
Looking at Multivariate Data: Visualisation

2.1 Introduction

According to [Chambers, Cleveland, Kleiner, and Tukey \(1983\)](#), “there is no statistical tool that is as powerful as a well-chosen graph”. Certainly graphical presentation has a number of advantages over tabular displays of numerical results, not least in creating interest and attracting the attention of the viewer. But just what is a graphical display? A concise description is given by [Tufte \(1983\)](#):

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

Graphs are very popular; it has been estimated that between 900 billion (9×10^{11}) and 2 trillion (2×10^{12}) images of statistical graphics are printed each year. Perhaps one of the main reasons for such popularity is that graphical presentation of data often provides the vehicle for discovering the unexpected; the human visual system is very powerful in detecting patterns, although the following caveat from the late Carl Sagan (in his book *Contact*) should be kept in mind:

Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.

Some of the advantages of graphical methods have been listed by [Schmid \(1954\)](#):

- In comparison with other types of presentation, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time since the essential meaning of large measures of statistical data can be visualised at a glance.

- Charts and graphs provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.
- Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

Schmid's last point is reiterated by the legendary John Tukey in his observation that "the greatest value of a picture is when it forces us to notice what we never expected to see".

The prime objective of a graphical display is to communicate to ourselves and others, and the graphic design must do everything it can to help people understand. And unless graphics are relatively simple, they are unlikely to survive the first glance. There are perhaps four goals for graphical displays of data:

- To provide an overview;
- To tell a story;
- To suggest hypotheses;
- To criticise a model.

In this chapter, we will be largely concerned with graphics for multivariate data that address one or another of the first three bulleted points above. Graphics that help in checking model assumptions will be considered in Chapter 8.

During the last two decades, a wide variety of new methods for displaying data graphically have been developed. These will hunt for special effects in data, indicate outliers, identify patterns, diagnose models, and generally search for novel and perhaps unexpected phenomena. Graphical displays should aim to tell a story about the data and to reduce the cognitive effort required to make comparisons. Large numbers of graphs might be required to achieve these goals, and computers are generally needed to supply them for the same reasons that they are used for numerical analyses, namely that they are fast and accurate.

So, because the machine is doing the work, the question is no longer "shall we plot?" but rather "what shall we plot?" There are many exciting possibilities, including interactive and dynamic graphics on a computer screen (see [Cook and Swayne 2007](#)), but graphical exploration of data usually begins at least with some simpler *static* graphics. The starting graphic for multivariate data is often the ubiquitous *scatterplot*, and this is the subject of the next section.

2.2 The scatterplot

The simple *xy* scatterplot has been in use since at least the 18th century and has many virtues—indeed, according to [Tufté \(1983\)](#):

The relational graphic—in its barest form the scatterplot and its variants—is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y .

The scatterplot is the standard for representing continuous *bivariate data* but, as we shall see later in this chapter, it can be enhanced in a variety of ways to accommodate information about other variables.

To illustrate the use of the scatterplot and a number of other techniques to be discussed, we shall use the air pollution in US cities data introduced in the previous chapter (see Table 1.5).

Let's begin our examination of the air pollution data by taking a look at a basic scatterplot of the two variables `manu` and `popul`. For later use, we first set up two character variables that contain the labels to be printed on the two axes:

```
R> mlab <- "Manufacturing enterprises with 20 or more workers"
R> plab <- "Population size (1970 census) in thousands"
```

The `plot()` function takes the data, here as the data frame `USairpollution`, along with a “formula” describing the variables to be plotted; the part left of the tilde defines the variable to be associated with the ordinate, the part right of the tilde is the variable that goes with the abscissa:

```
R> plot(popul ~ manu, data = USairpollution,
+       xlab = mlab, ylab = plab)
```

The resulting scatterplot is shown in Figure 2.2. The plot clearly uncovers the presence of one or more cities that are some way from the remainder, but before commenting on these possible *outliers* we will construct the scatterplot again but now show how to include the marginal distributions of `manu` and `popul` in two different ways. Plotting marginal and joint distributions together is usually good data analysis practise. In Figure 2.2, the marginal distributions are shown as rug plots on each axis (produced by `rug()`), and in Figure 2.3 the marginal distribution of `manu` is given as a histogram and that of `popul` as a boxplot. And also in Figure 2.3 the points are labelled by an abbreviated form of the corresponding city name.

The necessary R code for Figure 2.3 starts with dividing the device into three plotting areas by means of the `layout()` function. The first plot basically resembles the `plot()` command from Figure 2.1, but instead of points the abbreviated name of the city is used as the plotting symbol. Finally, the `hist()` and `boxplots()` commands are used to depict the marginal distributions. The `with()` command is very useful when one wants to avoid explicitly extracting variables from data frames. The command of interest, here the calls to `hist()` and `boxplot()`, is evaluated “inside” the data frame, here `USairpollution` (i.e., variable names are resolved within this data frame first).

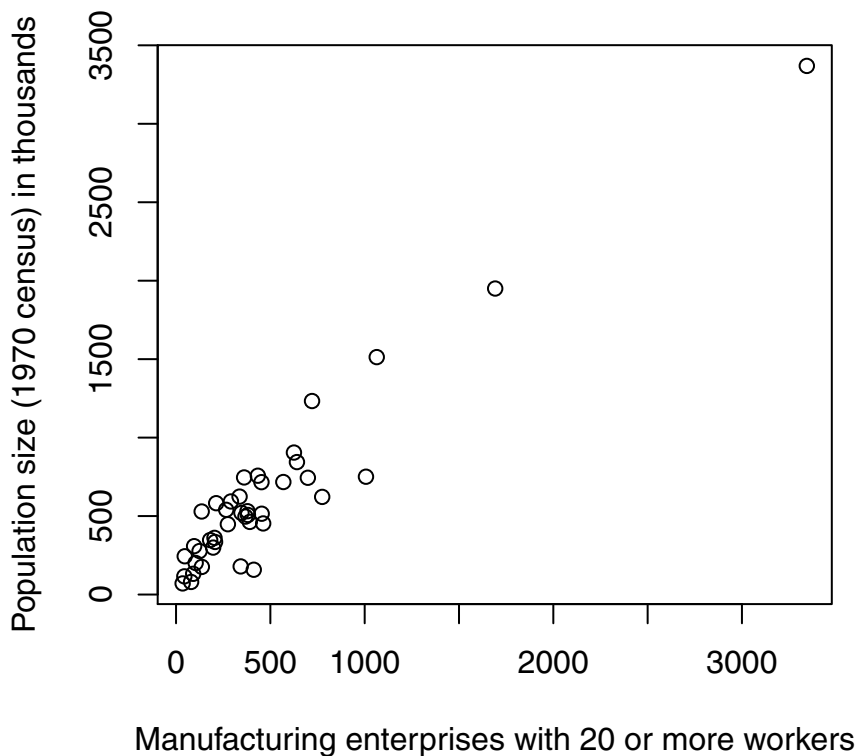


Fig. 2.1. Scatterplot of `manu` and `popul`.

From this series of plots, we can see that the outlying points show themselves in both the scatterplot of the variables *and* in each marginal distribution. The most extreme outlier corresponds to Chicago, and other slightly less extreme outliers correspond to Philadelphia and Detroit. Each of these cities has a considerably larger population than other cities and also many more manufacturing enterprises with more than 20 workers.

2.2.1 The bivariate boxplot

In Figure 2.3, identifying Chicago, Philadelphia, and Detroit as outliers is unlikely to invoke much argument, but what about Houston and Cleveland? In many cases, it might be helpful to have a more formal and objective method for labelling observations as outliers, and such a method is provided by the *bivariate boxplot*, which is a two-dimensional analogue of the boxplot for univariate data proposed by [Goldberg and Iglewicz \(1992\)](#). This type of graphic

```
R> plot(popul ~ manu, data = USairpollution,
+       xlab = mlab, ylab = plab)
R> rug(USairpollution$manu, side = 1)
R> rug(USairpollution$popul, side = 2)
```

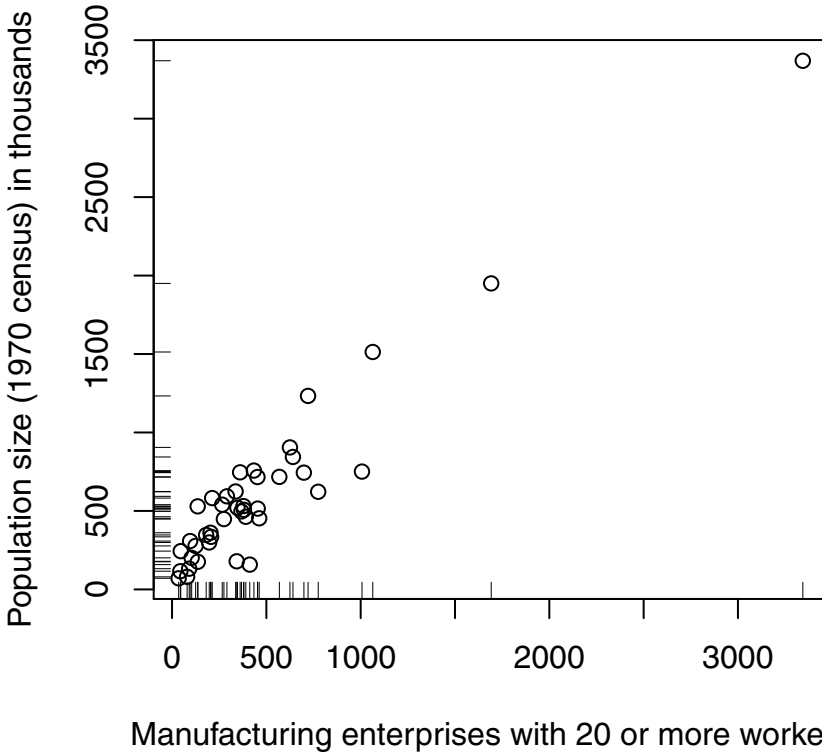


Fig. 2.2. Scatterplot of `manu` and `popul` that shows the marginal distribution in each variable as a rug plot.

may be useful in indicating the distributional properties of the data and in identifying possible outliers. The bivariate boxplot is based on calculating “robust” measures of location, scale, and correlation; it consists essentially of a pair of concentric ellipses, one of which (the “hinge”) includes 50% of the data and the other (called the “fence”) of which delineates potentially troublesome outliers. In addition, resistant regression lines of both y on x and x on y are shown, with their intersection showing the bivariate location estimator. The acute angle between the regression lines will be small for a large absolute value of correlations and large for a small one. (Using robust measures of location, scale, etc., helps to prevent the possible “masking” of multivariate outliers if

```

R> layout(matrix(c(2, 0, 1, 3), nrow = 2, byrow = TRUE),
+         widths = c(2, 1), heights = c(1, 2), respect = TRUE)
R> xlim <- with(USairpollution, range(manu)) * 1.1
R> plot(popul ~ manu, data = USairpollution, cex.lab = 0.9,
+       xlab = mlab, ylab = plab, type = "n", xlim = xlim)
R> with(USairpollution, text(manu, popul, cex = 0.6,
+       labels = abbreviate(row.names(USairpollution))))
R> with(USairpollution, hist(manu, main = "", xlim = xlim))
R> with(USairpollution, boxplot(popul))

