Behav Ecol Sociobiol (2011) 65:103–116
DOI 10.1007/s00265-010-1044-7

ORIGINAL PAPER

Model averaging, missing data and multiple imputation: a case study for behavioural ecology

Shinichi Nakagawa · Robert P. Freckleton

Received: 1 January 2010 / Revised: 9 August 2010 / Accepted: 9 August 2010 / Published online: 4 September 2010

Abstract Model averaging, specifically information theoretic approaches based on Akaike's information criterion (IT-AIC approaches), has had a major influence on statistical practices in the field of ecology and evolution. However, a neglected issue is that in common with most other model fitting approaches, IT-AIC methods are sensitive to the presence of missing observations. The commonest way of handling missing data is the complete-case analysis (the complete deletion from the dataset of cases containing any missing values). It is well-known that this results in reduced estimation precision (or reduced statistical power), biased parameter estimates; however, the implications for model selection have not been explored. Here we employ an example from behavioural ecology to illustrate how missing data can affect the conclusions drawn from model selection or based on hypothesis testing. We show how missing observations can be recovered to give accurate estimates for IT-related indices (e.g. AIC and Akaike weight) as well as parameters (and their standard errors) by utilizing 'multiple imputation'. We use this paper to illustrate key concepts from missing data theory and as a basis for discussing available methods for handling missing data. The example is intended to serve as a practically oriented case study for behavioural ecologists deciding on how to handle missing data in their own datasets and also as a first attempt to consider the problems of conducting model selection and averaging in the presence of missing observations.

Keywords Data augmentation · Data deletion · Estimation bias · The rate of missing information · Expectation maximization · QAIC · EPP · MCMC · House sparrows

Communicated by L. Garamszegi

This contribution is part of the Special Issue "Model selection, multimodel inference and information-theoretic approaches in behavioural ecology" (see Garamszegi 2010).

Electronic supplementary material The online version of this article (doi:10.1007/s00265-010-1044-7) contains supplementary material, which is available to authorized users.

S. Nakagawa
Department of Zoology, University of Otago,
P.O. Box 56, Dunedin, New Zealand

S. Nakagawa (✉) · R. P. Freckleton
Department of Animal and Plant Sciences,
University of Sheffield,
Sheffield S10 2TN, UK
e-mail: shinichi.nakagawa@otago.ac.nz

Introduction

Over recent years, information theoretic (IT) approaches (Burnham and Anderson 2002) have had far-reaching effects on the way that statistical analysis is conducted in ecology and evolutionary biology (Strimmer and Rambaut 2002; Rushton et al. 2004; Johnson and Omland 2004; Stephens et al. 2005, 2007; Lukacs et al. 2007; Garamszegi et al. 2009a, b; Garamszegi 2010). The reason for this has been an increased focus on modelling of data, rather than binary hypothesis testing (Hilborn and Mangel 1997; Bolker 2008), and the consequence has been a sharpening and refocusing of statistical practice and philosophy among ecologists and evolutionary biologists, even if not everyone uses IT methods (Stephens et al. 2005, 2007; Whittingham et al. 2006; Link and Barker 2006).

IT methods allow a number of important statistical problems to be dealt with. Most prominently these include model selection bias and data dredging (Burnham and Anderson 2002; Whittingham et al. 2006). The process termed 'model averaging' plays a central role in resolving these problems. Model averaging obtains weighted parameter estimates from models in a model set rather than relying on parameter estimates from a single model. Furthermore, model averaging provides uncertainty estimates for parameters (i.e. unconditional variance or standard errors), which incorporate both sampling variance for a given model and a variance component for model selection uncertainty (Burnham and Anderson 2002). As a result, model averaging offers more reliable and robust point and uncertainty estimation of parameters (for more detailed descriptions model averaging, see Burnham and Anderson 2002; Richards et al. 2010; Symonds and Moussalli 2010). Such robustness is even true in complex cases, for instance when there is collinearity among predictors (Freckleton 2010).

There are clearly other general statistical issues that need to be addressed, and one important issue, which affects almost all datasets, is how to deal with the problem of missing data (Nakagawa and Freckleton 2008). This problem has been largely ignored in the ecological and evolutionary literature. However, the problem of handling missing data is important because all information criterion-based approaches require complete cases (i.e. no missing data) if information criterion values of all models within a given set are to be compared. Statistics such as the IT criteria, e.g. the Akaike's information criterion (AIC) or AIC with a second-order bias correction, as well as measures of fit such as R^2 are not comparable within a model set if any of variables in the dataset include one or more missing observations.

The commonest approach for dealing with missing data is to delete cases containing missing observations. Then, model selection and goodness of fit statistics become comparable among models in a model set, and the problem seems to be solved. However, such an approach (termed case-wise data deletion or complete-case analysis) raises two problems. The first is data or information loss (resulting in reductions in estimation precision and thus in statistical power). For example, imagine that we have ten predictors with 5% of cases of each predictor missing. In order to ensure that all cases are complete for all variables, we would have to delete as much as 40% of the cases from the whole dataset on average. On its own this is severe enough a problem, given a typical dataset in behavioural studies is relatively small in general (Still 1992; Nakagawa 2004) and would greatly detract from statistical power. However, in addition to this, the second problem is that missing data result in biased estimates of parameters (e.g. intercepts and slopes in regression problems; Rubin 1976). This becomes particularly important if data are missing because of some underlying biological reason. For example, recent studies of behavioural syndromes revealed that behavioural observations are likely to be missing for 'shy' and inactive individuals (Biro and Dingemanse 2009; Garamszegi et al. 2009a, b), which in turn results in datasets in which biased estimates will occur if cases containing missing observations (shy and inactive individuals) are deleted.

Although the importance of handling missing data is only beginning to be realized in ecology and evolutionary biology (Hadfield 2008; Nakagawa and Freckleton 2008; Hadfield and Nakagawa 2010), the implications for model selection and averaging have not yet been considered. In this paper, we consider the effects that missing data can have when comparing models within an IT-AIC framework. We begin by reviewing the key concepts and methods for dealing with missing data (note that in this paper we provide a more comprehensive overview of missing data theory than Nakagawa and Freckleton 2008). We then illustrate the possible extent of problems in model selection procedures, using a case study based on a behavioural ecological dataset. We show the degree of biases that may result in IT-related indices and parameters when ignoring missing data. We demonstrate how to alleviate some of these biases using a readily implementable method (Rubin 1987, 1996; Schafer 1999; Schafer and Graham 2002). By describing the details of processes required to 'recover' missing data, we hope that this paper become a useful case study for behavioural ecologists to follow when they deal with missing data in their own datasets.

Missing data types, problems and methods

Missing data classification

Rubin (1976) first formalized classifications for the mechanism (distribution patterns) of missing data. He identified three classes, which were termed 'missing completely at random' (MCAR), 'missing at random' (MAR) and 'missing not at random' (MNAR). The distinction between these is very important and has consequences for the degree of bias expected and the effectiveness of different strategies for dealing with missing data.

We begin by providing more formal definitions of these terms and then give an illustrative example in the next subsection. The important concepts in Rubin's classification system are Y and R and their relationships where Y is a matrix of original data and can be broken into Y_{obs} and Y_{mis} (observed and missing parts, respectively) and R is a matrix of binary recoding of distributions of missing data with 0 and 1, referred to as 'missingness' (Table 1).

In simple terms, if data are MCAR, then those cases containing missing data are completely random with respect to the other variables in the dataset. Specifically, for missing data to be MCAR, there is no systematic relationship between R and either Y_{obs} and Y_{mis} (e.g. missing observations in variable 2 does not depend on observed values in $V1$).

