Chapter 2

Missing Values

2.1 Background

When there are missing values in multivariate data we need to explore the nature of missingness, especially in terms of the joint distribution. Missing at random makes it easier to justify a full case analysis, but more often there is some dependence between variables in terms of missing values. Exploratory visual methods can be used for extracting information about the distribution of missing values. In addition, if we wish to impute missing values, visual methods can be used to evaluate the results, and compare imputation methods.

Traditionally the classical approach to plotting missing values is to assign them a fixed value just below the minimum data value, so the missings a displayed as a line of points in the margins. This approach works well for pairwise analysis, but fails completely in multivariate graphics. New approaches to handling multivariate missing values graphically are described in (Swayne & Buja 1998). The methods are implemented in XGobi.

In XGobi input files, missing values can be marked in two different ways. In the actual data file they can be represented as “NA”, “na”, or “.”. To accommodate more complex types of missing structure the alternative method is to create a separate file, with the extension .missing, that contains a matrix which mimics the size of the original matrix, but the elements are 0, 1, 2, 3, .... The value 0 indicates the data value is present, other values indicate missing, and they can be used to indicate different types of missingness, such as censoring. There are some simple imputation methods included with the missing values control panel, but more complex imputation needs to be done outside XGobi. Several different computed imputations can be read into XGobi using the file .imp, and then within XGobi the user can switch between the different imputation schemes to evaluate their effectiveness.

When XGobi finds missing values indicated in the input file, it provides two options in the Tools menu: “Impute missing values”, “Launch missing values XGobi”. The first option brings up a control panel for imputing values. The
The panel contains facilities for random imputation, and assigning a fixed value. In addition if there is a .imp file then there is a menu which allows selection between the imputations provided in the file. Launching a missing values XGobi brings up a separate XGobi window on the matrix of missings indicators. This allows linked brushing between the two XGobi’s, with missings easily identified in the missing values XGobi. In addition, the full multivariate visual tools available in XGobi can be used to explore the distribution of missings in the matrix of missing indicators.

2.2 Case Study: Tropical Atmosphere Ocean (TAO) array

2.2.1 Data description

This data comes from an array of buoys moored in the Pacific Ocean that continuously record temperature, humidity and winds to monitor El Niño events. The first buoys were operational in March 1986, and more have been added since. There are about 70 buoys in total and more are planned. Figure 2.1 displays the location variables of the buoys in the south-eastern Pacific. It is interesting to note that the buoys drift around some, some more than others. The web site that keeps daily updates on the data is http://www.pmel.noaa.gov/toga-tao/home.html.

![Figure 2.1: Latitude vs longitude of buoys in the south-eastern Pacific, showing how some drift around a lot.](image)


2.2.2 Exploring Missingness

The data used in this case study has monthly measurements from March 1980 to May 1998 for 13 buoys. There are 5 variables: zonal winds, meridian winds, relative humidity, air temperature, sea surface temperature. There are 2184 data points, and many missing values. The missing values are coded in the .dat file as “NA”. Figure 2.2 displays plots of the variables in this data with missing values imputed in different ways.

Upon starting up XGobi on this data look at the scatterplot of air temperature against humidity. When missing values are marked in the data files, XGobi assigns them a numerical value 0, but then also creates a separate file .missing of the binary map of the location of missing in the data matrix. The missings appear as a line of points at 0 in each variable which is well out of range of the non-missing data values. To change the numerical value of the missings, select the Impute missing values option on the Tools menu. Then on the control panel that pops up select “All variables” in the “Variables to impute” menu. Click on “% below minimum”. This will give you the top left plot in Figure 2.2. The missings are neatly displayed in the margins of the plot. To examine the nature of missings as compared to non-missings, it is probably necessary to use the jitter tool to assess the overplotting of missings. Add a small amount of jitter. The marginal distribution of air temperature for missings and non-missings on humidity appear a little different: the missings appear to have a lower mean value. Now examine what happens to the display of missings when rotation and tour are used. Change into either of the two modes, and select humidity, air temperature and sea surface temperature as the active variables. The missing values appear as bands of points that float around, which is quite distracting and not helpful in understanding the nature of missingness (Figure 2.2 top right).

Change to a univariate plot of humidity, select “Currently selected variables” from the “Variables to Impute” menu. Enter 80 in the Specify box, and click on “Specify”. Now switch to a univariate plot of air temperature and enter 25 in the Specify box, and click on “Specify”. Lastly switch to a univariate plot of sea surface temperature, enter 25 in the the Specify box, and click on “Specify”. This approximates imputing using the variable mean value. Now change to a scatterplot of sea surface temperature against humidity. The missings appear as a lovely cross shape in the plot (Figure 2.2 middle left). If you change to rotation or tour of the three variables, you will see a nice 3D cross structure.

Change to a scatterplot of sea surface temperature. On the imputation control panel select “All variables” in the “Variables to impute” menu. Then click on “Perform random imputation” which gives a picture like that shown in Figure 2.2 bottom left. This type of imputation completely ignores the correlation structure between the two variables. The destruction of correlation structure is even more pronounced in multivariate space (Figure 2.2 bottom right).

To examine the conditional distributions of missings in a more multivariate sense, brush the points that are missing in humidity. Examine the plot of air temperature against sea surface temperature. Figure 2.3 displays the results, missings on humidity are displayed as yellow crosses. The distribution is similar
Figure 2.2: Plots of variables with different methods of imputing missing values.
Figure 2.3: Plot of air temperature against sea surface temperature with point missing on humidity highlighted as yellow crosses.
to the non-missings, especially in the correlation structure, although the missings do look like they have a slightly different mean. This approach can be easily extended to higher dimensions. Erase all the cases missing on the variables to be viewed, and then examine the tour plots. So in this case if we wanted to examine the distribution of zonal winds, meridian winds, air and sea surface temperature conditional on whether humidity is missing or not, we first enter univariate plots of each of the 4 variables erase the missings. Then watch the tour of the 4 variables with the cases that are missing in humidity highlighted. Similarly we could look at plots with missings on two variables highlighted, etc.

2.2.3 Exercises

1. Explore the nature of missings in the variables zonal winds and meridian winds.

2. Explore the distribution of the two wind variables conditional on missing in humidity.

3. Explore the distribution of the four variables, winds, temperatures, conditional on missing in humidity, using brushing and the tour.