```

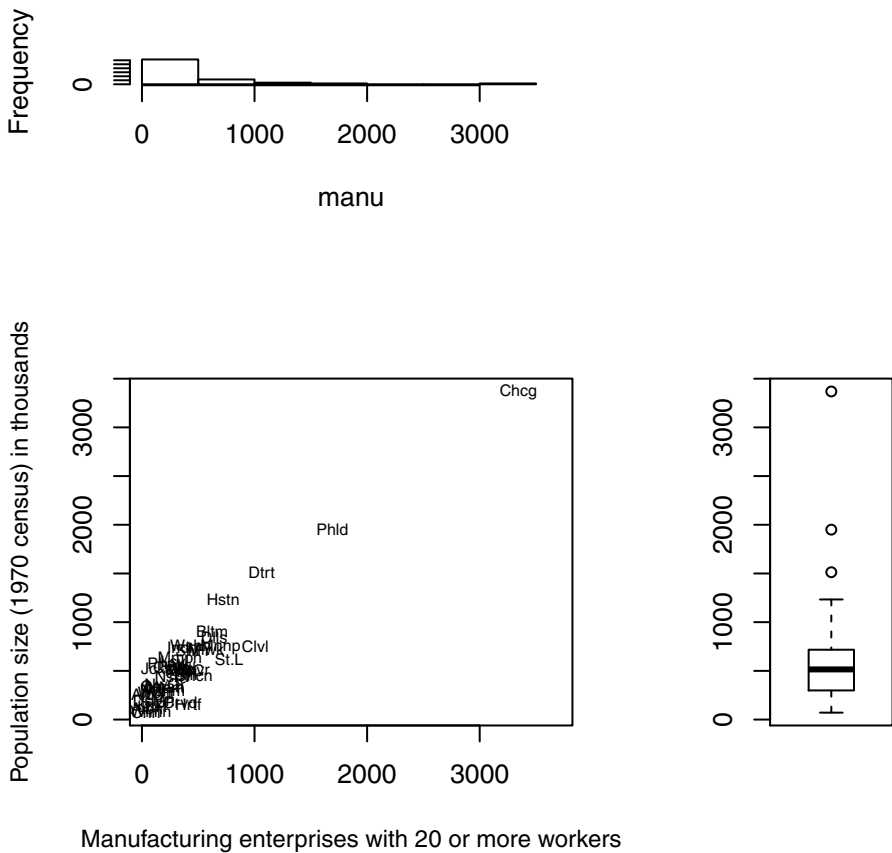


Fig. 2.3. Scatterplot of `manu` and `popul` that shows the marginal distributions by histogram and boxplot.

the usual measures are employed when these may be distorted by the presence of the outliers in the data.) Full details of the construction are given in [Goldberg and Iglewicz \(1992\)](#). The scatterplot of `manu` and `popul` including the bivariate boxplot is shown in Figure 2.4. Figure 2.4 clearly tells us that Chicago, Philadelphia, Detroit, and Cleveland should be regarded as outliers but not Houston, because it is on the “fence” rather than outside the “fence”.

```
R> outcity <- match(lab <- c("Chicago", "Detroit",
+   "Cleveland", "Philadelphia"), rownames(USairpollution))
R> x <- USairpollution[, c("manu", "popul")]
R> bvbox(x, mtitle = "", xlab = mlab, ylab = plab)
R> text(x$manu[outcity], x$popul[outcity], labels = lab,
+   cex = 0.7, pos = c(2, 2, 4, 2, 2))
```

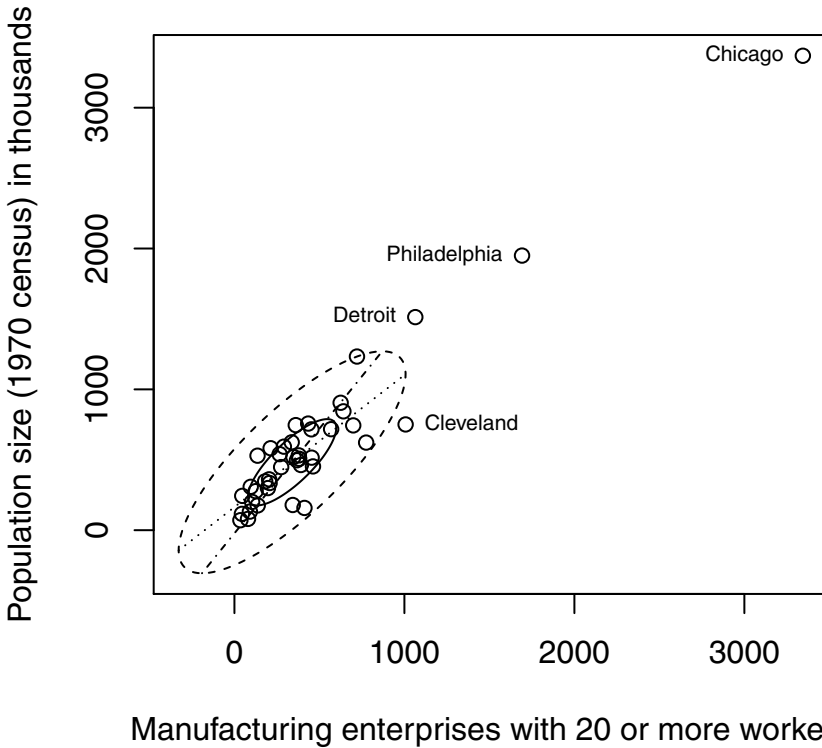


Fig. 2.4. Scatterplot of `manu` and `popul` showing the bivariate boxplot of the data.

Suppose now that we are interested in calculating the correlation between `manu` and `popul`. Researchers often calculate the correlation between two vari-

ables *without* first looking at the scatterplot of the two variables. But scatterplots should *always* be consulted when calculating correlation coefficients because the presence of outliers can on occasion considerably distort the value of a correlation coefficient, and as we have seen above, a scatterplot may help to identify the offending observations particularly if used in conjunction with a bivariate boxplot. The observations identified as outliers may then be excluded from the calculation of the correlation coefficient. With the help of the bivariate boxplot in Figure 2.4, we have identified Chicago, Philadelphia, Detroit, and Cleveland as outliers in the scatterplot of `manu` and `popul`. The R code for finding the two correlations is

```
R> with(USairpollution, cor(manu, popul))

[1] 0.9553

R> outcity <- match(c("Chicago", "Detroit",
+                    "Cleveland", "Philadelphia"),
+                  rownames(USairpollution))
R> with(USairpollution, cor(manu[-outcity], popul[-outcity]))

[1] 0.7956
```

The `match()` function identifies rows of the data frame `USairpollution` corresponding to the cities of interest, and the subset starting with a minus sign removes these units before the correlation is computed. Calculation of the correlation coefficient between the two variables using all the data gives a value of 0.96, which reduces to a value of 0.8 after excluding the four outliers—a not inconsiderable reduction.

2.2.2 The convex hull of bivariate data

An alternative approach to using the scatterplot combined with the bivariate boxplot to deal with the possible problem of calculating correlation coefficients without the distortion often caused by outliers in the data is *convex hull trimming*, which allows *robust estimation* of the correlation. The convex hull of a set of bivariate observations consists of the vertices of the smallest convex polyhedron in variable space within which or on which all data points lie. Removal of the points lying on the convex hull can eliminate isolated outliers without disturbing the general shape of the bivariate distribution. A robust estimate of the correlation coefficient results from using the remaining observations. Let's see how the convex hull approach works with our `manu` and `popul` scatterplot. We first find the convex hull of the data (i.e., the observations defining the convex hull) using the following R code:

```
R> (hull <- with(USairpollution, chull(manu, popul)))

[1] 9 15 41 6 2 18 16 14 7
```



```
R> with(USairpollution,
+       plot(manu, popul, pch = 1, xlab = mlab, ylab = plab))
R> with(USairpollution,
+       polygon(manu[hull], popul[hull], density = 15, angle = 30))
```

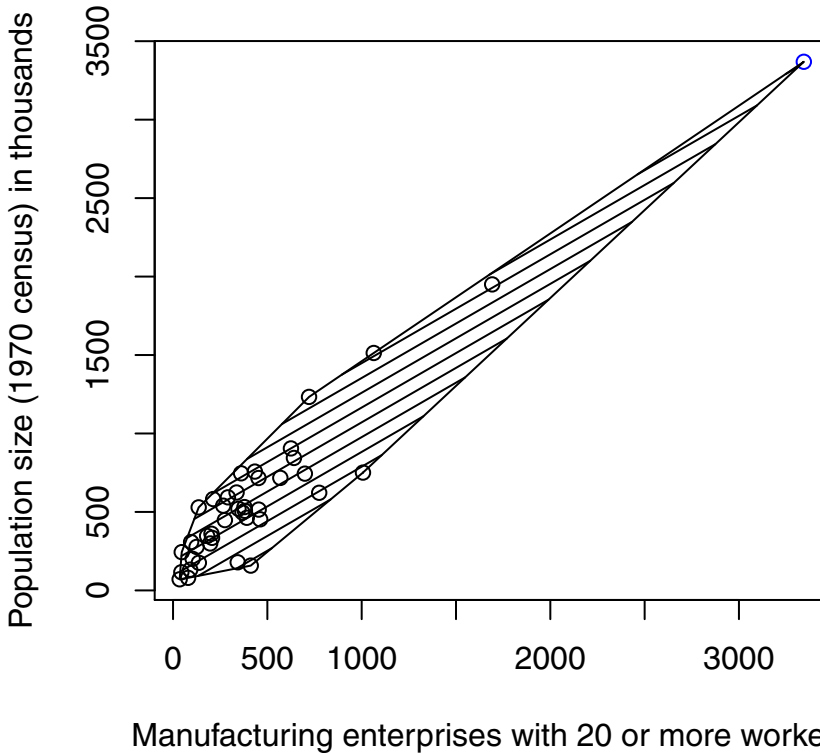


Fig. 2.5. Scatterplot of `manu` against `popul` showing the convex hull of the data.

Now we can show this convex hull on a scatterplot of the variables using the code attached to the resulting Figure 2.5.

To calculate the correlation coefficient after removal of the points defining the convex hull requires the code

```
R> with(USairpollution, cor(manu[-hull],popul[-hull]))

[1] 0.9225
```

The resulting value of the correlation is now 0.923 and thus is higher compared with the correlation estimated after removal of the outliers identified by using the bivariate boxplot, namely Chicago, Philadelphia, Detroit, and Cleveland.