There are two fundamentally different ways in which data can be missing and not be MCAR. The distinction is

Table 1 An illustrative example of a data set, Y with three variables ($V1$–3; note $V3$ is an unobserved variable) and its missingness, R (the recording of $V1$–3 or $M1$–3)

Case	Original data [Y=(Y_{obs}, Y_{mis})]			Missingness (R)		
	$V1$	$V2$	$V3$	$M1$	$M2$	$M3$
1	Obs	Obs	Mis	0	0	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

Mis missing observations, *Obs* observed values

subtle, but important. For missing data to be MAR, there is a relationship between R and Y_{obs} but not between R and Y_{mis} (e.g. missing observations in $V2$ depends on observed values in $V1$). When there is a relationship between R and Y_{mis} (not necessary between R and Y_{obs}), missing data are MNAR (e.g. missing observations in variable 2 depends on these observations themselves and/or on an unobserved third variable). Also, MAR is termed 'ignorable' while MNAR is described as 'non-ignorable'. However, 'ignorable' missing data do not mean that we can discard cases with missing data; instead, the term 'ignorable' means that we do not need to make any particular assumptions about how data are missing in order to 'recover' missing data (and by this definition, MCAR data are also termed ignorable). A graphical summary of these concepts is in Fig. 1.

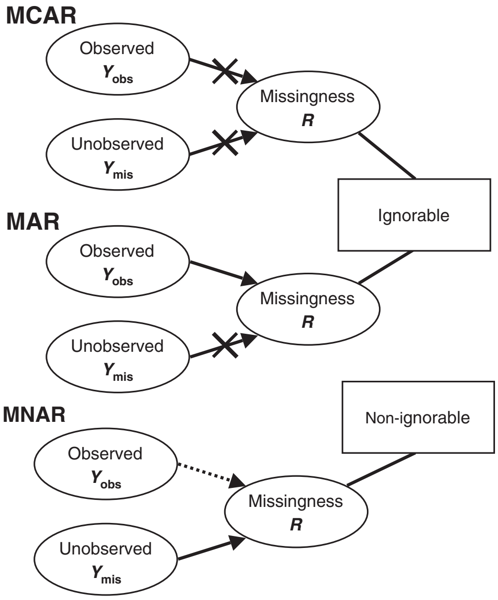


Fig. 1 A schematic representation of the three classes (mechanisms) of missing data (i.e. MCAR, MAR and MNAR) in relation to the observed values/variables (Y_{obs}), unobserved values/variables (Y_{mis}), missingness (R) and ignorability (modified from McKnight et al. 2007); see the text for the details

An illustration of missing mechanisms and problems

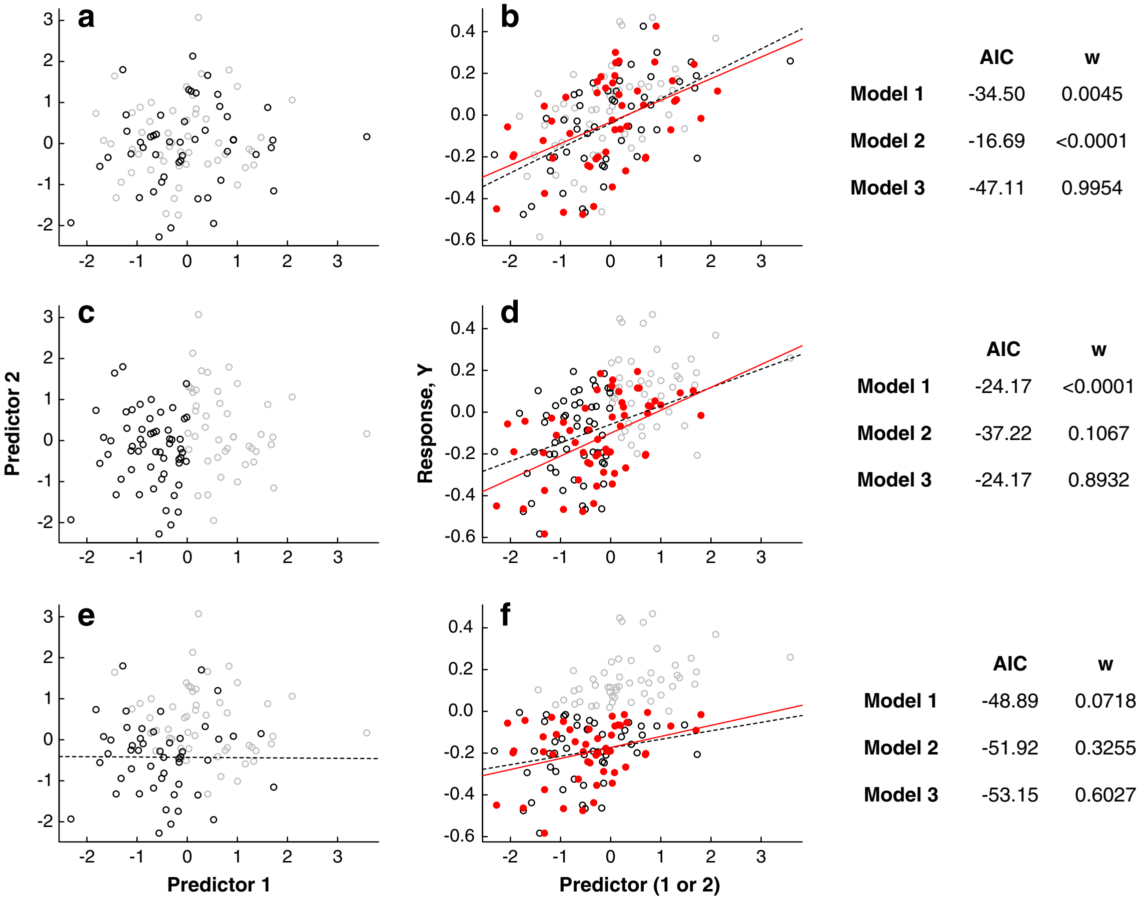
The three missing mechanisms (MCAR, MAR and MNAR) as well as their associated problems can be best illustrated by graphical examples; Nakagawa and Freckleton (2008) present examples for the case of simple bivariate regressions (in this previous paper, we only made suggestions regarding possible effects of missing data on IT model selection procedures but we did not provide any concrete evidence or examples). Here we consider how missing data affect model selection and IT indices using slightly more complex models.

When considering a more complex model selection/fitting problem, there are numerous aspects of the data that may be affected that will influence model selection. These can include the possible correlations between the variables, which will be determined by the pattern of missingness, with consequences for parameter estimates and variances (Freckleton 2010). To illustrate the type of problems that may arise in such analyses, consider the hypothetical example shown in Fig. 2. These data were generated according to a simple linear equation. The response variable was y. This is a function of two predictors (x_1 and x_2) and an error term (ε), all of which are independent of each other and drawn from a standard normal distribution with zero mean and unit standard deviation. Observation i was given by:

$$y_i = 0.1x_{1,i} + 0.1x_{2,i}\sqrt{0.02}\varepsilon_i$$

One hundred observations were used in total. These values were chosen so as to generate data that are typical of those used in many studies in behavioural ecology; that is, the effects of the predictors, relative to the error, are moderate with the parameters set so that the R^2 of the whole model should be 50%. The two predictors were given the same coefficient so that in model selection, models containing either variable should be weighted equally. Half of the data were assumed to be missing.

