

# Visualization for Classification and Clustering

Brian D. Ripley

*Professor of Applied Statistics  
University of Oxford*

ripley@stats.ox.ac.uk  
<http://www.stats.ox.ac.uk/~ripley>

## Needles in Haystacks

The analogy is rather helpful. We probably know what the ‘hay’ looks like. If we really know what ‘needles’ look like, the task is purely computational pattern matching. But in practice we either

- have seen some past examples of needles: **supervised** pattern recognition, or
- are interested in anything which is not hay: **unsupervised** pattern recognition.

or just possibly both. The second is much harder, and really we are only interested in some departures (e.g. in fraud detection in those which may lose us income).

In the examples of this talk we will know about some patterns, but be interested to see if we can find others.

## ‘Data Mining’

Challenge is to explore data in many dimensions, often with thousands to millions of cases.

*Classification* has two senses:

- ‘to arrange in classes or categories’
- ‘assign (a thing) to a class or category’

and data mining is concerned with both but especially the first. It is a re-emphasised version of what has been discussed in statistics as *multivariate analysis* and *pattern recognition*.

Fifteen years ago *data mining* was a pejorative phrase amongst statisticians, but the English language evolves and that sense is now encapsulated in the phrase *data dredging*.

## A Forensic Example

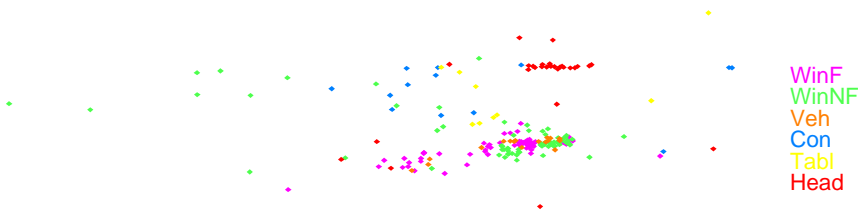
Data on 214 fragments of glass collected at scenes of crimes. Each has a measured refractive index and composition (weight percent of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe).

Grouped as window float glass (70), window non-float glass (76), vehicle window glass (17) and other (containers, tableware, headlamps) (22).

## Statistical Data Mining

We will always need to bear in mind the ‘data dredging’ aspect of the term. When (literally) mining or dredging, the proportion of good material to dross is usually very low, and when mining for minerals can often be too low to cover the costs of extraction.

Exactly the same issues occur in looking for structure in ‘small’ data: it is all too easy to find structure that is only characteristic of the particular set of data to hand. We want *generalization* in the terminology of the psychologists, that is to find structure that will help with future examples too.



MDS representation of the glass fragments.  
Note there appear to groupings beyond the *a priori* specified ones.

## ‘Large Databases’

What is ‘large’ about large databases 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 SPM map of  $t$  statistics for 100,000 voxels (per session, with less than 100 sessions).

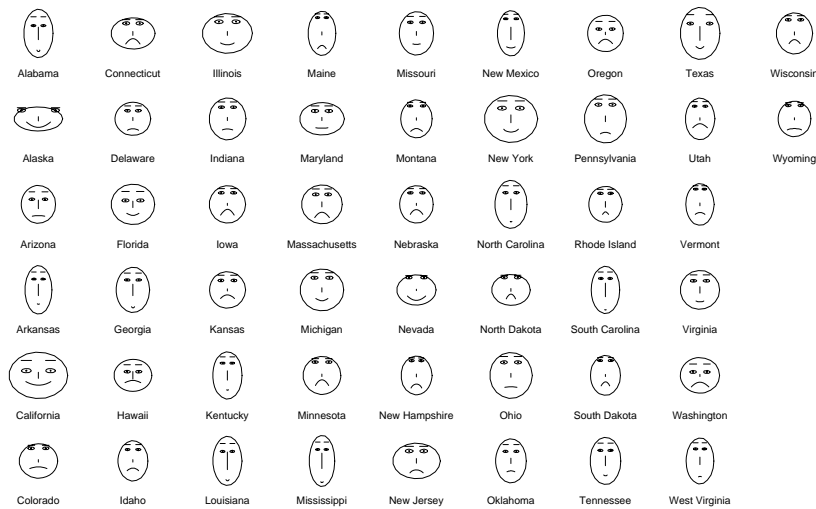
An unusual example which has both is so-called CRM, e.g. supermarket sales records.