2.2.3 The chi-plot

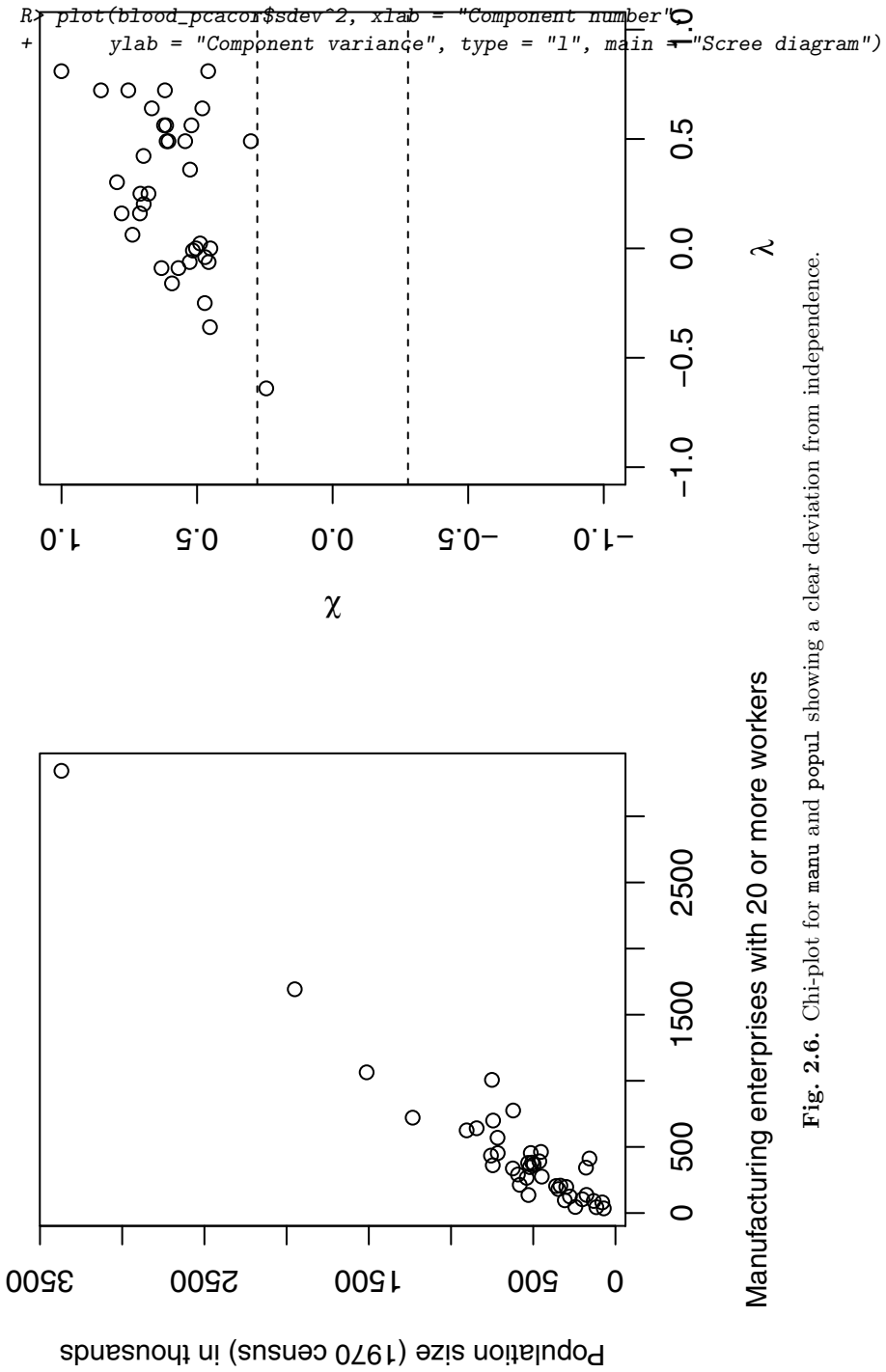
Although the scatterplot is a primary data-analytic tool for assessing the relationship between a pair of continuous variables, it is often difficult to judge whether or not the variables are independent—a random scatter of points is hard for the human eye to judge. Consequently it is sometimes helpful to augment the scatterplot with an auxiliary display in which independence is itself manifested in a characteristic manner. The *chi-plot* suggested by Fisher and Switzer (1985, 2001) is designed to address the problem. Under independence, the joint distribution of two random variables X_1 and X_2 can be computed from the product of the marginal distributions. The chi-plot transforms the measurements (x_{11}, \dots, x_{n1}) and (x_{12}, \dots, x_{n2}) into values (χ_1, \dots, χ_n) and $(\lambda_1, \dots, \lambda_n)$, which, plotted in a scatterplot, can be used to detect deviations from independence. The χ_i values are, basically, the root of the χ^2 statistics obtained from the 2×2 tables that are obtained when dichotomising the data for each unit i into the groups satisfying $x_{.1} \leq x_{i1}$ and $x_{.2} \leq x_{i2}$. Under independence, these values are asymptotically normal with mean zero; i.e., the χ_i values should show a non-systematic random fluctuation around zero. The λ_i values measure the distance of unit i from the “center” of the bivariate distribution. An R function for producing chi-plots is `chiplot()`. To illustrate the chi-plot, we shall apply it to the `manu` and `popul` variables of the air pollution data using the code

```
R> with(USairpollution, plot(manu, popul,
+                             xlab = mlab, ylab = plab,
+                             cex.lab = 0.9))
R> with(USairpollution, chiplot(manu, popul))
```

The result is Figure 2.6, which shows the scatterplot of `manu` plotted against `popul` alongside the corresponding chi-plot. Departure from independence is indicated in the latter by a lack of points in the horizontal band indicated on the plot. Here there is a very clear departure since there are very few of the observations in this region.

2.3 The bubble and other glyph plots

The basic scatterplot can only display two variables. But there have been a number of suggestions as to how extra variables may be included on a scatterplot. Perhaps the simplest is the so-called *bubble plot*, in which three variables are displayed; two are used to form the scatterplot itself, and then the values of the third variable are represented by circles with radii proportional to these values and centred on the appropriate point in the scatterplot. Let's begin by taking a look at the bubble plot of `temp`, `wind`, and `S02` that is given in Figure 2.7. The plot seems to suggest that cities with moderate annual temperatures and moderate annual wind speeds tend to suffer the greatest air



pollution, but this is unlikely to be the whole story because none of the other variables in the data set are used in constructing Figure 2.7. We could try to include all variables on the basic `temp` and `wind` scatterplot by replacing the circles with five-sided “stars”, with the lengths of each side representing each of the remaining five variables. Such a plot is shown in Figure 2.8, but it fails to communicate much, if any, useful information about the data.

```
R> ylim <- with(USairpollution, range(wind)) * c(0.95, 1)
R> plot(wind ~ temp, data = USairpollution,
+       xlab = "Average annual temperature (Fahrenheit)",
+       ylab = "Average annual wind speed (m.p.h.)", pch = 10,
+       ylim = ylim)
R> with(USairpollution, symbols(temp, wind, circles = S02,
+                               inches = 0.5, add = TRUE))
```

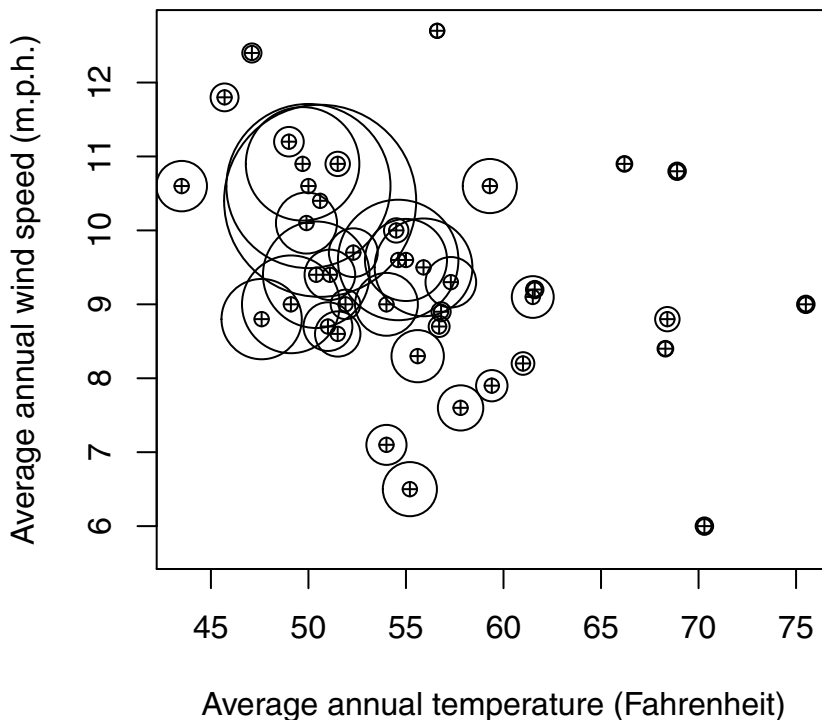


Fig. 2.7. Bubble plot of temp, wind, and S02.

```
R> plot(wind ~ temp, data = USairpollution,
+       xlab = "Average annual temperature (Fahrenheit)",
+       ylab = "Average annual wind speed (m.p.h.)", pch = 10,
+       ylim = ylim)
R> with(USairpollution,
+       stars(USairpollution[, -c(2,5)], locations = cbind(temp, wind),
+             labels = NULL, add = TRUE, cex = 0.5))
```

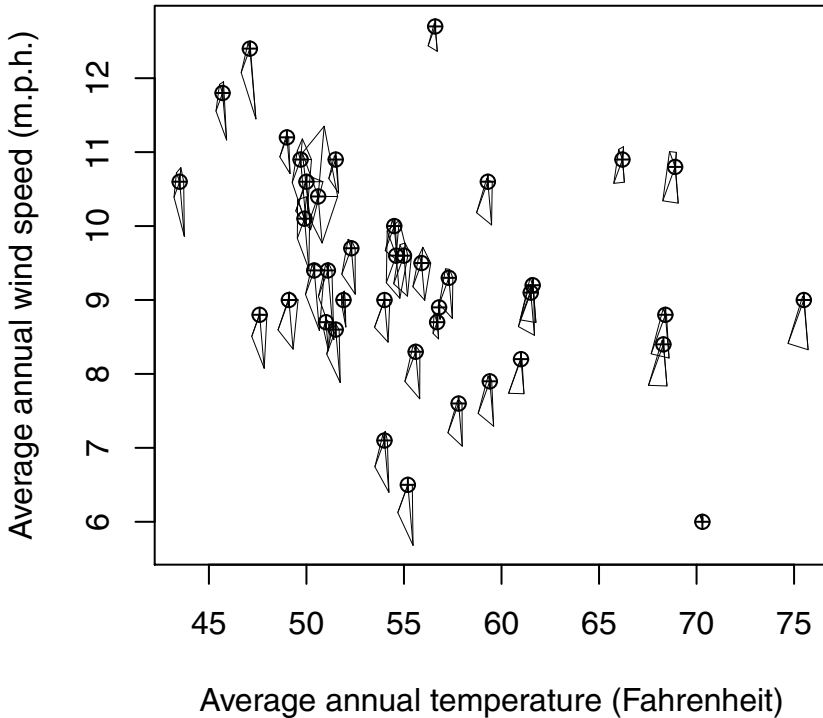


Fig. 2.8. Scatterplot of `temp` and `wind` showing five-sided stars representing the other variables.

