How Computing has Changed Statistics

(and is changing . . .)

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

ripley@stats.ox.ac.uk
http://www.stats.ox.ac.uk/∼ripley

The Pace of Change

The Pace of Change

Even the meaning of the word *computer* has changed: my predecessor at the University of Strathclyde (Rupert Leslie who retired in 1983) still used ‘computer’ to refer to a human being.

Someone like

The Pace of Change

David Cox's professional life started when programmable electronic computers were still in research laboratories (and military establishments). I gather it was not until the 1960's that (a few) statisticians got their 'own' computers — e.g. Biometry in Oxford acquired one in 1962 (and was the first Oxford unit to do so).

One of the most famous of all failures of foresight is the quoted remark of Thomas Watson, then Chairman of IBM, in 1943 that

'I think there is a world market for maybe five computers'

(I am sure *he* meant machines.) If you want to follow that up, make use of one of the revolutionary changes wrought by computers: *Google* it.

But as we are statisticians we can be quantitative.

Moore's 'Law'

Gerald Moore[1] made a statement that has become folk-law. In the words of a glossy Intel flyer of the late 1980s

> 'Moore's Law was driving a doubling of computing performance every 18 months.'

This almost violates Stigler's Law of Eponomy:[2] Moore did say something in 1965, but it was not that.

What Moore said was that number of transistors on an integrated circuit will increase exponentially fast, and he projected to 1975 at the rate of a doubling each year.

In 1975 he amended the prediction to a doubling every two years.

By 1997 he thought it would hold good for another 20 years (and then hit physics-based limits).

---

[1] one of the founders of Intel.

[2] Trans NY Acad Sci 1980 – *nothing is due to the person it is named after*

Does Moore's 'Law' fit the data?

First collect some data . . . .

When I moved to Oxford in 1990 my home computer was a 25MHz Sun workstation with 12Mb of RAM and 400Mb of disc space.

My current home computer bought for Christmas 2003 has a 2.6GHz processor with 1Gb RAM and 160Gb of disc space (at a quarter of the price). [Note that in the last seven months the predicted change is 30%.]

The folk version of Moore's Law predicts a 400-fold increase.

The processor speed has increased less than that, but what the processor can do per clock cycle has increased, to several integer operations and a couple of floating-point ones.

A more extreme test

How would Moore's 'Law' predict computing speeds have changed over the length (so far) of David's career. A factor of 10^{12}, a *billion* in English back then.

If we have 4GHz chips now, that would be 1/250Hz back then. That's not too far off: using a Brunsviga for a single floating-point multiply and add would take tens of seconds (but I gather they used fixed point wherever possible).

Computers have got relentlessly faster and cheaper, so much so that we can each have (at least) several[3] of them, and we can expect that pace of change to continue for the foreseeable future.

[3] one of my colleagues has five for personal use and 110 in a compute cluster.

How have Statisticians Reacted?

- Most of us use a PC several times a working day. Maybe even David.
- We have become more impatient and expect an instant answer.
- With the possibility of a near-instant response we try out many more ideas when faced with a set of data.

  There is a tendency to replace thinking by trying, and sensible conclusions are not reached all that much faster.

  At least, that is our experience with teaching *applied statistics* to MSc students.
- We now teach practicals on parts of statistics that were inaccessible for lack of software and computing power, e.g. time series and multivariate analysis.

How have Statisticians Reacted?

Another aspect is the amount of storage available.

John Tukey developed EDA (1977) essentially for hand calculation.
JWT was a ‘early adopter’ of computing, so I once asked him why.
I was told he only had his HP calculator with him on plane trips to consulting jobs.

As all plane travellers know, businessmen have for a decade or so been crunching their spreadsheets *en route* to sales opportunities.

I think it was in 1998 I first encountered people who carried their ‘life’s work’ around in a small pack of CD-ROMs.

Today we have DVDs holding 4.5Gb, and my career’s statistics work, including all the data, fits on a couple of those.

We think of writing a book as a large amount of work but mine amount to just 1–2Mb each. The whole DRC canon will probably fit on a key-fob.

What can we do with all that power?

‘Computer-Intensive Statistics’

One sense of “computer-intensive” statistics is just statistical methodology which makes use of a large amount of computer time — examples include the bootstrap, smoothing, image analysis and many uses of the ‘EM algorithm’.

The term is usually used for methods which go beyond the minimum of calculations needed for an illuminating analysis.

