**THE UNIVERSITY OF CHICAGO**
**Booth School of Business**
Business 41912, Spring Quarter 2020, Mr. Ruey S. Tsay

**Lecture: Visualization of Multivariate Data**

This note provides some visualization methods for multivariate data analysis. For convenience, some data sets from the R package **MVA** are used. Besides **MVA**, the following packages are also used: **lattice**, **scatterplot3d**, **rgl**, **KernSmooth** and their related packages. You can find information of R packages from R CRAN under Packages.
There are several books on graphics in R. For multivariate data analysis, Chapter 2 of *An Introduction to Applied Multivariate Analysis with R* by B. Everitt and T. Hothorn (2011, Springer) is a good reference. We adopt some methods from that chapter. [The book might be available online.]
The methods discussed include

1. boxplot and bvbox:
2. matrix scatterplot: `pairs`
3. convex hull:
4. chi-plot:
5. bubble and glyph plots:
6. contour and perspective plots
7. Trellis graphics
8. scatterplot3d
9. stalactite plot:

# 1 Examples used

The data sets used are from either the **MVA** package or the textbook. They are briefly described below:

1. Paper quality measurements: Table 1-2 of the textbook. 41 observations and three variables, namely density, machine-direction and cross-direction.

2. USairpollution: U.S. air pollution data in the **MVA** package. 41 observations (cities) and 7 variables (SO2, temp, manu, popul, wind, precip, predays). Use the command ?USairpollution to obtain description of the data. For instance, predays is the average number of days with precipitation per year and manu is the number of manufacturing enterprises employing 20 or more workers.

3. COGOB1: Energy output and surface temperature for Star cluster CYG OB1 of package **HSAUR2** and available in **MVA**. 47 observations and 2 variables (logst and logli), log surface temperature and log light intensity.

4. quakes: Locations of 1000 seismic events of MB $>$ 4.0. The events occurred in a cube near Fiji since 1964. Available in **MVA**. The variables are lat (latitude), long (longitude), depth (km), mag (Richter magnitude), and stations (number of stations reporting).

## 2 The boxplot and bvbox

`boxplot` is used to summarize the basic features of the empirical density function of a scalar variable. `bvbox` is bivariate boxplot. For the paper-quality data, Figure 1 shows the boxplot of the density, Figure 2 shows the boxplots of all three variables, and Figure 3 shows the bivariate boxplot of paper strength (machine-direction and cross-direction).

**R commands used**:

```
> setwd("C:/Users/rst/teaching/ama/sp2020")  ## set my working directory
> require(MVA)
> y <- read.table("T1-2.DAT")
> colnames(y) <- c("density","mach-dir","cross-dir")
> boxplot(y[,1])
> boxplot(y)
> bvbox(y[,2:3],xlab="mach-dir",ylab="cross-dir")
```

## 3 Matrix scatterplot

Scatterplot is useful to show the relationship between two variables. The plot can be enhanced by adding some marginal information such as marginal distributions or marginal histograms. Figure 4 shows the scatterplot of `popul` versus `manu` of the US air pollution data. In addition, marginal distributions of the variables are also shown using the command `rug`. Figure 5 shows the scatterplot of `popul` versus `manu` as before. However, we use the command `layout` to create space for marginal histogram and marginal boxplot.

Matrix scatterplot provides an overall view of the relationships between all pairs of variables. For a moderate dimension $p$, the plot is informative. Consider the US air pollution data,