In fact, both the bubble plot and “stars” plot are examples of *symbol* or *glyph plots*, in which data values control the symbol parameters. For example, a circle is a glyph where the values of one variable in a multivariate observation control the circle size. In Figure 2.8, the spatial positions of the cities in the scatterplot of `temp` and `wind` are combined with a star representation of the five other variables. An alternative is simply to represent the seven variables for each city by a seven-sided star and arrange the resulting stars in

a rectangular array; the result is shown in Figure 2.9. We see that some stars, for example those for New Orleans, Miami, Jacksonville, and Atlanta, have similar shapes, with their higher average annual temperature being distinctive, but telling a story about the data with this display is difficult.

Stars, of course, are not the only symbols that could be used to represent data, and others have been suggested, with perhaps the most well known being the now infamous Chernoff’s faces (see [Chernoff 1973](#)). But, on the whole, such graphics for displaying multivariate data have not proved themselves to be effective for the task and are now largely confined to the past history of multivariate graphics.

```
R> stars(USairpollution, cex = 0.55)
```



Fig. 2.9. Star plot of the air pollution data.

2.4 The scatterplot matrix

There are seven variables in the air pollution data, which between them generate 21 possible scatterplots. But just making the graphs without any coordination will often result in a confusing collection of graphs that are hard to integrate visually. Consequently, it is very important that the separate plots be presented in the best way to aid overall comprehension of the data. The *scatterplot matrix* is intended to accomplish this objective. A scatterplot matrix is nothing more than a square, symmetric grid of bivariate scatterplots. The grid has q rows and columns, each one corresponding to a different variable. Each of the grid's cells shows a scatterplot of two variables. Variable j is plotted against variable i in the ij th cell, and the same variables appear in cell ji , with the x - and y -axes of the scatterplots interchanged. The reason for including both the upper and lower triangles of the grid, despite the seeming redundancy, is that it enables a row and a column to be visually scanned to see one variable against all others, with the scales for the one variable lined up along the horizontal or the vertical. As a result, we can visually *link* features on one scatterplot with features on another, and this ability greatly increases the power of the graphic.

The scatterplot matrix for the air pollution data is shown in Figure 2.10. The plot was produced using the function `pairs()`, here with slightly enlarged dot symbols, using the arguments `pch = "."` and `cex = 1.5`.

The scatterplot matrix clearly shows the presence of possible outliers in many panels and the suggestion that the relationship between the two aspects of rainfall, namely `precip`, `predays`, and `S02` might be non-linear. Remembering that the *multivariable* aspect of these data, in which sulphur dioxide concentration is the response variable, with the remaining variables being explanatory, might be of interest, the scatterplot matrix may be made more helpful by including the linear fit of the two variables on each panel, and such a plot is shown in Figure 2.11. Here, the `pairs()` function was customised by a small function specified to the `panel` argument: in addition to plotting the x and y values, a regression line obtained via function `lm()` is added to each of the panels.

Now the scatterplot matrix reveals that there is a strong linear relationship between `S02` and `manu` and between `S02` and `popul`, but the (3, 4) panel shows that `manu` and `popul` are themselves very highly related and thus predictive of `S02` in the same way. Figure 2.11 also underlines that assuming a linear relationship between `S02` and `precip` and `S02` and `predays`, as might be the case if a multiple linear regression model is fitted to the data with `S02` as the dependent variable, is unlikely to fully capture the relationship between each pair of variables.

In the same way that the scatterplot should always be used alongside the numerical calculation of a correlation coefficient, so should the scatterplot matrix always be consulted when looking at the correlation matrix of a set of variables. The correlation matrix for the air pollution data is

```
R> pairs(USairpollution, pch = ".", cex = 1.5)
```

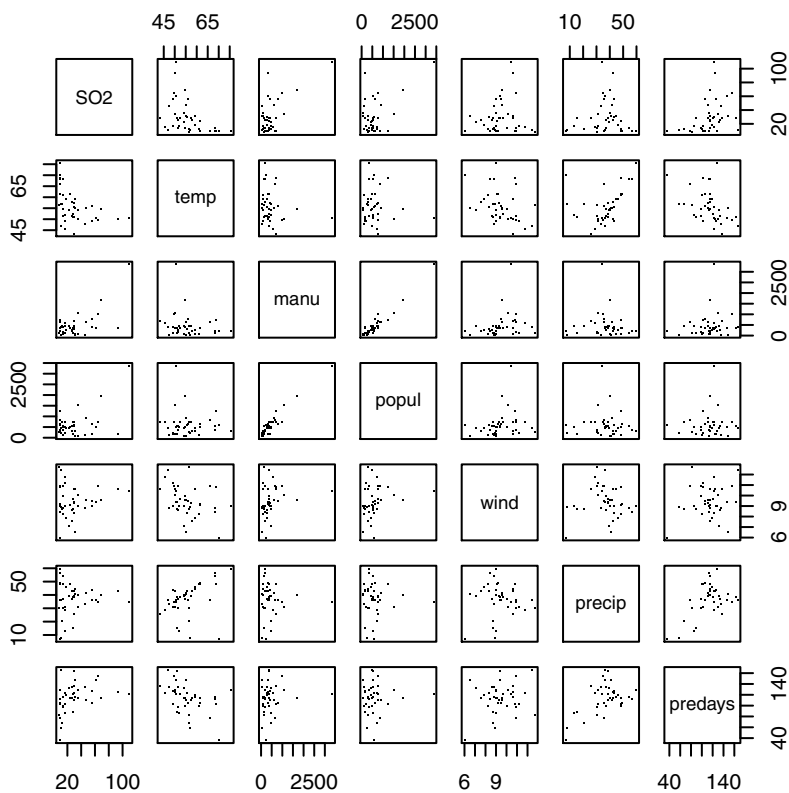


Fig. 2.10. Scatterplot matrix of the air pollution data.

```
R> round(cor(USairpollution), 4)
```

	SO2	temp	manu	popul	wind	precip	predays
SO2	1.0000	-0.4336	0.6448	0.4938	0.0947	0.0543	0.3696
temp	-0.4336	1.0000	-0.1900	-0.0627	-0.3497	0.3863	-0.4302
manu	0.6448	-0.1900	1.0000	0.9553	0.2379	-0.0324	0.1318
popul	0.4938	-0.0627	0.9553	1.0000	0.2126	-0.0261	0.0421
wind	0.0947	-0.3497	0.2379	0.2126	1.0000	-0.0130	0.1641
precip	0.0543	0.3863	-0.0324	-0.0261	-0.0130	1.0000	0.4961
predays	0.3696	-0.4302	0.1318	0.0421	0.1641	0.4961	1.0000

Focussing on the correlations between SO2 and the six other variables, we see that the correlation for SO2 and precip is very small and that for SO2 and predays is moderate. But relevant panels in the scatterplot indicate that the correlation coefficient that assesses only the linear relationship between


```
R> pairs(USairpollution,
+       panel = function (x, y, ...) {
+         points(x, y, ...)
+         abline(lm(y ~ x), col = "grey")
+       }, pch = ".", cex = 1.5)
```

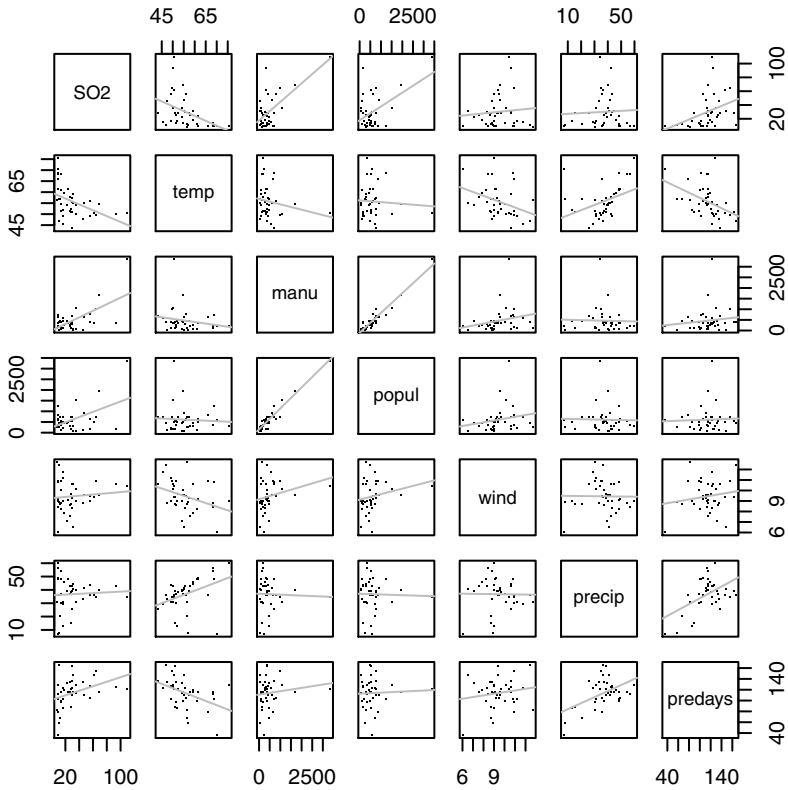


Fig. 2.11. Scatterplot matrix of the air pollution data showing the linear fit of each pair of variables.

two variables may not be suitable here and that in a multiple linear regression model for the data quadratic effects of `predays` and `precip` might be considered.

2.5 Enhancing the scatterplot with estimated bivariate densities

As we have seen above, scatterplots and scatterplot matrices are good at highlighting outliers in a multivariate data set. But in many situations another aim in examining scatterplots is to identify regions in the plot where there are high or low densities of observations that may indicate the presence of distinct groups of observations; i.e., “clusters” (see Chapter 6). But humans are not particularly good at visually examining point density, and it is often a very helpful aid to add some type of *bivariate density estimate* to the scatterplot. A bivariate density estimate is simply an approximation to the bivariate probability density function of two variables obtained from a sample of bivariate observations of the variables. If, of course, we are willing to assume a particular form of the bivariate density of the two variables, for example the bivariate normal, then estimating the density is reduced to estimating the parameters of the assumed distribution. More commonly, however, we wish to allow the data to speak for themselves and so we need to look for a *non-parametric estimation* procedure. The simplest such estimator would be a two-dimensional histogram, but for small and moderately sized data sets that is not of any real use for estimating the bivariate density function simply because most of the “boxes” in the histogram will contain too few observations; and if the number of boxes is reduced, the resulting histogram will be too coarse a representation of the density function.

