CHAPTER 4

Missing data: mechanisms, methods, and messages

Shinichi Nakagawa

4.1 Introduction to dealing with missing data

In an ideal world, your data set would always be perfect without any missing data. But perfect data sets are rare in ecology and evolution, or in any other field. Missing data haunts every type of ecological or evolutionary data: observational, experimental, comparative, or meta-analytic. But this issue is rarely addressed in research articles. Why? Researchers often play down the presence of missing data in their studies, because it may be perceived as a weakness of their work (van Buuren 2012); this tendency has been confirmed in medical trials (Wood et al. 2004), educational research (Peugh and Enders 2004), and psychology (Bodner 2006). I speculate that many ecologists also play down the issue of missing data.

The most common way of handling missing data is called *list-wise deletion*: researchers delete cases (or rows/lists) containing missing values and run a model, e.g., a GLM (chapter 13) using the data set without missing values (known as *complete case analysis*). While common, few researchers explicitly state that they are using this approach. Another common practice that is usually not explicit involves statistics performed on pairs of points, like correlation analysis. For example, in analyzing the correlations among x, y, and z, we may be missing some data for each variable. Missing a value for x in some cases still allows one to use y and z from those cases. This is called *pair-wise deletion*, and it can often be noticed by seeing that there are different sample sizes for different correlations. List-wise and pair-wise deletion are often the default procedures used by statistical software.

What is wrong with deletion? The problems are twofold: (1) loss of information (i.e., reduction in *statistical power*) and (2) potential *bias* in parameter estimates under most circumstances (*bias* here means systematic deviation from population or true parameter values; Nakagawa and Hauber 2011).

To see the impact on statistical power, imagine a data set with 12 variables. Say only 5% of each variable is missing, without any consistent patterns. Using complete case analysis, we would lose approximately 43% of all cases. The resulting reduction in statistical power is fairly substantial. To ameliorate the reduction in power, some researchers use stepwise regression approaches called *available case analysis*, where cases are deleted if they are missing values needed to estimate a model, but the same cases are included for simpler models not requiring those values. For example, a full model would contain 12 variables with

Ecological Statistics: Contemporary Theory and Application. First Edition. Edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa.
Published in 2015 by Oxford University Press.

~43% of cases missing, while a reduced model might have 3 variables with ~10% missing. Are parameter estimates or indices like R^2 from these different models comparable? Certainly one cannot use information criteria such as AIC (Akaike Information Criteria; see chapter 3) for model selection procedures because these procedures require a complete case analysis. Such model selection can only be done using *available variable analysis* that only considers variables with complete data. However, this approach can exclude key information (e.g., Nakagawa and Freckleton 2011).

With regard to the bias problem in deleting missing data, cases are often missing for underlying biological reasons, so that parameter estimates from both complete case and available case analyses are often biased. For example, older or "shy" animals are difficult to catch in the field (Biro and Dingemanse 2009), so that their information may be systematically missing, leading to biased parameter estimates.

Some researchers "fill-in" or impute missing values to circumvent these problems. You may be familiar with filling missing values with the sample mean value (*mean imputation*). Indeed, in comparative phylogenetic analysis it has been common to replace missing values with taxon means (see Freckleton et al. 2003). Alternatively, missing data imputation can be slightly more sophisticated, using regression predictions to fill in missing cases (*regression imputation*). However, these methods, known as *single imputation* techniques, result in uncertainty estimates that do not account for the uncertainty that the missing values would have contributed (e.g., too small a standard error, or too narrow a confidence interval; McKnight et al. 2007; Graham 2009, 2012; Enders 2010). Thus, the rate of Type I error (chapter 2) increases; I call this phenomenon *biased uncertainty estimates*. These simple fixes using single imputation will yield biased parameter estimates.

The good news is that we now have solutions that combat missing data problems. They come in two forms: *multiple imputation* (MI), and *data augmentation* (DA; in the statistical literature, the term data augmentation is used in different ways, but I follow the usage of McKnight et al. 2007). The bad news is that very few researchers in ecology and evolution use such statistical tools (Nakagawa and Freckleton 2008). MI and DA have been available to us since the late 1980s, with some key publications in 1987 (Allison 1987; Tanner and Wong 1987; Little and Rubin 1987; Rubin 1987). In the beginning, few of us could use such techniques as they were not implemented in statistical packages or programs until the late 1990s. There are now *R* libraries (e.g., `norm` and `pan`; Schafer 1997, 2001) that make MI and DA relatively easy to use for many analyses (for reviews of statistical software for treating missing data, see Horton and Kleinman 2007; Yucel 2011). Why the lag in using such important statistical tools? Many of us may have never heard about *missing data theory* until now because it is not a part of our general training as ecologists and evolutionary biologists. However, the main reason may be psychological. It certainly feels a bit uneasy for me to "make up" data to fill in gaps! We are not alone: medical and social scientists have also been exposed to methods for handling missing data, but they have also been slow to adopt them (Raghunathan 2004; Graham 2009; Sterne et al. 2009; Enders 2010). Researchers often may see procedures such as data imputation and augmentation as cheating, or even as something akin to voodoo. It turns out that our current quick fixes are a lot more like voodoo! As Todd Little (cited in Enders 2010) puts it: "For most of our scientific history, we have approached missing data much like a doctor from the ancient world might use bloodletting to cure disease or amputation to stem infection (e.g., removing the infected parts of one's data by using list-wise or pair-wise deletion)." It is high time for us to finally start using missing data procedures in our analyses. This is especially so given the recent growth in the number of *R* libraries that can handle missing data appropriately using MI and DA (Nakagawa and Freckleton 2011; van Buuren 2012).

In this chapter, I explain and demonstrate the powerful missing data procedures now available to researchers in ecology and evolution (e.g., Charlier et al. 2009; González-Suárez et al. 2012). I first describe the basics and terminology of missing data theory, particularly the three different classes of missing data (*missing data mechanisms*). I then explain how different missing data mechanisms can be detected and, at least for some of the classes, how to prevent it in the first place. The main section will cover three types of methods for analyzing missing data (deletion, augmentation, and imputation), with emphasis on MI, practical issues associated with missing data procedures, guidelines for the presentation of results, and the connection between missing data issues and other chapters in this book.

4.2 Mechanisms of missing data

4.2.1 *Missing data theory, mechanisms, and patterns*

Rubin (1976) and his colleagues (e.g., Little and Rubin 1987, 2002; Little 1992, 1995) established the foundations of missing data theory. Central to missing data theory is his classification of missing data problems into three categories: (1) missing completely at random (*MCAR*), (2) missing at random (*MAR*), and (3) missing not at random (*MNAR*). These three classes of missing data are referred to as *missing data mechanisms* (for a slightly different classification, see Gelman and Hill 2007). Despite the name, these are not causal explanations for missing data. Missing data mechanisms represent the statistical relationship between observations (variables) and the probability of missing data. The other term easy to confuse with missing data mechanisms is *missing data patterns*; these are the descriptions of which values are missing in a data set (see section 4.3.1).

4.2.2 *Informal definitions of missing data mechanisms*

Here, I use part of a data set from the house sparrow (*Passer domesticus*) population on Lundy Island, UK (Nakagawa et al. 2007a; Schroeder et al. 2012; I will use a full version of this data set in section 4.4.5). Male sparrows possess what is termed a "badge of status," which has been shown to reflect male fighting ability (Nakagawa et al. 2007b). The badge of status is a black throat patch, which substantially varies in size, with larger badges resenting superior fighters. Table 4.1 contains the information on badge size (Badge) and male age (Age) from 10 males, and also on the three missing data mechanisms in the context of this data set. The MCAR mechanism occurs when the probability of missing data in one variable is not related to any other variable in the data set. The variable, $\text{Age}_{[\text{MCAR}]}$ in table 4.1 is missing completely at random (MCAR) because the probability of missing data on Age is not related to the other observed variable, Badge.

The MAR mechanism is at work when the probability of missing data in a variable is related to some other variable(s) in the data set. If you are wondering "So how is this missing at random?" you are not alone: the term confuses many people. It is helpful to see MAR as "conditionally missing at random"; that is, missing at random after controlling for all other related variables (Graham 2009). In our sparrow example, $\text{Age}_{[\text{MAR}]}$ is missing at random (MAR) because the missing values are associated with the smallest three values of Badge. Once you control for Badge, data on Age are missing completely at random, MCAR. This scenario may happen, for example, if immigrants to this sparrow population (whose ages are unknown to the researcher) somehow have a smaller badge size.

Table 4.1 Badge size (mm) and Age (yr) information for 10 house sparrow males. Age consists of 4 different types of data sets according to the mechanism of missing values (–): Complete data, MCAR data, MAR data, and MNAR data

Bird	Badge	Age			
(Case)	Complete	Complete	MCAR	MAR	MNAR
1	31.5	1	1	–	1
2	33.5	2	–	–	2
3	34.4	3	3	–	3
4	35.1	1	–	1	1
5	35.4	2	2	2	2
6	36.7	4	4	4	–
7	37.8	2	2	2	2
8	38.8	4	4	4	–
9	40.3	3	3	3	3
10	41.5	4	–	4	–

The MNAR mechanism happens when the probability of missing data in a variable is associated with this variable itself, even after controlling for other observed (related) variables. $\text{Age}_{[\text{MNAR}]}$ is missing not at random because the three missing values are 4-year old birds, and it is known that older males tend to have larger badge sizes. Such a scenario is plausible if a study on this sparrow population started 3 years ago, and we do not know the exact age of older birds.

4.2.3 *Formal definitions of missing data mechanisms*

Now to provide more formal/mathematical definitions of the missing data mechanisms, I introduce relevant notation and terminology from missing data theory.