- Working with (much) larger datasets.
- Using more realistic models and better ways to fit models.
- Exploring a (much) larger class of models.
- Attempting a more realistic analysis of existing simple models.
- Better visualization of data or fitted models or their combination.

Data Mining

Data mining is currently a popular term for exploring large datasets, although one of my favourite quotes is

> Cynics, looking wryly at the explosion of commercial interest (and hype) in this area, equate data mining to statistics plus marketing.

(from Witten & Franke, 2000).

‘Large Datasets’

What is ‘large’ about large datasets as used in data mining?
Normally just one of two aspects

- Many cases
  - motor insurance database with 66 million drivers (about 1/3 of all US drivers).
  - Sales data from Amazon, or an airline.
  - Credit-card transactions.
- Many observations
  - screening 10,000+ genes.
  - fMRI maps of t statistics for 100,000 voxels (per session, with less than 100 sessions).

An unusual example which has both is so-called CRM, e.g. supermarket sales records. Note the predominance of discrete observations.

However, many datasets are already close to the maximal possible size.

In 1993 Patty Solomon and I worked on a dataset of all 2,843 pre-1990 AIDs patients in Australia. That was a large dataset in survival analysis then, and it still is.

Over the last three years my D.Phil. student Fei Chen has been looking at data mining in the insurance industry—motor insurance actuaries already have databases of 66 million motor insurance proposals, some one third of all drivers in the USA. There are around 30 items for each driver, and that is not going to increase much as potential customers will not be prepared to answer more questions (and the more questions they are asked the less reliable their answers will become).

There *are* fields in which there is the potential to collect substantially more data on common activities.

- So-called *Customer Relationship Management* uses loyalty cards to track the shopping habits by individual customer in, for example, supermarkets.
- Fei is now employed to do data mining for fraud detection, looking for unusual patterns of activity in, say, credit-card transactions.

But even these fields have limits that are not so far off given the changes predicted by Moore’s law, and it seems that working with all the available data will be the norm in almost all fields within a few years.

Computational Complexity

In the theory of computational complexity an exponential growth is regarded as very bad indeed, and most lauded algorithms run in at most polynomial time in the size of the problem (in some suitable sense).

When the available resources are growing exponentially the constants do matter in determining for polynomial-time algorithms when they will become feasible, and for exponential algorithms if they ever will.

Complexity of Linear Regression

Consider linear regression, with n cases and p regressors. The time taken for the least-squares fitting problem is $O(np^2)$.

For a fixed set of explanatory variables this is linear in n. If we had the ability to collect a large amount of data, how large should we take n?

About 1997 some software packages started to boast of their ability to fit regressions with at least 10,000 cases, and Bill Venables and I discussed one evening if we had ever seen a regression that large [no] and if we ever would.

We 'know' that theory suggests that the uncertainty in the estimated coefficients goes down at rate $O(1/\sqrt{n})$.

It is easy to overlook the conditions attached. The most important are probably

- The data are collected by a process close to independent identically distributed sampling and
- The data were actually generated by the linear regression being fitted for some unknown set of coefficients.

Neither of these is likely to be realistic.

Large datasets are rarely homogeneous and often include identifiable subgroups that might be expected to behave differently. A formal analysis might well make use of *mixed models*, models including random effects for different subgroups. These require orders of magnitude more computation, and under reasonable assumptions may be quadratic in the number of groups. (There are examples from educational testing which probably exceed current software's limits.)

A failure of the second assumption will lead to systematic errors in prediction from the model, and it is very likely that systematic errors will dwarf random errors before n reaches 10,000. As another famous quotation goes

All models are false, but some are useful

(G. E. P. Box, 1976) and as n increases the less falsehood we will tolerate for a model to be useful. So for large n we find ourselves adding more explanatory variables, adding interactions, non-linear terms or even applying non-linear regression models such as neural networks.

It seems that in practice p increases at roughly the same rate as n so we really have a $O(n^3)$ methodology.

On the other hand, the number of possible submodels of a regression model with p regressors is $2^p - 1$, so exhaustive search remains prohibitive as a method of model selection for p above about 70.

Statistical Software

The Role of Software

It is not really the change in computational speed as predicted by Moore's Law that has affected the way we do things.

If computers were still accessed by the interfaces of the 1960's they would be the preserve of a specialist cadre of operators/programmers, and it is the software which they run which has made computers so widely acceptable.

Point-and-click interfaces are now so commonplace that we encounter graduate students who have never seen a command line.