Other non-parametric density estimators attempt to overcome the deficiencies of the simple two-dimensional histogram estimates by “smoothing” them in one way or another. A variety of non-parametric estimation procedures have been suggested, and they are described in detail in [Silverman \(1986\)](#) and [Wand and Jones \(1995\)](#). Here we give a brief description of just one popular class of estimators, namely *kernel density estimators*.

2.5.1 Kernel density estimators

From the definition of a probability density, if the random variable X has a density f ,

$$f(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P(x - h < X < x + h). \quad (2.1)$$

For any given h , a naïve estimator of $P(x - h < X < x + h)$ is the proportion of the observations x_1, x_2, \dots, x_n falling in the interval $(x - h, x + h)$,

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^n I(x_i \in (x - h, x + h)); \quad (2.2)$$

i.e., the number of x_1, \dots, x_n falling in the interval $(x - h, x + h)$ divided by $2hn$. If we introduce a weight function W given by

$$W(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{else,} \end{cases}$$

then the naïve estimator can be rewritten as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} W\left(\frac{x - x_i}{h}\right). \quad (2.3)$$

Unfortunately, this estimator is not a continuous function and is not particularly satisfactory for practical density estimation. It does, however, lead naturally to the kernel estimator defined by

$$\hat{f}(x) = \frac{1}{hn} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad (2.4)$$

where K is known as the *kernel function* and h is the *bandwidth* or *smoothing parameter*. The kernel function must satisfy the condition

$$\int_{-\infty}^{\infty} K(x) dx = 1.$$

Usually, but not always, the kernel function will be a symmetric density function; for example, the normal. Three commonly used kernel functions are rectangular,

$$K(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{else.} \end{cases}$$

triangular,

$$K(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{else,} \end{cases}$$

Gaussian,

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The three kernel functions are implemented in R as shown in Figure 2.12. For some grid \mathbf{x} , the kernel functions are plotted using the R statements in Figure 2.12.

The kernel estimator \hat{f} is a sum of “bumps” placed at the observations. The kernel function determines the shape of the bumps, while the window width h determines their width. Figure 2.13 (redrawn from a similar plot in Silverman 1986) shows the individual bumps $n^{-1}h^{-1}K((x - x_i)/h)$ as well as the estimate \hat{f} obtained by adding them up for an artificial set of data points

```

R> rec <- function(x) (abs(x) < 1) * 0.5
R> tri <- function(x) (abs(x) < 1) * (1 - abs(x))
R> gauss <- function(x) 1/sqrt(2*pi) * exp(-(x^2)/2)
R> x <- seq(from = -3, to = 3, by = 0.001)
R> plot(x, rec(x), type = "l", ylim = c(0,1), lty = 1,
+       ylab = expression(K(x)))
R> lines(x, tri(x), lty = 2)
R> lines(x, gauss(x), lty = 3)
R> legend("topleft", legend = c("Rectangular", "Triangular",
+                               "Gaussian"), lty = 1:3, title = "kernel functions",
+       bty = "n")

```

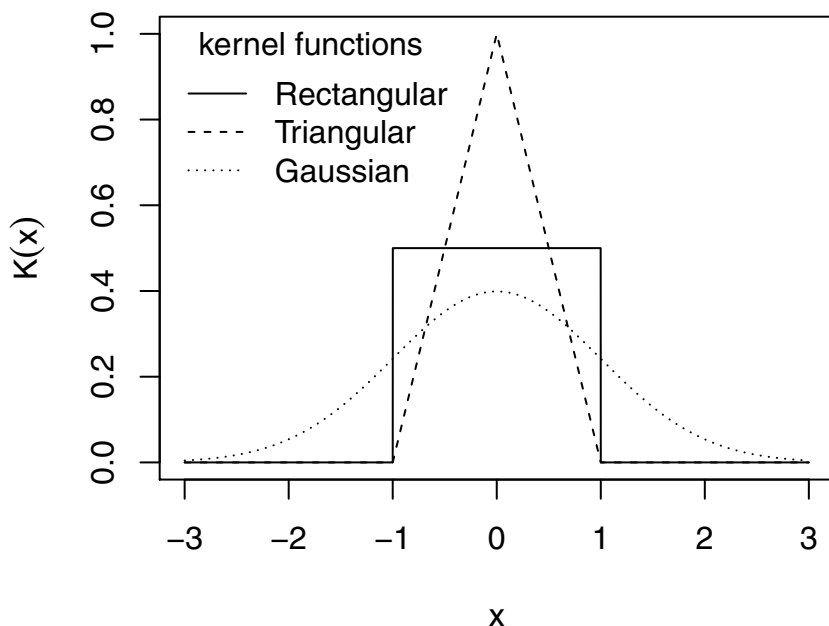


Fig. 2.12. Three commonly used kernel functions.

```

R> x <- c(0, 1, 1.1, 1.5, 1.9, 2.8, 2.9, 3.5)
R> n <- length(x)

```

For a grid

```

R> xgrid <- seq(from = min(x) - 1, to = max(x) + 1, by = 0.01)

```

on the real line, we can compute the contribution of each measurement in \mathbf{x} , with $h = 0.4$, by the Gaussian kernel (defined in Figure 2.12, line 3) as follows:

```
R> h <- 0.4
R> bumps <- sapply(x, function(a) gauss((xgrid - a)/h)/(n * h))
```

A plot of the individual bumps and their sum, the kernel density estimate \hat{f} , is shown in Figure 2.13.

```
R> plot(xgrid, rowSums(bumps), ylab = expression(hat(f)(x)),
+       type = "l", xlab = "x", lwd = 2)
R> rug(x, lwd = 2)
R> out <- apply(bumps, 2, function(b) lines(xgrid, b))
```

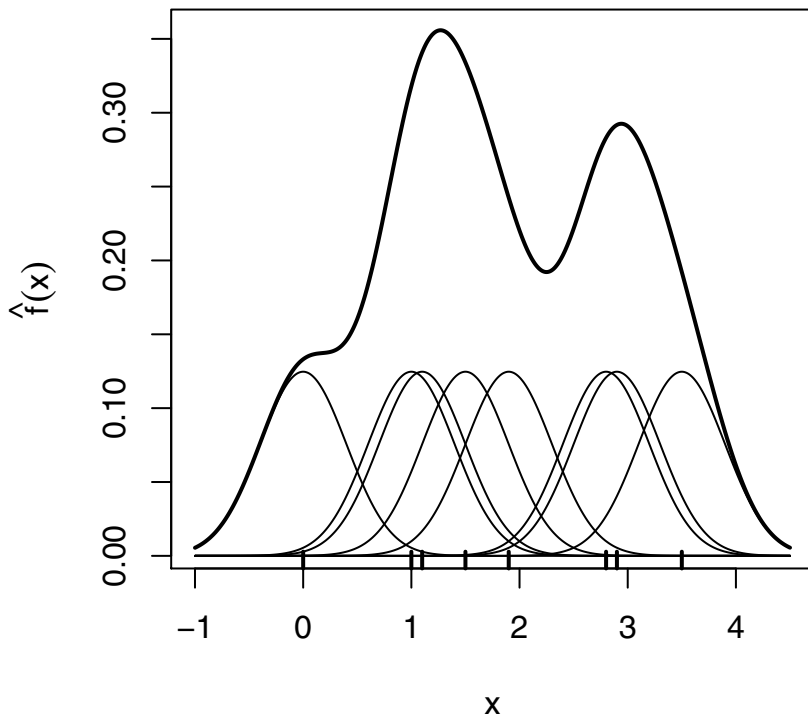


Fig. 2.13. Kernel estimate showing the contributions of Gaussian kernels evaluated for the individual observations with bandwidth $h = 0.4$.

The kernel density estimator considered as a sum of “bumps” centred at the observations has a simple extension to two dimensions (and similarly for more than two dimensions). The bivariate estimator for data (x_1, y_1) , (x_2, y_2) , \dots , (x_n, y_n) is defined as

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^n K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right). \quad (2.5)$$

In this estimator, each coordinate direction has its own smoothing parameter, h_x or h_y . An alternative is to scale the data equally for both dimensions and use a single smoothing parameter.

For bivariate density estimation, a commonly used kernel function is the standard bivariate normal density

$$K(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)}.$$

Another possibility is the bivariate Epanechnikov kernel given by

$$K(x, y) = \begin{cases} \frac{2}{\pi}(1 - x^2 - y^2) & x^2 + y^2 < 1 \\ 0 & \text{else,} \end{cases}$$

which is implemented and depicted in Figure 2.14 by using the `persp` function for plotting in three dimensions.

According to [Venables and Ripley \(2002\)](#), the bandwidth should be chosen to be proportional to $n^{-1/5}$; unfortunately, the constant of proportionality depends on the unknown density. The tricky problem of bandwidth estimation is considered in detail in [Silverman \(1986\)](#).

Our first illustration of enhancing a scatterplot with an estimated bivariate density will involve data from the Hertzsprung-Russell (H-R) diagram of the star cluster CYG OB1, calibrated according to [Vanisma and De Greve \(1972\)](#). The H-R diagram is the basis of the theory of stellar evolution and is essentially a plot of the energy output of stars as measured by the logarithm of their light intensity plotted against the logarithm of their surface temperature. Part of the data is shown in Table 2.1. A scatterplot of the data enhanced by the contours of the estimated bivariate density ([Wand and Ripley 2010](#), obtained with the function `bkde2D()` from the package **KernSmooth**) is shown in Figure 2.15. The plot shows the presence of two distinct clusters of stars: the larger cluster consists of stars that have high surface temperatures and a range of light intensities, and the smaller cluster contains stars with low surface temperatures and high light intensities. The bivariate density estimate can also be displayed by means of a perspective plot rather than a contour plot, and this is shown in Figure 2.16. This again demonstrates that there are two groups of stars.

Table 2.1: CYGOB1 data. Energy output and surface temperature of star cluster CYG OB1.

logst logli		logst logli		logst logli	
4.37	5.23	4.23	3.94	4.45	5.22

Table 2.1: CYGOB1 data (continued).

logst logli		logst logli		logst logli	
4.56	5.74	4.42	4.18	3.49	6.29
4.26	4.93	4.23	4.18	4.23	4.34
4.56	5.74	3.49	5.89	4.62	5.62
4.30	5.19	4.29	4.38	4.53	5.10
4.46	5.46	4.29	4.22	4.45	5.22
3.84	4.65	4.42	4.42	4.53	5.18
4.57	5.27	4.49	4.85	4.43	5.57
4.26	5.57	4.38	5.02	4.38	4.62
4.37	5.12	4.42	4.66	4.45	5.06
3.49	5.73	4.29	4.66	4.50	5.34
4.43	5.45	4.38	4.90	4.45	5.34
4.48	5.42	4.22	4.39	4.55	5.54
4.01	4.05	3.48	6.05	4.45	4.98
4.29	4.26	4.38	4.42	4.42	4.50
4.42	4.58	4.56	5.10		

For our next example of adding estimated bivariate densities to scatterplots, we will use the body measurement data introduced in Chapter 1 (see Table 1.2), although there are rather too few observations on which to base the estimation. (The gender of each individual will not be used.) And in this case we will add the appropriate density estimate to each panel of the scatterplot matrix of the **chest**, **waist**, and **hips** measurements. The resulting plot is shown in Figure 2.17. The **waist/hips** panel gives some evidence that there might be two groups in the data, which, of course, we know to be true, the groups being men and women. And the Waist histogram on the diagonal panel is also *bimodal*, underlining the two-group nature of the data.

