3

Missing Values

Values are often missing in data, for several reasons. Measuring instruments fail, samples are lost or corrupted, patients do not show up to scheduled appointments, and measurements may be deliberately censored if they are known to be untrustworthy above or below certain thresholds. When this happens, it is always necessary to evaluate the nature and the distribution of the gaps, to see whether a remedy must be applied before further analysis of the data. If too many values are missing, or if gaps on one variable occur in association with other variables, ignoring them may invalidate the results of any analysis that is performed. This sort of association is not at all uncommon and may be directly related to the test conditions of the study. For example, when measuring instruments fail, they often do so under conditions of stress, such as high temperature or humidity. As another example, a lot of the missing values in smoking cessation studies occur for those people who begin smoking again and silently withdraw from the study, perhaps out of discouragement or embarrassment.

In order to prepare the data for further analysis, one remedy that is often applied is imputation, that is, filling in the gaps in the data with suitable replacements. There are numerous methods for imputing missing values. Simple schemes include assigning a fixed value such as the variable mean or median, selecting an existing value at random, or averaging neighboring values. More complex distributional approaches to imputation start with the assumption that the data arises from a standard distribution such as a multivariate normal, which can then be sampled to generate replacement values. See Schafer (1997) for a description of multiple imputation and Little & Rubin (1987) for a description of imputation using multivariate distributions.

In this chapter, we will discuss the power of visual methods at both of these stages: diagnosing the nature and seriousness of the distribution of the missing values (Sect. 3.2) and assessing the results of imputation (Sect. 3.3). The approach follows those described in Swayne & Buja (1998) and Unwin et al. (1996).

3.1 Background

Missing values are classified by Little & Rubin (1987) into three categories according to their dependence structure: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Only if values are MCAR is it considered safe to ignore missing values and draw conclusions from the set of complete cases, that is, the cases for which no values are missing. Such missing values may also be called *ignorable*.

The classification of missing values as MCAR means that the probability that a value is missing does not depend on any other observed or unobserved value; that is, $P(missing|observed, unobserved) = P(missing)$. This situation is ideal, where the chance of a value being missing depends on nothing else. It is impossible to verify this classification in practice because its definition includes statements about unobserved data; still, we assume MCAR if there is dependence between missing values and observed data.

In classifying missing values as MAR, we make the more realistic assumption that the probability that a value is missing depends only on the observed variables; that is, $P(missing|observed, unobserved) = P(missing|observed)$. This can be verified with data. For values that are MAR, some structure is allowed as long as all of the structure can be explained by the observed data; in other words, the probability that a value is missing can be defined by conditioning on observed values. This structure of missingness is also called *ignorable*, since conclusions based on likelihood methods are not affected by MAR data.

Finally, when missing values are classified as MNAR, we face a difficult analysis, because $P(missing|observed, unobserved)$ cannot be simplified and cannot be quantified. *Non-ignorable* missing values fall in this category. Here, we have to assume that, even if we condition on all available observed information, the reason for missing values depends on some unseen or unobserved information.

Even when missing values are considered ignorable we may wish to replace them with imputed values, because ignoring them may lead to a non-ignorable loss of data. Consider the following constructed data, where missings are represented by the string NA, meaning “Not Available.” There are only 5 missing values out of the 50 numbers in the data:

Case	X_1	X_2	X_3	X_4	X_5
1	NA	20	1.8	6.4	−0.8
2	0.3	NA	1.6	5.3	−0.5
3	0.2	23	1.4	6.0	NA
4	0.5	21	1.5	NA	−0.3
5	0.1	21	NA	6.4	−0.5
6	0.4	22	1.6	5.6	−0.8
7	0.3	19	1.3	5.9	−0.4
8	0.5	20	1.5	6.1	−0.3
9	0.3	22	1.6	6.3	−0.5
10	0.4	21	1.4	5.9	−0.2

Even though only 10% of the numbers in the table are missing, 100% of the variables have missing values, and so do 50% of the cases. A complete case analysis would use only half the data.

Graphical methods can help to determine the appropriate classification of the missing structure as MCAR, MAR or MNAR, and this is described in Sect. 3.2. Section 3.3 describes how the classification can be re-assessed when imputed values are checked for consistency with the data distribution.