## Visualization

Challenge is to explore data in more than two or perhaps three dimensions. The talk is almost entirely about continuous data, but visualization for discrete data is also important.

via glyph representations

There are many ways to represent each case by a small diagram, of which Chernoff’s faces are the most (in)famous. Only viable for modest numbers of cases and variables. Similarly for parallel coordinate plots.



## via projections

Principal components is the most obvious technique:  $k$ D projection of data with largest variance matrix (in several senses). Usually ‘shear’ the view to give uncorrelated axes.

Lots of other projections looking for ‘interesting’ views, for example groupings, outliers, clumping. Known as (exploratory) *projection pursuit*.

‘Random’ searching (so-called *grand tours*) are not viable even in 5D.

Examples from GGobi (<http://www.ggobi.org>).

## via squeezing

Multidimensional scaling aims to represent distances between points well. Think of a point cloud in  $p$ D connected by springs, being squeezed down to  $k$ D for  $k = 1, 2, 3$ .

The ‘distances’ can be a *dissimilarity* matrix, and so derived from discrete or continuous data, e.g. from proportion of object types in common in neolithic graves.

Classical MDS plots the first  $k$  principal components, and minimizes

$$\sum_{i \neq j} d_{ij}^2 - \tilde{d}_{ij}^2$$

where  $(\tilde{d}_{ij})$  are the Euclidean distances in the  $k$ D space.

More interested in getting small distances right. Sammon (1969) proposed

$$\min E(d, \tilde{d}) = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \tilde{d}_{ij})^2}{d_{ij}}$$

Shepard and Kruskal (1962–4) proposed only to preserve the ordering of distances, minimizing

$$STRESS^2 = \frac{\sum_{i \neq j} [\theta(d_{ij}) - \tilde{d}_{ij}]^2}{\sum_{i \neq j} \tilde{d}_{ij}^2}$$

over both the configuration of points and an increasing function  $\theta$ .

The optimization task is quite difficult and this can be slow. There will be multiple local minima.

## Leptograpsus variegatus Crabs

200 crabs from Western Australia. Two colour forms, blue and orange; collected 50 of each form of each sex. Are the colour forms species?

Measurements of carapace (shell) length CL and width CW, the size of the frontal lobe FL, rear width RW and body depth BD.