Statistical Software

It is statistical software that has revolutionized the way we approach data analysis, replacing the calculators used by earlier generations.

Remember that data analysis is not done only, or even mainly, by statisticians and spreadsheets (notably Excel) are almost certainly the dominant tools in data analysis.

Software has often been an access barrier to statistical methods. Many times over the years I would have liked to try something out or give my students a chance to try a method out, and have been frustrated by the inaccessibility of software—for reasons of expense, usage conditions or machine requirements.

There is an old adage (from the days of VisiCalc) that runs

> *one should choose one's hardware and operating system to run the software one needs*

but many users do not have that degree of choice.

Part of an Advertisement

'I'm a LECTURER IN STATISTICS – responsible for ensuring that good statistical practise becomes the norm with new generations of analysts.

Which is why I chose Xxxxx.'

- This does seems an ambitious goal for one lecturer or one piece of software.
- Software is important, but teaching the right mix of methodology and how to use it well is far more important.
- Xxxxx describes itself as 'a cutting-edge statistical software package'. One of the most difficult tasks in training the data analysts of the future is predicting what it will be important for them to know. Having software available biases that choice.

Standards – Real and Imaginary

Beware proprietary ‘standards’. People claim Microsoft Office is ‘standard’.

- How do we know that different versions of Excel behave in the same way?
- Do they behave the same way on Windows and MacOS?
- How do we even know how they are intended to behave?
- What about the ‘clones’ such as StarOffice / OpenOffice?

At least some things are now standard. Thanks to IEC60559 (also, incorrectly, known as IEEE754) we can reasonably assume that computer arithmetic will work to the same precision and more-or-less the same way everywhere.

We can hope the following will never be emulated:

> Whilst I was at Imperial College, UCL Computer Centre put out an announcement that a bug had been discovered in their systems' floating point unit and
>
> *'any important numerical work should be repeated on some other machine'*.

But what about the implementation of `arcsin` or `pnorm` . . . ?

Is My Statistical Software Reliable?

It probably is the case that (by far) most incorrect statistical analyses result from user error rather than incorrect software, but the latter is often not even considered. The issue was highlighted in 2002 by a controversy over the link between 'sooty air pollution' and higher death rates, which made the *New York Times*.

Here is the Johns Hopkins' press release:

> Their work appeared in the *New England Journal of Medicine* and other peer-reviewed publications. While updating and expanding their work, the investigators recognized a limitation of the S-plus statistical program used to analyze data for the study. The S-plus program is standard statistical software used by many researchers around the world for time-series and other analyses. The Hopkins investigators determined that the default criteria in one regression program used for examining patterns and fitting the statistical model (referred to as *convergence criteria*) were too lenient for this application, resulting in an upward bias in the estimated effect of air pollution on mortality.

A better summary, courtesy of Bert Gunter, then a senior statistician at Merck:

> *Data analysis is a tricky business—a trickier business than even tricky data analysts sometimes think.*

This was a case of users blaming their tools with only a little cause (and the need to change this default is in a certain well-known[4] book I co-author). But all credit to them for actually checking.

But what if the software really had been faulty?

[4] maybe even well-read

Nowadays we rely on the ability to fit a Poisson log-linear model accurately as much as we rely on our calculators' ability to multiply.

I suspect few of us will have consulted a book of statistical tables in the last year, instead using the equivalents built into statistical packages.

Beware: they are found wanting alarmingly frequently.

The issue raised is trust in software, which is not peer-reviewed in general, and may well be understood by no one.

‘Open Source’ and ‘Free’ Software

These are emotive terms, coined by zealots.

Richard Stallman’s (RMS) Free Software Foundation is ‘free as in speech, not free as in beer’. The GNU project was set up to provide a ‘free’ Unix but made slow progress. In the early 1990s Linus Torvalds came along with the missing piece, a kernel, and *Linux* was born. However, well over half a ‘Linux distribution’ is from GNU, and RMS and others (e.g. the Debian distribution) refer to GNU-Linux.

There are other free licences (X, BSD, Apache, Artistic, . . .), and the term ‘Open Source’ was coined for the concept, with a legalistic definition.

The freedom to know how things work may be equally important.

The R Project – Open Source Statistics

R is an Open Source statistics project. It may not be nirvana, and it may not be suitable for everyone, but it is an conscious attempt to provide a high-quality environment for leading-edge statistics which is available to everyone.

