information 'pops out', with an emphasis on qualitative understanding (e.g. 'Wow! Only a fraction of the soldiers who departed actually returned'), which, if interest persists, can be converted to quantitative understanding by close examination of numbers. Similarly, the 'circle size' representation (Figure 3.31) of the effect of components in an electronic circuit was chosen primarily because a designer is, at least in the initial stages, principally interested in whether some effect is large or small. A different consideration is associated with the multi-dimensional icon employed to represent the attributes of a house and which we discuss later in this chapter. Here we encounter *iconic encoding*, and this introduces yet another factor to be taken into consideration in the choice of encoding for a multi-attribute representation.

3.1.4 Hypervariate data

The challenge of representing hypervariate (also termed multivariate) data is substantial and continues to stimulate invention. It is an important challenge because so many real problems that are potentially amenable to at least partial solution via information visualization are of high dimensionality. The design of even the simplest silicon chip, for example, can involve over 100 components whose value has to be chosen by the designer in such a way that over 200 performance limits are satisfied. A decision regarding an investment portfolio is equally complex, as is the decision whether or not to continue with the development of a new drug. Even a basis for representing only eight scholastic achievements of a pupil can be challenging, as Exercise 3.3 will reveal.

It should first be mentioned that some of the representation techniques already discussed can be scaled, though sometimes to a limited extent, to handle hypervariate data. Thus, the TileBars scheme for the representation of text can handle more than three keywords, interactive histograms (as we shall see below) can be extended to many attributes and, as shown in Chapter 2, bargrams can be effective in the representation of many attributes of a collection of objects.

Coordinate plots

Parallel coordinate plots

One of the most popular and valued techniques for the representation of hypervariate data has a very simple basis and is called the method of parallel coordinate plots (Inselberg, 1985, 1997; Wegman, 1990). To explain the underlying principle we consider a simple case of bivariate data for which the details of two houses can be represented within a scatterplot (Figure 3.46), each house being represented by a single point. Imagine now the two axes to be detached and placed parallel to each other (Figure 3.47). Necessarily, each house will have to be represented by a point on both axes, thereby doubling the number of points required.

A further disadvantage would appear to be the need to show the relation between the two points characterizing a given house, here achieved (Figure 3.48) for each house by a straight line together with an identifying label. What is to be gained? For the case of two attributes, nothing. However, if we are dealing with objects characterized by more than two or three attributes the so-called parallel coordinate plot offers many advantages. Figure 3.49 shows the parallel

FIGURE 3.46
A simple scatterplot representing the price and number of bedrooms associated with two houses

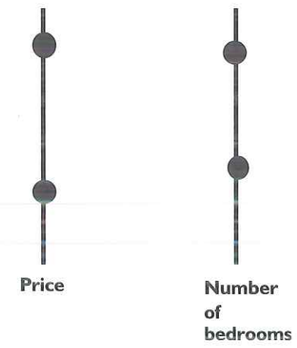


FIGURE 3.47
An alternative representation to the scatterplot in which the two attribute scales are presented in parallel, thereby requiring two points to represent each house

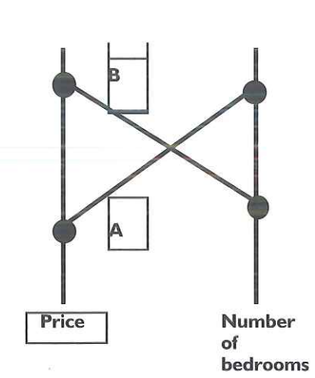
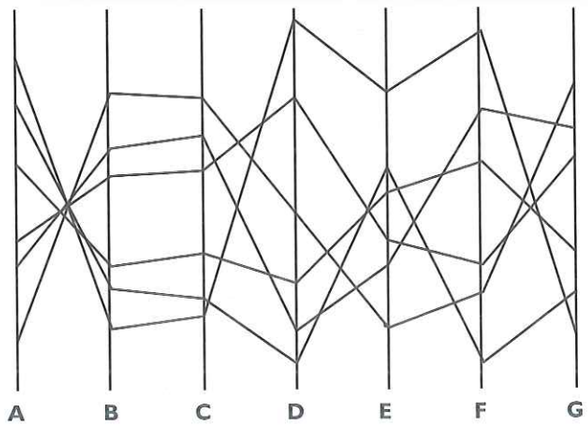


