Article

Technology-enabled advance in the worlds of statistics, machine learning and data mining

John H Maindonald*
Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200

Advances in digital computing continue to have large effects on all aspects of life and society, including science. These advances are possible because we have computer languages that translate directly into computational steps that can be implemented in computer hardware. Here, I draw attention to changes that are affecting the theory and practice of data analysis, with a focus on methodologies that feature in expositions of data mining and machine learning. The R language and system is playing an increasingly important role in making the new abilities readily accessible at the scientific workbench.

The computer language revolution

Human language makes possible the rich fabric of human culture, of which mathematics and science are a part. Computer language provides a powerful mechanism for describing computational talks, now with the bonus that talk translates directly into action. These tasks may now, with the software and hardware that is available in 2011, include text processing, mathematical tasks, image and auditory processing, communication, and much else besides. The scientific and mathematical imagination has been stimulated to conceive and carry out tasks of previously unimagined complexity.

Application oriented language

There are huge advantages in working with a language or languages whose terminology closely mirrors what specialists find appropriate when describing a computation. Here, note the language implemented by the R system, which has become the environment of choice for implementing new statistical methodology, and for much else besides. Figure 1 demonstrates the use of R code for plotting and for fitting a regression line. The inset shows R code that gives a simplified version of the figure. Relative to languages such as Fortran, C and Java, R is very high level. Many of the R packages rely heavily on code that has been written in C or Fortran. These earlier languages remain important, but their role has changed.

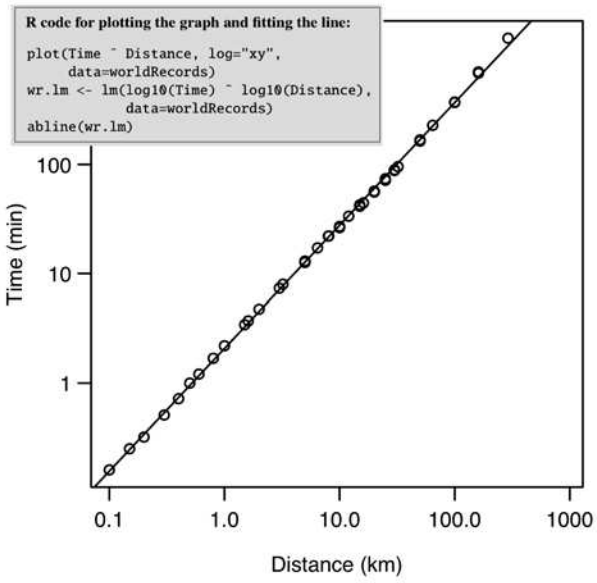
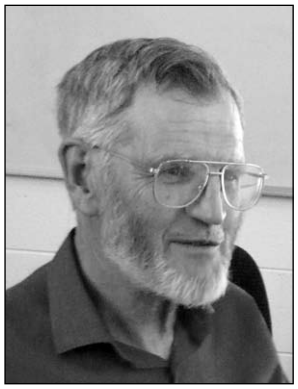


Figure 1. Record times for road and track races, as at August 2006, are plotted against distances. Logarithmic scales have been used on both axes, with equal distances showing a change by a factor of 10. The slope of the line is 1.125, indicating that the relative rate of increase of Time is 1.125 times that for Distance. The data (in the dataset worldRecords) can be made available, from an R session, by installing the DAAG package for R and typing library(DAAG). Code is shown that gives a simplified version of the graph. The supplementary materials include further investigation of these data. (Although the line appears a good fit, two of the points deviate by more than 12% from it. Relative to times that vary by a factor of around 9000, a change by a factor of 1.12 appears slight.)

The R system has many attractions. It is free. It is open source, so that anyone can inspect the underlying code and check that its commands do what they claim. It is readily and

*Correspondence: JHMaindonald@gmail.com



John Maindonald is a Visiting Fellow at Australian National University. He has had wide experience as a quantitative problem solver, working closely with researchers in diverse areas of science, with industrial consulting as a sideline. In 1996, John moved from New Zealand to Australia, then taking a position at ANU in 1998. He is the author of a book on Statistical Computation, and the senior author of a widely used book, now in its third edition, that demonstrates use of the open source R system for data analysis and graphics. Now in semi-retirement, he does occasional consulting, fronts workshops on the R system, and continues to write.