Figure 2 illustrates how observations of this relationship could be affected by the three different mechanisms of missing data. In Fig. 2a, b, provided that all measurements of y are known, the data are MCAR: This is because unknown values of x_1 or x_2 (or missingness) were chosen 'at random', both with respect to y and with respect to x_1 and x_2. Thus, the underlying relationships between the predictors (Fig. 2a) and between the response and predictors (Fig. 2b) are unaffected



	AIC	w
Model 1	-34.50	0.0045
Model 2	-16.69	<0.0001
Model 3	-47.11	0.9954

	AIC	w
Model 1	-24.17	<0.0001
Model 2	-37.22	0.1067
Model 3	-24.17	0.8932

	AIC	w
Model 1	-48.89	0.0718
Model 2	-51.92	0.3255
Model 3	-53.15	0.6027

Fig. 2 Effects of different mechanisms of missingness on model estimation and selection in a hypothetical dataset. The data were generated according to the model described in the text. Regression models were fit to the data to predict the response variable using ordinary least squares, with three models being contrasted: model 1, predictor 1 only used as a predictor; model 2, predictor 2 only used; and model 3, both predictors 1 and 2 included. For each model, AIC values and likelihood weights were calculated (see text for details). For the complete dataset, the AIC weight for model 3 was >0.99, those for models 1 and 2 were<0.01. In all figures, grey points (*empty circles*) indicate missing values; in **b**, **d** and **f** black points (*empty circles*) represent the relationship between the response and predictor 1, red points (*solid circles*) indicate the relationship for predictor 2. Data for **a** and **b** are MCAR; in **b**, the fitted relationship is $y = 0.11(\pm 0.02)x_1 + 0.08(0.02)x_2$. Data for **c** and **d** are MAR; in **d**, the fitted relationship is $y = 0.10(\pm 0.04)x_1 + 0.11(0.02)x_2$. Data for **e** and **f** are MNAR; in **f**, the fitted relationship is $y = 0.04(\pm 0.02)x_1 + 0.05(0.02)x_2$

by missingness. The parameter estimates in this situation are unbiased and close to the true values. Analysis of the data using model selection methods (AIC-IT and Akaike weights, *w*, see below for a more detailed description of methodology) unequivocally leads to the correct model being identified as being the best model.

The data in Fig. 2c, d are MAR because in this case missingness depends on the value of one of the predictors, x_1. In this example, all values of x_1 greater than zero are systematically missing. In this situation, although the range of values of x_1 is truncated, the relationship between the predictors is unaffected by missingness, there being no relationship between the observed values of x_1 and x_2 (Fig. 2c); the estimated relationship between y and both x_1 and x_2 is also not affected by missingness, with the estimated parameters being unbiased (Fig. 2d). However, because of the truncation of the range of x_1, the variance of the slope for x_1 is estimated to be higher than for x_2. One consequence is that the effect size for x_2 is estimated to be higher than for x_1 (4.67 versus 2.50), whereas they should be identical. Model selection again correctly identifies model 3 as being the best model: However, the AIC for model 2, containing x_2 only, is much lower than that of model 1 containing x_1 only, again indicating that the effect of x_2 is stronger than that of x_1.

If data are MNAR, the consequences may be complex. In Fig. 2e, f, missingness depends on y, with all values greater than zero missing. This type of missingness has consequences for the whole dataset: A first consequence is that, although x_1 and x_2 are unrelated in the complete dataset, when missingness depends on y, the remaining values are correlated with each other, as shown in Fig. 2e. This happens because the missing values of y correspond to observations in which both x_1 and x_2 are large. The outcomes of this are that it could be mistakenly concluded that these two variables are related to each other, as well as issues with collinearity when analysing the determinants of y. Because missingness depends on y, the relationship between y and x_1 and x_2 is incorrectly estimated: The slopes are under-estimated relative to their true values. As a consequence of this combination of effects, model selection is equivocal in this case. Although a high weight (0.60) is given to the true model, model 2 arbitrarily gets a relatively high weighting (0.33). The p value for the coefficient for x_1 in an ordinary least square regression is equivocal (0.07); hence, it cannot be concluded with any certainty whether both x_1 and x_2 should be included in a best fit model.

In summary, the example shows that, in theory, missingness in data can yield a variety of effects. Below we describe an example with real data and how the issues of missing data can be addressed. However, even without performing a more systematic analysis, the lessons from Fig. 2 are that missingness in sets of models can yield (a) biased estimates of parameters, (b) incorrect estimates of effect sizes, (c) incorrect rankings of effect sizes and (d) lead to equivocal model comparisons. Finally, in the specific example we used, it is evident that the correlation structure of the data could be affected by the nature of the missingness.

Telling missing mechanisms apart

The common practice of complete-case analysis assumes data are MCAR; this is a strong and probably frequently incorrect assumption. A more reasonable assumption is MAR (i.e. the ignorable assumption; the basis of many missing data handling methods; see below). Detection of non-MCAR missingness in datasets is a relatively easy process. Little (1988) proposed a simple method in which one conduct a series of t tests between 'observed' and 'missing' groups, defined by a variable with missing observations and comparing mean differences in other variables, while controlling for type I errors. Notably, some researchers in the field of behavioural ecology have employed such procedures to compare these two groups (e.g. Garamszegi et al. 2004, 2005). Another simple and more practical method is the use of binary-recorded variables in relation to missing data, i.e. missingness R (Table 1). The recorded variables (e.g. $M2$ in Table 1) can be analysed as the response in logistic regression models to see whether the other variables (the original scale) can predict the patterns of observed and missing values. Also, one can visually assess (non-)MCAR missingness by drawing a series of bivariate plots between recorded variables (in R) and variables on original scale (in Y).

On the other hand, the distinction between MAR and MNAR is very difficult in many cases because we usually do not have information about missing values, which is required to determine if missing observations are MNAR (Schafer 1997; McKnight et al. 2007). However, when missingness is identified as MNAR (non-ignorable), the process of handling is rather complex (as we need to specifically model how missing values occurred). Biological examples of modelling non-ignorable missing data were discussed in a recent paper (Hadfield 2008). In our paper, we do not consider the modelling processes for MNAR although below we consider a dataset with non-ignorable missingness.

Missing data handling

Three broad categories of how to handle missing data are deletion, imputation and augmentation. Deleting missing data is the simplest method and is widely used. However, this is obviously problematic when missing data are not MCAR. The terms data imputation and augmentation are sometime used interchangeably in the statistical literature. This conflation is probably due to the fact that both methods can use exactly the same mathematical and computational methods to 'recover' missing data (for overviews of the procedures, see Schafer 1997; Allison 2002; Little and Rubin 2002; McKnight et al. 2007).

In understanding the difference between imputation and augmentation, it is useful to follow the classification by McKnight et al. (2007) who distinguish data imputation from data augmentation in terms of missing value replacement. For data imputation, one can 'see' imputed values, with missing values being replaced by imputed ones (i.e. one will have a filled-in dataset). On the other hand, for data augmentation, one does not see explicit 'augmented' values because data augmentation occurs simultaneously with data analysis. In this article, we focus on data imputation methods not only because available statistical packages for data augmentation are limited, compared to data imputation counterparts, but also because data augmentation involves complex implementations and thus is often not practical (at least currently).

In terms of imputation of missing data, behavioural ecologists are already familiar with the procedure termed single imputation. For example, substituting missing values with the mean is a fairly common practice (e.g. Nakagawa et al. 2001). If we have a model for the data, then it may be possible to use this to generate predictions for the missing data. For instance, this can be done in comparative analysis by using phylogenetic similarity to predict values for species with missing data (Garland and Ives 2000). However, single imputation completely ignores the uncertainty regarding imputed values, which results not only in overconfident precision (i.e. inappropriately small standard errors (SE)) but also often in biased parameter estimates

(McKnight et al. 2007). Multiple imputation (MI; used correctly) overcomes such problems. We introduce statistical software for MI and a general procedure involved in MI along with an overview of MI, using an example in the next section. We emphasize that this section is not a comprehensive introduction to the methods for handling missing data, but rather we aim to demonstrate the potential approaches. For more extensive but accessible treatments of the methods, we refer readers to Allison (2002) and McKnight et al. (2007) and, for more advanced treatments of this topic, to Schafer (1997) and Little and Rubin (2002).