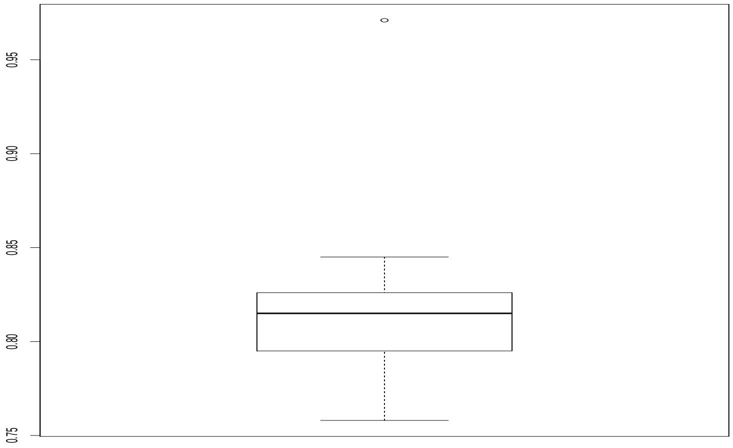

Figure 1: Boxplot of paper density

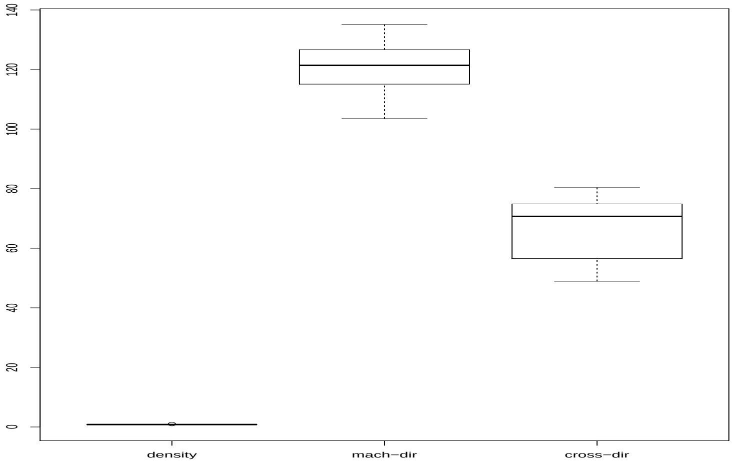

Figure 2: Boxplots of paper quality measurements

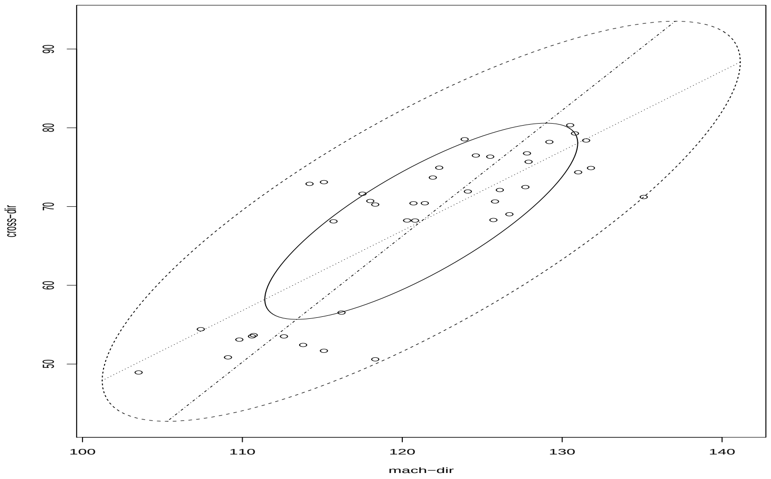

Figure 3: Bivariate boxplot of paper strength

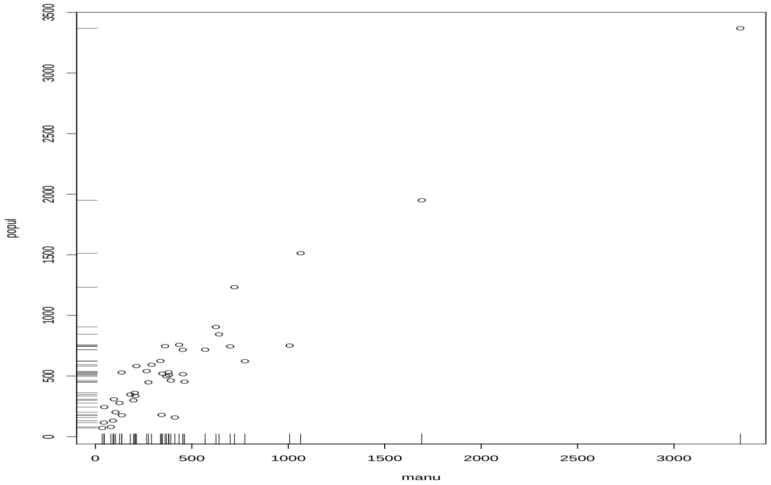

Figure 4: Scatterplot of `popul` versus `manu` with marginal distributions for U.S. air pollution data

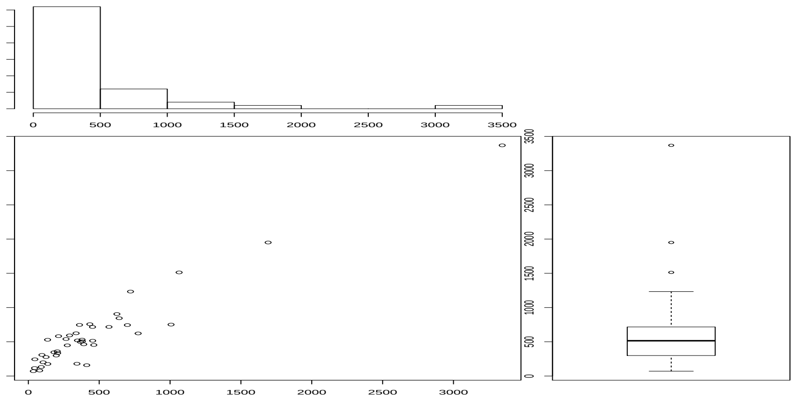

Figure 5: Scatterplot of `popul` versus `manu` with marginal histogram (manu) and boxplot (popul) for U.S. air pollution data