- $\mathbf{Y}$ is a matrix of the entire data set (including response and predictor variables) that can be decomposed into $\mathbf{Y_{obs}}$ and $\mathbf{Y_{mis}}$ (the observed and missing parts of the data);
- $\mathbf{R}$ is a *missingness* matrix—these are indicators of whether the corresponding locations in $\mathbf{Y}$ are observed (0) or missing (1); and
- $\mathbf{q}$ is a vector of parameters describing the relationship between missingness, $\mathbf{R}$ and the data set, $\mathbf{Y}$ (table 4.2; see Little and Rubin 2002; McKnight et al. 2007; Molenberghs and Kenward 2007; Enders 2010; Graham 2012). Importantly, $\mathbf{q}$ is known as the mechanism of missing data and provides the basis for distinguishing between MCAR, MAR, and MNAR. An intuitive interpretation of $\mathbf{q}$ is that the content of $\mathbf{q}$ indicates one of the three missing data mechanisms.

It is the easiest to begin with the description of MNAR data because it includes all these mathematical terms. The three mechanisms, in relation to the concepts of missingness and ignorability (discussed later in this section), are summarized in figure 4.1.

Following Enders (2010), the probability distribution for MNAR can be written as:

$$p(\mathbf{R}|\mathbf{Y_{obs}}, \mathbf{Y_{mis}}, \mathbf{q}). \tag{4.1}$$

This says that the probability of whether a position in $\mathbf{R}$ is 0 or 1 depends on both $\mathbf{Y_{obs}}$ and $\mathbf{Y_{mis}}$, and this relationship is governed by $\mathbf{q}$. In table 4.2, $\mathbf{v_2}$ is MNAR, if missing

Table 4.2 An illustrative example of a data set $\mathbf{Y_{obs}}$ with three variables ($\mathbf{v_1}$–$\mathbf{v_3}$; Mis = missing observations and Obs = observed values) and its missingness, **R** (the recording of $\mathbf{v_1}$–$\mathbf{v_3}$ into binary variables, $\mathbf{m_1}$–$\mathbf{m_3}$); modified from Nakagawa and Freckleton (2011). Note that $\mathbf{v_3}$ is not measured; it is included here for illustrative purposes but would not usually be a part of **Y** and **R**

Case	Data [$\mathbf{Y} = (\mathbf{Y_{obs}}, \mathbf{Y_{mis}})$]			Missingness [**R**]		
	$\mathbf{v_1}$	$\mathbf{v_2}$	$\mathbf{v_3}$	$\mathbf{m_1}$	$\mathbf{m_2}$	$\mathbf{m_3}$
1	Obs	Mis	Mis	0	1	1
2	Obs	Obs	Mis	0	0	1
3	Obs	Obs	Mis	0	0	1
4	Obs	Mis	Mis	0	1	1
5	Obs	Obs	Mis	0	0	1
6	Obs	Obs	Mis	0	0	1
7	Obs	Obs	Mis	0	0	1
8	Obs	Mis	Mis	0	1	1
9	Obs	Mis	Mis	0	1	1
10	Obs	Obs	Mis	0	0	1

values depend on $\mathbf{v_2}$ itself. Such missing values can (but need not) be related to $\mathbf{v_1}$, a completely observed variable. In this particular case, the probability of MNAR missingness depends completely on $\mathbf{Y_{mis}}$, i.e., $p(\mathbf{R} \mid \mathbf{Y_{mis}}, \mathbf{q})$, which is a special case of missing values that are related to both $\mathbf{v_1}$ and $\mathbf{v_2}$ (i.e., equation 4.1). Another more complicated, form of MNAR is when $\mathbf{v_2}$ depends on a completely unobserved variable, for example $\mathbf{v_3}$ in table 4.2. For a concrete example in table 4.1, the MAR missing values in Age would become MNAR, if we had no measurement of badge size (Badge). In practice one can only suspect or assume MNAR, because it depends on the unobserved values in $\mathbf{Y_{mis}}$ (but see section 4.4.5).

The probability distribution for MAR can be written as:

$$p(\mathbf{R} \mid \mathbf{Y_{obs}}, \mathbf{q}). \tag{4.2}$$

This means that missingness depends only on $\mathbf{Y}_{\boldsymbol{obs}}$, and this relationship is governed by $\mathbf{q}$. In table 4.2, $\mathbf{v_2}$ is MAR if its missing values depend on the observed variable $\mathbf{v_1}$.

Finally, the probability distribution for MCAR is expressed as:

$$p(\mathbf{R} \mid \mathbf{q}). \tag{4.3}$$

This says that the probability of missingness does not depend on the data (neither $\mathbf{Y_{obs}}$ nor $\mathbf{Y_{mis}}$), but that whether positions in **R** take 0 or 1 is still governed by $\mathbf{q}$. So $\mathbf{v_2}$ is MCAR if its missing values do not depend on an observed variable ($\mathbf{v_1}$) or the values of v_2 itself.

Ignorability is another important concept; note that the same word is used with other meanings in other statistical contexts (for examples, see Gelman and Hill 2007). MNAR missingness is "non-ignorable" whereas MAR and MCAR are "ignorable" (Little and Rubin 2002). Ignorability refers to whether we can ignore the way in which data are missing when we impute or augment missing data; it does not imply that one can remove missing data! In the MAR and MCAR mechanisms, imputation and augmentation do not require that we make specific assumptions about how data are missing. On the other hand, non-ignorable MNAR missingness requires such assumptions to build a model to fill in missing values (section 4.4.8).

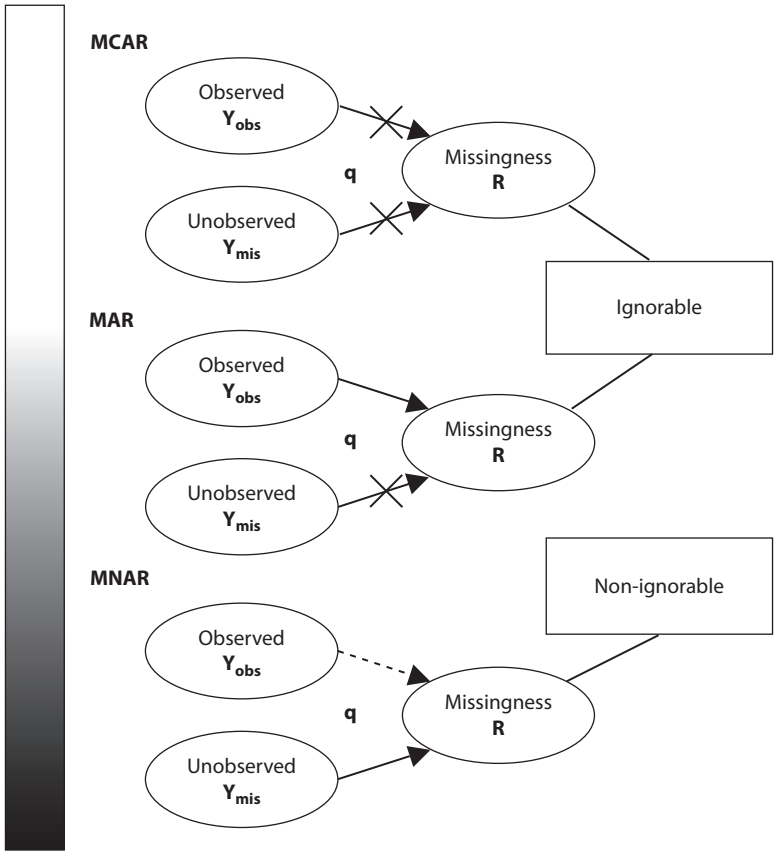

Fig. 4.1 The three missing data mechanisms (MCAR, MAR, and MNAR) and ignorability (whether we need to model the mechanism of missing data) in relation to observed data ($\mathbf{Y_{obs}}$), missing data ($\mathbf{Y_{mis}}$), the missingness matrix (**R**), and their relationships (**q**; parameters that explain missingness, i.e., mechanism). The solid arrows, dotted arrows, and arrows with crosses represent "connection," "possible connection," and "no connection," respectively. The lines connecting ignorability and missingness group the three mechanisms into the two ignorability categories. Also no pure forms of MCAR, MAR, and MNAR exist, and all missingness can be considered as a form of MAR missingness; this is represented by the shaded continuum bar on the left. Modified from Nakagawa and Freckleton (2011).

4.2.4 *Consequences of missing data mechanisms: an example*

Figure 4.2 shows the three different mechanisms of missing data in a bivariate example in two situations where % missing values are different (40% and 80% from the sample size of 200). The missing values are all in Variable 2 (plotted as a response variable; analogous to $\mathbf{v_2}$ in table 4.2) but not in Variable 1 (analogous to $\mathbf{v_1}$ in table 4.2). The population true mean (μ) and standard deviation (σ) for Variable 2 are 0 and 1.41 (variance, $\sigma^2 = 2$), respectively, while the true intercept (α), slope (β) and residual variance (σ_e^2) for the linear relationship between Variable 1 and Variable 2, are 0, 1, and 1, respectively. Parameter estimates from analysis from "observed" data of three missing data mechanisms (i.e., complete case analysis) are summarized in table 4.3.