2.6 Three-dimensional plots

The scatterplot matrix allows us to display information about the univariate distributions of each variable (using histograms on the main diagonal, for example) and about the bivariate distribution of all pairs of variables in a set of multivariate data. But we should perhaps consider whether the use of three-dimensional plots offers any advantage over the series of two-dimensional scatterplots used in a scatterplot matrix. To begin, we can take a look at the three-dimensional plot of the body measurements data; a version of the plot that includes simply the points along with vertical lines dropped from each point to the x - y plane is shown in Figure 2.18. The plot, produced with the **scatterplot3d** package (Ligges 2010), suggests the presence of two relatively separate groups of points corresponding to the males and females in the data.

```

R> epa <- function(x, y)
+   ((x^2 + y^2) < 1) * 2/pi * (1 - x^2 - y^2)
R> x <- seq(from = -1.1, to = 1.1, by = 0.05)
R> epavals <- sapply(x, function(a) epa(a, x))
R> persp(x = x, y = x, z = epavals, xlab = "x", ylab = "y",
+       zlab = expression(K(x, y)), theta = -35, axes = TRUE,
+       box = TRUE)

```

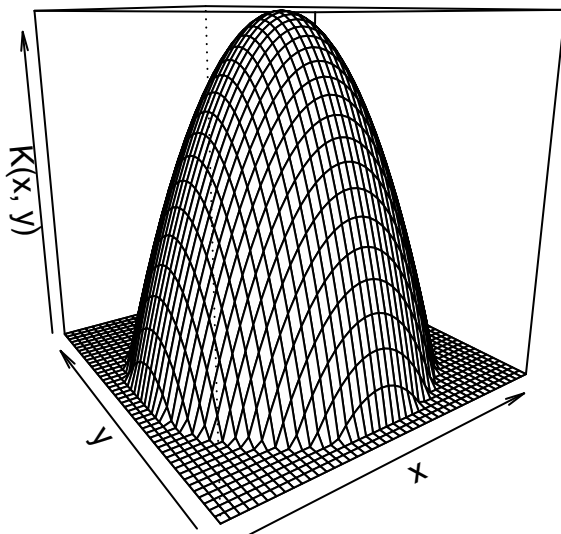


Fig. 2.14. Epanechnikov kernel for a grid between $(-1.1, -1.1)$ and $(1.1, 1.1)$.

As a second example of using a three-dimensional plot, we can look at `temp`, `wind`, and `S02` from the air pollution data. The points and vertical lines versions of the required three-dimensional plot are shown in Figure 2.19. Two observations with high `S02` levels stand out, but the plot does not appear to add much to the bubble plot for the same three variables (Figure 2.7).

Three-dimensional plots based on the original variables can be useful in some cases but may not add very much to, say, the bubble plot of the scatterplot matrix of the data. When there are many variables in a multivariate


```
R> library("KernSmooth")
R> CYGOB1d <- bkde2D(CYGOB1, bandwidth = sapply(CYGOB1, dpik))
R> plot(CYGOB1, xlab = "log surface temperature",
+       ylab = "log light intensity")
R> contour(x = CYGOB1d$x1, y = CYGOB1d$x2,
+         z = CYGOB1d$fhat, add = TRUE)
```

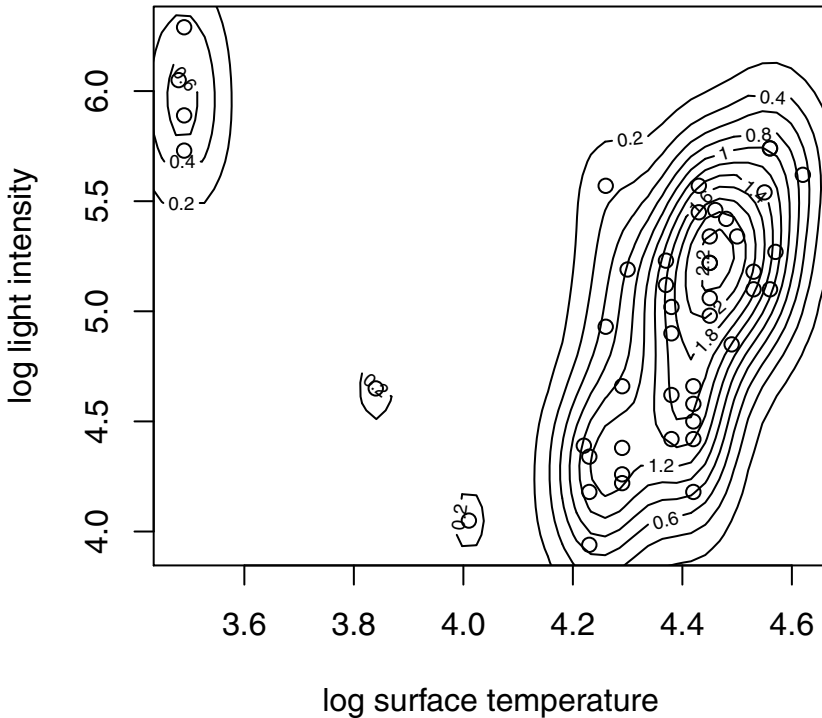


Fig. 2.15. Scatterplot of the log of light intensity and log of surface temperature for the stars in star cluster CYG OB1 showing the estimated bivariate density.

data set, there will be many possible three-dimensional plots to look at and integrating and linking all the plots may be very difficult. But if the dimensionality of the data could be reduced in some way with little loss of information, three-dimensional plots might become more useful, a point to which we will return in the next chapter.

```
R> persp(x = CYGOB1d$x1, y = CYGOB1d$x2, z = CYGOB1d$fhat,
+        xlab = "log surface temperature",
+        ylab = "log light intensity",
+        zlab = "density")
```

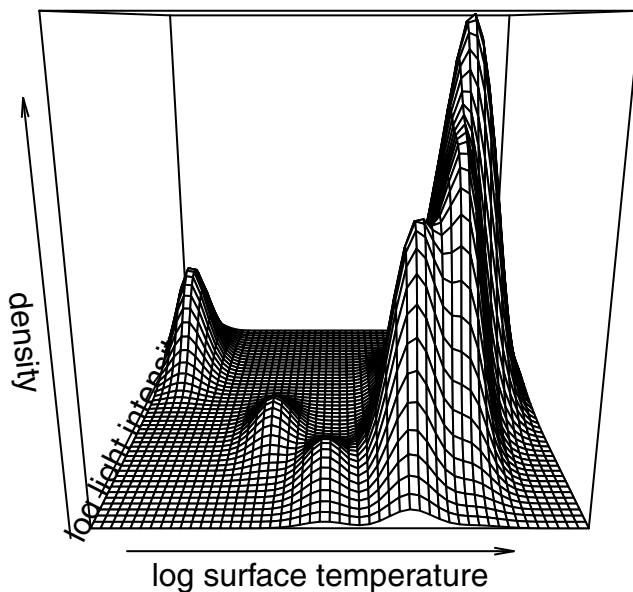


Fig. 2.16. Perspective plot of estimated bivariate density.

2.7 Trellis graphics

Trellis graphics (see [Becker, Cleveland, Shyu, and Kaluzny 1994](#)) is an approach to examining high-dimensional structure in data by means of one-, two-, and three-dimensional graphs. The problem addressed is how observations of one or more variables depend on the observations of the other variables. The essential feature of this approach is the *multiple conditioning* that allows some type of plot to be displayed for different values of a given variable (or variables). The aim is to help in understanding both the structure of the data and how well proposed models describe the structure. An example of the application of trellis graphics is given in [Verbyla, Cullis, Kenward, and](#)


```
R> library("scatterplot3d")
R> with(measure, scatterplot3d(chest, waist, hips,
+                             pch = (1:2)[gender], type = "h", angle = 55))
```

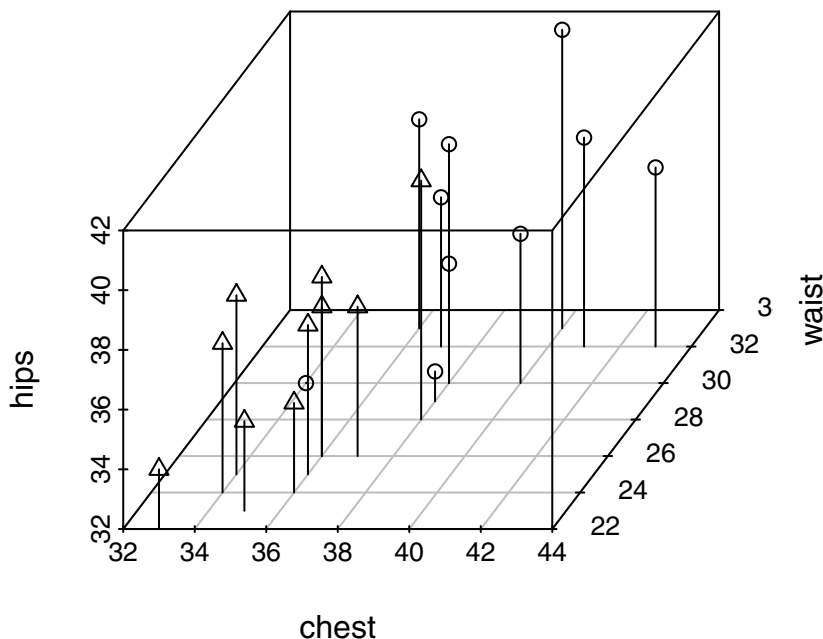


Fig. 2.18. A three-dimensional scatterplot for the body measurements data with points corresponding to male and triangles to female measurements.

the data? Probably not, as there are few points in each of the three, three-dimensional displays. This is often a problem with multipanel plots when the sample size is not large.

For the last example in this section, we will use a larger data set, namely data on earthquakes given in [Sarkar \(2008\)](#). The data consist of recordings of the location (latitude, longitude, and depth) and magnitude of 1000 seismic events around Fiji since 1964.