seamlessly extensible, so that its abilities can be the basis for computations that are tailored to the demands of pretty much any area of science or commerce. Witness its use by Google and others for mining web data. Its 3300+ 'packages' extend the base system to give access to an unsurpassed and widening range of abilities. The R Task Views web page (http://cran.r-project.org/web/views/) gives an idea of the range. Extensive tutorial and expository material is available on the internet. There are more than 100 books that expound R, or that describe its use for a particular area of application. A Python specialist could perhaps use Python to illustrate similar points. There is, however, no direct equivalent of the R Task Views web page, suggesting a more limited penetration into field and laboratory bench science.

Abilities that are immediately available in R make it possible to bring together data from different sources, do preliminary checks, extract analysis output in various alternative forms, apply checks to the output, provide graphs and tables, and use output (perhaps combined with other data) for further analysis. Access can be provided, from R, to programs that were initially designed to run as independent programs. Thus, note the R interface to the Weka data mining software (Witten & Frank 2005). Note also the extensive spatial analysis abilities, most of them added since 2003. These have relied heavily on interfaces to other systems, including the widely used GRASS system that must be installed outside of R. Bivand *et al.* (2008) give an overview, now somewhat dated, of what is available. There is currently no R equivalent of the impressive graphical interface to GRASS that the QGIS system (www.qgis.org/) provides.

Point and click interfaces are well suited to some tasks. In general, however, analyses that meet high professional standards will require some use of steps that are spelled out in computer language. The grunts and gestures on which non-human apes rely severely limit what can be communicated. Point and click interfaces overcome these limitations, to some extent, by using language to indicate where and to what end the user can click or type. For use for statistics courses at school and beginning university level, note the R-based iNZight GUI, developed by Christopher Wild and Dineika Chandra at Auckland University (http://www.stat.auckland.ac.nz/~wild/iNZight/).

The focus of the discussion will now narrow somewhat, to consider statistical or (though I have reservations about the term) data mining analyses.

Data analysis challenges

Data, data everywhere

Alongside advances in computing hardware and software there has been, over the past decade, a steady growth in the scope and detail of datasets that are available for scientific use, in large part because of advances in automatic data capture. This is not without problems. The size of the collection does not guarantee quality, or relevance to some particular question, or representativeness. Nonetheless, the massive datasets now available open new vistas, and will be a large part of the future of science.

Genomic data provide much more detailed information about some parts of the genome than about others, much more information about some species than about others. Some of this has to do with ease of collection, some with perceived relevance to questions of biological interest, and some with accidents of circumstance. This matters more for some purposes, less for others. How should the different pieces of evidence be weighted for purposes of taxonomic classification? Is a taxonomic tree the right way to characterise biological relationship. What of horizontal gene transfer? Is there risk that traditional tree-structured classification systems will force the data into an alien mould?

An interesting development, with large potential implications for the handling of data analysis, has been the development of the Kaggle platform (http://www.kaggle.com/) for data prediction competitions. This allows organisations to post their data and have it scrutinised by teams that relish such a challenge. Maindonald (2005) argued for making it standard practice to expose to open scrutiny all datasets that are the basis for scientific claims. The kaggle initiative may be even more effective in serving the same purposes.

Extensive computation, and large datasets

Increasingly, advances in science seem likely to rely on a mix of extensive computation that brings together existing scientific theory in new ways, and the use of very large datasets. The Lytro camera (http://www.lytro.com/), due to come on to the market later this year, demonstrates how well-understood physical processes, combined with the power of modern computation, can be marshalled to create a radical innovation in the marketplace. With this camera, the picture is taken first and focused later, so that the only shutter lag is that due to the human operator. Global Climate Models provide another example. They use extensive computation to account for many different physical processes, different in their importance and in the precision with which their effects can be modelled. They rely on data from many different sources.

New traditions of data analysis

The invention of new names that reflect specific application areas has a long tradition – theory of errors, psychometrics, biometrics, biostatistics, geostatistics, chemometrics, and so on. The word *statistics*, used to describe the theory and methodology that underpins the analysis of data, is perhaps 200 years old. Problems in robotics, in speech and image recognition, and in related areas of engineering have spawned the discipline of *machine learning*. The term *data mining* has come from the computing community. Machine learning, prior to about 1980 on the fringes of Artificial Intelligence, has moved to occupy a central place. It has moved from an initial focus on symbolic logic to use a theory and methodology that are thoroughly statistical. There is nothing in Bishop (2006) that would be markedly out of place in an advanced statistics course. Its traditional focus has been robotics and pattern recognition in an engineering context, but that may be changing.