3.2 Exploring missingness

One of our first tasks is to explore the distribution of the missing values, seeking to understand the nature of “missingness” in the data. Do the missing values appear to occur randomly, or do we detect a relationship between the missing values on one variable and the recorded values for some other variables in the data? If the distribution of missings is not random, this will weaken our ability to infer structure among the variables of interest. It will be shown later in the chapter that visualization is helpful in searching for the answer to this question.

3.2.1 Shadow matrix

As a first step in exploring the distribution of the missing values, consider the following matrix, which corresponds to the data matrix above and has the same dimensions:

Case	X_1	X_2	X_3	X_4	X_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	0	0	1
4	0	0	0	1	0
5	0	0	1	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

In this binary matrix, 1 represents a missing value and 0 a recorded value. It is easier to see the positions of the missing values in this simple version and to consider their distribution apart from the data values. We like to call this the "shadow matrix."

Sometimes there are multiple categories of missingness, in which case this matrix would not simply be binary. For example, suppose we were conducting a longitudinal study in which we asked the same questions of the same subjects over several years. In that case, the missing values matrix might include three values: 0 could indicate an answered question, 1 that a survey respondent failed to answer a question, and 2 that a respondent died before the study was completed.

As an example for working with missing values, we use a small subset of the TAO data: all cases recorded for five locations (latitude 0°with longitudes 110°W and 95°W, 2°S with 110°W and 95°W, and 5°S with 95°W) and two time periods (November to January 1997, an El Niño event, and for comparison, the period from November to January 1993, when conditions were considered normal). Load the data into R and GGobi to investigate missingness:

```
> library(norm)
> library(rggobi)
> d.tao <- read.csv("tao.csv", row.names=1)
> d.tao.93 <- as.matrix(subset(
  d.tao,year==1993,select=sea.surface.temp:vwind))
> d.tao.97 <- as.matrix(subset(
  d.tao,year==1997,select=sea.surface.temp:vwind))
> d.tao.nm.93 <- prelim.norm(d.tao.93)
> d.tao.nm.93$nmis
sea.surface.temp         air.temp         humidity
               3                4               93
           uwind            vwind
               0                0
> d.tao.nm.97 <- prelim.norm(d.tao.97)
> d.tao.nm.97$nmis
sea.surface.temp         air.temp         humidity
               0               77                0
           uwind            vwind
               0                0
```

There are 736 data points, and we find missing values on three of the five variables (Table 3.1).

Table 3.1. Missing values on each variable

Variable	Number of missing values	
	1993	1997
sea surface temp	3	0
air temp	4	77
humidity	93	0
uwind	0	0
vwind	0	0

We are also interested in tabulating missings by case:

```
> d.tao.nm.93$r
    [,1] [,2] [,3] [,4] [,5]
274    1    1    1    1    1
  1    0    0    1    1    1
 90    1    1    0    1    1
  1    1    0    0    1    1
  2    0    0    0    1    1
> d.tao.nm.97$r
    [,1] [,2] [,3] [,4] [,5]
291    1    1    1    1    1
 77    1    0    1    1    1
```

From Table 3.2, we can see that most cases have no missing values (74.5% in 1993, 79.1% in 1997), and less than a quarter of cases have one missing value (24.5% in 1993, 20.9% in 1997). In 1993 two cases are missing two values and two cases have missing values on three of the five variables, sea surface temp, air temp and humidity.

To study the missingness graphically, load the data and set up the colors and glyphs to reflect the year, which corresponds to two different climate conditions.

```
> gd <- ggobi(d.tao)[1]
> glyph_color(gd) <- ifelse(gd[,1]==1993,5,1)
> glyph_type(gd) <- ifelse(gd[,1]==1993,6,4)
```

Table 3.2. Distribution of the number of missing values on a case.

No. of missings	1993		1997	
on a case	No. of cases	%	No. of cases	%
3	2	0.5	0	0
2	2	0.5	0	0
1	90	24.5	77	20.9
0	274	74.5	291	79.1

3.2.2 Getting started: missings in the "margins"

The simplest approach to drawing scatterplots of variables with missing values is to assign to the missings some fixed value outside the range of the data, and then to draw them as ordinary data points at this unusual location. It is a bit like drawing them in the margins, which is an approach favored in other visualization software. In Fig. 3.1, the three variables with missing values are shown. The missings have been replaced with a value 10% lower than the minimum data value for each variable. In each plot, missing values in the horizontal or vertical variable are represented as points lying along a vertical or horizontal line, respectively. A point that is missing on both variables appears as a point in the lower left corner; if multiple points are missing on both, this point is simply over-plotted.