Figure 6 shows the simple matrix scatterplot. The plotting character and size can be changed by using the subcommands `pch` and `cex`. From the plots, it is seen, for instance, that `manu` and `popul` are positively correlated and there exists a large value in `manu`. One can impose a linear regression line on each of the scatterplot to aid visualization. See Figure 7.

**R commands used**:

```
> plot(popul˜manu,data=USairpollution)
> rug(USairpollution$manu,side=1)
> rug(USairpollution$popul,side=2)
> par(mar=c(2,1,1,1))  ## set margins for plots
> layout(matrix(c(2,0,1,3),nrow=2,byrow=T),widths=c(2,1),heights=c(1,2),respect=T)
> plot(popul˜manu,data=USairpollution)
> with(USairpollution,hist(manu,main=""))
> with(USairpollution,boxplot(popul))

> pairs(USairpollution)
> pairs(USairpollution,pch=’’.’’,cex=0.5) ## plot not shown
> pairs(USairpollution,panel=function(x,y){points(x,y); abline(lm(y˜x),col="red")})
```

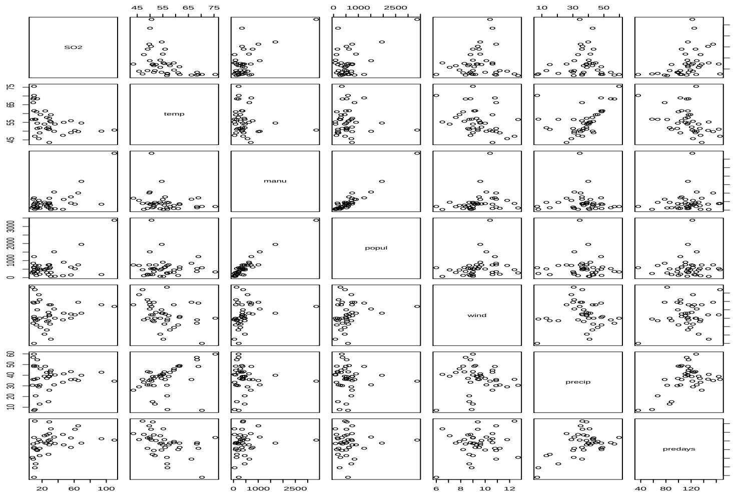

Figure 6: Matrix scatterplot of U.S. air pollution data

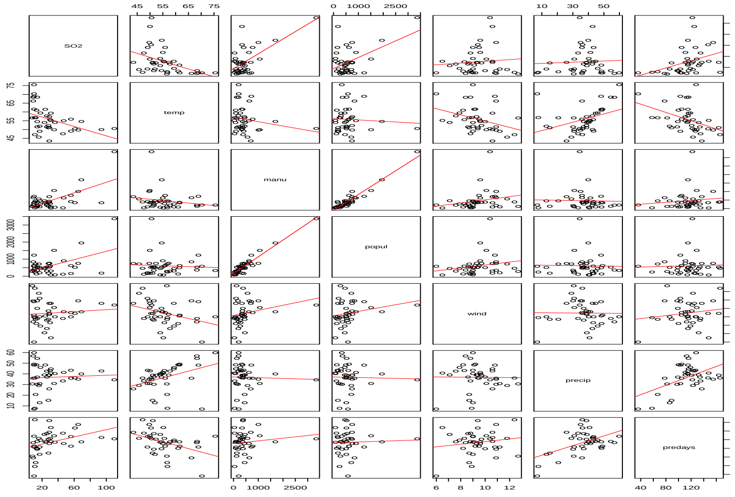

Figure 7: Matrix scatterplot of U.S. air pollution data with least squares lines

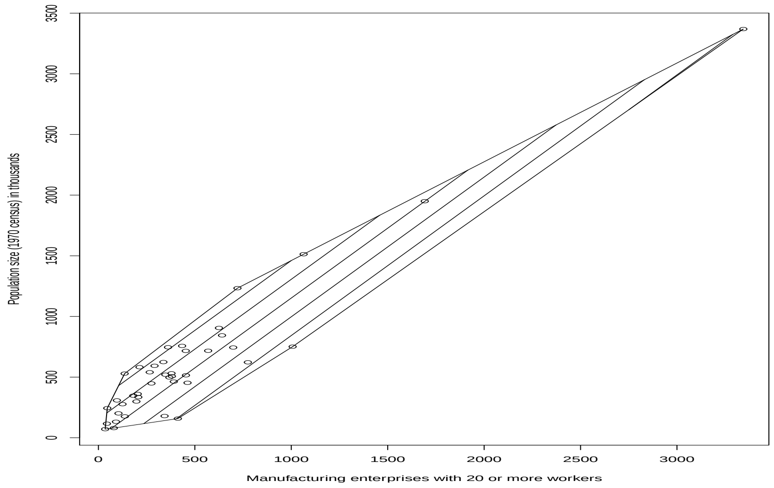

Figure 8: Convex hull of `manu` and `popul` of U.S. air pollution data

# 4 Convex hull

In some applications, one might be interested in seeing the convex hull of two variables generated by the observed data. This can be done by a two-step procedure. First, observations defining the hull can be identified via the command `chull`. Obviously, those points are the boundary points in the scatterplot. In the second step, one can draw the convex hull via scatterplot. For example, Figure 8 shows the convex hull of `manu` and `popul` of the air pollution data.

**R commands used**:

```
> hull <- with(USairpollution,chull(manu,popul))
> hull
[1]  9 15 41  6  2 18 16 14  7
> maxis <- c("Manufacturing enterprises with 20 or more workers") #define axises