Expositions of data mining often place emphasis on the methods, or algorithms, that it offers. Its only theoretical basis is that of the statistical theory to which some data mining texts make vague reference. It may be best seen as a name that emphasises the new challenges that arise from the very large datasets that are now presenting themselves for management and analysis. Data mountaineering might now be a better description. See Maindonald (2006).

Features of data analysis challenges

However described, the analysis challenges have common features. I will illustrate with a subset of a dataset that has been widely used for demonstration in statistics and data mining texts.

It relates to glass fragments that were collected in the course of forensic work. Numbers of pieces of glass of each of the glass types that are included in Figure 2 are:

Window float (70)	Window non-float (76)
Headlamps (29)	Containers (13)

Variables are percentages of Na, Mg, . . ., plus refractive index. In all there are 214 rows of data (observations) by 10 columns (variables). The aim is to find a rule that predicts the type of any new piece of glass. Figure 2 is a visual summary of the result from the use of a simple form of classification methodology, with the name *linear discriminant analysis*.

First, two points about the graph:

- It reflects the performance of the methodology for classifying the data used to develop the model. This may lead to an overly favourable view of its performance.
- As there are four groups, there are three dimensions of separation. Separation in the third dimension requires a second graph. In this respect, use of Figure 2 on its own gives an overly unfavourable view of performance.

Ideally, the classification accuracy should be estimated for new data that reflects the context in which results will be used. With data that accumulate over time, historical accuracies for forecasts that were made one year ahead may give a good indication of the accuracy of prediction for the following year. Here, it is necessary to make do with the data that we have, noting the accompanying caveats. With such a small dataset, a split between training and test (and perhaps validation) sets would be a poor use of the available data. Hence the use of cross-validation, which uses repeated splits into training and test data.

A simple version of cross-validation leaves data values out one at a time, fits (trains) the model using the remaining data, and makes a prediction for the omitted point. When the process is complete, predictions are available for all points that are independent of the data for the point. Use of such a *leave-one-out* cross-validation process gives, for the present data, a 70% accuracy. The accuracy is, however, different for the different glass types. Table 1 tells a more complete story. The classification accuracy is highest (86%) for headlamp glass, as Figure 2 might suggest.

Questions, for any use of the results (e.g. to identify glass on a suspect), are:

- How/when were the data generated? (1987)
- Are the samples truly representative of the various categories of glass? (To make this judgement, we need to know how data were obtained.)
- Are they relevant to current forensic use? (Glass manufacturing processes and materials have surely changed since 1987.)
- What are the prior probabilities? (Would one expect to find headlamp glass on the suspect's clothing?)

Table 1. Different accuracies for different types of glass.

Actual	Predicted (cv)			
	WinF	WinNF	Con	Head
WinF	0.71	0.29	0	0
WinNF	0.26	0.67	0.07	0
Con	0	0.46	0.46	0.08
Head	0.03	0.07	0.03	0.86

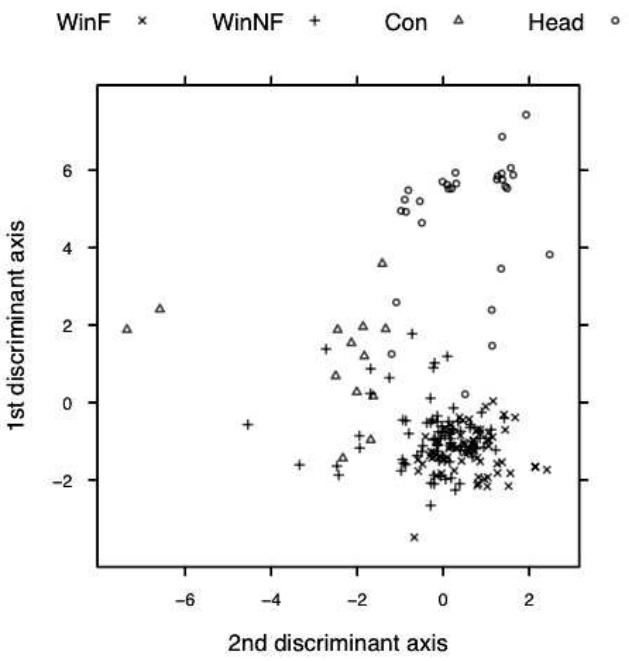


Figure 2. Visual summary of the result when the linear discriminant analysis methodology is applied to a forensic glass dataset, as described in the text. As there are four types of glass, there are three dimensions of separation, and the figure shows the two-dimensional projection that best distinguishes the four types.