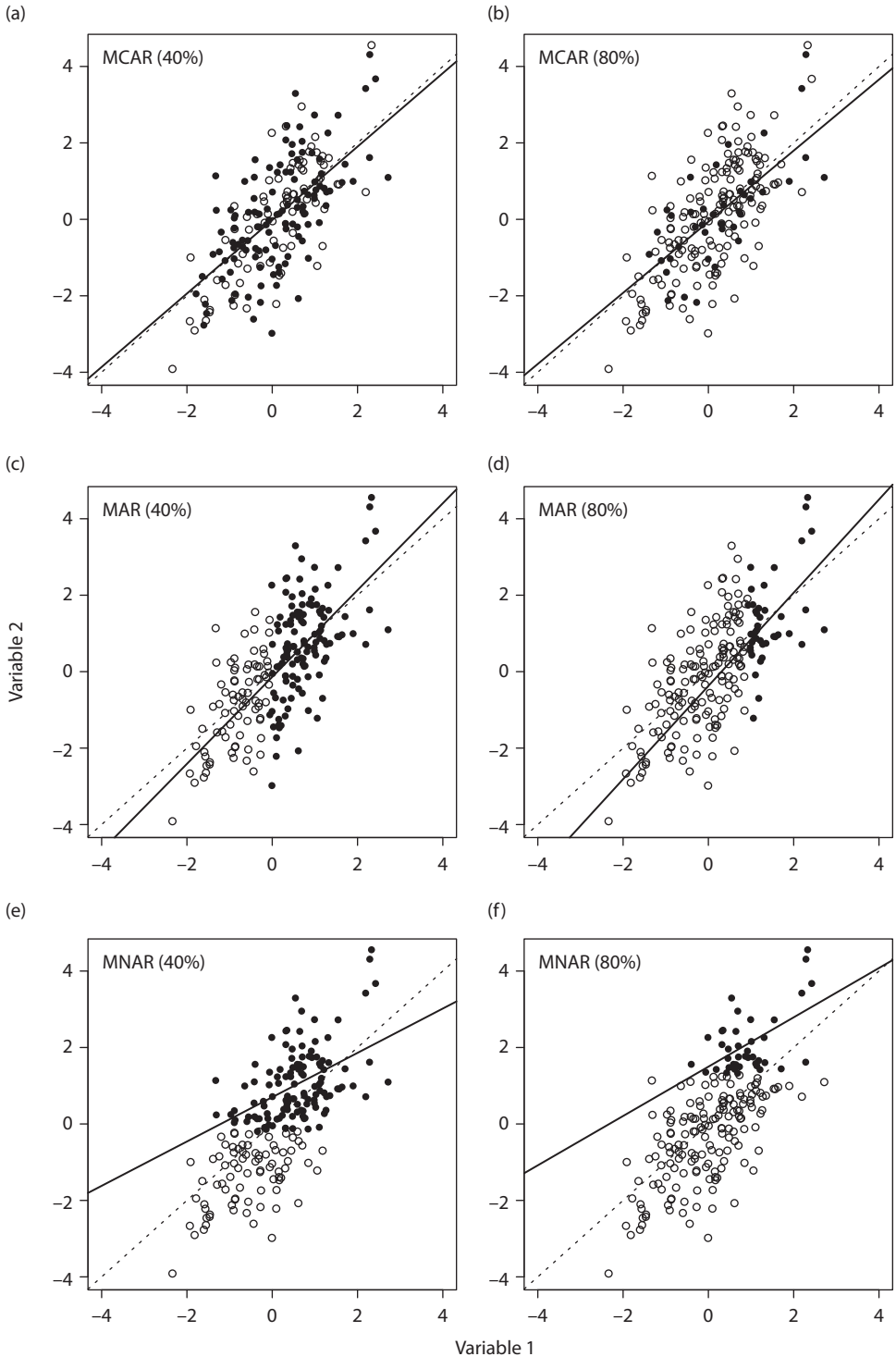

Fig. 4.2 Bivariate illustrations of the three missing data mechanisms and consequences for a) MCAR with 40% missing values (40%), b) MCAR with 80% missing values (80%), c) MAR (40%), d) MAR (80%), e) MNAR (40%), and f) MNAR (80%). Solid circles are observed data and empty circles are missing data; dotted lines represent "true" slopes while solid lines were estimated from observed data.

Table 4.3 The estimates of descriptive statistics for Variable 2 (see the main text) and the estimates from regression analysis of Variable 2 against Variable 1 (complete case analysis), using the complete data set and the three types of data sets with missing values (MCAR, MAR, and MNAR) in two scenarios where 40% or 80% of Variable 2 are missing (the total sample size, $n = 200$; no missing values in Variable 1). The true value for each parameter is $\mu = 0$, $\sigma = 1.414$, $\alpha = 0$, $\beta = 1$, and $\sigma_e = 1$; the mean (μ) and standard deviation (σ) are for Variable 2, α and β are the intercept and slope respectively, and σ_e is the residual standard deviation. For corresponding plots, see figure 4.2

Missing data mechanisms (% missing data)	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\alpha}$	*s.e.*	$\hat{\beta}$	*s.e.*	$\hat{\sigma}_e^2$
No missing data	0.091	1.464	–0.052	0.075	1.069	0.079	1.055
MCAR (40%)	0.129	1.415	–0.019	0.101	0.961	0.106	1.092
MCAR (80%)	0.189	1.351	–0.063	0.155	0.930	0.145	0.950
MAR (40%)	0.723	1.308	–0.139	0.170	1.136	0.177	1.131
MAR (80%)	1.355	1.185	–0.374	0.510	1.219	0.341	1.038
MNAR (40%)	1.040	0.942	0.700	0.095	0.580	0.098	0.831
MNAR (80%)	2.093	0.811	1.499	0.186	0.645	0.163	0.691

As we would expect, parameter estimates from the regression, using the complete data set are close to population true values (table 4.3). As theory suggests, no obvious bias in the parameter estimates from the MCAR data sets can be detected, although standard errors for regression estimates increased (i.e., there is less statistical power). In general, many parameter estimates from the MAR data sets seem to be biased to some certain extent. Noticeably, many parameter estimates from the MNAR data sets seem to be severely biased. In the data sets of all the three mechanisms, deviations from true estimates usually increase when the percentage of missing values is raised, i.e., from 40% to 80% (all relevant *R* code is provided in appendix 4A).

In real data sets, the consequences of missing data will be further complicated by the existence of more than two variables and the presence of missing values in more than one variable. Furthermore, it is usually impossible to unambiguously classify cases into the three mechanisms (Graham 2009, 2012). For example, it is hard to imagine missing data that are entirely unrelated to other variables in the data set, i.e., purely MCAR. Missing data in real data sets are somewhere on a continuum from MCAR through MAR and to MNAR, as depicted in figure 4.1. In a sense, it may be easiest to think of all missing data as belonging to MAR to some degree because MAR resides in the middle of this continuum. Further details can be found in Nakagawa and Freckleton (2008, 2011).

4.3 Diagnostics and prevention

4.3.1 *Diagnosing missing data mechanisms*

In this and the next section (section 4.4), I will use snippets of *R* code along with example data sets. The full *R* code, related data sets, and more detailed explanations of these are all found in appendix 4B.

It is straightforward to visualize missing data patterns with the aid of *R* functions. As an example, I again use a part of the Lundy male sparrow data (table 4.1). The `missingmap` function in the `Amelia` library (Honaker et al. 2011) produces figure 4.3, which is a visual representation of missing data patterns or in fact, a matrix, **R** (missingness). Plotting

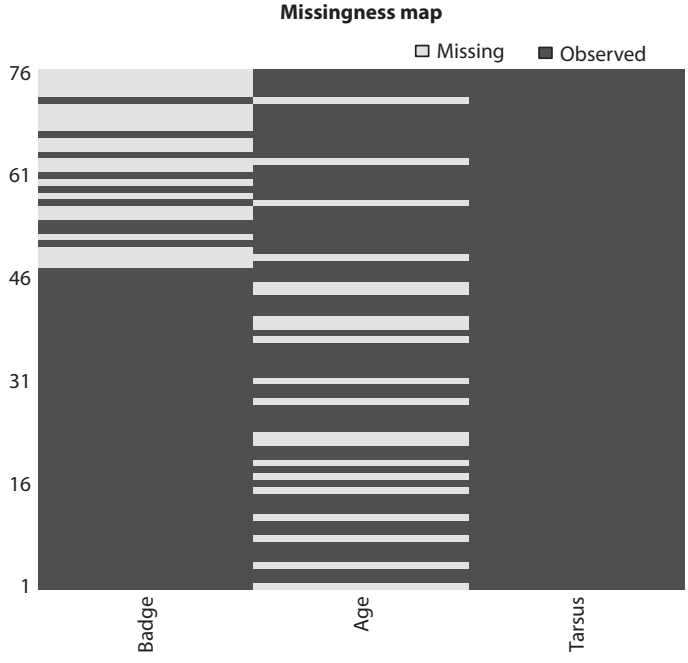

Fig. 4.3 A plot of missing data patterns of the three variables (Badge, Age, and Tarsus), produced by the `missmap` function in the `Amelia` library (Honaker et al. 2011). See text for more details.

missing data patterns can sometimes reveal unexpected patterns such as a cluster of missing values, which were not noticeable during the data collection stages. Then we can ask why such patterns exist. However, missing data patterns alone do not tell us about which missing data mechanism(s) underlie our data.

By deleting cases where missing values exist (complete case analysis), we implicitly assume MCAR. There are a number of ways to diagnose whether or not missing data can be classified as MCAR (reviewed in McKnight et al. 2007). However, as we have learned, MCAR is an unrealistic assumption because such precise missingness is implausible (Little and Rubin 2002; Graham 2009, 2012; see figure 4.1) and also because biological and/or practical reasons generally underlie missingness (Nakagawa and Freckleton 2008). MAR—for which the pattern of missingness is ignorable—is a more realistic assumption. In fact, the MAR assumption is central to many missing data procedures (section 4.4). My main recommendation is to deal with missing values under the assumption of MAR even when all missing data are diagnosed as MCAR (see Schafer and Graham 2002; Graham 2009, 2012; Enders 2010).

When is it really useful to identify missing data mechanisms? You may want to see MCAR diagnostics if you have to resort to missing data deletion. The simplest method is to conduct a series of *t* tests on values between observed and missing groups in each variable (0 being the one group and 1 the other in missingness **R**; see table 4.2), which assess mean difference in the other variables in the data set. If all *t*-tests are non-significant, then you can say missing values in that data set are MCAR; if not, they are MAR or MNAR. However, as the size of the matrix grows, performing and assessing multiple *t*-tests gets tedious very quickly and also may result in Type I errors. Little (1988) proposed a multivariate

version of this procedure, which produces one statistic (a χ^2 value) for the entire data set (for details, see Little 1988; McKnight et al. 2007; Enders 2010). This extension of the *t*-test approach can be carried out by the `LittleMCAR` function in the `BaylorEdPsych R` library (Beaujean 2012).