> paxis <- c("Population size (1970 census) in thousands")
> with(USairpollution,plot(manu,popul,xlab=maxis,ylab=paxis))
> with(USairpollution,polygon(manu[hull],popul[hull],density=15,angle=30))
```

# 5 Chi-plot

The `chi-plot` suggested by Fisher and Switzer (1985, Biometrika) is designed to check the independence between two continuous random variables. The basic idea is that the joint distribution of two independent random variables is the product of the marginal distributions.

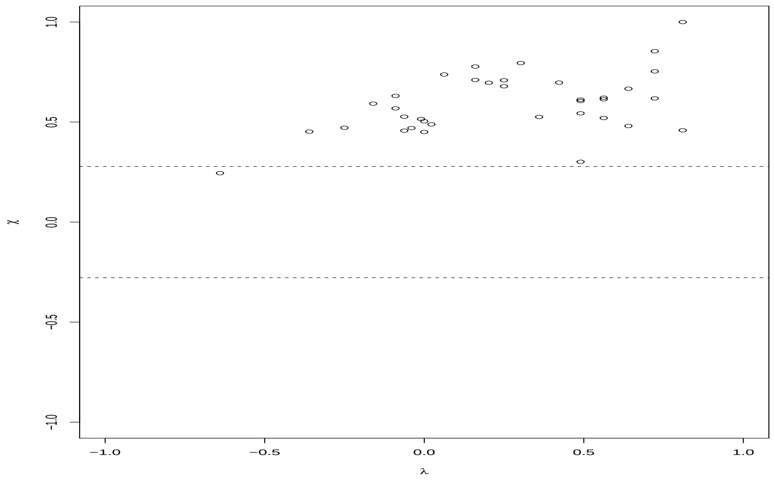

Figure 9: Chi-plot of `manu` and `popul` of U.S. air pollution data

This fact can be checked using the idea of chi-square test for a $2 \times 2$ table. Consider a random sample of $n$ observations, say $\{(x_i, y_i)|i = 1, \ldots, n\}$. For each data point $(x_i, y_i)$, define

$$\begin{aligned}
\chi_i &= \frac{H_i - F_i G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}, \\
F_i &= \frac{1}{n-1}\sum_{j \neq i} I(x_j \leq x_i), \\
G_i &= \frac{1}{n-1}\sum_{j \neq i} I(y_j \leq y_i), \\
H_i &= \frac{1}{n-1}\sum_{j \neq i} I(x_j \leq x_i,\ y_j \leq y_i), \\
\lambda_i &= 4\mathrm{sign}_i \max\{(F_i - 0.5)^2, (G_i - 0.5)^2\},
\end{aligned}$$

where $I(A)$ is the indicator function of $A$ and $\mathrm{sign}_i = \mathrm{sign}(F_i - 0.5)(G_i - 0.5)$. The chi-plot is the scatterplot of $(\lambda_i, \chi_i)$ for all $|\lambda_i| < 4[1/(n-1) - 0.5]^2$, with confidence interval given by

$$\left(-\frac{c_p}{\sqrt{n}}, \frac{c_p}{\sqrt{n}}\right),$$

where $c_p = 1.54, 1.78$ and $2.18$ corresponding to $p = 0.9$, $0.95$, and $0.99$, respectively. In short, under independence, we expect roughly $p$ proportion of points are to be within the two horizontal bounds. See Fisher and Switzer (2001, American Statistician) for details. Figure 9 shows the chi-plot for `manu` and `popul` of the air pollution data. Clearly, the independence assumption is rejected as most of the points are outside the two bounds.