It is free even 'as in beer'. You can download the source code (at `www.r-project.org`), as well as binary versions. People in the Third World can afford it and teach with it.

The only barrier to understanding how it works, precisely, is skill.

Case Studies

- Classification trees – CART
- Characterizing Alzheimer’s Disease

Classification Trees — CART

Classification trees is one area which illustrates the importance of software.

They have been (fairly) independently developed in machine learning, electrical engineering and statistics from the mid 70s to the end of the 80s.

Classification and Regression Trees by Breiman, Friedman, Olshen & Stone (1984) was a seminal account. Unusually for statisticians, they marketed their software, CART®.

The other communities also marketed their software. Ross Quinlan even wrote a book about his, *C4.5: Programs for Machine Learning*, containing the source code *but not allowing* readers to use it. The C code could be bought separately, for restricted[5] use.

The net effect is that classification trees did not enter the mainstream of statistical methodology. Neither CART nor C4.5 had a user-friendly interface.

---

[5] 'may not be used for commercial purposes or gain'

Classification Trees — in S

The advent of classification and regression trees in S in 1991 made the technique much more accessible.

Unfortunately the implementation was bug-ridden.

Eventually I decided to write my own implementation to try to find out what the correct answers were.

Terry Therneau had re-implemented CART (the book) during his Ph.D. and his code formed the basis of `rpart`.

Classification Trees — Lessons

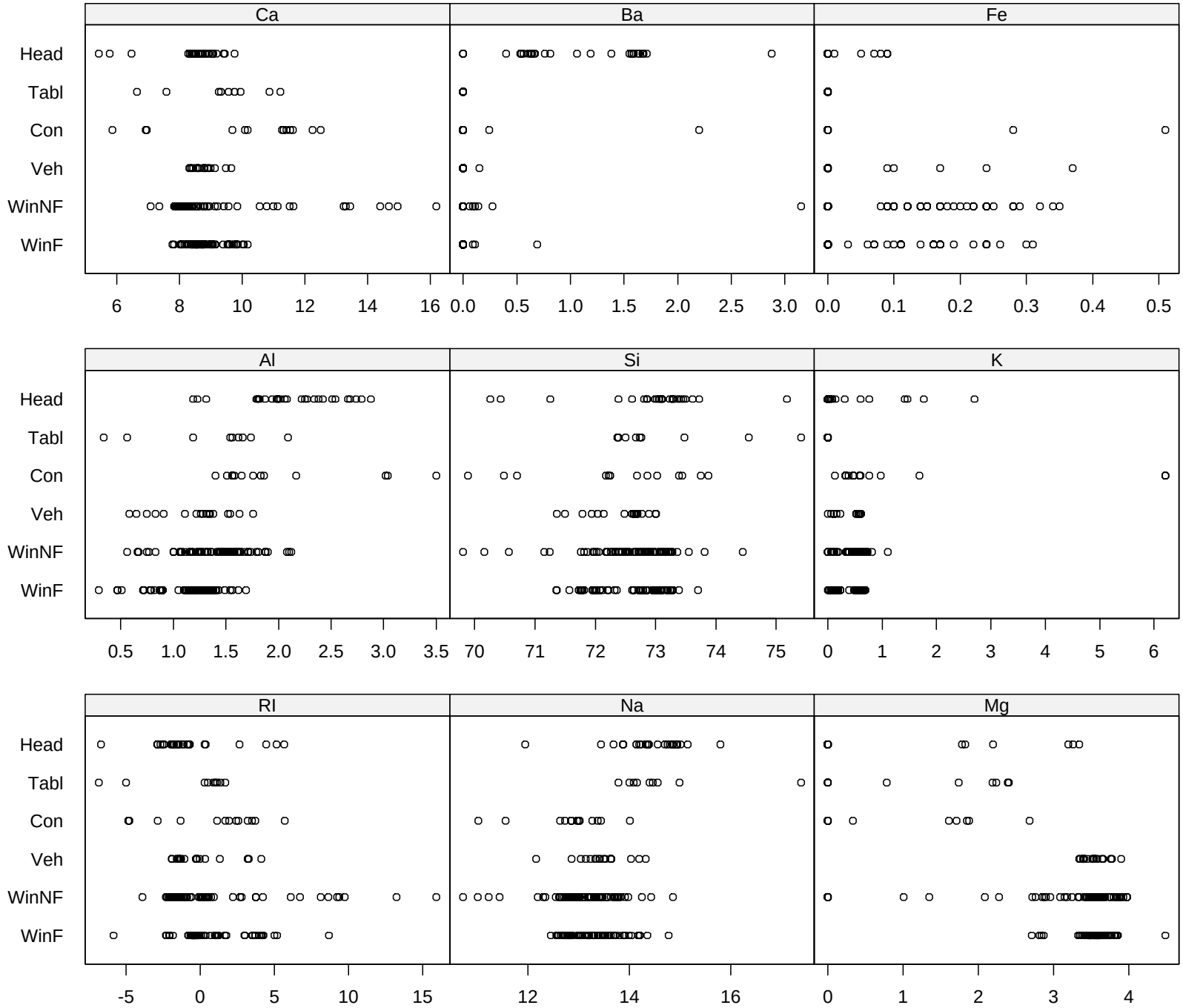
- Having the source code available makes it *much* easier to find out what is actually done.
- Having independent open implementations increases confidence in each.
- People keep on reporting discrepancies between the implementations. Almost inevitably these are not using comparable ‘tuning’ parameters, and people never appreciate how important these are.