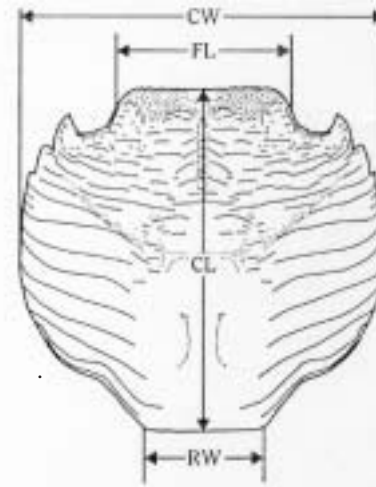
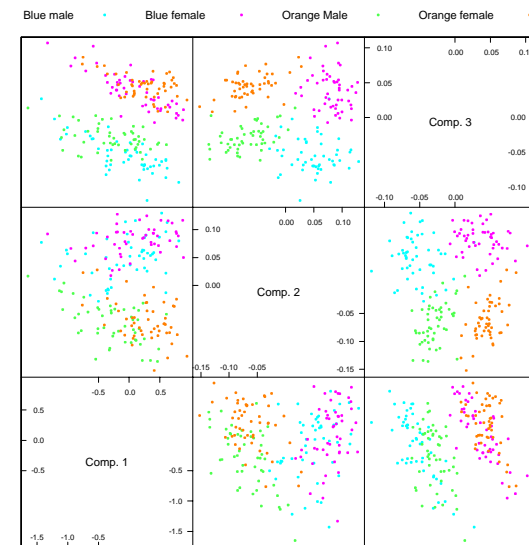
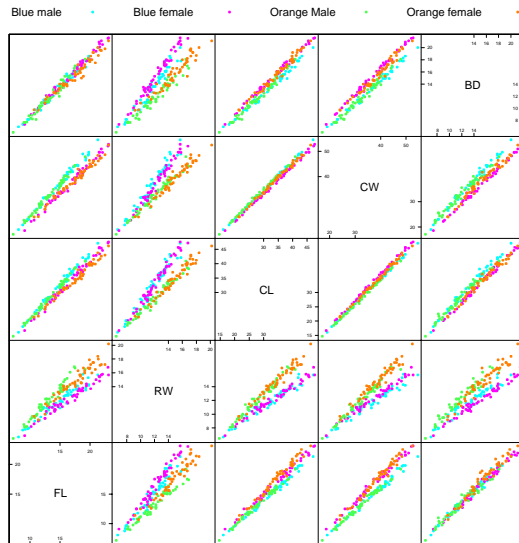
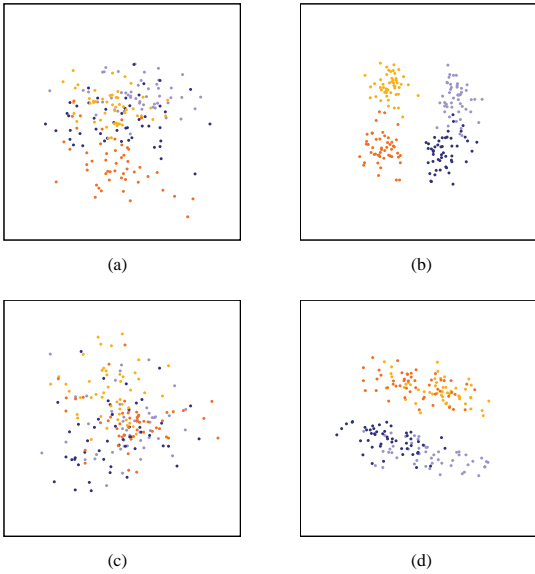


Fig. 1. Dorsal view of carapace of *Leptograpsus*, showing measurements taken. *FL*, width of frontal region just anterior to frontal tubercles, *RW*, width of posterior region. *CL*, length along midline. *CW*, maximum width. The body depth was also measured; in females but not in males the abdomen was first displaced.

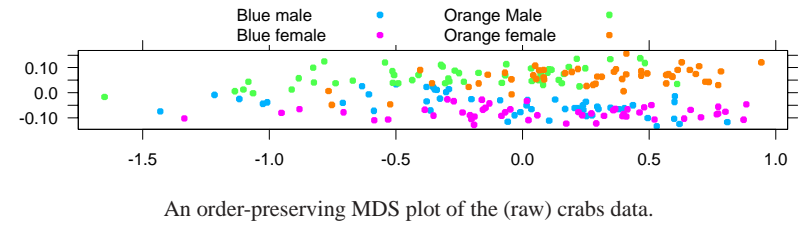


First three principal components on log scale.

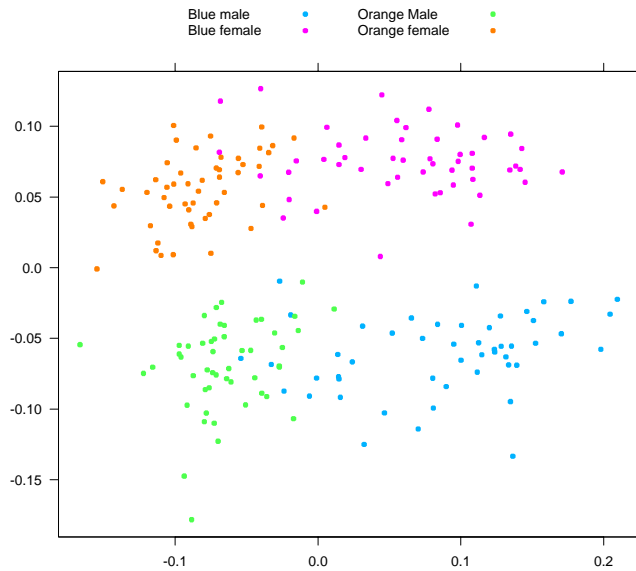


Projections of the *Leptograpsus* crabs data found by projection pursuit. View (a) is a random projection. View (b) was found using the natural Hermite index, view (c) by the Friedman–Tukey index and view (d) by Friedman’s (1987) index.