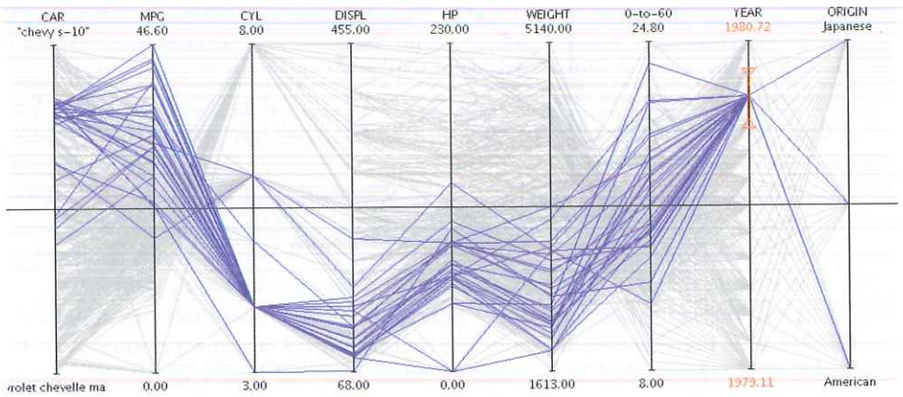
FIGURE 3.48
To avoid ambiguity the pair of points representing a house are joined and labelled

coordinate plot for six objects, each characterized by seven attributes A to G. Each object is represented by a point on each axis and hence by the piecewise linear line ('polyline') joining them. It is immediately (i.e. pre-attentively) apparent that there is a 'trade-off' between attributes A and B, as well as a strong

correlation between B and C. Nevertheless, even though a parallel coordinate plot facility is available in many commercial information visualization packages, limitations can be identified. For example, for the data shown in Figure 3.49, it is not apparent that there is also a 'trade-off' between B and E and a strong correlation between C and G; in other words, the ordering of the attributes can significantly affect the ease with which relationships can be identified.



As with many other visualization tools, the potential offered by interaction is considerable (Siirtola, 2000). For example, a range of one attribute can be identified, thereby highlighting all the object lines which pass through that range (Figure 3.50); furthermore, that range can be dynamically explored manually, allowing a user to gain quick insight into relationships between different attributes. Other facilities are normally available, including averages, standard deviations and Tukey box plots. Selected ranges of two different attributes can, for example, be 'ANDed' or 'ORed' highlighting, respectively, only those object lines that pass through *both*



selected ranges and those which pass through either or both.[3] The parallel coordinate plot technique of representation has found a wide range of application. Inselberg (1997), for example, has shown how such plots can be used to enhance the manufacturing yield of a process of making silicon chips.

It is useful to be able to characterize the sort of insight that can readily be gleaned from a parallel coordinate plot. Whereas with the scatterplot each object was easily identified and discriminable in the company of other objects, that is not the case with parallel coordinate plots because each object is now represented by a polyline which usually intersects with many other such curves. What *is* particularly visible from a parallel coordinate plot are the characteristics of the separate attributes and, in some cases, the nature of the relation between them. Thus, remarks such as 'the majority of cars appear to have four cylinders' and 'it seems that low MPG inevitably means high price' can be based on rapidly acquired 'pop-out' insight. Thus, we may observe that the parallel coordinate plot technique can support **attribute visibility**. By visibility we mean the ability to gain insight pre-attentively or without involving a great detail of cognitive effort. We return later to the concept of visibility, and the associated concept of correlation, when more encoding examples have been accumulated.