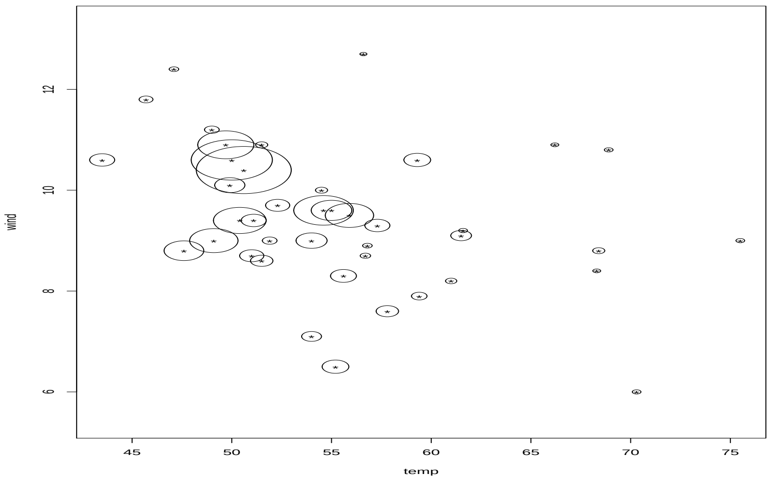

Figure 10: Bubble of `wind` versus `temp` with `SO2` of U.S. air pollution data

**R command used**:

```
> with(USairpollution,chiplot(manu,popul))
```

# 6 Bubble and other glyph plots

Bubble and glyph plots are designed to show multiple variables in a scatterplot. Suppose that we have $p > 2$ variables of interest, say $(X_1, \ldots, X_p)$. We can consider the scatterplot of $X_1$ versus $X_2$. To add information of variables $X_3, \ldots, X_p$, we can impose an object on each point of the scatterplot. The number of edges of the object is $p-2$ with length of each edge representing one of the variables $X_3, \ldots, X_p$. For instance, if $p = 3$, one can use a cycle to represent $X_3$ with the radius of the circle proportion to the magnitude of $X_3$. Such a plot is called `bubble plot`. The choices of object include face (Chernoff, 1973, JASA), star, etc. Figure 10 shows a bubble plot of average annual wind speed (m.p.h.) versus average annual temperature (Fahrenheit) with `SO2` showing as bubbles. For the plot, it is seen that large measurements of SO2 seem to occur around temperature between 47 and 55 and wind speed around 10. Figure 11 shows a star plot of the US air pollution data. Note that columns 2 and 5 are removed from the `stars` command because they are used in the scatterplot.

**R commands used**:

```
> ylim <- with(USairpollution,range(wind))*c(0.9,1.05)
> plot(wind˜temp,data=USairpollution,xlab="temp",ylab="wind",ylim=ylim,pch="*")
> with(USairpollution,symbols(temp,wind,circles=SO2,inches=0.5,add=TRUE))
> plot(wind˜temp,data=USairpollution,xlab="temp",ylab="wind",ylim=ylim,pch="*")
```

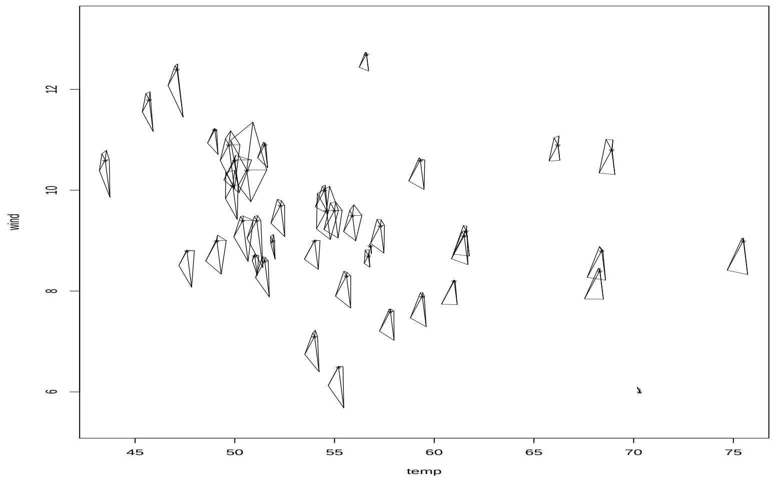

Figure 11: Star plot of `wind` versus `temp` with other variables forming five-sided stars for U.S. air pollution data

```
>with(USairpollution,stars(USairpollution[,c(2,5)],locations=cbind(temp,wind),
     labels=NULL,add=TRUE,cex=0.5))
```