For the example data set (`PdoDataPart`, see appendix 4B), the test produces $\chi^2_5 = 36.65$ and $p < 0.0001$. We can conclude that this data set contains non-MCAR missingness. This test has the advantage of being simple, but has two major shortcomings: (1) the data set may often have weak statistical power, especially when the observed and missing groups are unbalanced and (2) a non-significant result can be obtained even if missingness is MAR or MNAR. This occurs when, for example, missing values in a variable are related to the high and low values of another variable.

There are neither statistical tests nor visual techniques to distinguish between MAR and MNAR (McKnight et al. 2007; van Buuren 2012). This is not surprising given that the probability distributions for MAR ($p(\mathbf{R} \mid \mathbf{Y_{obs}}, \mathbf{q})$) and MNAR ($p(\mathbf{R} \mid \mathbf{Y_{obs}}, \mathbf{Y_{mi}}, \mathbf{q})$) differ only in that MNAR depends on $\mathbf{Y_{mi}}$ (unobserved values), and we have no way of knowing what unobserved values were. Rather, we need to ascertain whether or not missing values are considered MNAR from our understanding of the biological systems under investigation. For example, in the MNAR example in table 4.1, age information was missing from the oldest birds, because of the limited duration of the study.

Graphical methods for diagnosing missingness are generally much more useful. Visualizations of the relationship between the original data set and missingness (e.g., $\mathbf{m_1}$ in table 4.2) is easily done in *R*, using built-in functions and the `pairs.panels` function from the `psych` library (Revelle 2012).

```
> Missingness <- ifelse (is.na (PdoPartData) == TRUE, 0, 1)
# create the missingness matrix
> MissData <- data.frame (PdoPartData, Missingness)
# combine the original dataset with the missingness matrix
> library (psych) # loading the psych library
> pairs.panels (MissData, ellipses = FALSE, method = "spearman")
```

The resulting figure (figure 4.4) contains visual information on all the original variables and missingness variables, as well as information about all the correlations among these variables. I encourage the reader to study this figure to identify non-MCAR missingness.

4.3.2 *How to prevent MNAR missingness*

As the father of modern statistics, Ronald A. Fisher is reported to have said, "the best solution to handling missing data is to have none," but this is probably not the easiest solution (McKnight et al. 2007). Missing data prevention requires careful planning and execution of studies and experiments, as well as a good understanding of the biological systems at hand, and even then, missing data are often unavoidable (Nakagawa and Freckleton 2008). However, there is a trick that you can use to make missing values much easier to handle. The trick is to begin your study with a data collection plan, wherein you will turn MNAR missingness into MAR missingness. In other words, this means altering non-ignorable missing values to make them ignorable; missing values can then be handled with ordinary missing data procedures such as multiple MI (or without making special assumptions due to MNAR; see Schafer and Graham 2002; Graham 2009, 2012).

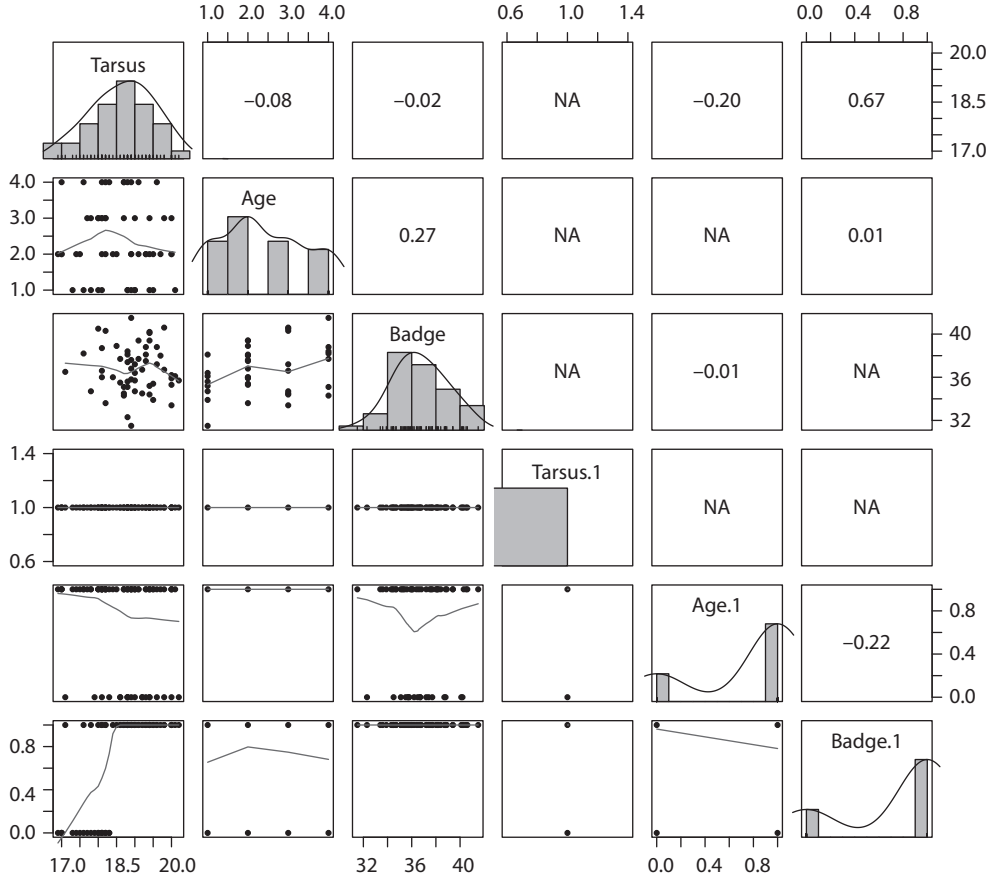

Fig. 4.4 Paired panel plots of the data matrix **Y** and missingness matrix **R** for the house sparrow data set, created by the `pairs.panels` function in the `psych` library (Beaujean 2012). Tarsus, Age, and Badge are numerical values in **Y**, while Tarsus.1, Age.1, and Badge.1 indicate missingness for these values, respectively. The upper triangle panels show Spearman correlations (NA means "not available"), while the lower triangle panels show scatterplots with lowess (locally weighted scatterplot smoothing) lines. The diagonals show histograms. There is some evidence for MAR because the correlation between Tarsus and Badge.1 is high (r_S = 0.67). Similarly, the moderate correlation between Tarsus and Age.1 (r_S = –0.20) suggests that we may have missing data in Age when birds have smaller tarsus size.

When you have a good understanding of your biological system, you usually know which variables will be likely to have missing values. If you collect data on known correlates of these missing-prone variables your missing values will be more likely to be MAR than MNAR. These correlates are called *auxiliary variables* in the missing data literature. An extension of this idea is the *planned missing data design*, in which you make the use of the MAR assumption to deliberately incorporate MAR missingness in your data collection. This may seem very strange at first, but think of a situation where Measurement A is very expensive to collect and is a variable of interest, while Measurement B is very cheap to measure but is not of interest (e.g., A may be a biochemical marker of oxidative stress while B is the color of a trait, which is correlated with this marker). If A and B are correlated, you can collect B for all subjects, while you can only collect A for a random subset

(i.e., creating missing values on purpose). Given missing values in A are MAR, missing data procedures can actually restore the statistical power of your statistical models as if you had collected A for all subjects! This design is called two-method measurement design (Graham et al. 2006; Enders 2010). Investigations into planned missing data design are relatively new and an active area of research (Baraldi and Enders 2010; Graham 2009, 2012; Rhemtulla and Little 2012), but I expect that developments will enormously benefit research planning in ecological and evolutionary studies in the near future.

4.4 Methods for missing data

4.4.1 *Data deletion, imputation, and augmentation*

Three broad categories of methods for handling missing data are: deletion, imputation, and augmentation (McKnight et al. 2007; see also Nakagawa and Freckleton 2008). Data imputation has two subcategories: single imputation and multiple imputation (MI). Schematics in figure 4.5 provide conceptual representations of the four ways of handling missing data (i.e., data deletion, single imputation, MI and DA).

Here I focus on MI under the MAR assumption, because I believe that MI methods are currently the most practical and useful for ecologists and evolutionary biologists. Further, many recent software developments have focused on MI methods (van Buuren 2012), so *R* has a number of libraries available. Despite this focus, I will also provide brief pointers for non-ignorable (MNAR) missing data and sensitivity analysis (in section 4.4.8).

4.4.2 *Data deletion*

Data deletion methods such as list-wise and pair-wise deletion (section 4.1) are efficient ways of dealing with missing data as long as missing data are MCAR (figure 4.5A). Then, relevant analysis (e.g., complete or available case analysis) will produce unbiased parameter estimates with tolerable reductions in statistical power (cf. figure 4.2). If, say, only 1% of cases have missing values, then deletion would certainly offer the quickest way to deal with missing data. However as the fraction of missing cases grows, problems will quickly arise. I would follow Graham's (2009) recommendation that, if 5% or more of cases are missing, one should use multiple imputation or data augmentation.

4.4.3 *Single imputation*

Single imputation (figure 4.5B) has often been used because this procedure will result in a complete data set. There are many commonly used methods for single imputation, such as mean imputation and regression imputation (section 4.1). Other single imputation methods include hot- and cold-deck, and last and next observation carried forward, to name a few (reviewed in McKnight et al. 2007; Enders 2010). These methods often result in severe bias in parameter estimates, especially when missing data are not MCAR, so I will not discuss them further. However, stochastic regression imputation is worth mentioning, as it forms the basis of some missing data procedures introduced below. Like regression imputation, this method uses regression predictions to fill in missing values in a variable by using observed variables, but it incorporates noise in each predicted value by adding error based on a residual term. Under the MAR assumption, parameter estimates from single imputation by stochastic regression are unbiased (for more details, see Gelman and Hill 2007; Enders 2010). Unfortunately, they suffer from biased uncertainty estimates—for example, *s.e.* values are too small or unrealistically precise.