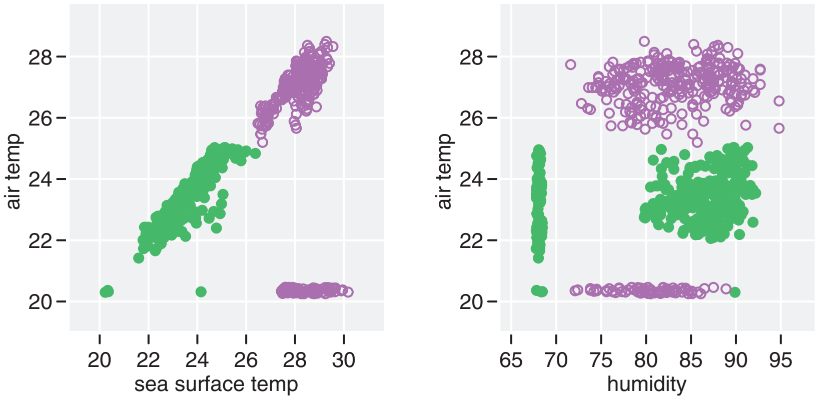

Fig. 3.1. Assigning constants to missing values. In this pair of scatterplots, we have assigned to each missing value a fixed value 10% below the variable minimum, so the "missings" fall along vertical and horizontal lines to the left and below the point scatter. The green solid circles (the cluster that has lower values of air temp) represent data recorded in 1993; the purple open circles show the 1997 data.

What can be seen? Consider the plot of air temp vs. sea surface temp. Not surprisingly, the two temperature values are highly correlated as indicated by the strong linear pattern in the plot; we will make use of this fact a bit later. We can also see that the missings in that plot fall along a horizontal line, telling us that more cases are missing for air temp than for sea surface temp. Some cases are missing for both, and those lie on the point in the lower left corner. The live plot can be queried to find out how many points are over-plotted there. To alleviate the problem of over-plotting, we have also jittered the values slightly; i.e., we have added a small random number to the missing values. In this plot, we also learn that there are no cases missing for sea surface temp but recorded for air temp — if that were true, we would see some points plotted along a vertical line at roughly sea surface temp = 20. The right-hand plot, air temp vs. humidity, is different: There are many cases missing on each variable but not missing on the other.

Both pairwise plots contain the same two clusters of data, one for 1993 records (green filled circles) and the other for 1997, an El Niño year (purple open circles). There is a relationship between the variables and the distribution of the missing values, as we can tell simply from the color of the missings. For example, all cases for which humidity was missing are green, so we know they were all recorded in 1993. The position of the missings on humidity tells the same story, because none of them lie within the range of air temp in 1997. We know already that, if we excluded these cases from our analysis, the results would be distorted: We would exclude 93 out of 368 measurements for 1993, but none for 1997, and the distribution of humidity is quite different in those two years.

The other time period, in 1997, is not immune from missing values either, because all missings for air temp are in purple.

In summary, from these plots we have learned that there is dependence between the missing values and the observed data values. We will see more dependencies between missings on one variable and recorded values on others as we continue to study the data. At best, the missing values here may be MAR (missing at random).

3.2.3 A limitation

Populating missing values with constants is a useful way to begin, as we have just shown. We can explore the data we have and begin our exploration of the missing values as well, because these simple plots allow us to continue using the entire suite of interactive techniques. Multivariate plots, though, such as the tour and parallel coordinate plots are not amenable to this method.

Using fixed values in a tour causes the missing data to be mapped onto artificial planes in p-space, which obscure each other and the main point cloud. Figure 3.2 shows two tour views of sea surface temp, air temp, and humidity with missings set to 10% below minimum. The missing values appear as clusters in the data space, which might be thought of as lying along three walls of a room with the complete data as a scattercloud within the room.

