Earlier Days of Computer Classification, by Gavin Ross

I welcome this opportunity to congratulate Fionn on his appointment here and his excellent and informative inaugural lecture.

He has talked about massive data sets. He has asked me to reminisce about the early days of computer classification.

I joined John Gower in the Statistics Department at Rothamsted Experimental Station in 1961, when Frank Yates was head of department. The department had acquired the first electronic computer used in civil research, the Elliott 401, which had been lying idle in Cambridge in 1954. The remit was to use the computer in whatever way we thought fit, as an aid to statistical computing and research.

Peter Sneath in Leicester had published a paper on the use of computers in bacterial classification, and Margaret Pleasance at the Low Temperature Research Station in Cambridge asked John Gower if the methods described could be programmed for the Elliott 401. Gower then wrote a suite of programs to read bacterial test data, to compute a similarity coefficient between each pair of strains, to perform single linkage cluster analysis, to compute mean similarities between and within user-defined groups, and to identify the most typical elements of each group.

The 401 was a valve machine, with paper tape input and output, and could do 10 divisions per second, with multiplication a little faster. It occupied a large room, and had about 2000 words of backing store. By judicious programming, packing information into bits within words, using logical operations to identify matching bits, and rewriting the similarity matrix on top of the data matrix, a maximum of 128 strains could be classified, the whole exercise taking hours rather than seconds.

With this program up and running we had several requests for classification jobs, mainly agricultural and biological at first, such as classification of nematode worms, bacterial strains, and soil profiles. On this machine and its faster successor, the Ferranti Orion, we performed numerous jobs, for archaeologists, linguists, medical research laboratories, the Natural History Museum, ecologists, and even the Civil Service Department.

On the Orion we could handle 600 units and 400 properties per unit, and we programmed several alternative methods of classification, ordination and identification, and graphical displays of the minimum spanning tree, dendrograms and data plots. My colleague Roger Payne developed a suite of identification programs which was used to form a massive key to yeast strains.

The world of conventional multivariate statistics did not at first know how to view cluster analysis. Classical discriminant analysis assumed random samples from multivariate normal populations. Cluster analysis mixed discrete and continuous variables, was clearly not randomly sampled, and formed non-overlapping groups where multivariate normal populations would always overlap. Nor was the choice of variables independent of the resulting classification, as Sneath had originally hoped, in the sense that if one performed enough tests on bacterial strains the proportion of

matching results between two strains would reflect the proportion of common genetic information. But we and our collaborators learnt a lot from these early endeavours.

John Gower was amused to read in Peter Wright's Spycatcher (1985) that the secret services had discovered a new and powerful technique for identifying wrongdoers in society: it was called Cluster Analysis!

Now we can compute in milliseconds what we could not compute in hours in 1960, and our storage capacity has increased by millions. But the human brain has still the same capacity to examine and understand information.

I welcome Fionn's contributions to this endeavour.

Gavin Ross
Rothamsted Research
February 2007