# 7 Contour and perspective plots

In some applications, one can enhance the scatterplot with empirical density of the variables of interest. The empirical density function can be obtained by kernel smoothing. In R, the package **KernSmooth** provides density estimation. We use the CYG OB1 dataset of **MVA** to demonstrate the plots. Figure 12 shows the contour plot of log light intensity versus log surface temperature. The contour is based on a 2-dimensional density estimated by Gaussian kernel with direct plug-in optimal bandwidth. See the subcommand `sapply(CYGOB1,dpik)` in the command `bkde2D`, bivariate kernel density estimation. The plot shows that the distribution is not unimodal, indicating there are two main clusters of stars. Figure 13 shows the associated perspective plot, which provides a 3-dimensional view of the empirical density function.

**R commands used**:

```
> require(KernSmooth)
### 2-dimensional density estimation
> cygob1d <- bkde2D(CYGOB1,bandwidth = sapply(CYGOB1,dpik))
> plot(CYGOB1,xlab="Log surface temperature",ylab="Log light intensity")
> contour(x=cygob1d$x1,y=cygob1d$x2,z=cygob1d$fhat,add=TRUE)
> persp(x=cygob1d$x1,y=cygob1d$x2,z=cygob1d$fhat,xlab="Log surface temperature",
```

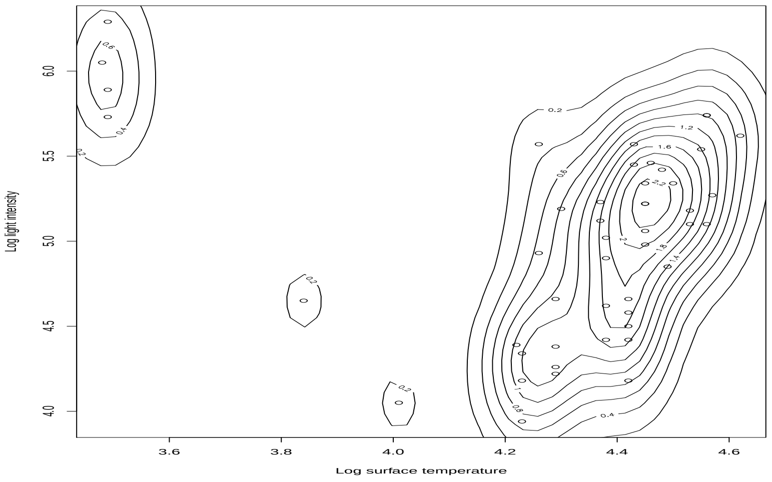

Figure 12: Contour plot for CYG OB1 data

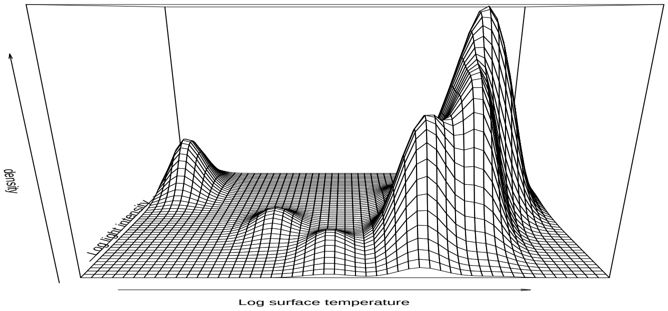

Figure 13: Perspective plot for CYG OB1 data

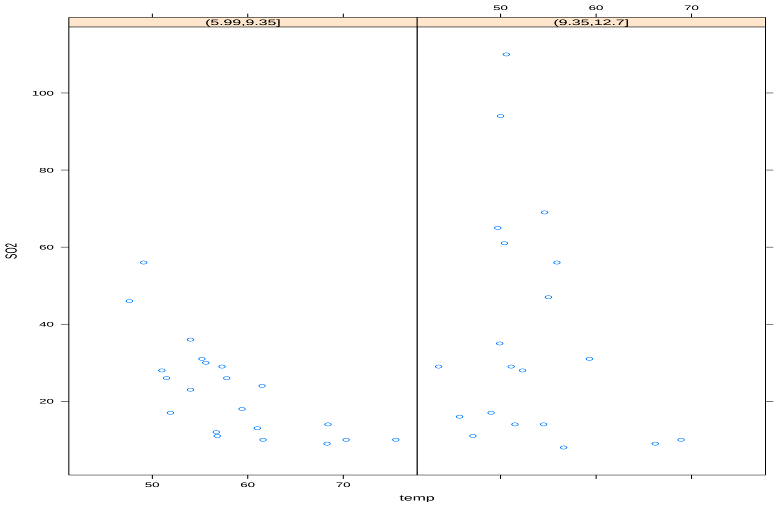

Figure 14: Scatterplot of SO2 versus temperature for light and high winds. Air pollution data