A major attraction of parallel coordinate plots, which they share with some but not all other techniques, is that their complexity (here, the number of axes) is directly proportional to the number of attributes. They also have the advantage, which is sometimes not present with other techniques (Feiner and Beshers, 1990), that all attributes receive uniform treatment. The literature (e.g. Bendix *et al.*, 2005; Siirtola, 2006)) provides many examples of the continual development of the parallel coordinate technique.

Star plots

A star plot (Coekin, 1969) has many features in common with parallel coordinate plots in that an attribute value is represented by a point on a coordinate axis and, for a given object, those points are joined by straight lines. The difference is that attribute axes now radiate from a common origin. Thus, a star plot of my school report (Figure 3.51) shows, relative to a class average indicated by the extremities of the grey region, good performance in mathematics and chemistry but very poor performance in sport and literature. To make a comparison with the talents of my friend Tony, a separate star plot (Figure 3.52) can be employed. The shape associated with a star plot can provide a reasonably rapid appreciation of the student's achievement and permit comparison with another student. It can be argued that a star plot offers '**object visibility**'. Thus, unlike parallel coordinate plots which are especially suited to the identification of relations between attributes, star plots are perhaps better for comparing specific objects. Star plots have been used to compare objects as different as police forces and mortgage options. Other encoding techniques, such as colour and thickness, can provide additional flexibility.

---

[3] The observant reader may have noticed cars apparently characterized by zero HP or zero miles per gallon, neither constituting a very desirable attribute of a car. In fact, these are misleading representations associated with missing data and illustrate a challenge to the designer of a visualization technique.

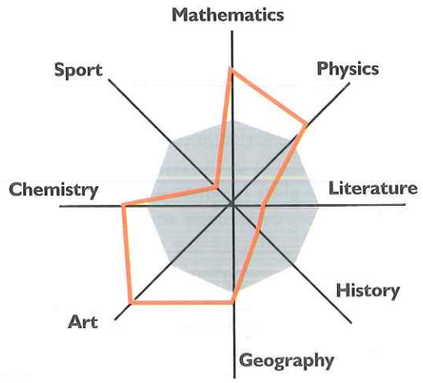


FIGURE 3.51 In a star plot attribute scales radiate from a common origin. Because shape can often effectively represent the combined attribute values of a single object, the points on each attribute scale can usefully be joined. Other useful information such as average values or thresholds can be encoded on the star plot

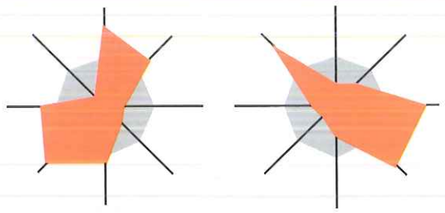


FIGURE 3.52 Star plots can be used to compare the attributes of two different objects, here the exam performance of two people in the subjects identified in Figure 3.51

Scatterplot matrix

The scatterplot matrix discussed in the context of trivariate data is equally applicable to higher dimensions, but with one major disadvantage arising from the fact that, as the number of attributes increases, the number of different *pairs* of attributes increases rapidly. With two attributes we needed one scatterplot, with three we needed three and with four attributes there are six unique pairs. Thus, for 100 houses each associated with four attributes, the scatterplot matrix would contain 600 points. This is in contrast to the linear increase in complexity associated with parallel coordinate plots and, as we shall see, the Attribute Explorer to be discussed immediately below. While there is no theoretical limit to the number of attributes a scatterplot can handle, this unwelcome dependence of complexity on dimension is always present.

The use of a single scatterplot together with other encoding techniques to represent hypervariate data was demonstrated very effectively by Ahlberg (Ahlberg *et al.,* 1992; Ahlberg and Shneiderman, 1994) with an interactive representation designed to allow a user to select a film to watch on video (Figure 3.53). On the main (scatterplot) display each coloured square identifies a film. Colour encodes type (horror, musical, etc.), horizontal position indicates the year of production and vertical position indicates duration. On the right, sliders can be used to specify other attributes of a film such as director or actor. Scroll bars can be used to confine attention to a particular span of years and film length, whereupon more detail can be displayed (Figure 3.54). The Film Finder, as the interface was called, provides a good illustration of the potential of combining different representation techniques.