Multiple imputation and a working example

An example: extra-pair paternity in house sparrows

The dataset we use is based on one breeding season taken from a long-term dataset on a house sparrow (*Passer domesticus*) population on Lundy Island, UK (for the descriptions of this study system and the previous research, see Griffith et al. 1999; Nakagawa et al. 2007a, b; Ockendon et al. 2009). A complete dataset, which was used for this paper, is a slightly modified version of the original data (S. Nakagawa and T. Burke, unpublished data).

The complete dataset we use for our analysis contains eight variables associated with 76 males: (a) the number of eggs which males sired outside of their social nests (extra-pair paternity (EPP) gain; range 0–13), (b) the number of fledglings which males raise with their social partners (fledgling; range 0–7), (c) standardized microsatellite-based heterozygosity (heterozygosity; range 0.60–1.19), (d) male age (age; range 1–4), (e) tarsus length (tarsus; range 16.9–20.2 mm), (f) wing length (wing; range 76–83 mm), (g) residual weight controlling for capture time and date (weight; range −2.76–4.05) and (h) the size of bibs, a secondary sexual trait, which is often referred to as a badge of status (badge; 31.5–41.5 mm). We use these data to analyse the factors influencing the success of EPP gain in male house sparrows in this population. EPP is therefore the response variable, and the other variables are used as predictors of EPP gain.

Given that previous work showed that the patterns of EPP can be influenced by factors such as male morphology, experience and genetic characteristics (Griffith et al. 2002), we consider that all of our predictors are biological meaningful. To illustrate the relationships between these variables, we present bivariate Spearman's correlations between the variables and their 95% confidence intervals in Table 2.

We conducted all statistical analysis and data manipulations in the R environment (version 2.8.2; R Development Core Team 2009). Our response, EPP gain was a count variable and, in general, count data are overdispersed. Our statistical models were run as generalized linear models with the quasi-Poisson error structure with the log-link function. We always used a dispersion estimate from the full model for fitting of all the models in a particular model set, according to Burnham and Anderson (2002). The AIC which incorporate overdispersion is known as quasi-AIC (QAIC), which is given by:

$$\mathrm{QAIC} = -[21\mathrm{n}L(\theta)/c] + 2K, \qquad (1)$$

where $L(\theta)$ is the likelihood of parameters given the data and the model, K is the number of parameter in the model and c is the overdispersion parameter (Burnham and Anderson 2002; Anderson 2008). Notably, when $c=1$, QAIC reduces to AIC.

For calculations of IT-related indices, i.e. QAIC, delta QAIC (Δ), Akaike weight (w) and parameters based on model averaging (for model averaging details, see Burnham and Anderson 2002; Anderson 2008; Richards et al. 2010; Symonds and Moussalli 2010), we used an R package 'MuMIn' (Barton 2009), which has a range of functions to automate an AIC-IT approach. We provide all the datasets

Table 2 Pair-wise Spearman's correlations for the eight variables (the upper triangle) and the 95% confidence intervals, CI (the lower triangle; statistically significant correlations and their CIs are in bold)

Variable	EPP gain	Fledgling	Heterozygosity	Age	Tarsus	Wing	Weight	Badge
EPP gain	–	0.144	0.131	**0.478**	0.159	0.211	0.269	0.080
Fledgling	−0.084 to 0.358	–	**0.283**	**0.309**	−0.139	0.059	−0.035	−0.066
Heterozygosity	−0.097 to 0.347	**0.061 to 0.478**	–	0.067	0.035	0.206	0.0780	−0.127
Age	**0.283 to0.635**	**0.090 to 0.500**	−0.161 to 0.288	–	−0.123	0.165	0.044	**0.294**
Tarsus	−0.069 to 0.371	−0.353 to 0.090	−0.192 to 0.259	−0.339 to 0.106	–	**0.401**	**0.631**	0.050
Wing	−0.016 to 0.416	−0.169 to 0.281	−0.021 to 0.412	−0.063 to 0.376	**0.193 to 0.574**	–	**0.411**	0.081
Weight	**0.046 to 0.469**	−0.259 to 0.192	−0.150 to 0.298	−0.183 to 0.267	**0.473 to 0.750**	**0.205 to 0.583**	–	0.199
Badge	−0.148 to 0.300	−0.287 to 0.162	−0.342 to 0.102	**0.073 to 0.487**	−0.178 to 0.272	−0.148 to 0.301	−0.028 to 0.406	–

and an R script prepared for this paper as electronic supplemental material (see S2).

Comparing IT indices and parameter estimates

We created three incomplete datasets in addition to the original complete dataset (referred to as complete dataset). The first incomplete dataset mimicked a MCAR process (incomplete MCAR). To do this, we randomly deleted six cases from each of the eight variables using a function in the R package 'mi', which resulted in only 38 complete cases out of 76 (50% of individuals had at least one missing observations). To create the second incomplete dataset, we deleted 38 age observations of male birds, which had the highest badge values, thus simulating MAR (incomplete MAR). The third incomplete dataset simulated MNAR (incomplete MNAR) by deleting 38 age observations of the oldest birds, except that the observations for one 4-year-old and one 3-year-old were left. This focus on age variable is not only because we know age is an important factor in relation to the response, i.e. EPP gain (Table 2) but also because the information on age is often missing in a real dataset. In all the incomplete datasets, the number of complete cases was 38 (only 50% of the original cases).

We considered three models as our model set for each of the four datasets (complete and incomplete MCAR, MAR and MNAR): (1) a model including all seven predictors (referred to as full model), (2) a model including the four morphological predictors (badge, tarsus, weight and wing; morphological model) and (3) a model including the three non-morphological predictors (age, heterozygosity and fledgling; non-morphological model). For all the datasets, IT indices and model-averaged parameter estimates were first calculated using 'complete-case' analyses, and the results are summarized in Tables 3 and 4. We used this simple model set including only the three models to highlight and to facilitate the understanding of problems relating missing data and model comparisons using IT-related indices, although we acknowledge that these three models would not necessarily constitute the most meaningful biological model set. Comparable results from a more realistic and practical model set are summarized in an electronic supplementary material (S1). Also our morphological model may suffer from the problem of collinearity given correlations among morphological variables; the issue of collinearity is beyond the scope of this paper (we refer readers to Freckleton 2010).

We use all the models and parameter estimates fitted to complete dataset as our baseline and for comparison with those obtained when data are missing. AIC values (including QAIC) cannot be compared between datasets with different dimensions. However, it is valid to compare Akaike weights within datasets because these are relative measures of the weight of evidence for different models.

Incidentally but importantly, there are two kinds of model-averaged parameter estimators, which were discussed in Burnham and Anderson (2002). The first kind of model-averaged estimate $\overline{\beta}_j$ where the parameter β_j is averaged over all models in which the predictor x_j appears is given by:

$$\overline{\beta}_j = \frac{\sum_{i=1}^{R} w_i I_j(g_i)\beta_{j,i}}{\sum_{i=1}^{R} w_i I_j(g_i)}, \tag{2}$$

where $I_j(g_i)$=1 if the predictor and 0 otherwise, $\beta_{j,i}$ is the regression coefficient estimate of the predictor x_j in the model g_i and the entire denominator term represents the sum of Akaike weights of the models in which the predictor x_j is found (Burnham and Anderson 2002). The second estimator of $\overline{\beta}_j$ is Eq. 2 without the denominator

Table 3 Results of model selection for the four dataset, listing QAIC, *K* (the number of parameters in the model including the intercept and the residual error estimates), delta QAIC (Δ) and Akaike weight (*w*; the best models are in bold)