```
    ylab='Log light intensity',zlab="density")
```

# 8 Trellis graphics

Trellis graphics of Becker, Cleveland, Shru and Kaluzny (1994) provides an approach to examining high-dimensional structure in data by using 1-, 2-, and 3-dimensional graphics. The plot can be thought of multiple conditionings to see the high-dimensional structure. The **lattice** package of R can be used to produce trellis graphics. Figure 14 shows the scatterplot of SO2 versus temperature for light and high winds of the US air pollution data. As a second example, consider the dependence of `precip` on `wind` and `temp` in the U.S. air pollution data conditioned on `SO2`. Figure 15 shows the 3-dimensional plots of `precip` versus (`wind`, `temp`) when `SO2` is divided into four groups of equal size. Finally, Figure 16 shows the scatterplots of earth quakes conditioned on depth.

**R commands used**:

```
> require(lattice)
> plot(xyplot(SO2~temp|cut(wind,2),data=USairpollution))
> pollu <- with(USairpollution,equal.count(SO2,4))
> plot(cloud(precip~temp*wind|pollu,panel.aspect=0.9,data=USairpollution))
> plot(xyplot(lat~long|cut(depth,3),data=quakes,layout=c(3,1),xlab="Longitude",ylab=Lati
```

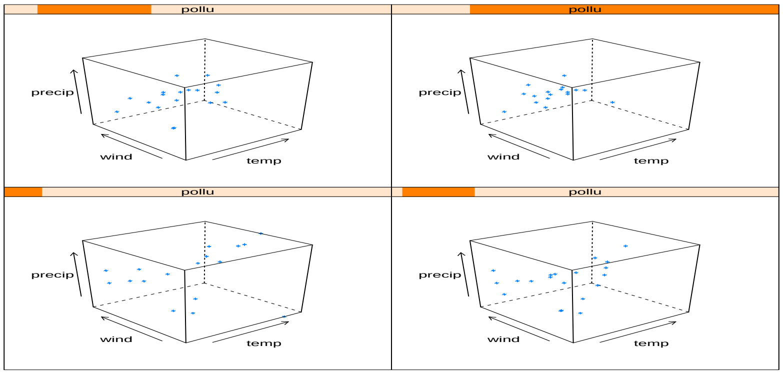

Figure 15: 3-D plot of `precip` versus (`wind`,`temp`) conditioned on size of `SO2`: U.S. air pollution data

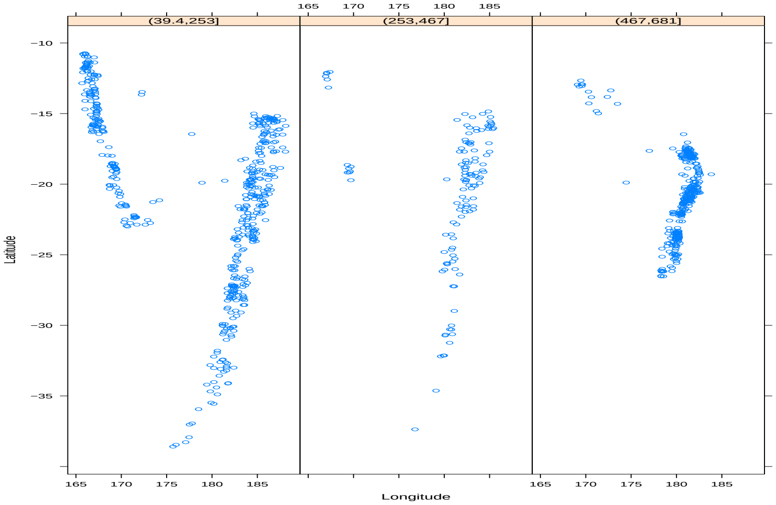

Figure 16: Scatterplots of earth quake locations conditioned on depth: earth quake data