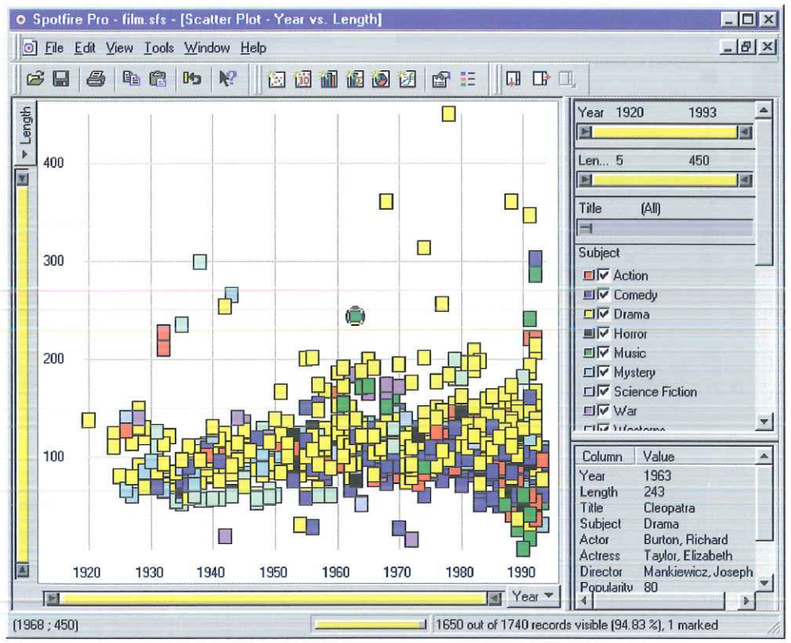


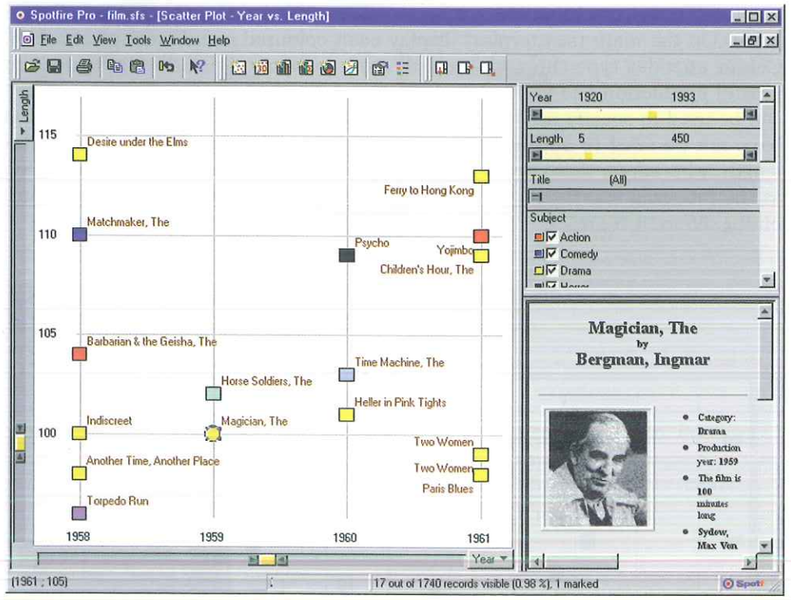
FIGURE 3.53 A scatterplot enhanced by additional and selective encoding, allowing the selection of a film on the basis of type, duration, year of production and other attributes

Linked histograms

The technique of linked histograms presented in the discussion of bivariate data (see Figure 3.20) can be extended to hypervariate data and, in the Attribute Explorer, can be considerably enhanced in value by additional encoding (Tweedie *et al.,* 1994; Spence and Tweedie, 1998). We shall first illustrate the technique in the context of buying a house, a task which, like that of buying a car, can usefully be generalized:

> *Given a collection of objects, each described by the values associated with a set of attributes, find the most acceptable such object or, perhaps, a small number of candidate objects worthy of more detailed consideration.*

FIGURE 3.54 The automatic display of additional detail following the selection of narrower limits on years of production and film length



However, in the course of illustrating the use of the Attribute Explorer, we shall see that an equally important task for which it (and many other techniques) is suited is:

> *the acquisition of insight into multivariate data.*

We begin by examining a histogram (Figure 3.55) of one attribute of a collection of houses, that of *Price*. Here, each house contributes one small rectangle to the histogram. Upper and lower limits to *Price* can easily be positioned (Figure 3.56) to identify a subset of houses that may initially and experimentally be regarded as affordable. The result is that the houses so identified are encoded green. For reasons soon to be apparent, houses outside the limits continue to be displayed. Many house attributes will normally be of interest and we shall consider just three to establish the concepts underlying the Attribute Explorer: *Price, Number of bedrooms* and *Garden Size*. Figure 3.57 shows the corresponding attribute histograms, but with the same limits on *Price* as in Figure 3.56. Those houses satisfying the limits on *Price* are now coded green not only on the *Price* histogram but also on the other two histograms, another example of brushing. Immediately, the consequences of the *Price* limits for the availability of houses with given garden sizes and numbers of bedrooms are apparent and can, in fact, be dynamically explored by moving the range bar between the *Price* limits to and fro: some idea of any correlation between *Price* and *Number of bedrooms* could easily be acquired. If limits are now additionally placed on the remaining attributes (Figure 3.58), green encoding applies only to those houses which satisfy *all* the attribute limits. Again, dynamic exploration achieved by adjusting either range positions or individual limits can help the user to gain insight into the house data and gradually come to a decision about which house or houses are worthy of more detailed consideration. It is because houses can be explored on the basis of their attributes that the technique described is called the Attribute Explorer.

Of major importance in the Attribute Explorer is the colour coding involved. Black houses are those which fail only one limit; therefore, if a black house is located outside a limit, that must be the limit it fails. This can be very useful

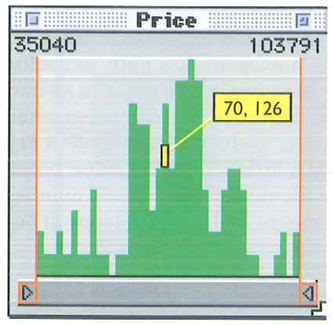


FIGURE 3.55 A histogram representing the prices of a collection of houses. The contribution of one house is shown in yellow

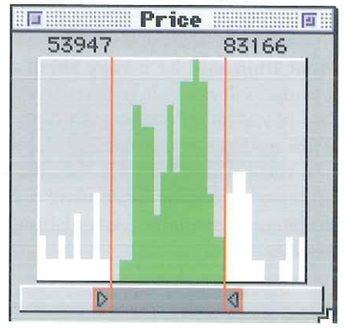


FIGURE 3.56 Limits on *Price* identify a subset of houses, coded green

FIGURE 3.57 Houses defined by the limits on *Price* in Figure 3.56 are coded green in other attribute histograms

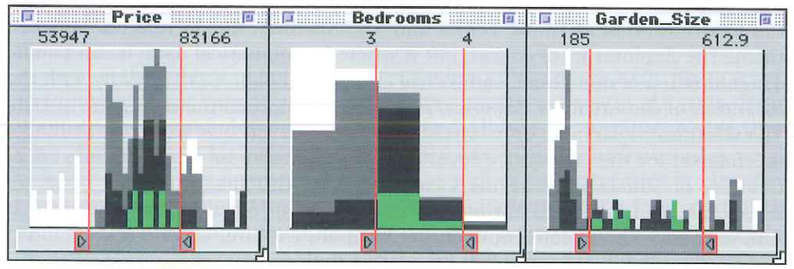


FIGURE 3.58 Green coding applies only to houses which satisfy all attribute limits. Houses which fail one limit are coded black, so if a black house is positioned outside a limit it will turn green if the limit is extended to include it