Dataset	Model (set)	QAIC	*K*	Δ	*w*
Complete (*N*=76)	Full	52.86	9	1.66	0.296
	Morphology	57.92	6	6.16	0.025
	Non-morphology	**51.25**	**5**	**0.00**	**0.679**
Incomplete MCAR (*N*=38)	Full	37.13	9	6.01	0.045
	Morphology	37.66	6	6.54	0.035
	Non-morphology	**31.12**	**5**	**0.00**	**0.919**
Incomplete MAR (*N*=38)	**Full**	**45.47**	**9**	**0.00**	**0.995**
	Morphology	86.36	6	40.9	<0.001
	Non-morphology	56.07	5	10.6	0.005
Incomplete MNAR (*N*=38)	**Full**	**43.41**	**9**	**0.00**	**0.673**
	Morphology	54.53	6	11.1	0.003
	Non-morphology	44.87	5	1.46	0.325

Table 4 Model-averaged parameter estimates for the four datasets, listing relative importance (∑), regression coefficient (*b*), unconditional SE and 95% CI for *b* (statistically significant predictors are in bold)

Dataset	Predictor	∑[a]	*b*	Unconditional SE[b]	95% CI for *b*
Complete	**Age**	0.975	**0.738**	**0.215**	**0.317 to 1.160**
	Badge	0.975	−0.076	0.119	−0.314 to 0.161
	Fledgling	0.975	0.001	0.106	−0.207 to 0.207
	Heterozygosity	0.321	−0.917	1.930	−4.705 to 2.870
	Tarsus	0.321	0.033	0.461	−0.800 to 1.006
	Weight	0.321	0.326	0.203	−0.072 to 0.724
	Wing	0.321	−0.024	0.207	−0.429 to 0.381
Incomplete MCAR	**Age**	0.965	**0.882**	**0.284**	**0.325 to 1.439**
	Badge	0.081	−0.020	0.205	−0.422 to 0.381
	Fledgling	0.965	−0.119	0.138	−0.389 to 0.152
	Heterozygosity	0.965	−0.521	2.167	−4.789 to 3.707
	Tarsus	0.081	−0.302	0.771	−1.814 to 1.209
	Weight	0.081	0.478	0.424	−0.353 to 1.309
	Wing	0.081	−0.158	0.248	−0.644 to 0.328
Incomplete MAR	**Age**	>0.999	**1.120**	**0.338**	**0.453 to 1.779**
	Badge	>0.999	−0.087	0.246	−0.572 to 0.396
	Fledgling	>0.999	**0.265**	**0.130**	**0.010 to 0.520**
	Heterozygosity	>0.999	−5.408	3.135	−11.55 to 0.737
	Tarsus	>0.999	**1.095**	**0.367**	**0.374 to 1.815**
	Weight	>0.999	0.147	0.148	−0.143 to 0.437
	Wing	>0.999	−0.181	0.198	−0.569 to 0.208
Incomplete MNAR	**Age**	0.997	**1.640**	**0.629**	**0.407 to 2.873**
	Badge	0.675	0.189	0.190	−0.183 to 0.560
	Fledgling	0.997	−0.146	0.281	−0.699 to 0.405
	Heterozygosity	0.997	9.756	5.044	−0.130 to 19.64
	Tarsus	0.675	0.632	0.832	−0.999 to 2.262
	Weight	0.675	0.366	0.462	−0.540 to 1.272
	Wing	0.675	−0.549	0.325	−1.188 to 0.089

[a] The relative importance for a predictor is the sum of Akaike weights of the models in which the predictor was present

[b] Unconditional SE incorporates model uncertainty and thus a more conservative and appropriate uncertainty estimator for parameters (see Burnham and Anderson 2002)

(which represents the sum of model weights the predictor x_i is included), and thus, the model-averaged estimate of this type is generally shrunk towards zero, especially when β_j are not included in models with high weightings. We used the first type in this paper because we wanted to avoid the shrinkage of the parameters, which only affects variables in relatively unimportant models (see the provided R script for the implementation, S2).

The first important result is that, as predicted by our hypothetical example, the nature of missingness can fundamentally alter the conclusions we can draw based on model comparisons. The weighting of the three models in the different datasets were remarkably different (Table 3). The best model for complete and incomplete MCAR is non-morphology model whereas for incomplete MAR and incomplete MNAR, it was full model (Table 3). As we warned in a recent article (Nakagawa and Freckleton 2008) and also as the illustrative example in Fig. 2 implies, the case-wise deletion of missing data (i.e. complete-case analysis) resulted in substantial differences in Akaike weights among the datasets.

In terms of parameter estimation and hypothesis testing, the form of missingness is also important. The parameter estimates for incomplete MCAR were relatively unbiased although SE generally increased, as expected. However, for incomplete MAR and MNAR, the estimates were biased for some parameters. Noticeably, the predictors fledgling and tarsus became spuriously significant in incomplete MAR while the predictor age was upwardly biased in incomplete MNAR (Table 4). We discuss more on the parameter estimates once we obtain comparable estimates via MI in the next subsection.

Multiple imputation

MI is a relatively new tool for handling missing data although MI is becoming the preferred approach (Rubin 1996; Schafer 1999; Allison 2002; Schafer and Graham

2002; McKnight et al. 2007). The process of MI usually refers to a three-step procedure. The first component is the imputation step in which missing data are imputed. This imputation step is repeated to create M datasets. Rubin (1987) showed that M between 3 and 10 are usually sufficient (we return to this issue below). The two most widely methods of data imputation are (1) model-based methods where certain distributional properties are assumed (e.g. most commonly multivariate normal is assumed between variables in a dataset) and (2) Markov chain Monte Carlo (MCMC) methods where distributional assumptions are not necessary (although MCMC methods are not widely implemented possibly due to their implementation difficulty and slow convergence speed; for more descriptions of the methods, see McKnight et al. 2007 and the references therein). Importantly, although a real dataset rarely confirms multivariate normality, simulations by Graham and Schafer (1999) showed that MI could perform robustly even for a set of highly non-normal variables under a multivariate normality assumption (for cases of a dataset containing nominal variables; see Schafer 1997).

The second component is the analysis step in which routine statistical analysis is conducted upon each of the M datasets, separately. This straightforward integration of the imputation and analysis steps is one of strengths of MI and makes MI much more practical and popular, compared to data augmentation procedures. The final component is the pooling step in which pooled parameter estimates of interest are calculated from M result sets. This pooling step is relatively straightforward and the process involved in this step can be automated by MI software packages (see below).

In the pooling step, parameter estimates and measure of variance may be calculated using the following formulae (Rubin 1987; Schafer 1999; McKnight et al. 2007):

$$\overline{Q} = \frac{1}{M}\sum_{j=1}^{M} Q_j, \tag{3}$$

$$\overline{U} = \frac{1}{M}\sum_{j=1}^{M} U_j, \tag{4}$$

$$B = \frac{1}{M-1}\sum_{j=1}^{M} (Q_j - \overline{Q}), \tag{5}$$

$$T = \overline{U} + \left(1 + \frac{1}{M}\right)B, \tag{6}$$

where $\overline{Q}$ is the mean of Q_j which is a parameter estimated from jth dataset (j=1, 2,…, M), $\overline{U}$ is the within-imputation variance calculated from the standard error associated with Q_j, B is the between-imputation variance estimates and T is the total variance (the square root of T is the overall standard error for $\overline{Q}$).

Statistical significance (via t tests) and confidence intervals (CIs) of pooled parameters are calculated using:

$$df = (M-1)\left(1 + \frac{M\overline{U}}{(M+1)B}\right)^2, \tag{7}$$

$$t_{df} = \frac{\overline{Q}}{\sqrt{T}}, \tag{8}$$

$$100(1-\alpha)\%\,\mathrm{CI} = \overline{Q} \pm t_{df,(1-\alpha/2)}\sqrt{T}, \tag{9}$$

where df is the degree of freedom used for t test or to obtain t values and CI calculations and α is the significance level (e.g. 95% CI, α=0.05).