## Multidimensional scaling



## A Forensic Example



After re-scaling to (approximately) constant carapace area.

More examples using GGobi, including in 3D.

## Crop Viruses

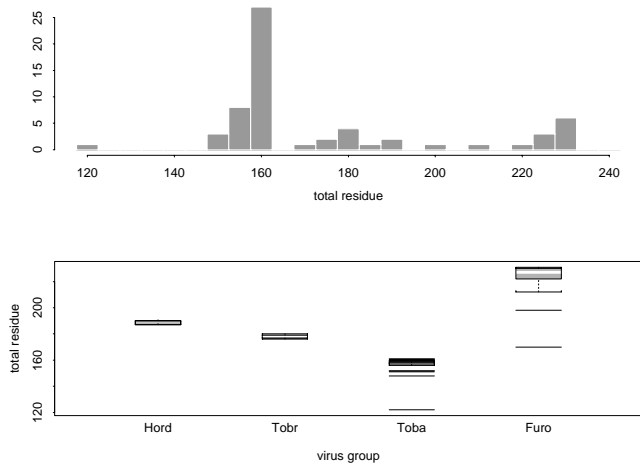
This is a dataset on 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) described by Fauquet *et al.* (1988) and analysed by Eslava-Gómez (1989). There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein.

There is an existing classification by the number of RNA molecules and mode of transmission, into

- 39 *Tobamoviruses* with monopartite genomes spread by contact,
- 6 *Tobraviruses* with bipartite genomes spread by nematodes,
- 3 *Hordeiviruses* with tripartite genomes, transmission mode unknown and
- 13 'furoviruses', 12 of which are known to be spread fungally.

## An initial look

Histogram and boxplot by group of the viruses dataset.



One initial question with this dataset is whether the numbers of residues are absolute or relative. The data are counts from 0 to 32, with the totals per virus varying from 122 to 231. The average numbers for each amino acid range from 1.4 to 20.3.

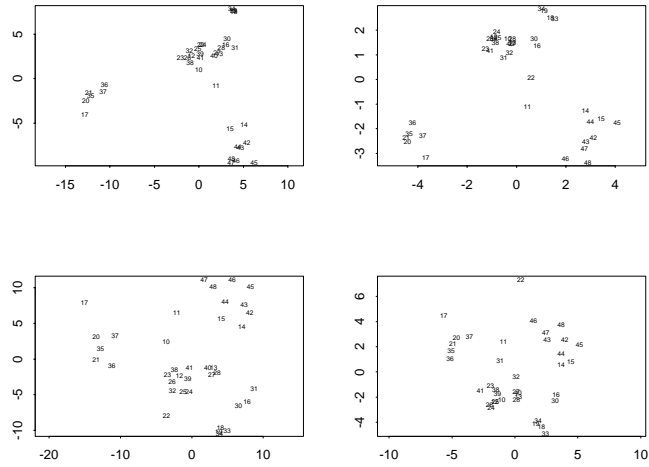
As a classification problem, this is very easy given the right visualization.

The histogram shows a multimodal distribution, and the boxplots show an almost complete separation by virus type.

The only exceptional value is one virus in the *furovirus* group with a total of 170; this is the only virus in that group whose mode of transmission is unknown and Fauquet *et al.* (1988) suggest it has been tentatively classified as a *Tobamovirus*. The other outlier in that group (with a total of 198) is the only beet virus. The conclusions of Fauquet *et al.* may be drawn from the totals alone.