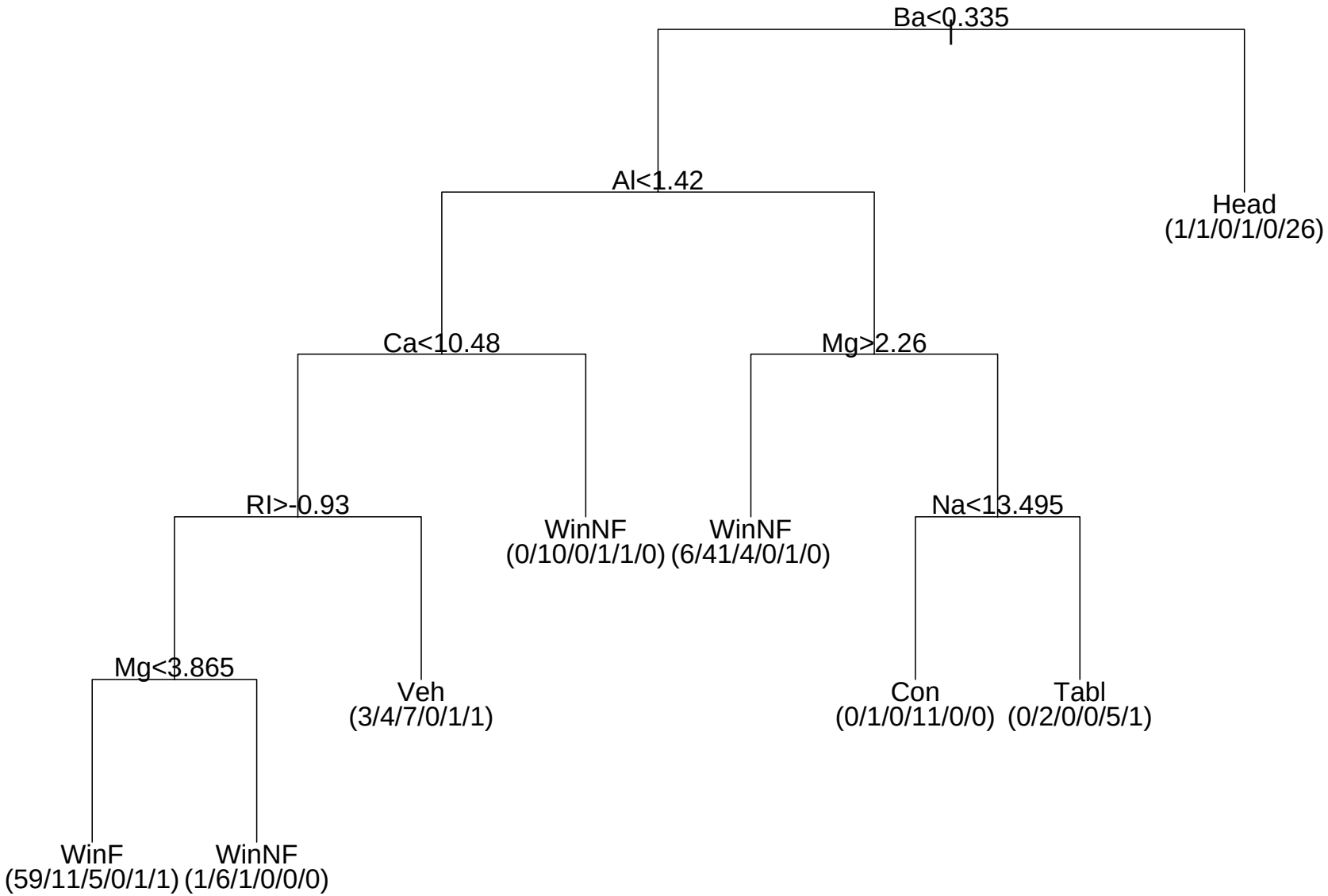
Classification Trees — Example

This dataset has 10 measurements on 214 fragments of glass from forensic testing, the measurements being of the refractive index and composition (percent weight of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe). The fragments have been classified by six sources.