Based on the quantities estimated in Eqs. 3, 4, 5, 6 and 7, we can calculate an important property of MI, namely the rate of missing information (γ), which is given by:

$$\gamma = \frac{r + 2/(df+3)}{r+1}, \tag{10}$$

$$r = \frac{(1+m^{-1})B}{U}, \tag{11}$$

where r represents the relative increase in variance (i.e. uncertainty) due to missing observations. The rate of missing information γ ranges from 0 to 1. Although we mentioned that the distinction between MAR and MNAR is difficult above (Fig. 1), γ can be an indicator of the difference between them. When missing data are non-ignorable (MNAR), γ will be large (McKnight et al. 2007). However, unfortunately, a general measure of effect size thresholds (e.g. analogous to Cohen's d and the correlation coefficient r; Cohen 1988) is not available for γ. Therefore, we cannot explicitly state what constitutes a small or large γ or what γ value represents non-ignorability. Nevertheless, γ is a useful index. For instance, the rate of missing information γ can be used to quantify the efficiency of MI, compared to an infinite number of imputed datasets given M datasets as:

$$\frac{1}{1+\frac{\gamma}{M}} \tag{12}$$

For example, at $M=5$ and $\gamma=0.5$, the efficiency is 91% compared to a case with M being infinite while at $M=10$ and $\gamma=0.5$, the efficiency is 95% (Allison 2002). This example seems to validate the soundness of Rubin's (1987) recommendations regarding M between 3 and 10. However, given current computational power, high M (e.g. $M=20$–30) can be easily achieved to obtain even higher efficiency. Equation 12 also shows that when at γ is large, M should be large.

This is a brief outline of the MI process. More complete accounts of MI can be found in Schafer (1999) and Schafer and Graham (2002). Additionally, practical information on MI can be found in the form of Frequently Asked Questions at http://www.stat.psu.edu/~jls/mifaq.html/. A summary of statistical packages for MI can be found in Horton and Kleinman (2007) and at http://www.multiple-imputation.com/. We present a list of R packages for MI and their capabilities in Table 5.

Implementation for the case study

We used the MI function 'amelia' in the R package 'Amelia' (the official programme name is Amelia II; Honaker et al. 2008; Table 5). Amelia II uses a type of expectation maximization (EM) algorithm in conjunction with a bootstrapping procedure, termed EMB algorithm (a type of model-based method). The EM algorithm is a general method of obtaining maximum likelihood (ML) estimates of means, standard deviations and correlations (i.e. a covariance matrix of variables) when a data matrix has missing values (Dempster et al. 1977). The name EM represents the two processes involved, the expectation step where missing data are replaced and the maximization step where the imputed data are adjusted. These two steps work together in an iterative fashion till they converge to ML estimates (for accessible accounts, see Allison 2002; McKnight et al. 2007). The EM algorithm requires distributional assumptions such as multivariate normality. The EMB algorithm of Amelia II uses bootstrapping (i.e. sampling with replacement) of data points from every variable in the original dataset (with missing data), which produces M datasets (without missing data). Then, each resultant dataset is subject to an EM procedure to obtain ML estimates of means and a covariance matrix, from which missing data are imputed (for more details of the EMB algorithm, see Honaker and King 2009).

In addition to its user-friendly design, Amelia II can provide a range of options, which can improve the recovery of missing data. Importantly, different types of prior information (priors) about missing data can be incorporated. For example, it is common that behavioural ecologists know the minimum age of animals (e.g. at least 3 yearsold) and such information can be integrated easily as priors. Other practical considerations for MI are (a) considering inclusions of new variables which are correlated with the variables with missing data in the data matrix (i.e. satisfying the ignorable assumption; Fig. 1) and (b) checking distributions for each variable (e.g. transform non-normal data if a MI procedure assumes a multivariate normal distribution; but see Graham and Schafer 1999).

We set $M=5$ for our three 'incomplete' datasets: incomplete MCAR, MAR and MNAR; we refer to each type of five imputed datasets as imputed MCAR, imputed MAR and imputed MNAR corresponding to our incomplete datasets with the three different mechanisms (note that

Table 5 An overview for capabilities of R packages including routines for multiple imputations; whether the package can handle continuous data, categorical data and hierarchical data (also referred to as repeated, nested cluster or multilevel data), time series data (a type of hierarchical data), whether the package include a function to analyse multiple-imputed datasets in either LM, GLM, LMM or GLMM and also whether the software is available as stand-alone (note that this is not an exhaustive list)

Package	Continuous	Categorical	Hierarchical[b]	Time Series	LM	GLM	LMM	GLMM	Stand-alone	Reference
mi	Yes	Yes	Probably[c]	Probably[c]	Yes	Yes	Yes	Yes	No	Gelman and Hill (2007)
Amelia	Yes	Yes	Probably[c]	Yes	No	No	No	No	Yes	Honaker et al. (2008)
mice	Yes	Yes	No	No	Yes	Yes	No	No	Yes	van Buuren and Groothuis-Oudshoorn (2009)
Hmisc	Yes	Yes	Probably[c]	No	No	No	No	No	No	Harrell et al. (2008)
norm	Yes	No	No	No	No	No	No	No	Yes	Schafer (1997)
cat	No	Yes	No	No	No	No	No	No	No	Schafer (1997)
mix	Yes	Yes	No	No	No	No	No	No	No	Schafer (1997)
pan[a]	Yes	Yes	Yes	Probably[c]	No	No	Yes	No	No	Schafer (1997)

LM linear modelling, *GLM* generalized linear modelling, *LMM* linear mixed modelling, *GLMM* generalized linear mixed modelling

[a] This package mainly includes functions for data augmentation rather than multiple imputations in our definitions (see the text)

[b] Some tips for multiple imputations of hierarchical data can be found in the page 541 of Gelman and Hill (2007)

[c] These packages are not specifically or explicitly designed to deal with these types of datasets although functions in these packages can probably be used

Table 6 A summary of mean QAIC with SE, parameter number (K; including the intercept and residual error), delta QAIC (Δ) and Akaike weights (w) from five imputed datasets (M=5) for the three incomplete datasets (the best models are in bold)

Dataset	Model (set)	Mean QAIC (SE)	K	Δ	w
Imputed MCAR (N=76)	Full	52.39 (1.13)	9	1.92	0.274
	Morphology	58.96 (0.29)	6	8.49	0.010
	Non-morphology	**50.47 (1.09)**	**5**	**0.00**	**0.715**
Imputed MAR (N=76)	Full	61.59 (2.12)	9	2.35	0.234
	Morphology	68.75 (NA[a])	6	9.51	0.007
	Non-morphology	**59.24 (1.95)**	**5**	**0.00**	**0.759**
Imputed MNAR (N=76)	**Full**	**63.63 (4.53)**	**9**	**0.00**	**0.902**
	Morphology	77.32 (NA[a])	6	13.69	0.001
	Non-morphology	68.08 (4.90)	5	4.45	0.098

[a] For the datasets incomplete MAR and MNAR, there are no missing observations for morphological data (only in age), and thus, all the imputed datasets for the morphological variables had exactly the same data

each type has M=5 or five datasets; for more details of our imputation settings, see the provided R script in the supplement, S2).