In Figure 2.22, scatterplots of latitude and longitude are plotted for three ranges of depth. The distribution of locations in the latitude-longitude space is seen to be different in the three panels, particularly for very deep quakes. In

```
R> with(USairpollution,
+       scatterplot3d(temp, wind, SO2, type = "h",
+                     angle = 55))
```

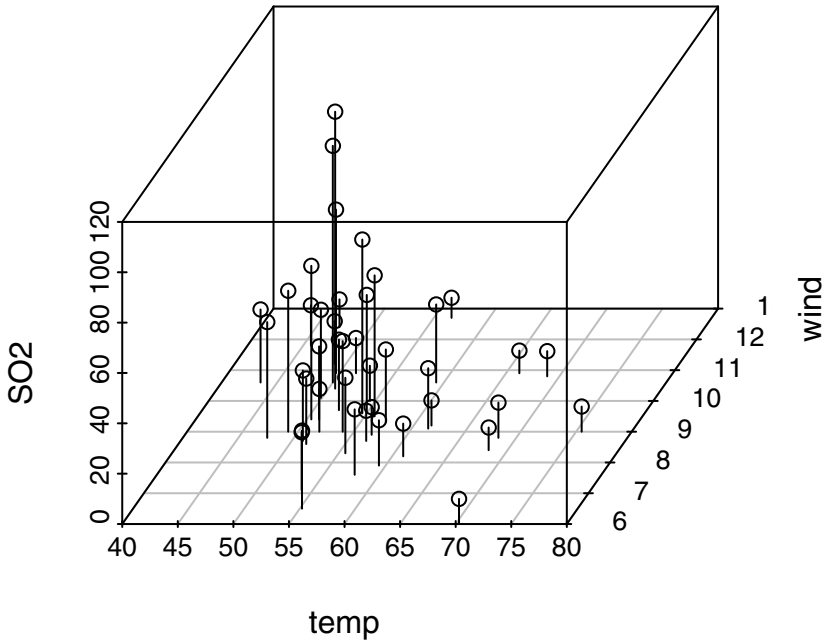


Fig. 2.19. A three-dimensional scatterplot for the air pollution data.

Figure 2.23 (a tour de force by Sarkar) the four panels are defined by ranges of magnitude and depth is encoded by different shading.

Finally, in Figure 2.24, three-dimensional scatterplots of earthquake epicentres (latitude, longitude, and depth) are plotted conditioned on earthquake magnitude. (Figures 2.22, 2.23, and 2.24 are reproduced with the kind permission of Dr. Deepayan Sarkar.)

2.8 Stalactite plots

In this section, we will describe a multivariate graphic, the stalactite plot, specifically designed for the detection and identification of multivariate out-

```
R> plot(xyplot(SO2 ~ temp| cut(wind, 2), data = USairpollution))
```

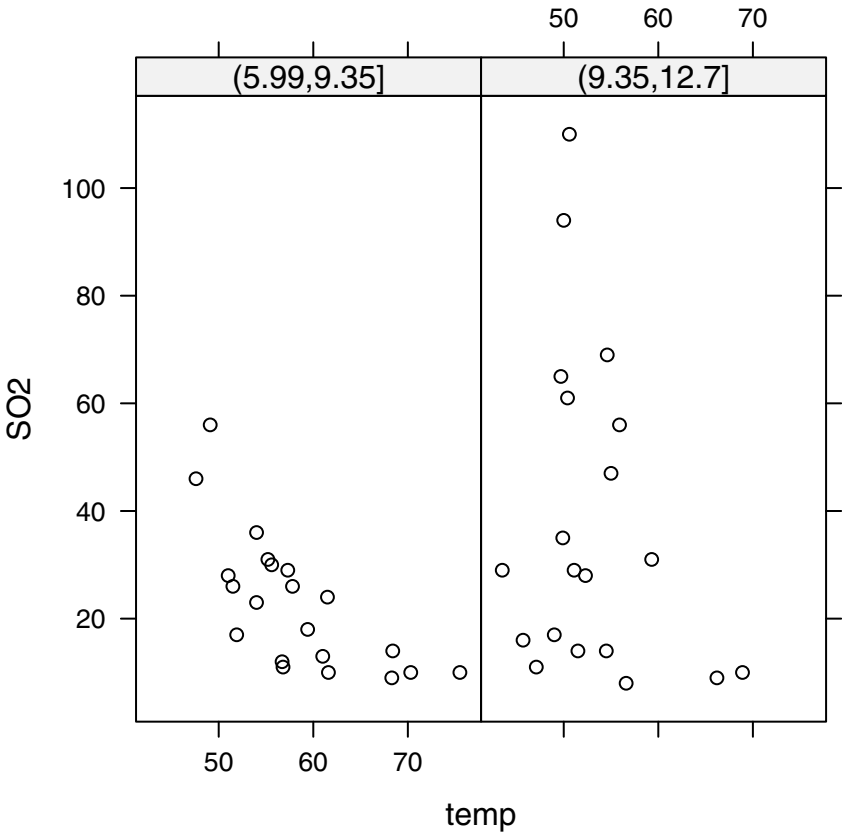


Fig. 2.20. Scatterplot of SO2 and temp for light and high winds.

liers. Like the chi-square plot for assessing multivariate normality, described in Chapter 1, the stalactite plot is based on the generalised distances of observations from the multivariate mean of the data. But here these distances are calculated from the means and covariances estimated from increasing-sized subsets of the data. As mentioned previously when describing bivariate boxplots, the aim is to reduce the masking effects that can arise due to the influence of outliers on the estimates of means and covariances obtained from all the data. The central idea of this approach is that, given distances using, say, m observations for estimation of means and covariances, the $m + 1$ observations to be used for this estimation in the next stage are chosen to be those with the $m + 1$ smallest distances. Thus an observation can be included in the subset used for estimation for some value of m but can later be excluded as m increases. Initially m is chosen to take the value $q + 1$, where q is the number of variables in the multivariate data set because this is the smallest number

```
R> pollution <- with(USairpollution, equal.count(SO2,4))
R> plot(cloud(precip ~ temp * wind | pollution, panel.aspect = 0.9,
+          data = USairpollution))
```

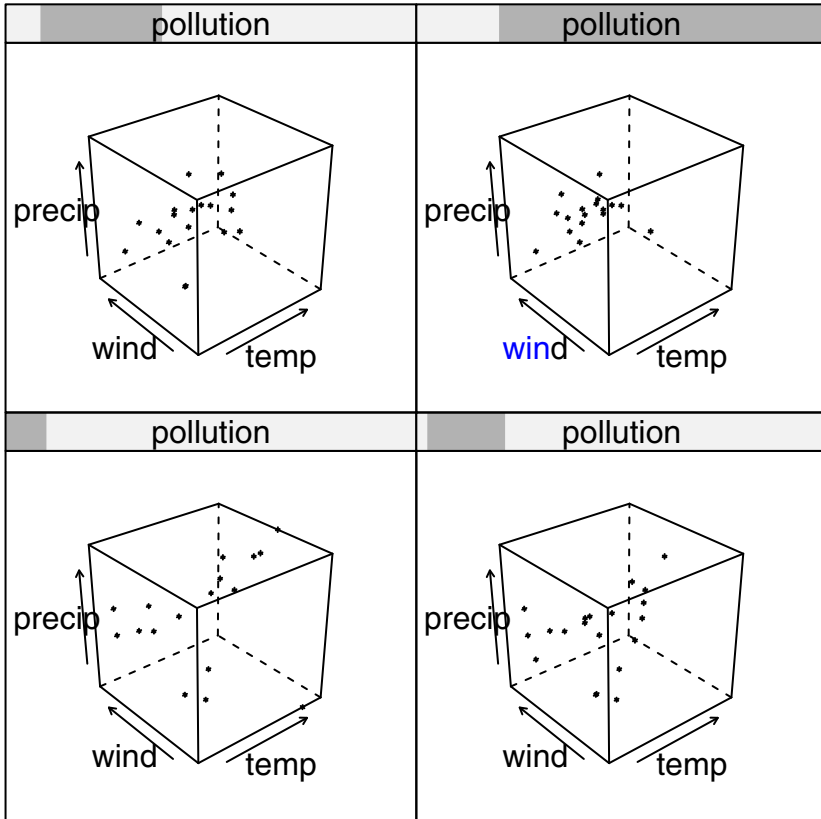


Fig. 2.21. Three-dimensional plots of `temp`, `wind`, and `precip` conditioned on levels of `SO2`.

allowing the calculation of the required generalised distances. The cutoff distance generally employed to identify an outlier is the maximum expected value from a sample of n random variables each having a chi-squared distribution on q degrees of freedom. The stalactite plot graphically illustrates the evolution of the outliers as the size of the subset of observations used for estimation increases. We will now illustrate the application of the stalactite plot on the US cities air pollution data. The plot (produced via `stalac(USairpollution)`) is shown in Figure 2.25. Initially most cities are indicated as outliers (a “*” in the plot), but as the number of observations on which the generalised distances are calculated is increased, the number of outliers indicated by the plot decreases. The plot clearly shows the outlying nature of a number of cities over

```
R> plot(xyplot(lat ~ long| cut(depth, 3), data = quakes,  
+           layout = c(3, 1), xlab = "Longitude",  
+           ylab = "Latitude"))
```

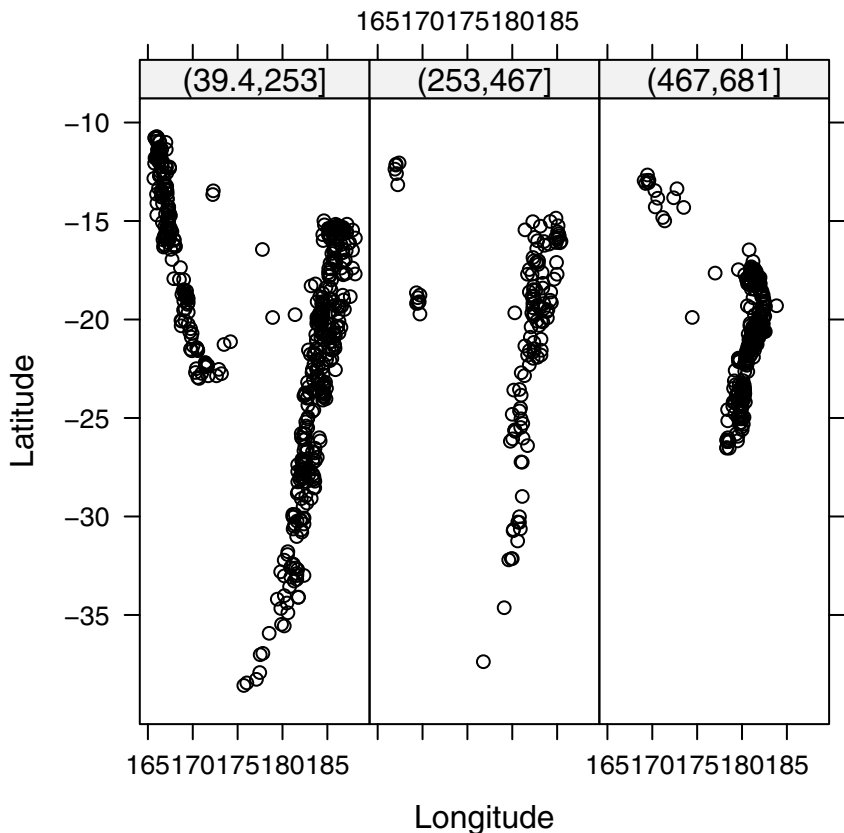


Fig. 2.22. Scatterplots of latitude and longitude conditioned on three ranges of depth.

nearly all values of m . The effect of masking is also clear; when all 41 observations are used to calculate the generalised distances, only observations Chicago, Phoenix, and Providence are indicated to be outliers.