These data are probably not a good basis for making judgements about glass fragments found, in 2011, on a suspect's clothing. Too much is likely to have changed since 1987. This 1987 source population is unlikely to be a good match for the glass fragments that one might expect to find now in 2011.

In practice, the only available data may be from a population that is a less than perfect match to the population to which results are to be applied. All available checks should be applied to investigate the closeness of the source/target match. Issues of this sort are crucial once one moves from such engineering applications as robotics where the data that are needed may be generated at the time of use, to an area such as forensic data analysis.

Structures of variation

Data often have a structure. For example, data on mortality rates of patients in critical care might be collected across some hundreds of hospitals. A result that generalises across hospitals must account for variation between hospitals. An algorithm that uses historic data to detect email spam becomes, unless regularly updated, increasingly less effective as time proceeds. In commerce, financial shocks wreak havoc with assessments that are based on pre-shock data. Taleb (2004) makes this point forcefully and at length. Issues of this type are widespread. None of the software I know that has a data mining focus addresses this issue, short of making summary information for each hospital the unit of analysis. The machine learning literature shows some awareness of such issues; cf. Bishop (2006). Books that have statistical learning or data mining in their title, whether written by statisticians or computer scientists, mostly ignore it. The otherwise excellent text by Clarke *et al.* (2009) gives the issue a passing mention that grossly downplays it, then proceeds to ignore it.

Are the new methods better?

Trees, neural nets and Support Vector Machines (SVM) have been the stock-in-trade of Data Mining and Machine Learning, for data such as were used to create Figure 2. Do they do better than the more traditional linear discriminant analysis approach that was used above? Sometimes! Beware though of exaggerated claims, such as have appeared in some of the Support Vector Machine literature. See Ambroise & McLachlan (2002) and Zhu *et al.* (2006).

A generally preferable alternative to leave-one-out cross-validation is *k*-fold validation, where $k = 10$ is a common choice. This splits the data into 10 parts, then leaves out each of the 10 parts in turn, fits the model to the remaining 9 parts, and makes predictions for the omitted data. At the end of the process, predictions are available for all the data. Different splits of the data into 10 parts will give different accuracies. This can be useful, because re-runs of the cross-validation process provide an indication of the statistical uncertainty in the accuracy estimate. R's rpart function for tree-based classification gave accuracies for the forensic glass data that varied between 71% and 76%. Support Vector Machines, used as implemented in the svd function in R's *e1071* package and without any tuning, gave accuracies between 76% and 81%. Note again that these accuracies are for the population from which the original sample was taken. The only obvious continuing relevance of the forensic glass data is to forensic archaeology!

Where there are large numbers of variables, some preliminary variable selection may be needed. As noted in Ambroise & McLachlan (2002), this complicates the estimation of accuracy. The same is true for the tuning that SVM commonly requires to work well. There must be new selection or tuning at each cross-validation fold.

Tree-based classification, which mimics the classification keys that are used by botanists, differs more radically from the mainstream of statistical methods than any of the other methods mentioned. Figure 3 is an example. Splits are optimised over existing nodes, over variables, and (for each variable) the threshold for the split. Because rpart has a built-in procedure that assesses the cross-validation accuracy following each new split, the tree that is formed will vary from one run to another. The tree that is shown had an estimated cross-validation accuracy of 74%.