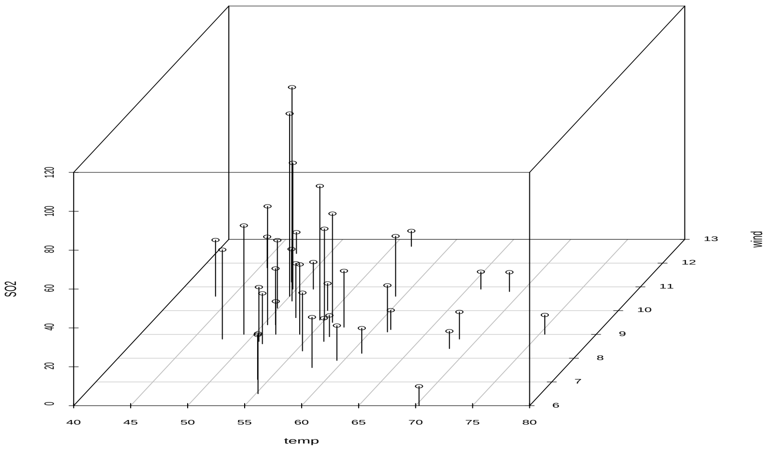

Figure 17: 3-dimensional scatterplot of `temp`, `wind`, and `SO2` of the U.S. air pollution data

# 9 3-dimensional scatterplot

We mentioned last week rotating a 3-dimensional plot by the package **rgl**. In this section, we use **scatterplot3d** to demonstrate 3-dimensional scatterplot in general. Figure 17 shows a 3-dimensional scatterplot of `SO2`, `wind` and `temp` of the U.S. air pollution data. The view is from angle 55. The vertical bars were produced by the subcommand `type=''h''`.

**R command used**:

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

# 10 Stalactite plots

The stalactite plot is multivariate graphics designed for the detection and identification of multivariate outliers. It is based on the generalized distances of observations from the multivariate mean of the data. But the distances are calculated from the means and covariances estimated from increasing-sized subsets of the data. For a $p$-dimensional data set, the mean and covariance are calculated starting with $p + 1$ observations. One can calculate the distances of observations based on the estimated mean and covariances. Using chi-square distribution for the distance, one can judge potential outlying observations. The plot is proceeded as follows. Suppose that the distances are calculated by $m$ data points. Then, one uses data with the smallest $m + 1$ distances to revise the estimates of mean and covariance. This process is continued until no further data is available. Figure 18 shows the stalactite plot for the U.S. air pollution data. The plot shows that with all 41 observations are used to

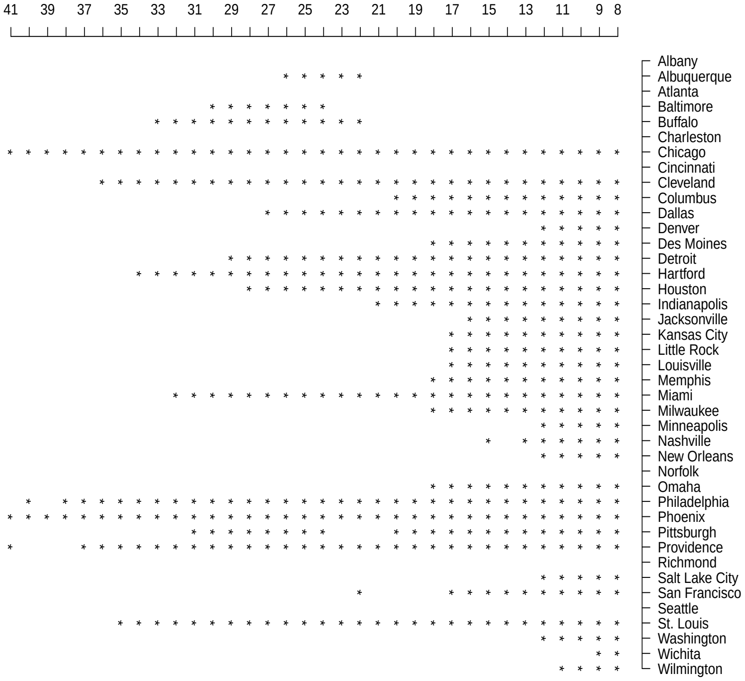

Figure 18: Stalactite plot of the U.S. air pollution data

calculate the generalized distances, only observations of Chicago, Phoenix, and Providence are identified as outliers.

## R command used

```
> stalac(USairpollution}
```