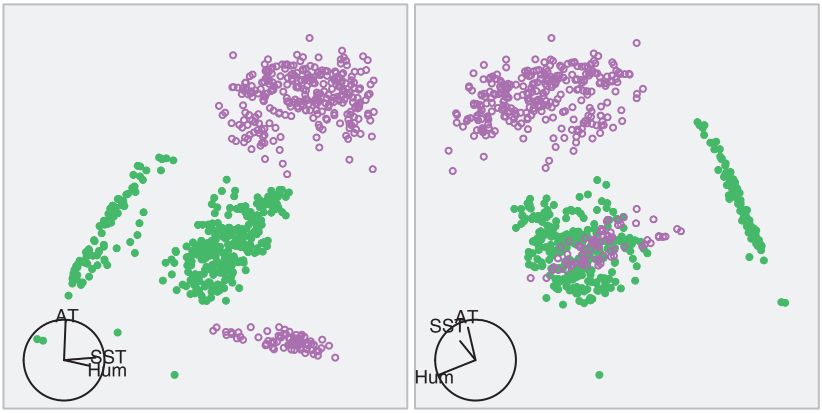

Fig. 3.2. Tour views of sea surface temp, air temp, and humidity with missings set to 10% below minimum. There appear to be four clusters, but two of them are simply the cases that have missings on at least one of the three variables.

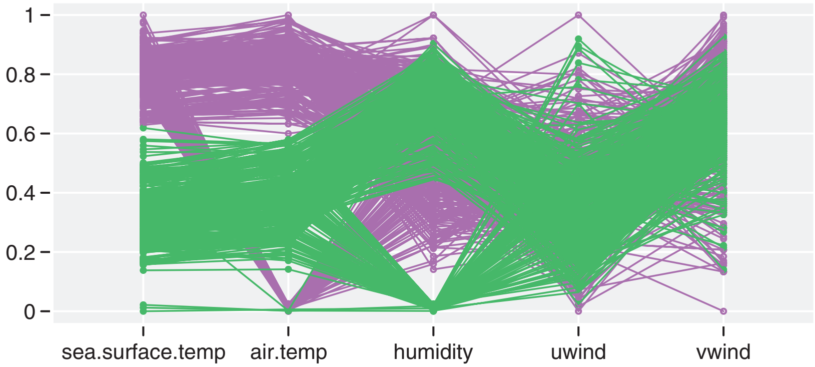

Fig. 3.3. Parallel coordinates of the five variables sea surface temp, air temp, humidity, uwind, and vwind with missings set to 10% below minimum. There are two groups visible for humidity in 1993 (green, the color drawn last), but that is because a large number of missing values are plotted along the zero line; for the same reason, there appear to be two groups for air temp in 1997 (purple).

Figure 3.3 shows the parallel coordinate plot of sea surface temp, air temp, humidity, uwind, and vwind with missings set to 10% below minimum. If we did not know that the points along the zero line were the missings, we could be led to false interpretations of the plot. Consider the values of humidity in 1993 (the green points, the color drawn last), where the large number of points drawn at the zero line look like a second cluster in the data.

When looking at plots of data with missing values, it is important to know whether the missing values have been plotted, and if they have, how they are being encoded.

3.2.4 Tracking missings using the shadow matrix

In order to explore the data and their missing values together, we will treat the shadow matrix (Sect. 3.2.1) as data and display projections of each matrix in linked windows, side by side. In one window, we show the data with missing values replaced by imputed values; in the missing values window, we show the binary indicators of missingness.

Although it may be more natural to display binary data in area plots, we find that scatterplots are often adequate, and we will use them here. We need to spread the points to avoid multiple over-plotting, so we jitter the zeros and ones. The result is a view such as the left-hand plot in Fig. 3.4. The data fall into four squarish clusters, indicating presence and missingness of values for the two selected variables. For instance, the top right cluster consists of the cases for which both variables have missing values, and the lower right cluster shows the cases for which the horizontal variable value is missing but the vertical variable value is present.

Figure 3.4 illustrates the use of the TAO dataset to explore the distribution of missing values for one variable with respect to other variables in the data. We have brushed in orange squares only the cases in the top left cluster, where air temp is missing but humidity is present. We see in the right-hand plot that none of these missings occur for the lowest values of uwind, so we have discovered another dependence between the distribution of missingness on one variable and the distribution of another variable.

We did not really need the missings plot to arrive at this observation; we could have found it just as well by continuing to assign constants to the missing values. In the next section, we will continue to use the missings plot as we begin using imputation.

3.3 Imputation

Although we are not finished with our exploratory analysis of this subset of the TAO data, we have already learned that we need to investigate imputation methods. We have already learned that we will not be satisfied with complete case analysis. We cannot safely throw out all cases with a missing value, because the distribution of the missing values on at least two variables (humidity and air temp) is strongly correlated with at least one other data variable (year).