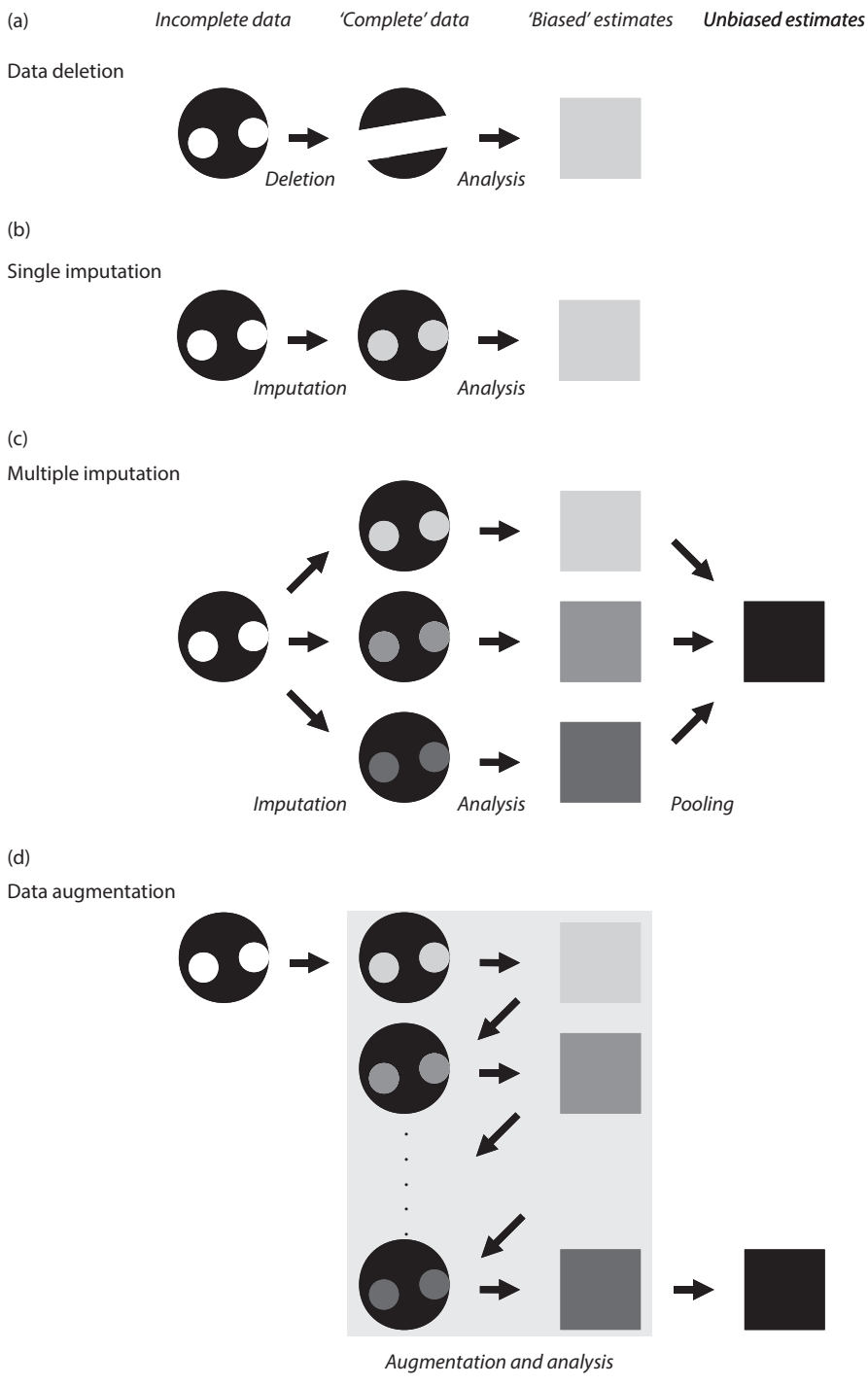

Fig. 4.5 Diagrams illustrating the process of a) data deletion, b) single imputation, c) multiple imputation, and d) data augmentation. "Biased" estimates mean biased parameter estimates, biased uncertainty estimates, or both. A circle represents a data set, and holes in the circle represent missing values. Such holes can be deleted (a) or filled in (b–d). A square represents a set of estimated parameters; the degree of bias in estimation is represented by a gray scale, with darker shades being less biased. See text for details (this figure was modified from Nakagawa and Freckleton 2008).

4.4.4 *Multiple imputation techniques*

Multiple imputation (MI) creates more than one filled-in completed data set. By doing so, MI, proposed by Rubin (1987), has solved the problem of biased uncertainty, which troubles all the available single imputation methods. MI has become the most practical and the best-recommended method in most cases (Rubin 1996; Schafer 1999; Allison 2002; Schafer and Graham 2002; McKnight et al. 2007; Graham 2009; Enders 2010; van Buuren 2012). Among imputation techniques that can generate unbiased parameter estimates under the MAR assumption, most relevant and useful are two methods, expectation maximization (EM) algorithms and Markov chain Monte Carlo (MCMC) procedures. These methods form the basis of multiple imputation.

EM (expectation maximization) algorithms are a group of procedures for obtaining maximum likelihood (ML; chapter 3) estimates of statistical parameters when there exist missing data and unobserved (unobservable underlying or latent; section 4.4.7) variables (for accessible descriptions, see McKnight 2007; Molenberghs and Kenward 2007; Graham 2009; Enders 2010: for more formal treatments, see Dempster et al. 1977; Schafer 1997; Little and Rubin 2002). The EM algorithm that estimates the descriptors of a multivariate matrix, a vector of means ($\mathbf{m}$), and a variance-covariance matrix ($\mathbf{V}$) consists of a two-step iterative procedure (E-step and the M-step). First, the E-step will use a very similar method to stochastic regression imputation to estimate $\mathbf{m}$ and $\mathbf{V}$ ($\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$) from observed values and then "expect" (or fill in) missing values. Next in the M-step, these complete data are used to estimate $\mathbf{m}$ and $\mathbf{V}$ and fill in missing values again. The two steps are repeated until $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$ converge to ML estimates. However, the EM algorithm does not provide uncertainty estimates (*s.e.*) for $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$. To obtain *s.e.*, bootstrapping (i.e., sampling observed data with replacement) can be combined with the EM algorithm to obtain frequency distributions for $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$. This combined procedure is termed the EMB algorithm (Honaker and King 2010; Honaker et al. 2011; see figure 4.6A). I note that the `Amelia` library mentioned above employs the EMB algorithm to conduct MI.

One restriction to the EM and EMB algorithms is the assumption of multivariate normality, or $\mathbf{Y} \sim \text{MVN}(\mathbf{m}, \mathbf{V})$, where all variables come from one distribution. That is why this type of approach is called *joint modeling*. MCMC procedures circumvent this restriction by using a *fully conditional specification* where each variable with missing values can be treated or imputed separately when it is conditioned on other values in the data set (i.e., using Gibbs sampling; van Buuren et al. 2006; van Buuren and Groothuis–Oudshoorn 2011; van Buuren 2012). In this process each variable can have a different distribution and different linear modeling. For example, the algorithm can apply a binomial and a Poisson generalized linear model (chapter 6) for a binary and count variable respectively. This type of procedure is also called *sequential regression imputation* (Enders 2010).

MCMC procedures (and also Gibbs sampling) are often called Bayesian methods (chapter 1) because their goal is to create the posterior distributions of parameters, but methods using MCMC have much wider applications than Bayesian statistics). The MCMC procedure, is akin to the EM algorithm (Schafer 1997) in that it uses a two-step iterative algorithm to find $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$. The imputation step (I-step) uses stochastic regression with observed data. Next, the posterior step (P-step) uses this filled-in data set to construct the *posterior distributions* of $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$. Then, it uses a Monte Carlo method to sample a new set of $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$ from these distributions. These new parameter estimates are used for the subsequent I-step. Iterations of the two steps create the Markov chain, which eventually converges into fully fledged posterior distributions of $\hat{\mathbf{m}}$ and $\hat{\mathbf{V}}$ (figure 4.6B). These

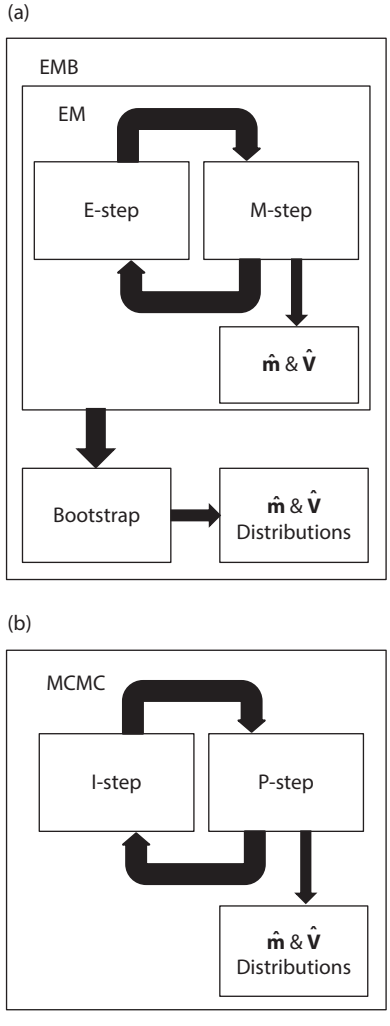

Fig. 4.6 Schematics illustrating the process of a) the EM (expectation maximization) and EMB (expectation maximization with bootstrapping) algorithm with the E-step (expectation) and M-step (maximization), and b) the MCMC procedure with the I-step (imputation) and the P-step (posterior). $\hat{\mathbf{m}}$ is a vector of the means and $\hat{\mathbf{V}}$ is a variance–covariance matrix. Thicker arrows represent iterative processes. See text for details.

distributions are, in turn, used for multiple imputation (for more details, see Schafer 1997; Molenberghs and Kenward 2007; Enders 2010). The two *R* libraries, `mice` (van Buuren and Groothuis–Oudshoorn 2011) and `mi` (Su et al. 2011), are notable here because they both implement MCMC procedures using a fully conditional specification, known as *multivariate imputation by chained equations* (MICE). In the statistical literature (e.g., Schafer 1997), this MCMC procedure is often referred to as data augmentation (see below).