2.9 Summary

Plotting multivariate data is an essential first step in trying to understand the story they may have to tell. The methods covered in this chapter provide just some basic ideas for taking an initial look at the data, and with software such as R there are many other possibilities for graphing multivariate obser-

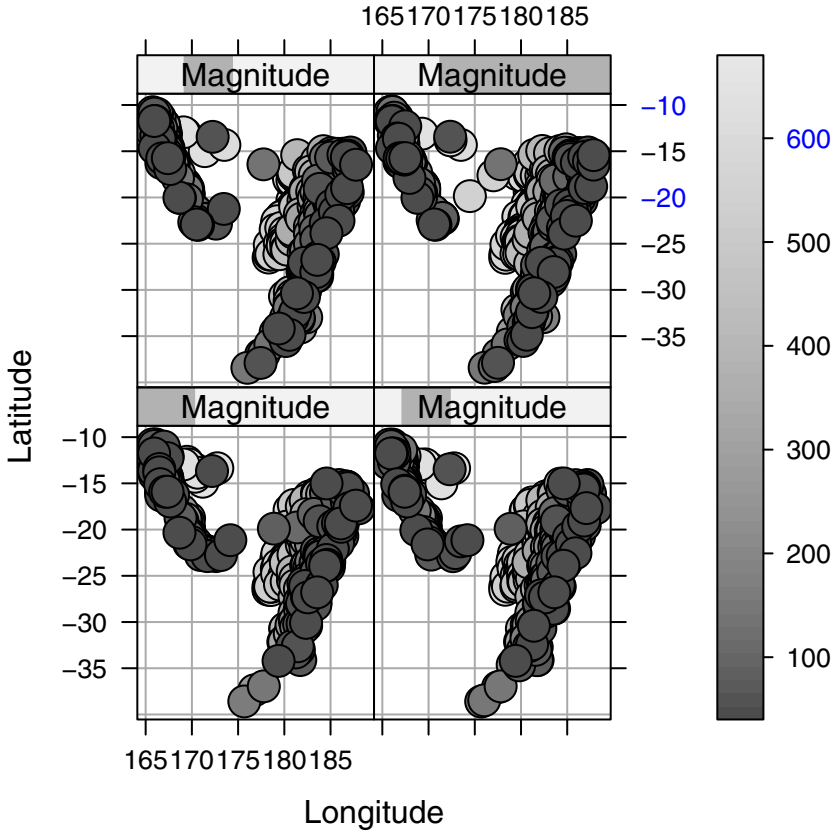


Fig. 2.23. Scatterplots of latitude and longitude conditioned on magnitude, with depth coded by shading.

variations, and readers are encouraged to explore more fully what is available. But graphics can often flatter to deceive and it is important not to be seduced when looking at a graphic into responding “what a great graph” rather than “what interesting data”. A graph that calls attention to itself pictorially is almost surely a failure (see [Becker et al. 1994](#)), and unless graphs are relatively simple, they are unlikely to survive the first glance. Three-dimensional plots and trellis plots provide great pictures, which may often also be very informative (as the examples in [Sarkar 2008](#), demonstrate), but for multivariate data with many variables, they may struggle. In many situations, the most useful graphic for a set of multivariate data may be the scatterplot matrix, perhaps with the panels enhanced in some way; for example, by the addition of bivariate density estimates or bivariate boxplots. And all the graphical approaches discussed in this chapter may become more helpful when applied to the data

```
R> plot(cloud(depth ~ lat * long / Magnitude, data = quakes,
+           xlim = rev(range(quakes$depth)),
+           screen = list(z = 105, x = -70), panel.aspect = 0.9,
+           xlab = "Longitude", ylab = "Latitude", zlab = "Depth"))
```

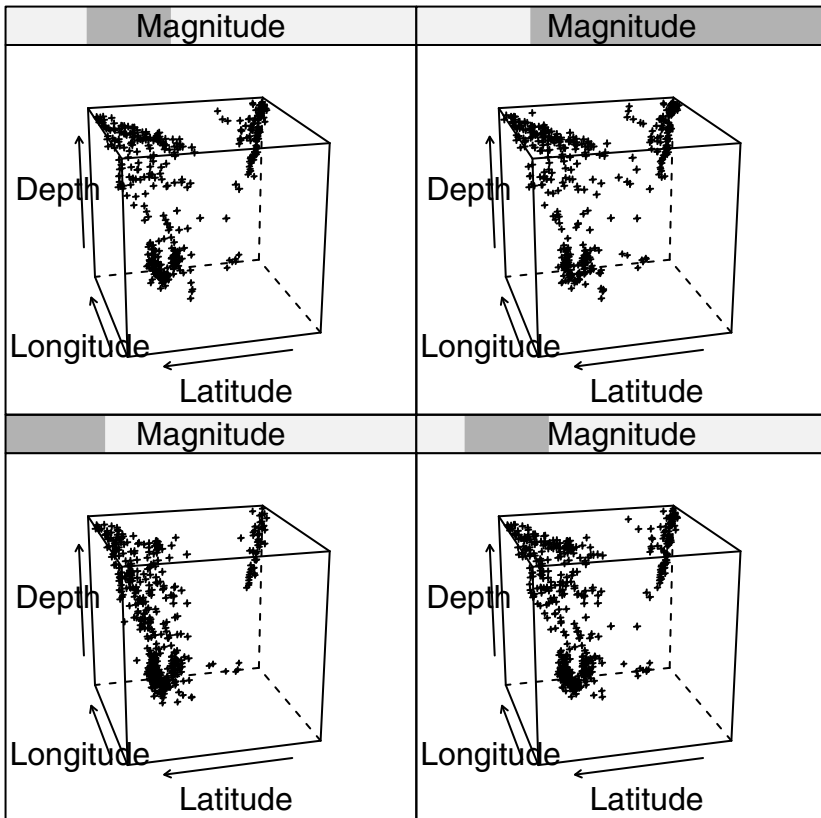


Fig. 2.24. Scatterplots of latitude and longitude conditioned on magnitude.

after their dimensionality has been reduced in some way, often by the method to be described in the next chapter.

Number of observations used for estimation

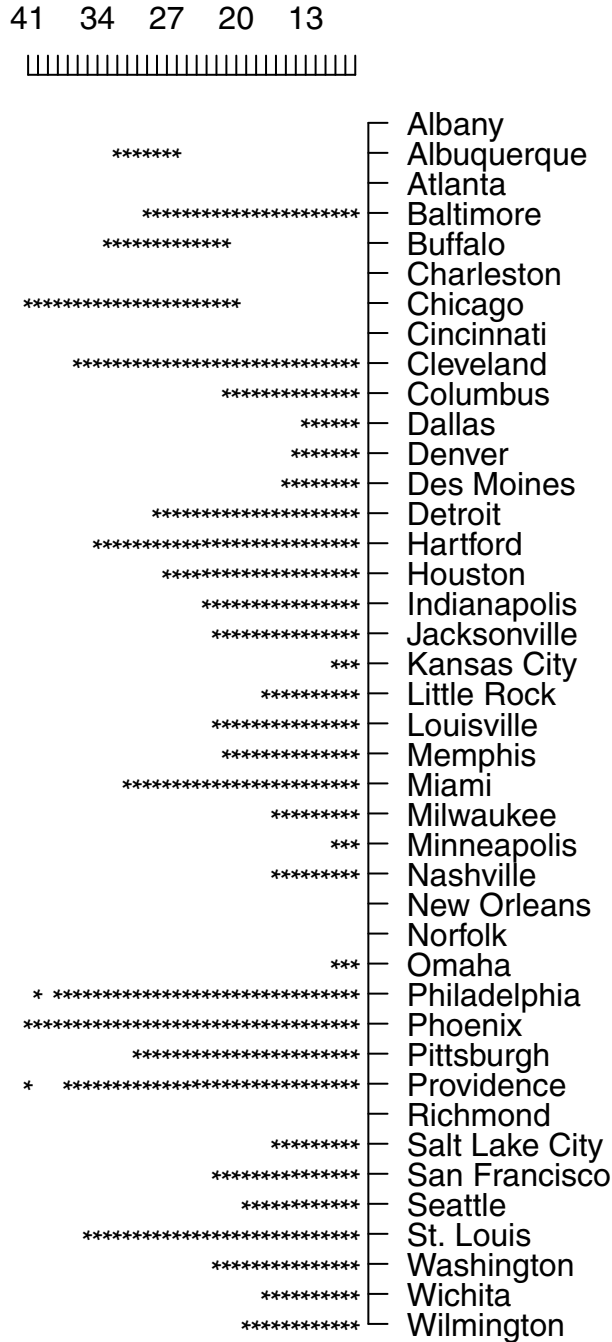


Fig. 2.25. Stalactite plot of US cities air pollution data.

2.10 Exercises

- Ex. 2.1 Use the bivariate boxplot on the scatterplot of each pair of variables in the air pollution data to identify any outliers. Calculate the correlation between each pair of variables using all the data and the data with any identified outliers removed. Comment on the results.
- Ex. 2.2 Compare the chi-plots with the corresponding scatterplots for each pair of variables in the air pollution data. Do you think that there is any advantage in the former?
- Ex. 2.3 Construct a scatterplot matrix of the body measurements data that has the appropriate boxplot on the diagonal panels and bivariate boxplots on the other panels. Compare the plot with Figure 2.17, and say which diagram you find more informative about the data.
- Ex. 2.4 Construct a further scatterplot matrix of the body measurements data that labels each point in a panel with the gender of the individual, and plot on each scatterplot the separate estimated bivariate densities for men and women.
- Ex. 2.5 Construct a scatterplot matrix of the chemical composition of Romano-British pottery given in Chapter 1 (Table 1.3), identifying each unit by its kiln number and showing the estimated bivariate density on each panel. What does the resulting diagram tell you?
- Ex. 2.6 Construct a bubble plot of the earthquake data using latitude and longitude as the scatterplot and depth as the circles, with greater depths giving smaller circles. In addition, divide the magnitudes into three equal ranges and label the points in your bubble plot with a different symbol depending on the magnitude group into which the point falls.

<http://www.springer.com/978-1-4419-9649-7>

An Introduction to Applied Multivariate Analysis with R

Everitt, B.; Hothorn, T.

2011, XIV, 274 p. 92 illus., Softcover

ISBN: 978-1-4419-9649-7