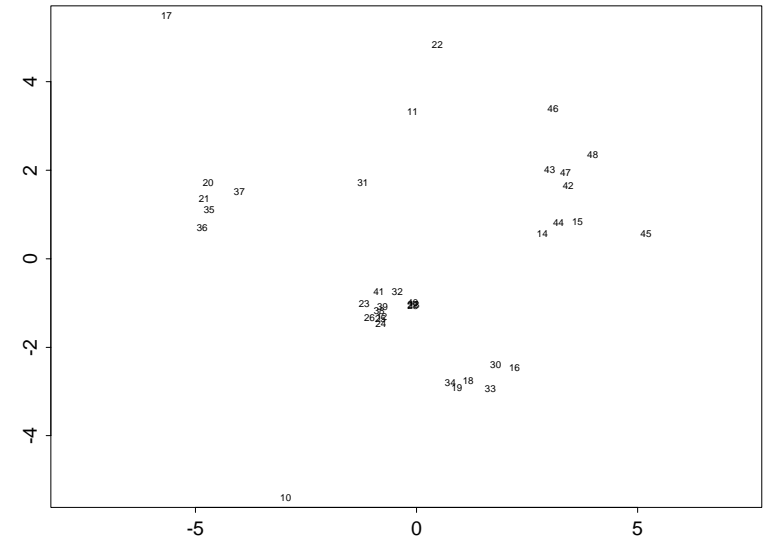
It is interesting to see if there are subgroups within the groups, so we will use this dataset to investigate further the largest group.

## Visualizing *Tobamoviruses*

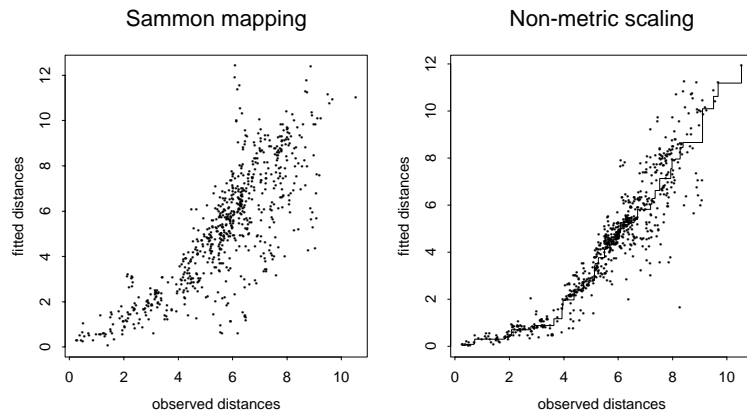


Principal component (top row) and Sammon mapping (bottom row) plots.  
Left: absolute values. Right: rescaled to have unit variance.

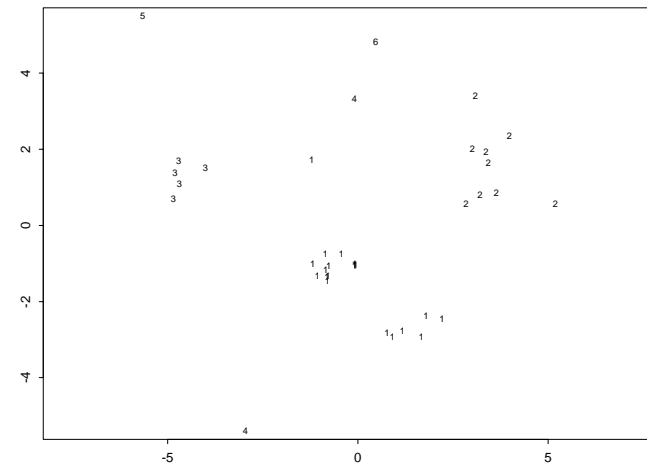
## Shepard–Kruskal MDS plot



## Shepard distortion plots

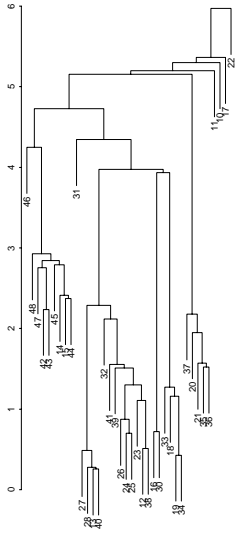


## Cluster Analysis

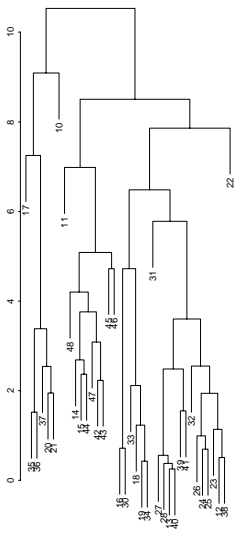


$k$ -means clustering for  $k = 6$ .

single-link



complete-link



group average