4.4.5 *Multiple imputation steps*

There are three main steps in MI: imputation, analysis, and pooling (figure 4.5C). In the imputation step, you create *m* copies of completed data set by using data imputation methods such as the EM/EMB algorithms or the MCMC procedure. In the analysis step, you run separate statistical analyses on each of *m* data sets. Finally, in the pooling step, you aggregate *m* sets of results to produce unbiased parameter and uncertainty estimates. This aggregation process is done by the following equations (which are automatically calculated in *R*):

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i, \tag{4.4}$$

$$v_W = \frac{1}{m}\sum_{i=1}^{m} s.e._i^2, \tag{4.5}$$

$$v_B = \frac{1}{m-1}\sum_{i=1}^{m} (b_i - \bar{b})^2, \tag{4.6}$$

$$v_T = v_W + v_B + \frac{v_B}{m}, \tag{4.7}$$

where $\bar{b}$ is the mean of b_i (e.g., regression coefficients), which is a parameter estimated from the ith data set (i = 1, 2, . . . , m), v_W is the within-imputation variance calculated from the standard error associated with b_i, v_B is the between-imputation variance estimates, and v_T is the total variance ($\sqrt{v_T}$ is the overall standard error for $\bar{b}$). This set of equations for combining estimates from m sets of results is often referred to as Rubin's rules, as it was developed by Rubin (1987).

Statistical significance and confidence intervals (CIs) of pooled parameters are obtained as:

$$df = (m-1)\left(1 + \frac{mv_W}{(m+1)v_B}\right)^2, \tag{4.8}$$

$$t_{df} = \frac{\bar{b}}{\sqrt{v_T}}, \tag{4.9}$$

$$100(1-\alpha)\%\ \text{CI} = \bar{b} \pm t_{df,(1-\alpha/2)}\sqrt{v_T}, \tag{4.10}$$

where df is the number of degrees of freedom used for t-tests or to obtain t values and CI calculations, and α is the significance level (e.g., 95% CI, $\alpha = 0.05$).

To illustrate the three steps in multiple imputation, I again use the house sparrow data set but this time with the seven variables (EPP, Age, Badge, Fledgling, Heterozygosity, Tarsus, Wing, and Weight). The question this time is which male non-morphological characteristics (i.e., Age, Fledgling, and Heterozygosity) best predict extra-pair paternity (EPP). EPP is a common phenomenon in the animal kingdom, especial among bird species, where males often have offspring outside their social bonds (Griffith et al. 2002). Nakagawa and Freckleton (2011) used the `Amelia` library (i.e., the EMB algorithm) for MI with this data set. Here I use the `mice` library (MCMC algorithm) to carry out the three steps of MI.

```
> library (mice) # loading the mice library
# the imputation step with 5 copies
> imputation <- mice (PdoData, m = 5, seed = 7777)
> analysis <- with (imputation, glm (EPP ~ Age + Fledgling
    + Heterozygosity, family = quasipoisson)) # the analysis step
    with a GLM (see chapters 6 & 12)
> pooling <- pool (analysis) # the pooling step
> summary (pooling)
```

With this three-step MI process, we obtain unbiased parameter and uncertainty estimates (table 4.4; for individual outputs, see the online appendix 4C).

Table 4.4 Results of analyses for the house sparrow data, using complete case analysis, `mice`, and `mi` (the latter two are multiple imputation via MCMC procedures). Estimates from `mice` and `mi` are the pooled model-averaged parameter estimates from the five imputed data sets ($m = 5$), pooled regression coefficients ($\bar{b}$), overall standard error, *s.e.* ($\sqrt{v_T}$), 95% confidence intervals (CI), and the fraction of missing information (γ). For details, see appendix 4C

Procedure	Predictor	Estimate	*s.e.*	Lower CI	Upper CI	γ
Complete case analysis	Intercept	–1.733	2.391	–7.009	2.466	–
	Age	0.479	0.273	–0.062	1.020	–
	Fledgling	0.090	0.132	–0.155	0.368	–
	Heterozygosity	0.167	2.232	–3.929	4.899	–
`mice`	Intercept	–3.389	2.406	–8.343	1.565	0.335
	Age	0.750	0.214	0.315	1.184	0.258
	Fledgling	–0.040	0.094	–0.227	0.148	0.099
	Heterozygosity	1.605	2.258	–3.102	6.312	0.392
`mi`	Intercept	–3.624	2.416	–	–	–
	Age	0.782	0.236	–	–	–
	Fledgling	–0.046	0.099	–	–	–
	Heterozygosity	1.844	2.221	–	–	–

In addition, you will get a value for each regression coefficient, labeled as "fmi," which stands for the *fraction (or rate) of missing information*, γ. This index γ varies between 0 and 1, and is a very important feature of MI, because it reflects the influence of missing data on uncertainty estimates for parameters. The fraction of missing information is defined by:

$$\gamma = \frac{v_B + v_B / m + 2/(df + 3)}{v_T}, \tag{4.11}$$

where all components are defined as in equations 4.5, 4.7, and 4.8. As you can see, the fraction of missing information, γ, reflects not only the fraction of missing values, but also the importance of missing values in relation to the complete information (McKnight et al. 2007; Enders 2010). There are two more indices in the missing data literature: ρ (the relative increase in variance due to missing data) and λ (the fraction of missing information assuming m is very large). They can be expressed as:

$$\rho = \frac{v_B + v_B / m}{v_W}, \tag{4.12}$$

$$\lambda = \frac{v_B + v_B / m}{v_T}. \tag{4.13}$$

Also, γ is often written using ρ, as:

$$\gamma = \frac{\rho + 2 / (df + 3)}{1 + \rho}. \tag{4.14}$$

The importance of γ can be more easily appreciated by examining λ (equations 4.7, 4.13) because λ is the ratio of variance due to missing data (between-imputation variance, v_B), in relation to the total variance (v_T).

This index γ has two practically useful properties. First, when missing data are non-ignorable (MNAR), γ will be large (McKnight et al. 2007), although there is no definite test to distinguish between MAR and MNAR (section 4.3.1). Li et al. (1991) proposed that γ up

to 0.2 can been seen as "modest," 0.3 as "moderately large" and 0.5 as "high." Although these benchmarks should not be used as absolute (analogous to Cohen's benchmarks, 1988), it is true that when $\gamma > 0.5$, the way missing data are handled will impact the final parameter estimates and statistical inferences (van Buuren 2012).

Second, γ can be used to quantify the efficiency of MI. The relative efficiency (ε) quantifies the errors due to MI, relative to its theoretical minimum (which occurs when $m = \infty$)

$$\varepsilon = \left(1 + \frac{\gamma}{m}\right)^{-1}. \tag{4.15}$$

For example, at $m = 3$ and $\gamma = 0.5$, the efficiency is 85.71 % while at $m = 10$ and $\gamma = 0.5$, the efficiency is 95.24%. So even in the latter case there is still much room for improvement in efficiency. Although Rubin (1987) suggested that m between 3 and 10 would be sufficient. Given that m can be easily increased with the use of *R*, we should aim for over 99% (m = 50 with $\gamma = 0.5$ produces $\varepsilon = 0.9901$). However, for practicality, we can use $m = 5$ during the analysis step, and only use high m for the "final" three steps of MI (van Buuren 2012). Other recommended rules of thumb or guidelines on the number of m can be found elsewhere (e.g., Graham et al. 2007; von Hippel 2009).

It is important to check the results from the MI models of your choice. One way of doing this (sensitivity analysis) is to run MI using a different library. The three-step MI process can be done using the `mi` library (Su et al. 2011), which uses a different version of MCMC procedure from the `mice` library.

```
> library (mi)
# get information on each variable
> info <- mi.info (PdoData)
# EPP and Fledgling are count data
> info <- update (info, "type", list (EPP = "count", Fledgling
    = "count"))
# the imputation step with 5 copies
> imputation <- mi (PdoData, info = info, n.imp = 5, seed = 777)
# the analysis step (with GLM) and the pooling step
> AandP <- glm.mi (EPP ~ Age + Fledgling + Heterozygosity,
    family = quasipoisson, mi.object = imputation)
> display (AandP)
```

The results are very similar for analyses using `mice` and `mi` (table 4.4).

The *R* code for both libraries gives the impression that MI procedures may be very simple and straightforward. In one sense, this is true, but there are many practical pitfalls, which need consideration before and during MI (e.g., convergence of the imputation steps and which variables should be included for MI). I will cover such practical considerations in section 4.5.1.

4.4.6 *Multiple imputation with multilevel data*

Multilevel structures in ecological and evolutionary data are common because biological processes by nature occur in hierarchies; therefore an ability to handle missing data for multilevel data sets will prove extremely useful. So-called multilevel or hierarchical data are modeled by linear and generalized linear mixed-effects models (LMM and GLMM respectively; chapter 13; Bolker et al. 2009; O'Hara 2009). However, proper missing data procedures for multilevel data are still in their infancy (van Buuren 2011, 2012). Available

R functions are currently very limited in both number and capacity. I will introduce some extensions of the above MI methods but great care needs to be taken when applying them.

Data are frequently arranged in clusters or groups (e.g., sibships, stands of trees, and the like), each of which has its own mean (and therefore intercept and sometimes slope). Handling of missing data in such cases is not straightforward because the imputation needs to account for this clustering (Graham 2009, 2012; van Buuren 2011, 2012). In other words, you have multiple levels of vectors of means and variance–covariance matrices (**m** and **V**; section 4.4.4).