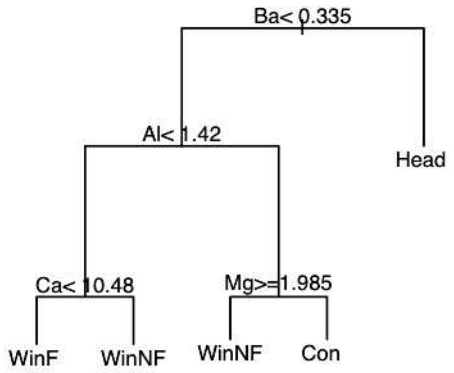


Figure 3. The inverted 'tree' is a visual representation of the classification rule given by R's rpart function (in the rpart package), for the forensic glass dataset. The tree has been pruned back to remove branches whose inclusion, as estimated by cross-validation, reduced classification accuracy. At each node, the left branch is taken if the condition is true, and otherwise the right branch. The tree that is shown gave an estimated cross-validation accuracy of 74%.

Random forests

The random forests method (Breiman 2001) warrants mention for two reasons – as a classification method it is hard to beat, and it introduces some novel ideas. It has relatively recently started to attract attention in the data mining literature. Its disadvantage is that it functions pretty much as a black box. Getting insight into why it delivers its results may not be easy.

When classification trees are formed, each individual split is optimal, given previous splits. The tree that is finally formed may be far from optimal. The random forests methodology aims to overcome this by simulating the taking of repeated random samples from the source population, with a tree formed for each such sample. More than 500 such samples might be taken. The classification is decided by a majority vote over all 500 or more trees.

The effect of taking repeated random samples from the source population is simulated by taking from the source sample repeated bootstrap samples that are of the same size as the source sample. In a bootstrap sample, each sampled observation is put back after it has been taken, so that it is available for selection when the next observation is taken. The end result is that some sample values, on average slightly less than 37% of the total in a large sample, are left out, while the same proportion of those that remain are repeats. For each split of bootstrap sample (called a *bag*), there is also a random sampling of variables – taking the square root of the total number of variables often works well. For each such sample, a tree model is fitted to the in-bag data and predictions are made for the *out-of-bag* (OOB) data. For the forensic glass data, this method gave accuracies in the range 85–87%.

Maindonald & Braun (2010) have an introduction to classification trees and random forests that is aimed at non-specialists.

Which method is best?

Predictive accuracy, as measured by cross-validation, estimates accuracy for the population from which the sample was derived. Differences of a few per cent between different methods are unlikely be of much practical consequence. This is especially true in the common situation where the source population is unlikely to be a very precise match to the target population. It is often hard to get a good handle on the differences that matter for the intended use of results! Comments in Nilsson (2010, p.425) do not go quite far enough:

> *Some methods work better for some problems than for others, but often these differences are only marginal, and most people in the field agree that having lots and lots of data is, in the end, more important than the particular machine learning algorithm used.*

Many analysts will find a choice between linear discriminant analysis and random forests all that they need. Even more important than having lots of data is to have data that are immediately relevant.

Statistical learning methods

Statistical learning methods automate the choice of an optimal model from some suitably large class of models. The random forests method is a good example. It can be extended for use with continuous outcome data also, but is not for this purpose a method of choice.

Consider now Generalised Additive Models. As implemented in R's *mgcv* package, these can fit smooth curves with automatic choice of smoothing parameter. Figure 4 was based on data from 155 sites in the flood plain of the river Meuse in the Netherlands. It shows contours of equal estimated lead contamination, averaged over effects from flooding frequency and soil type, as a smooth function of distance from river and height above river. The methodology does not completely protect against over-fitting, so that checks are desirable. Based on the use of ordinary cross-validation, accuracy is about 14% less than for a model that fits log(zinc) as the sum of smooth functions of elevation and distance, plus effects due to flooding frequency and soil type. Figure 4 may thus be an over-interpretation of the data. See Wood (2006) for extended discussion of the methodology.

Note that the methodology tries to find a fitted surface such that deviations from the surface appear as close as possible to statistical noise. If the residual variation can indeed be reduced to what looks like noise, the fitted surface should be effective for spatial interpolation. If there is remaining spatial pattern, some form of *kriging* may give improved spatial predictions. For an account of kriging as available in R packages, see Bivand *et al.* (2008).

Resampling methods

A feature of the discussion to date has been the heavy reliance on cross-validation or similar assessments of accuracy. These can be used when theoretically based assessments are not available or are of doubtful validity. The role of bootstrap sampling in random forests was noted. In modern statistical methodology, various forms of bootstrap sampling have wide-ranging applications, providing alternatives to methods that rely more strongly on theoretical assumptions.