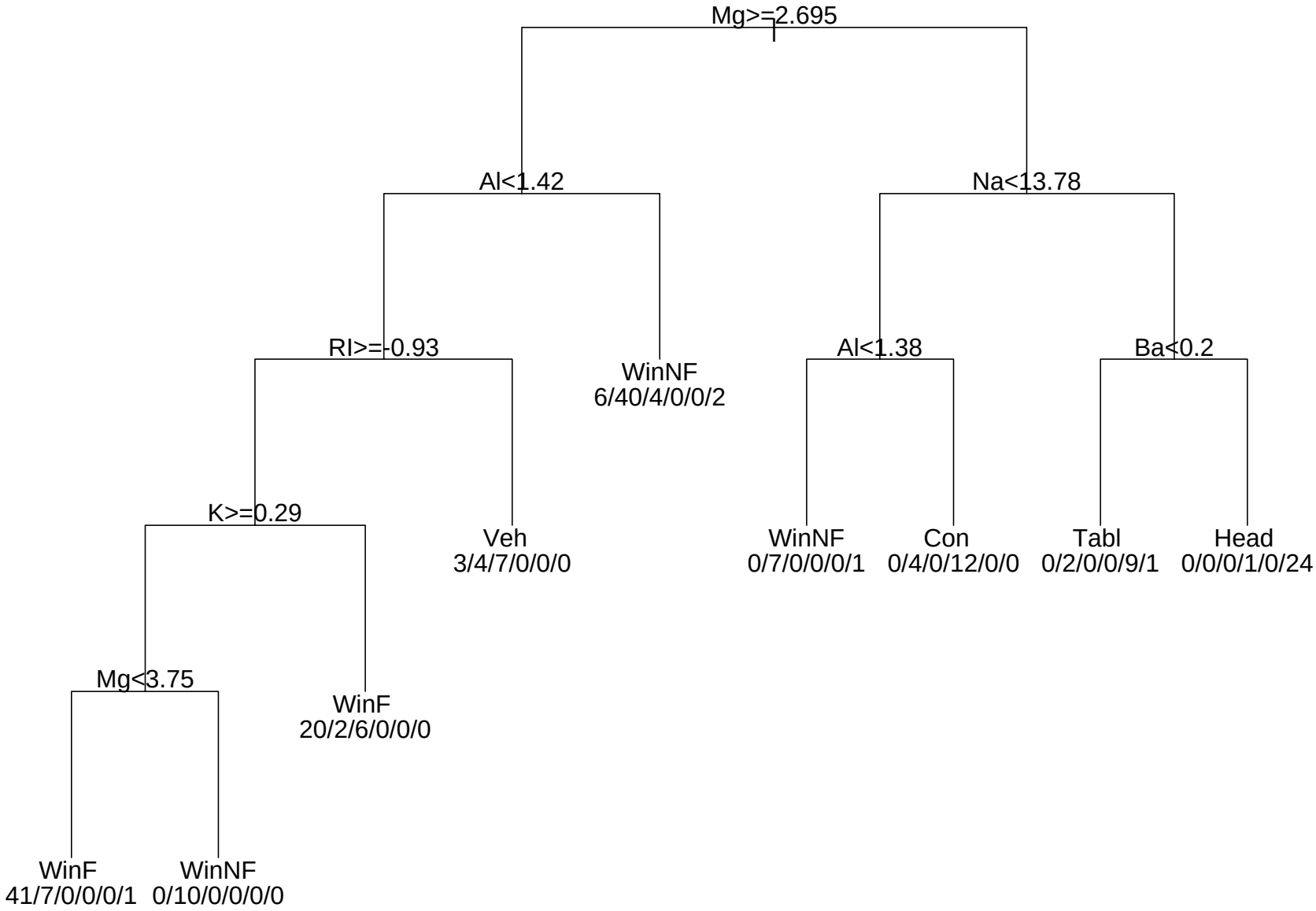
This data set is hard to visualize.