Longitudinal data are a case in point; imagine growth data of house sparrow chicks. Half the broods are fed extra food every second day (this was our treatment and what we were interested in); tarsus measurements (a good size indicator) of chicks were taken at 6 different time points (2, 4,. . .12 days after hatching). Here, each chick is a cluster and also, each brood acts as a higher-level cluster (usually 3–5 chicks). Typical to such data, some tarsus measurements are missing because some chicks died/disappeared due to adverse weather, predation etc. This data set of the seven variables (ChickID, Treatment, Age, Tarsus JulianDate, BroodID, and Year) includes 273 chicks from 76 broods, with 403 measurements missing out of 1638 (see Cleasby et al. 2011 and appendix 4C). Let us see how a normal MI procedure performs using the `mi` library. The coding will be exactly the same as the previous example, but I will introduce a LMM in the analysis step using the function `lmer.mi`.

```
# get information on each variable
> info <- mi.info (PdoGrowthData)
> imputation <- mi (PdoGrowthData, info = info, n.imp = 5, seed = 777)
# the imputation step with 5 copies (the default)
> AandP <- lmer.mi(Tarsus ~ Treatment + I (Age - 12) + (I (Age - 12)|
    ChickID) + (1 | BroodID), mi.object = imputation)
# the analysis step (with LMM; see Chapter 13) and the pooling step;
# note that I(Age-12) makes treatment effect be assessed at 12 days
# after hatching
> display (AandP)
```

This process gives us some (sensible) results (table 4.5; detailed results are in appendix 4C) and similar approaches have been often used. However, the validity of performance without explicitly specifying clustering and its consequences are not well studied (van Buuren 2011). In the `mice` library, we can actually specify grouping by incorporating the `pan` library, which uses a special MI procedure designed for two-level clustered data (Schafer 2001; Schafer and Yucel 2002). A current limitation is that only one grouping variable is allowed.

```
> preparation <- mice (PdoGrowthData1, maxit = 0)
# running an empty imputation for the two objects below as preparation.
#   Also variables in PdoGrowthData were turned numerical
# the predictor matrix
> predictor <- preparation $ predictorMatrix
# the vector of imputation methods
> imputation <- preparation $ method
# specify ChickID as a grouping factor
> predictor ["Tarsus", "ChickID"] <- -2
> imputation ["Tarsus"] <- "2l.pan"
```

```
# using the 2-level mixed modeling method from the pan library
> imputation <- mice (PdoGrowthData1, m = 5, seed = 7777)
# the imputation step with 5 copies
> analysis <- with (imputation, lmer (Tarsus ~ Treatment + I
(Age - 12)  + (I (Age - 12) | ChickID) + (1 | BroodID))
# the analysis step with a LMM (see Chapter 13)
> pooling <- pool (analysis) # the pooling step
> summary (pooling)
```

The preparation is a little involved, but the three-step MI process is the same as above. The results from `mice` specifying grouping in this data set resemble those from `mi` (table 4.5). This is encouraging, but recall that we were unable to include brood identities (i.e., correlated structure) as a grouping factor, so one should draw conclusions cautiously.

There is another important problem in multilevel data: there are multiple levels of predictors, so missing data processes can operate at different levels. Consider two-level data; if the response is weight at time t_i, predictors can be height at time t_i (level 1) and sex (level 2). If weights are taken at 6 different occasions (t_1–t_6), missing data on sex for one individual can appear as missing values in 6 cells. If we subject this data set to normal MI procedures, these 6 cells may be assigned different sexes! Where multiple types of predictors are present, Gelman and Hill (2007) suggest data imputation should be carried out separately for each level (e.g., time and sex). The `mice` library has this capability, but it is currently limited to only two levels (i.e., only one clustering variable is allowed).

Table 4.5 Results of analyses for house sparrow data, treated as a multilevel data set. Estimates are from four procedures: complete case analysis, MI using both the mice and mi libraries, and DA using MCMCglmm. Estimates from mice and mi are the pooled model-averaged parameter estimates from the five imputed data sets ($m = 5$), pooled regression coefficients ($\bar{b}$), overall standard error, *s.e.* ($\sqrt{v_T}$), 95% confidence intervals (CI), and the fraction of missing information (γ). For MCMCglmm, the estimates are posterior means, *s.e.* are standard deviation of the posterior distributions of the estimates, and CI represents credible intervals. Only the results from the fixed factors are presented. For details, see appendix 4C

Procedure	Predictor	Estimate	*s.e.*	Lower CI	Upper CI	γ
Complete case analysis	Intercept	18.110	0.152	17.807	18.413	–
	Treatment	0.167	0.160	–0.15	0.486	–
	Age	1.143	0.011	1.122	1.165	–
`mice`	Intercept	17.880	0.362	17.168	18.592	0.020
	Treatment	0.316	0.229	–0.133	0.766	0.028
	Age	1.157	0.009	1.139	1.174	0.097
`mi`	Intercept	18.147	0.208	–	–	–
	Treatment	0.259	0.241	–	–	–
	Age	1.147	0.010	–	–	–
`MCMCglmm`	Intercept	18.015	0.402	17.171	18.720	–
	Treatment	0.355	0.247	–0.106	0.837	–
	Age	1.169	0.011	1.149	1.192	–

Other issues associated with imputation in multilevel data are described in van Buuren (2011; see also Raudenbush and Bryk 2002; Daniels and Hogan 2008; Enders 2010; Graham 2012).

4.4.7 *Data augmentation*

The processes and results of data augmentation (DA; Graham 2009, 2012; Enders 2010) are similar to those of MI. The main difference is that in MI, the user will see the replaced missing values, while DA internalizes the three-step procedures, including Rubin's rules with $m = \infty$, and feedback between the imputation and analysis steps (figure 4.5D; *sensu* McKnight et al. 2007). DA is superior to MI because a DA procedure is akin to the number of data imputations (or augmentations) being infinite, and also because there is a feedback process between missing data and parameter estimation (Nakagawa and Freckleton 2008). However, MI has an advantage: DA can only use variables that are in the model, while MI can include auxiliary variables, which may often be required to convert MNAR missingness into MAR (section 4.3.2). Therefore, in most cases, MI procedures are recommended over DA (Graham 2009, 2012).

In the case of multilevel data, DA procedures may sometimes be preferable. If the response variable is the only variable with missing data, as is the case with the sparrow growth data used in section 4.4.3, DA can treat such missing values appropriately by taking all the clustering groups (e.g., individuals, broods, and families) into account. In Bayesian statistical packages, such features are usually included as the default. Here, I use the `MCMCglmm` library (Hadfield 2010).

```
> library (MCMCglmm)
# run a Bayesian LMM (see Chapter 13)
> model <- MCMCglmm (Tarsus ~ Treatment + I (Age - 12), random = ~ us
    (I (Age - 12)) : ChickID + BroodID, data = PdoGrowthData,
    verbose = FALSE)
> summary (model)
```

In this case, the results are very similar to those from `mi` and `mice` (table 4.5). Note that the `MCMCglmm` function will not tolerate missing values in predictors. However, if multiple variables with missing data are all entered as responses (i.e., multi-response models; Hadfield 2010), DA will handle missing values for all these response variables. As an example in which we used this strategy in a bi-response/bivariate meta-analysis, see Cleasby and Nakagawa (2012). It is worth mentioning that multi-response (or multivariate) models are closely related to structural equation modeling (SEM), which is sometimes referred to as latent variable modeling, path analysis, or causal modeling (chapter 8). Missing data in such models are briefly discussed later (section 4.5.3).

4.4.8 *Non-ignorable missing data and sensitivity analysis*

As mentioned above, there are no tests to detect MNAR (non-ignorable) missingness, so we need to rely on our understanding and knowledge of the biological systems at hand. We can, however, suspect that MNAR missingness is possible, especially when the fraction of missing information (γ) is high ($\gamma > 0.5$). Two main methods exist for non-ignorable (MNAR) missingness: *selection models* and *pattern-mixture models*. The details of the MNAR methods are beyond the scope of this chapter, so I refer readers to accessible accounts elsewhere (Allison 2002; Molenberghs and Kenward 2007; Enders 2010). However, I will

mention some main aspects of these models. Both models require constructing specific assumptions with regard to MNAR missingness. If these assumptions are incorrect, these non-ignorable models may perform worse than the models for ignorable missingness (i.e., MI and DA). To put it simply, a good MAR model may be better than a bad MNAR model (Schafer 2003; Demirtas and Schafer 2003).

The main problem of non-ignorable missing data is that there are an infinite number of ways in which such missingness can occur. Naturally, very few generally applicable software implementations are able to cope with infinitely different manifestations of non-ignorable missingness (Allison 2002). However, there is an *ad hoc* sensitivity analysis to explore the possible impacts of non-ignorable missingness on the pooled estimates from MI under MAR (Rubin 1987). For example, you might suspect the age variable in the sparrow data to be MNAR rather than MAR (section 4.3.1). It is possible younger birds (or older birds) are selectively missing. Such MNAR missingness can be explored by first adding (or subtracting) imputed values under MAR. We can then compare pooled estimates from this sensitivity analysis (a MNAR model) to the original estimates under MAR. Rubin (1987) suggested a 20% decrease or increase in imputed values would be a sufficient sensitivity test, but this is an arbitrary suggestion. Enders (2010) suggests ± 0.5 standard deviation of the variable should be added. This sensitivity method can be easily implemented using the `mice` library. You will find an example analysis in appendix 4C.

4.5 Discussion

4.5.1 *Practical issues*

There are several practical considerations to consider prior to using MI or DA procedures, and I discuss five of them here. First, is there a minimum requirement for sample size? This question is hard to answer. Of course, larger samples are desirable, because missing values in a small data set further decrease the amount of information, which is already limited (Graham 2009). However, Graham and Schafer (1999) conducted a simulation study where they showed that a MI procedure, which assumes multivariate normality, performed very well with up to 18 predictors and 50% missing data; this means that the data set only had around 15 degrees of freedom. They also demonstrated that a joint modeling approach with the multivariate normal assumption did well with non-normal data (a version of the `norm` library was used in this study; Schafer 1997) although such an approach would be limited compared to sequential regression imputation (used in the `mi` and `mice` libraries).