Finally, note the use of simulation. As the term is used here, this refers to the generation of repeated simulations of data that follow a theoretical model. The model is fitted to each set of simulated data. The results give insight into the distribution of fitted model statistics under the theoretical assumptions. Simulation is sometimes called the parametric bootstrap, reflecting the fact that the resamples are taken from a theoretical distribution rather than (as with the bootstrap) from available sample data. It gives information on the properties of the theoretical model, where cross-validation and bootstrap methodology provide information on the behaviour of the fitted model under repeated sampling.

Careful analysts will use simulation to check out the properties of any methodology that departs from the strict assumptions of the classical theory, as reflected for example in the output from regression software (including R's lm function). The classical theory assumes a single known model. If the model is selected from a wide class of models, or there is extensive variable selection, there may be serious bias in the choice of model and/or the model fit. Figure 5 uses extensive simulation to illustrate the extent of such effects. Data are pure noise; there is no relationship between explanatory variables and the dependent variable. When three variables are taken out of three, the nominal p-values for the three coefficients are spread out around 0.5. The solid line is designed to go through the median of the p-values. Notice that when the number of variables is around 18 or greater, the median nominal p-value will on average be around 0.05. These nominal p-values thus become seriously misleading.

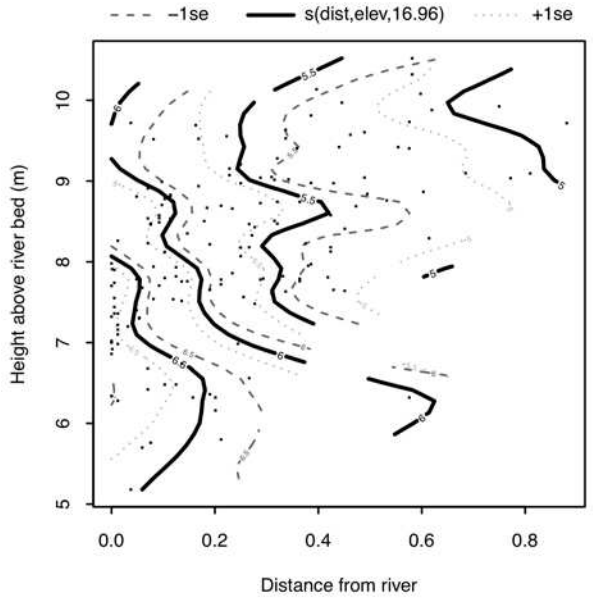


Figure 4. Contours of equal estimated lead contamination in the floodplain of the river Meuse in the Netherlands, averaged over effects from flooding frequency and soil type, as a smooth function of distance from river (scaled to lie between 0 and 1) and height above river. The contours, based on data from 155 sites, were derived using R's gam function, in the package mgcv.

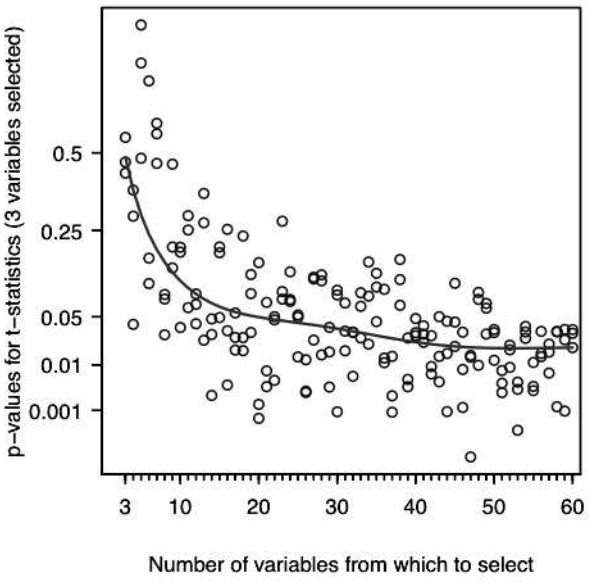


Figure 5. p-values, versus number of variables available for selection, when the 'best' 3 variables were selected by exhaustive search. The fitted line estimates the median p-value.

This illustrates well the light that simulation can shed on a methodology. The backward and forward and other variable selection methods that have been widely used for several decades are well designed to give specious results, unless simulation or another method that achieves the same effect is used to adjust for model selection bias.