Examples are from `rpart`.


Ca
Ba
Fe
Al
Si
K
RI
Na
Mg
Head
Tabl
Con
Veh
WinNF
WinF
6 8 10 12 14 16
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.0 0.1 0.2 0.3 0.4 0.5
0.5 1.0 1.5 2.0 2.5 3.0 3.5
70 71 72 73 74 75
0 1 2 3 4 5 6
-5 0 5 10 15
12 14 16
0 1 2 3 4



Classification tree using information index



Classification tree using Gini index

Characterizing Alzheimer's Disease

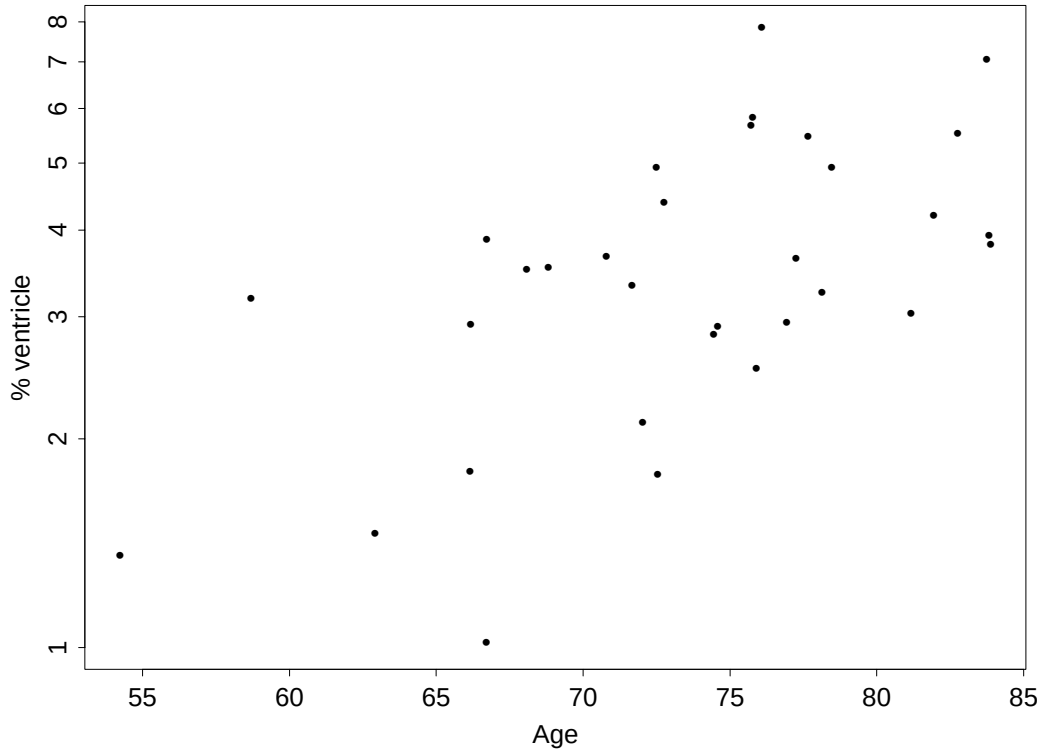
Joint work with Kevin Bradley, Radiologist at OPTIMA (Oxford Project to Investigate Memory and Ageing).

Published in *British Journal of Radiology*.

Structural MRI of Ageing and Dementia

Everyone's brain shrinks with age (0.4% per year), and not uniformly.

Disease processes, for example Alzheimer's Disease (AD), change both the overall rate and the differences in rates in different parts of the brain.



Use serial structural MRI, probably of two measurements n months apart.