This leads to my second point. For MI procedures assuming a multivariate normal distribution such as in the `norm` and `Amelia` libraries, non-normal data should be transformed first. Indeed, the `Amelia` library comes with various transformation options (Honaker et al. 2011). Back-transformation can be used to recover the original scale. A related issue is whether imputed data should be rounded when the original data are integers. Generally it is not a good idea to do so, unless an imputed variable is a response variable to which a Poisson regression (chapter 6) will be applied (Graham 2009, 2012; Enders 2010; van Buuren 2012). Furthermore, if you are using MI procedures with the multivariate normal assumption, categorical variables should probably be turned into binary variables using *dummy coding*. For example, if you have a categorical variable with four levels, this variable can be recoded into three binary (dummy) variables. More generally, p levels in a categorical variable can be turned into $(p-1)$ dummy variables. Note that coding dummy

variables from a categorical variable can be easily done in *R* using `dummy.code` in the `psych` library (Revelle 2012; see an example in the online appendix 4C). If you are using sequential regression imputation such as in the `mi` and `mice` libraries, you need to make sure missing values in categorical data are imputed with techniques for categorical data (e.g., logistic and multinomial regression).

Third, for MI, it is important to check for convergence in the imputation step. Convergence here means that an imputation step reaches a set of stable values for a vector of means (**m**) and a variance–covariance matrix (**V**) (section 4.4.4). There are graphical functions to assess convergence in the two *R* libraries mentioned (`Amelia` and `mi`; see appendix 4C). If you have trouble with convergence in MI, transformation of skewed data may help, as skewed data could be slowing down imputation processes (Graham 2009, 2012).

Fourth, when your statistical models include interaction terms, such terms should be included in the imputation step in MI procedures (von Hippel 2009; Graham 2009, 2012; Enders 2010; van Buuren 2012). Interaction terms usually come in two forms: the product of two continuous variables, or the product of one continuous variable and one categorical (dummy) variable (e.g., males and females). When creating interaction terms, a continuous variable needs to be centered (i.e., subtracting the mean from each value). In fact, centering or scaling (i.e., *z*-transformation) of all continuous variables is very frequently a good idea in regression modeling because this process can make linear models more interpretable (e.g., the intercept will be located at the means of predictors; Schielzeth 2010). Inclusion of interactions in the imputation step is necessary, because if you do not consider a particular interaction in the imputation step, the effect of this interaction can be lost even when missing data are MCAR. This is because data imputation is carried out assuming such an interaction does not exist (Enders 2010; Enders and Gottschall 2011). The same applies to a quadratic term, as it can be seen as an interaction with itself. These derived terms (i.e., terms created by existing variables) should be handled by *passive imputation* rather than included as extra variables in the data matrix. Passive imputation maintains relationships between original and derived variables during the imputation process (von Hippel 2009; van Buuren and Groothuis–Oudshoorn 2011). Examples for these processes are found in appendix 4C and in van Buuren and Groothuis–Oudshoorn (2011).

Fifth, our "expert" knowledge is useful during MI. The ranges, or possible maxima and minima, for variables with missing data can be included as *ridge priors* in a MI procedure, such as that in the `Amelia` library (Honaker et al. 2011; see Nakagawa and Freckleton 2011). This process potentially reduces bias, especially when the fraction of missing information (γ) is at least moderately large. Unfortunately, the ridge prior functionality is not implemented in the `mice` and `mi` libraries (but see the argument `squeeze` in `mice`). I recommend more than one library be used to run MI for a data set as a form of sensitivity analysis (section 4.4.5). If the results from different libraries disagree, one likely explanation is that the imputation step did not converge.

4.5.2 *Reporting guidelines*

For publication, it is advisable to provide details and rationale of your missing data procedures, because such procedures will probably look foreign and even outlandish to potential editors and reviewers. Here, I will present the reporting guidelines for missing data analysis from van Buuren (2012). His list consists of 12 items that should be included, when reporting results obtained from MI procedures.

(1) *Amount of missing data*: Give the ranges of % missing values in all variables and the average % in your data set.
(2) *Reasons for missingness*: Give reasons why such missing values were present.
(3) *Consequences*: Report known differences between subjects with and without missing values.
(4) *Method*: Describe which method was used, and under what assumptions (e.g., a MCMC procedure for MI under MAR).
(5) *Software*: Name the software libraries (e.g., `Amelia`) along with descriptions of the important settings.
(6) *Number of imputed data sets*: This is m in the imputation step (see section 4.4.5).
(7) *Imputation model*: Report the variables included in the imputation step (i.e., the imputation model) and whether any transformations were applied.
(8) *Derived variables*: Mention what kind of derived variables (e.g., interaction and quadratic terms) were included in the imputation step.
(9) *Diagnostics*: Report on diagnostics for convergence of the methods used, methods (section 4.5.1), and for checking whether imputed data are plausible.
(10) *Pooling*: Explain how pooling of results was done (usually pooling of m estimates by Rubin's rules; section 4.4.5), if possible along with related indices including, most importantly, the fraction of missing information (γ) and the relative efficiency (ε).
(11) *Complete case analysis*: Report results from complete case analysis, and compare with those from proper missing data procedures (i.e., MI and DA).
(12) *Sensitivity analysis*: Conduct sensitivity analyses, and report the results. Sensitivity analysis can be in the forms of Rubin's ad hoc adjustment or the use of different software packages.

van Buuren (2012) considers items 1, 2, 3, 4, 6, and 11 are essential, but that the others can be reported in an appendix or online materials. Although his list was tailored for MI, I believe that following his guidelines will be useful even when using DA and other missing data procedures. Such details will be certainly helpful for editors and reviewers who are unfamiliar with missing data methods.

4.5.3 *Missing data in other contexts*

Here I provide you with connections between this chapter and other chapters in this book. As mentioned in section 4.1, missing data procedures may be essential for model selection (chapter 3; Nakagawa and Freckleton 2011), although there is surprisingly little research on this relationship (but see Claeskens and Hjort 2008). Some procedures for censored or truncated data (chapter 5) involve an imputation step. In addition to the imputation used in the `NADA` library discussed in that chapter, the `kmi` library (Allignol 2012) uses a Kaplan–Meier estimator to impute missing censoring times.

Different forms of linear models, such as GLMs (chapter 6), models with overdispersion (chapter 12), and mixed models (chapter 13), can be integrated within MI procedures at the analysis step. However, special care is required for multilevel data (Raudenbush and Bryk 2002; van Buuren 2011). All the regression models can be seen as special cases of structural equation modeling, SEM (causal modeling or mediation analysis; chapter 8). SEM has a long history of missing data methods (reviewed in Allison 2002; Enders 2010), and the majority of stand-alone SEM software libraries (e.g., `Mplus` and `AMOS`) come with missing data procedures (MI or DA). There are a number of other *R* libraries available for missing data in SEM, including `bmem` (Yuan and Zhang 2012) and `rsem` (Zhang and

Wang 2012). Meta-analysis (chapter 9) is a type of weighted regression model. Therefore, missing data procedures described in this chapter are applicable for, at least, predictors (called moderators in the meta-analysis literature; Pigott 2009, 2012). However, treating potential missing data in the response variable (i.e., effect size statistics) has attracted much research, and has its own unique techniques, some of which are akin to selection models for MNAR missingness (Sutton 2009).

Hadfield (2008) utilizes missing data theory in evolutionary quantitative genetic contexts. He showed that MNAR missingness could be converted to MAR missingness using pedigree information, which can be included as a correlation matrix in mixed models. Genetic relatedness can act as a kind of auxiliary variable; siblings must share similar morphological characters. In a similar manner, spatial correlation (chapter 10) and phylogenetic correlation (chapter 11) can inform missing values in associated models because these different types of correlations are, in fact, the same (or very similar) mathematically in terms of specifying relationships among data points in the response variable (Ives and Zhu 2006; Hadfield and Nakagawa 2010). Interestingly, phylogenetic comparative analysis by Fisher et al. (2003) was the very first case of using MI in evolutionary biology, but few followed their initiative. The shortcomings of ignoring missing data are now, however, starting to be recognized in comparative analysis (Garamszegi and Møller 2011; González-Suárez et al. 2012), with some implementations of missing data procedures appearing (e.g., `PhyloPars`; Bruggeman et al. 2009). We can expect a rapid future integration and development of missing data procedures in this and related areas of research.

4.5.4 *Final messages*

Missing data are pervasive, and pose problems for many statistical procedures. I hope I have convinced you that we all should be using methods that treat missing data properly (i.e., MI or DA), rather than deleting data or using single imputation. Importantly, it is not difficult to implement these missing data procedures (in particular, MI) with the aid of *R*. I also hope that you will now think about the missingness mechanisms when planning studies (i.e., collecting auxiliary variables). Especially, I think that ecologists and evolutionary biologists can probably benefit a lot from learning the planned missing design (Baraldi and Enders 2010; Graham et al. 2006; Graham 2009, 2012; Rhemtulla and Little 2012), although such a concept is nearly unheard of in our field.

I also presented you with some current difficulties associated with missing data. There are no easy solutions for missing values in multilevel data, especially when missing values occur in multiple levels and when clustering occurs at more than two levels. Nor is the implementation of MNAR models straightforward. But missing data theory is an active area of research, so who knows what the future will bring to us and to *R*? Enders (2010) comments that "Until more robust MNAR analysis models become available (and that may never happen), increasing the sophistication level of MAR analysis may be the best that we can do."

Acknowledgments

I thank Losia Lagisz for help with figure preparation. I also thank Shane Richards, Gordon Fox, Simoneta Negrete, Vini Sosa, and Alistair Senior for their very useful and constructive comments on earlier versions of this chapter. I gratefully acknowledge support from the Rutherford Discovery Fellowship (New Zealand).