The wider world of statistical methodology

The methods that have been described in this paper are a small part of what is available in R. They have been chosen for attention because they are widely used in the data mining and machine learning literature, because it is easy to illustrate their use and usefulness, and because they provide a good context in which to demonstrate the importance of computationally intensive methods. Resampling and other computationally intensive methods have moved into the statistical mainstream, reducing somewhat the former reliance on theory.

While describing those methods, I have tried to give a sense of the power that the high-level commands of the R language puts in the hands of researchers who have the skills needed to use them. There is every reason why scientists whose work involves substantial statistical analysis or other computation should start using R, or something better when it comes along, early in their education. The ideal place to start is at senior secondary school level. There is a wider educational value. Anyone who claims to be well-educated should have some sense of the extent to which advances in science and the technology are a result of the new power that computer language has placed at the fingertips of those who are suitably skilled. I find support for this general view in Bishop (2010). Bishop contrasts Information and Communication Technology (ICT), widely taught in British schools, with Computer Science, using the car as an analogy. ICT, which focuses on spreadsheets and word processing and other such applications, is analogous to learning to drive, while "computer science would be the equivalent of understanding how the engine and other elements of the car work, as well as how to design new cars". Actually there are a large number of places, increasingly important in science, that a driver who knows only spreadsheets and word processing is unable to go.

The history of R

In 2008, Associate Professor Ross Ihaka from the University of Auckland was awarded the Royal Society of New Zealand's Pickering Medal for his work on the development of R, undertaken in collaboration with Robert Gentleman while he also was at the University of Auckland. It implements a dialect of the S language that was developed by John Chambers and others at Bell Laboratories. The introduction to Chambers (2008) has a good summary of the history.

Supplementary materials

The website http://www.maths.anu.edu.au/~johnm/nzsr/taws.html will have links to information and references that are relevant to this paper, including R code for all the graphs, supplementary graphs and calculations, and links to further relevant web pages.

Executables that will install R can be downloaded from http://cran.r-project.org (in New Zealand, use the mirror site http://cran.stat.auckland.ac.nz/). Lillis (2011) has extensive further details on R. For citation, refer to the current version of R Development Core Team (2011).

References

Ambroise, C.; McLachlan, G.J. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS (Proceedings of the National Academy of Sciences of the USA) 99*: 6262–6266.

Bishop, C.M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Bishop, C.M. 2010. Does not compute? *Public Service Review: UK Science and Technology (1)*: 26–27.

Bivand, R.S.; Pebesma, E.J.; Gómez-Rubio, V. 2008. *Applied Spatial Data Analysis with R*. Springer, New York.

Breiman, L. 2001. Random Forests. *Machine Learning 45*: 5–32.

Chambers, J.M. 2008. *Software for Data Analysis: Programming with R.* Springer, New York.

Clarke, B.; Fokoue, E.; Zhang, H.H. 2009. *Principles and Theory for Data Mining and Machine Learning*. Springer, New York.

Lillis, D. 2011. Use R for data analysis and research. *New Zealand Science Review 68*: 73–79.

Maindonald, J.H. 2005. Data, science and new computing technology. *New Zealand Journal of Science 62*: 126–128.

Maindonald, J.H. 2006. Data Mining Methodological Weaknesses and Suggested Fixes. Proceedings of Australasian Data Mining Conference (Aus06), Sydney, 2006. [http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.pdf]

Maindonald, J.H.; Braun, W.J. 2010. *Data Analysis and Graphics Using R – An Example-Based Approach*. 3rd edn, Cambridge University Press, Cambridge, UK. [http://www.maths.anu.edu.au/~johnm/r-book.html]

Nilsson, N.J. 2010. *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* Cambridge University Press, Cambridge, UK.

R Development Core Team, 2011. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Taleb, Naseem. 2004. *Fooled By Randomness: The Hidden Role of Chance in Life and in the Markets*. 2nd edn, Random House, New York.

Witten, I.H.; Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edn, Morgan Kaufmann, San Francisco.

Wood, S.N. 2006. *Generalized Additive Models. An Introduction with R.* Chapman & Hall/CRC.

Zhu, X.; Ambroise, C.; McLachlan, G.J. 2006. Selection bias in working with the top genes in supervised classification of tissue samples. *Statistical Methodology 3*: 29–41.