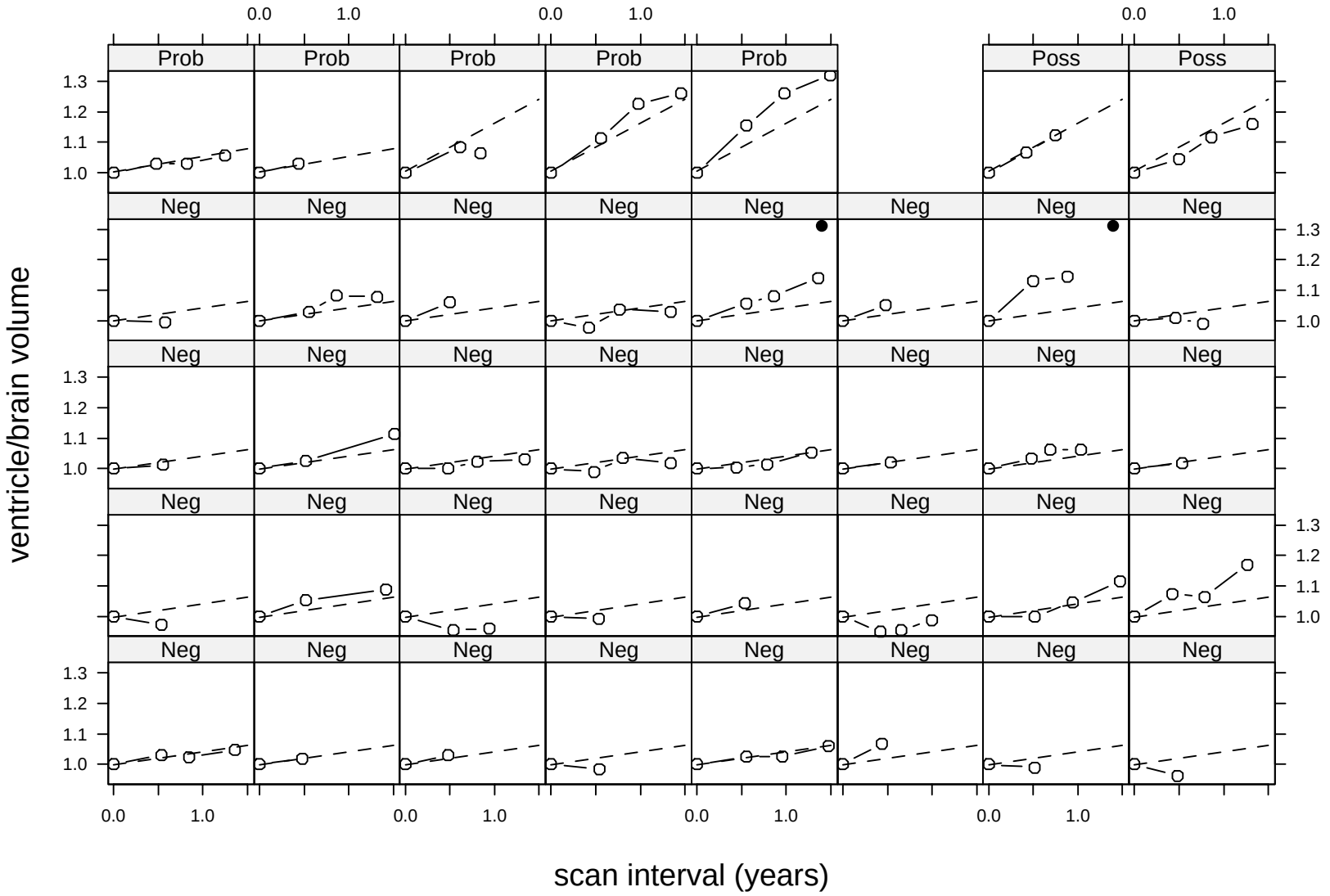
How large should n be?

How many patients are needed? (Parallel study by Fox *et al*, 2000, *Archives of Neurology*.)

Study with 39 subjects, most imaged 3 or 4 times over up to 15 months.

Three groups, ‘normal’ (32), ‘possible’ (2) and ‘probable (5).

Given the ages, expect a substantial fraction of ‘normals’ to have pre-clinical AD.


Prob
Poss
Neg
ventricle/brain volume
scan interval (years)
0.0
1.0
1.0
1.1
1.2
1.3

Statistical Analysis

Major source of variation is between subjects. Not many ‘abnormals’, and usually the diseased group is more variable than the normals.

Choose to use linear mixed-effects models (NLME of Pinheiro & Bates).

- The Trellis plot here really helps in visualizing the data.
- Longitudinal data like this are common, and here subject-specific random effects really help.
- Given the estimates of the variance components, we can answer the questions of ‘how far apart?’ and ‘how many patients?’.

Conclusions

- Better statistical computing allows analyses not dreamt of even a decade ago.
- It's not just more powerful computers and bigger datasets.
- Finding ways to visualize datasets can be as important as ways to analyse them.
- We can use more realistic models and better ways to fit them.
- Software availability now drives what we do, probably much more than we consciously realize.

The End