Because of this correlation, we need to investigate imputation methods. As we replace the missings with imputed values, though, we do not want to lose track of their locations. We want to use visualization to help us assess imputation methods as we try them, making sure that the imputed values have nearly the same distribution as the rest of the data.

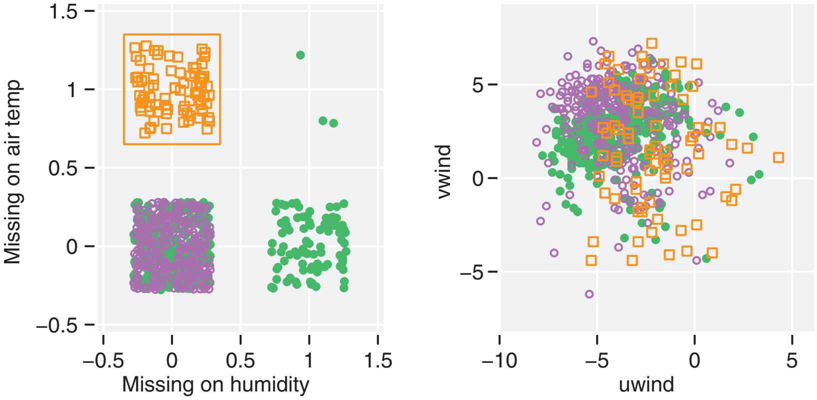

Fig. 3.4. Exploring missingness in the TAO data. The “missings” plot **(left)** for air temp vs. humidity is a jittered scatterplot of zeroes and ones, where one indicates a missing value. The points that are missing only on air temp have been brushed in orange. In a scatterplot of vwind vs. uwind **(right)**, those same missings are highlighted. There are no missings for the very lowest values of uwind.

3.3.1 Mean values

The most rudimentary imputation method is to use the variable mean to fill in the missing values. In the middle row of plots in Fig. 3.5, we have substituted the mean values for missing values on sea surface temp, air temp, and humidity. Even without highlighting the imputed values, some vertical and horizontal lines are visible in the scatterplots. This result is common for any imputation scheme that relies on constants. Another consequence is that the variance–covariance of the data will be reduced, especially if there are a lot of missing values.

3.3.2 Random values

It is clear that a random imputation method is needed to better distribute the replacements. The simplest method is to fill in the missing values with some value selected randomly from among the recorded values for that variable. In the bottom row of plots in Fig. 3.5, we have substituted random values for