Model averaging and MI

In our literature search, we found very little about how we should proceed with the summary of IT indices and model averaging in conjunction with MI. Therefore, we decided to run the model averaging procedure for each imputed dataset. Then, the pooling process for parameters was conducted using model-averaged estimates (i.e. each of our final parameter estimates is a pooled estimate combining five model-averaged estimates and their unconditional standard errors; using Eqs. 3, 4, 5 and 6). We used the function ‘mi.inference’ in the R package ‘norm’ (Schafer 1997; Table 5), which calculated pooled estimates, the increase in variance due to missing values (r) and the rate of missing information (γ). The results for IT-related indices, parameter estimates and MI-related indices are summarized in Tables 6 and 7. Alternatively, one could conduct the MI pooling step for each imputed dataset and

Table 7 Pooled model-averaged parameter estimates from the five imputed datasets (M=5) of each of the three incomplete datasets, listing relative importance ($\sum$), regression coefficient (b), overall SE and 95% CI for b, the increase in variance due to missing values (r) and the rate of missing information (γ; statistically significant predictors are in bold)

Dataset	Predictor	$\sum$	Mean b[a]	Overall SE[b]	95% CI for b[c]	r	γ
Imputed MCAR	**Age**	0.955	**0.734**	**0.226**	**0.286 to 1.182**	0.255	0.219
	Badge	0.284	−0.096	0.119	−0.330 to 0.139	0.087	0.083
	Fledgling	0.955	−0.022	0.096	−0.210 to 0.166	0.041	0.040
	Heterozygosity	0.955	−0.183	1.697	−3.512 to 3.147	0.059	0.057
	Tarsus	0.284	0.229	0.399	−0.554 to 1.011	0.020	0.020
	Weight	0.284	0.303	0.186	−0.061 to 0.667	0.012	0.012
	Wing	0.284	−0.105	0.192	−0.483 to 0.273	0.100	0.095
Imputed MAR	**Age**	0.993	**0.826**	**0.270**	**0.281 to 1.371**	0.455	0.344
	Badge	0.241	−0.049	0.113	−0.272 to 0.173	0.155	0.142
	Fledgling	0.993	0.008	0.102	−0.192 to 0.209	0.111	0.104
	Heterozygosity	0.993	−0.791	1.747	−4.237 to 2.655	0.170	0.154
	Tarsus	0.241	0.087	0.435	−0.775 to 0.949	0.233	0.203
	Weight	0.241	0.235	0.215	−0.194 to 0.664	0.303	0.253
	Wing	0.241	0.071	0.188	−0.301 to 0.443	0.235	0.205
Imputed MNAR	**Age**	0.999	**0.893**	**0.361**	**0.103 to 1.683**	1.426	0.644
	Badge	0.903	−0.131	0.117	−0.360 to 0.098	0.095	0.090
	Fledgling	0.999	−0.010	0.133	−0.281 to 0.260	0.520	0.378
	Heterozygosity	0.999	−0.604	2.020	−4.594 to 3.386	0.188	0.168
	Tarsus	0.903	−0.059	0.470	−1.015 to 0.896	0.516	0.376
	Weight	0.903	0.464	0.241	−0.032 to 0.960	0.674	0.446
	Wing	0.903	−0.057	0.261	−0.626 to 0.511	1.389	0.638

[a] These values were calculated as in Eq. 3

[b] These values were calculated as in Eq. 6

[c] These values were calculated as in Eq. 9

model-average across M datasets using mean QAIC. However, such a sequence would not be suitable because variation in QAIC values between imputed datasets can be large (e.g. imputed MNAR in Table 6).

In terms of including overdispersion, we note that for obtaining mean QAIC values from five model sets for each collection of imputed datasets, the overdispersion parameters (the term c in Eq. 1) were averaged from five full models, and this mean c was used to obtain QAIC for each model of each model set (e.g. imputed MCAR had one c values used for QAIC calculation). Alternatively, the overdispersion c could be calculated for each imputed dataset (i.e. obtaining c from a full model of each dataset) although this procedure should make little differences in resulting parameter estimates.

The Akaike weights of all models (full, morphology, non-morphology) for imputed MCAR and imputed MAR (Table 6) were very similar to those from complete dataset (Table 3) whereas the weights for imputed MNAR were quite different with full model being the best model (Table 6). Therefore, via the MI procedures, missing observations in incomplete MCAR and MAR were successfully recovered to give 'accurate' (original) Akaike weights (in imputed MCAR and MAR) although this was not the case for imputed MNAR. This result might be expected because MI, like other missing data handling methods, usually assumes that missingness is ignorable (but MNAR is non-ignorable; Fig. 1). Also, the rate of missing information (γ) is relatively large in imputed MNAR compared to the other two (e.g. γ (age)=64.4% for imputed MNAR, 21.9% for imputed MCAR and 34.4% for imputed MAR; note that one should usually focus on the largest γ value within a set of γ for regression coefficients in a dataset; Table 7). Thus, a larger number of imputations (e.g. M=10–30) could help for the incomplete MNAR dataset, although increasing M will not lead to dramatic improvements (the efficiency of MI in relation to γ is discussed above; Eq. 12).

It has been noted that MI can produce satisfactory results with minor deviations from MAR (Rubin 1996; Schafer and Graham 2002), implying that MI can provide good parameter estimates even when missingness is MNAR. To prove this point, the pooled estimates for the eight variables from all the types of imputed datasets including imputed MNAR (Table 7) were very similar to the original

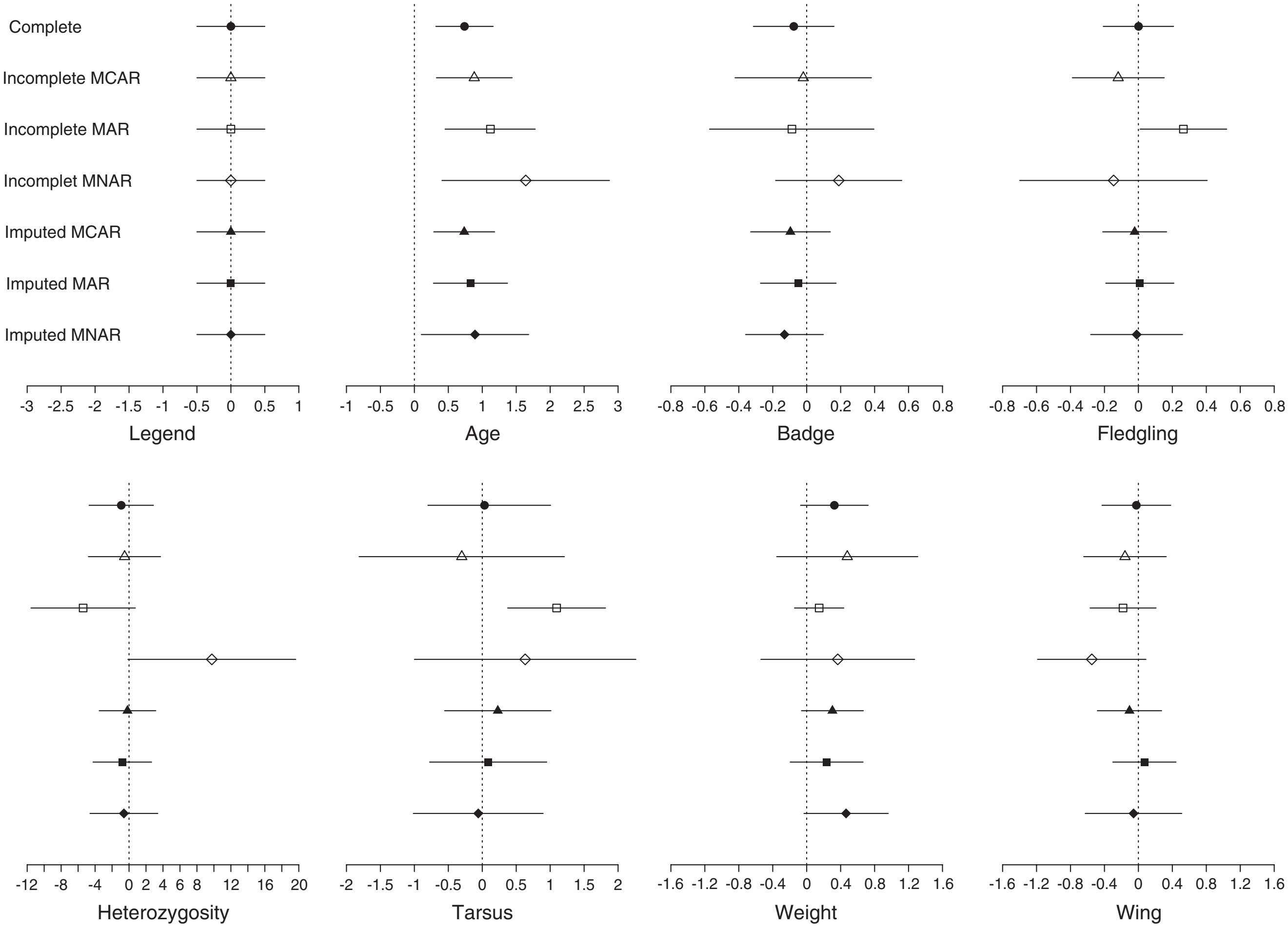


Fig. 3 A visual representation of the parameter estimates for the seven predictors and the 95% confidence intervals from the seven different types of datasets

