mailto:daa@daa.com.au)

Newsletters are archived online at <http://www.daa.com.au/newsletters/>