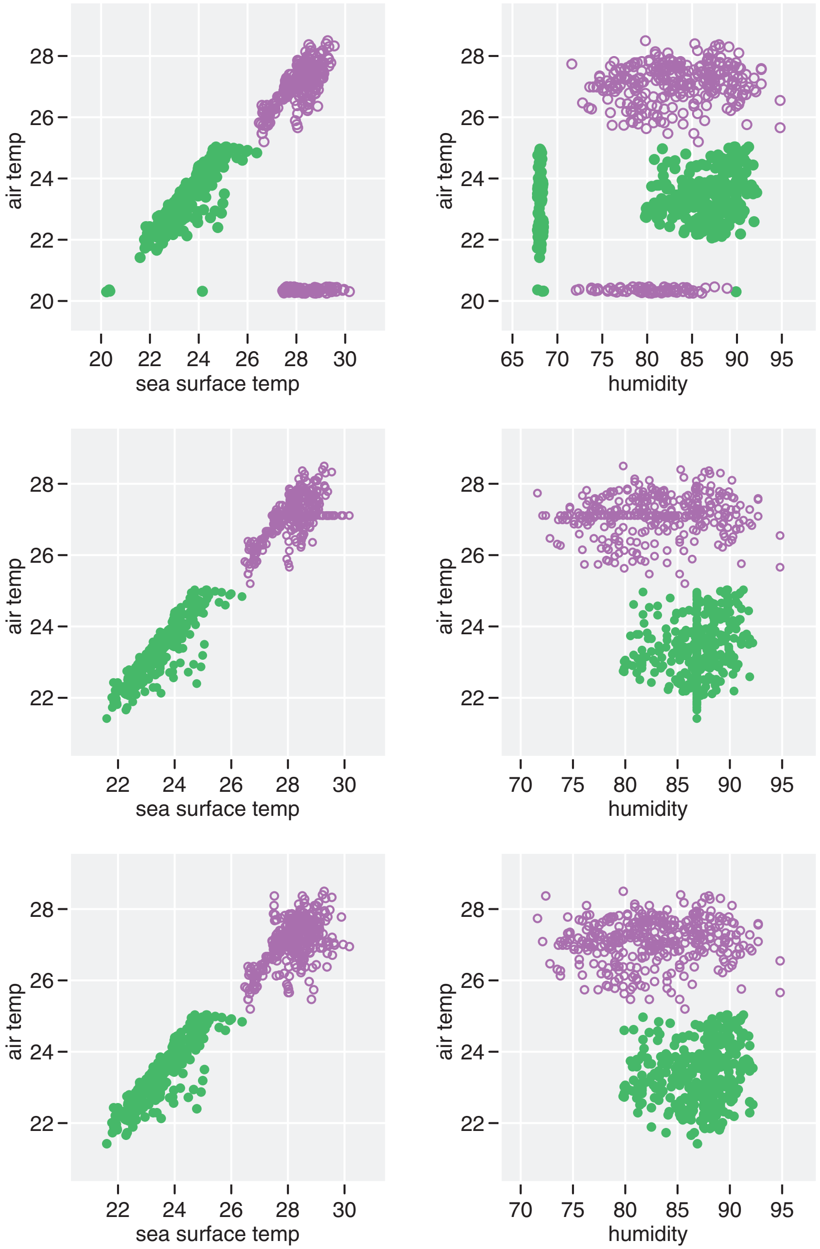

Fig. 3.5. Comparison of simple, widely used imputation schemes. Missings in the margin, as in Fig. 3.1 **(Top row)**. Missing values have been replaced with variable means, conditional on year, producing vertical and horizontal stripes in each cluster **(Middle row)**. Missing values have been filled in by randomly selecting from the recorded values, conditional on year **(Bottom row)**. The imputed values are a little more varied than the recorded data.

missing values on sea surface temp, air temp, and humidity. The match with the data is much better than when mean values were used: It is difficult to distinguish imputed values from recorded data! However, taking random values ignores any association between variables, which results in more variation in values than occurs with the recorded data. If you have a keen eye, you can see that in these plots. It is especially visible in the plot of sea surface temp and air temp, for the 1997 values (in purple): They are more spread, less correlated than the complete data.

The imputed values can be identified using linked brushing between the missings plot and the plot of sea surface temp vs. air temp (Fig. 3.6). Here the values missing on air temp have been brushed (orange rectangles) in the missings plot (left), and we can see the larger spread of the imputed values in the plot at right.

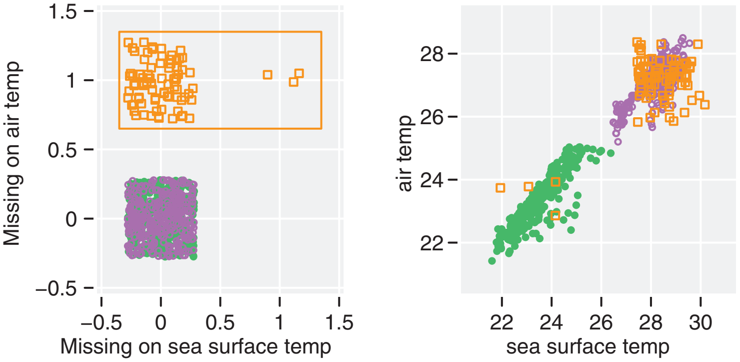

Fig. 3.6. Conditional random imputation. Missing values on all variables have been filled in using random imputation, conditioning on drawing symbol. The imputed values for air temp show less correlation with sea surface temp than do the recorded values.

3.3.3 Multiple imputation

A more sophisticated approach to imputation is to sample from a statistical distribution, which may better reflect the variability in the observed data. Common approaches use regression models or simulation from a multivariate distribution.

To use regression, a linear model is constructed for each of the variables containing missings, with that variable as the response, and the complete data variables as explanatory variables. The distribution of the residuals is used to simulate an error component, which is added to the predicted value for each missing, yielding an imputed value. Many models may be required to impute all missing values. For example, in the TAO data, we would need to fit a model for sea surface temp, air temp, and humidity separately for each year. This can be laborious!

Simulating from a multivariate distribution yields imputed values from a single model. For the TAO data, it might be appropriate to simulate from a multivariate normal distribution, separately for each year.

With either approach, it is widely acknowledged that one set of imputed values is not enough to measure the variability of the imputation. Multiple imputed values are typically generated for each missing value with these simulation methods, and this process is called *multiple imputation*.

R packages, such as norm by Novo & Schafer (2006) or Hmisc by Dupont & Harrell (2006), contain multiple imputation routines. To view the results in GGobi, we can dynamically load imputed values into a running GGobi process. This next example demonstrates how to impute from a multivariate normal model using R and how to study the results with GGobi. Values are imputed separately for each year.

To apply the light and dark shades used in Fig. 3.7 to the GGobi process launched earlier, select Color Schemes from GGobi's Tools menu and the Paired 10 qualitative color scheme before executing the following lines:

```
> gcolor <- ifelse(gd[,1]==1993,3,9)
> glyph_color(gd) <- gcolor
> ismis <- apply(gd[,4:8], 1, function(x) any(is.na(x)))
> gcolor[ismis] <- gcolor[ismis]+1
> glyph_color(gd) <- gcolor
```

Make a scatterplot of sea surface temp and air temp. At this stage, the missing values are still shown on the line below the minimum of the observed points. In the next step, the missing values are imputed multiply from separate multivariate normal distributions for each of the two years.

```
> rngseed(1234567)
> theta.93 <- em.norm(d.tao.nm.93, showits=TRUE)
Iterations of EM:
1...2...3...4...5...6...7...8...9...10...11...12...13...14...
15...16...17...18...19...20...21...22...23...
> theta.97 <- em.norm(d.tao.nm.97, showits=TRUE)
Iterations of EM:
1...2...3...4...5...6...7...8...9...10...11...12...13...14...
> d.tao.impute.93 <- imp.norm(d.tao.nm.93, theta.93,
    d.tao.93)
> d.tao.impute.97 <- imp.norm(d.tao.nm.97, theta.97,
    d.tao.97)
```

```
> gd[,"sea.surface.temp"] <- c(
    d.tao.impute.97[,"sea.surface.temp"],
    d.tao.impute.93[,"sea.surface.temp"])
> gd[,"air.temp"] = c(
    d.tao.impute.97[,"air.temp"],
    d.tao.impute.93[,"air.temp"])
> gd[,"humidity"] = c(
    d.tao.impute.97[,"humidity"],
    d.tao.impute.93[,"humidity"])
```

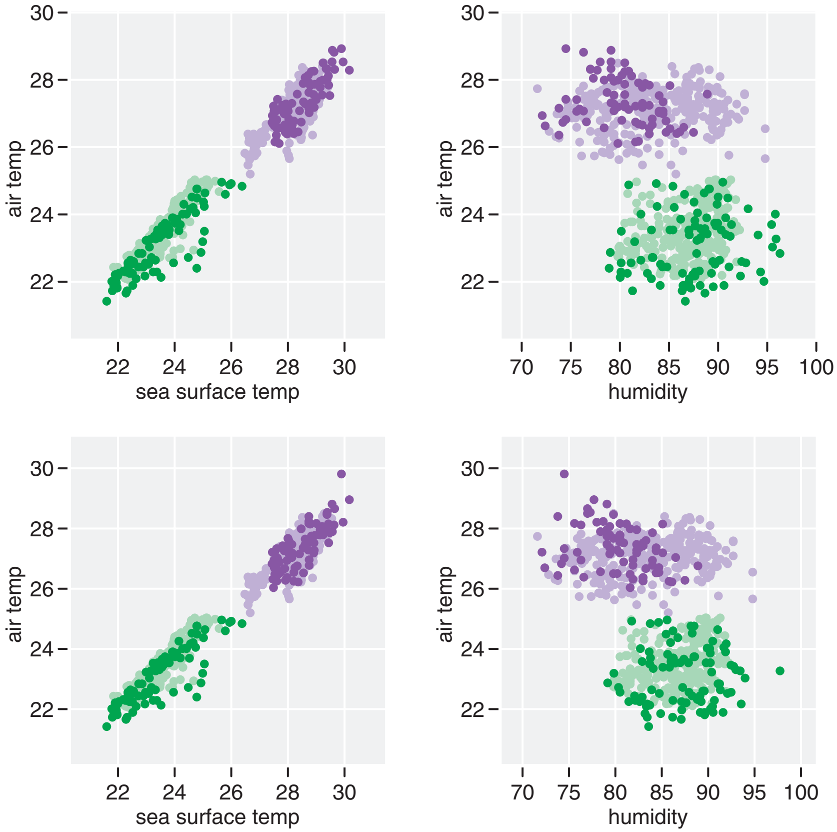

Fig. 3.7. Two different imputations using simulation from a multivariate normal distribution of all missing values. In the scatterplot of air temp vs. sea surface temp the imputed values may have different means than the complete cases: higher sea surface temp and lower air temp. The imputed values of humidity look quite reasonable.

The missings are now imputed, and the scatterplot of sea surface temp and air temp should look like the one in Fig. 3.7. The imputation might make it necessary to re-scale the plot if values have fallen outside the view; if so, use the Rescale button in the Missing Values panel.

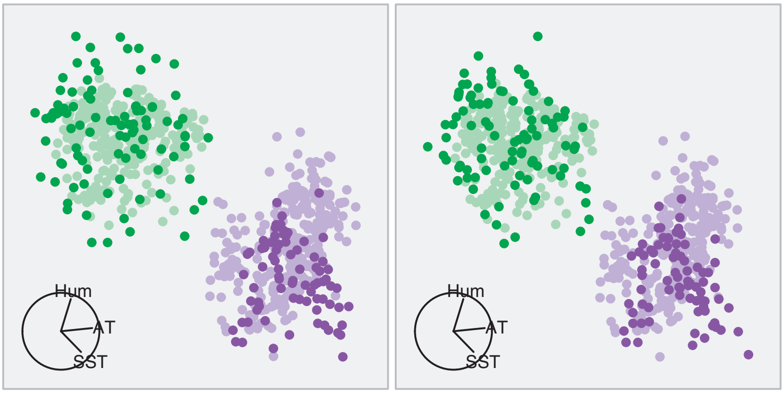

Fig. 3.8. Tour projection of the data after multiple imputation of sea surface temp, air temp, and humidity.

Figures 3.7 and 3.8 show plots of the data containing imputed values resulting from two separate simulations from a multivariate normal mixture. In this coloring, green and purple still mean 1993 and 1997, but now light shades represent recorded values and dark shades highlight the missing values — now imputed.

The imputed values look reasonably good. There are some small differences from the recorded data distribution: Some imputed values for sea surface temp and air temp in 1997 are higher than the observed values, and some imputed values for humidity in 1993 are higher than the observed values.

3.4 Recap

In this chapter, we showed how to use graphical methods to develop a good description of missingness in multivariate data. Using the TAO data, we were able to impute reasonable replacements for the missing values and to use graphics to evaluate them.

The data has two classes, corresponding to two distinct climate patterns, an El Niño event (1997) and a normal season (1993). We discovered the dependence on year as soon we started exploring the missingness, using missings plotted in the margins. Later we discovered other dependencies among the missings and the wind variables using linked brushing between the missings plot (shadow matrix) and other views of the data. These suggest the missing values should be classified as MAR and therefore ignorable, which means that imputation is likely to yield good results.

It was clear that we had to treat these two classes separately in order to get good imputation results, and we imputed values using multiple imputation, simulating from two multivariate normal distributions.

After studying the imputed values, we saw that they were not perfect. Some of the imputed values for air temp and sea surface temp were higher than the observed values. This suggests that the missingness is perhaps MNAR, rather than MAR. Still, the imputed values are close to the recorded values. For practical purposes, it may be acceptable to use them for further analysis of the data.

Exercises

1. Describe the distribution of the wind and temperature variables conditional on the distribution of missing values in humidity, using brushing and the tour.
2. For the Primary Biliary Cirrhosis (PBC) data:
   a) Describe the univariate distributions of complete cases for chol, copper, trig, and platelet. What transformations might be used to make the distributions more bell-shaped? Make these transformations, and use the transformed data for the rest of this exercise.
   b) Examine a scatterplot matrix of chol, copper, trig, platelet with missing values plotted in the margins.
      i. Describe the pairwise relationships among the four variables.
      ii. Describe the distribution between missings and non-missings for trig and platelet.
   c) Generate the shadow matrix, and brush the missing values a different color.
   d) Substitute means for the missing values, and examine the result in a tour. What pattern is obvious among the imputed values?
   e) Substitute random values for the missings, and examine the result in a tour. What pattern is obvious among the imputed values?
   f) In R, generate imputed values using multiple imputation. Examine different sets of imputed values in the scatterplot matrix. Do these sets of values look consistent with the data distribution?
   g) Using spine plots, examine each categorical variable (status, drug, age, sex), checking for associations between the variable and missingness.