parameter estimates from complete dataset (i.e. our baseline estimates; Table 4). The remarkable recovery of uncertainty estimates (SE) as well as parameter estimates in the imputed datasets is easy to see in Fig. 3, which depicts all the parameter estimates with 95% confidence intervals for the seven sets of parameter estimates. It should be reminded that MNAR missingness (R) depends on unobserved data (Y_{mis}) but also can depend on observed data (Y_{obs}; Fig. 1). Therefore, observed data in the data matrix can be related to unobserved and can contribute to recovering MNAR missingness (although limited; see Collins et al. 2001). Interestingly, our results in relation to Imputed MNAR suggest that bias in parameter estimates (including SE) could be independent from estimation bias in IT-related indices because our MI procedure resolved the estimation bias in parameters but not the bias in Akaike weights. How such a discrepancy could happen should require future investigation.

Discussion

In common with most other approaches, IT-AIC approaches require complete datasets for model comparisons. Therefore, unless methods such as MI are employed, the deletion of incomplete cases is unavoidable in cases where datasets include missing values. However, removal of cases in this way is known to result in increased uncertainty (i.e. reduced power) and biased parameter estimates. We illustrated the severity of these two problems using an example and also that how MI could alleviate these problems (see Fig. 3). The problem of biased model-averaged parameter estimates was apparent in incomplete MAR and MNAR. This paper also showed IT-related indices such as Akaike weights were incorrectly estimated in incomplete datasets. MI was efficient in recovering Akaike weights for datasets with MCAR and MAR missingness (incomplete MCAR and MAR), although MI failed to restore Akaike weights of the dataset with MNAR missingness (incomplete MNAR).

In behavioural ecology, it is probably reasonable to assume that missing data are not MCAR because there are often biological reasons behind the pattern of missingness (the aforementioned example of animal personality; Hadfield 2008; Nakagawa and Freckleton 2008). Even if missingness conforms to MCAR, we recommend researchers use MI (or other appropriate methods) partially because better estimates of Akaike weights as well as of standard errors can be obtained via MI (Tables 3, 4, 6 and 7). Another common method of handling missing data, which we did not mention so far, is the 'complete-variable' analysis in which only variables without missing data are included. The complete-variable analysis cannot be recommended because researchers will omit biological important variables by doing so. In our example, age was the only consistently important predictor, but to conduct complete-variable analysis, we would have to remove this variable. We feel that IT-AIC approaches, which are not amiable to missing values, have unintentionally been encouraging practices such as complete-case and complete-variable analysis (see Anderson 2008).

Finally, we note that the scarcity of the literature simultaneously dealing with MI and IT-AIC approaches (but see Claeskens and Hiort 2009). Therefore, guidelines for how we combine these two techniques may be necessary and such a task can be an area of future research. However, we hope that our work here to be a call for integration between MI and IT-AIC approaches and a guide for behavioural ecologist to conduct MI with their own datasets, whose missing data should not miss out an opportunity for recovery.

Acknowledgement We thank Laszlo Garamszegi for the invitation to this special issue and the comments for a previous version of the manuscript. We are also grateful for the constructive comments by two anonymous referees. SN is grateful to Losia Lagisz and Mihoko Nakagawa for their support during the manuscript writing and is supported by Marsden Fund (UOO0812). RPF is funded by a Royal Society University Research Fellowship.

References

Allison PD (2002) Missing data. Sage, Thousand Oaks

Anderson DR (2008) Model based inference in the life sciences: a primer on evidence. Springer, New York

Barton K (2009) MuMIn: multi-model inference. In: R package version 0.12.0. http://r-forge.r-project.org/projects/mumin/

Biro PA, Dingemanse NJ (2009) Sampling bias resulting from animal personality. Trends Ecol Evol 24:66–67

Bolker BM (2008) Ecological models and data in R. Princeton University Press, Princeton

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, Berlin

Claeskens G, Hiort NL (2009) Model selection and model averaging. Cambridge University Press, Cambridge

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale

Collins LM, Schafer JL, Kam CM (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Meth 6:330–351

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via EM algorithm. J R Stat Soc B Methodol 39:1–38

Freckleton RP (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol doi:10.1007/s00265-010-1045-6

Garamszegi LZ (2010) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol. doi:0.1007/s00265-010-1028-7

Garamszegi LZ, Moller AP, Torok J, Michl G, Peczely P, Richard M (2004) Immune challenge mediates vocal communication in a passerine bird: an experiment. Behav Ecol 15:148–157

Garamszegi LZ, Eens M, Hurtrez-Bousses S, Moller AP (2005) Testosterone, testes size, and mating success in birds: a comparative study. Horm Behav 47:389–409

Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jorgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009a) Changing philosophies and tools for statistical inferences in behavioral ecology. Behav Ecol 20:1363–1375

Garamszegi LZ, Eens M, Janos T (2009b) Behavioural syndromes and trappability in free-living collared flycatchers, *Ficedula albicollis*. Anim Behav 77:803–812

Garland T, Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155:346–364

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Graham JW, Schafer JL (1999) On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle R (ed) Statistical strategies for small sample research. Sage, Thousand Oaks

Griffith SC, Owens IPF, Burke T (1999) Environmental determination of a sexually selected trait. Nature 400:358–360

Griffith SC, Owens IPF, Thuman KA (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. Mol Ecol 11:2195–2212

Hadfield JD (2008) Estimating evolutionary parameters when viability selection is operating. Proc R Soc B Biol Sci 275:723–734

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 23:494–508

Harrell FEJ, with contributions from many other users (2008) Hmisc: Harrell miscellaneous. In: R package version 3.5-2. http://biostat.mc.vanderbilt.edu/s/Hmisc

Hilborn R, Mangel M (1997) The ecological detective: confronting models with data. Princeton University Press, Princeton

Honaker J, King G (2009) What to do about missing data values in time series cross-section data. http://gking.harvard.edu/files/abs/pr-abs.shtml

Honaker J, King G, Blackwell M (2008) Amelia: Amelia II: a program for missing data. In: R package version 1.1-33. http://gking.harvard.edu/amelia

Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61:79–90

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Link WA, Barker RJ (2006) Model weights and the foundations of multimodel inference. Ecology 87:2626–2635

Little RJA (1988) A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 83:1198–1202

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Lukacs PM, Thompson WL, Kendall WL, Gould WR, Doherty PF, Burnham KP, Anderson DR (2007) Concerns regarding a call for pluralism of information theory and hypothesis testing. J Appl Ecol 44:456–460

McKnight PE, McKnight KM, Sidani S, Figueredo AJ (2007) Missing data: a gentle introduction. Guilford, New York

Nakagawa S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15:1044–1045

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23:592–596

Nakagawa S, Waas JR, Miyazaki M (2001) Heart rate changes reveal that little blue penguin chicks (*Eudyptula minor*) can use vocal signatures to discriminate familiar from unfamiliar chicks. Behav Ecol Sociobiol 50:180–188

Nakagawa S, Gillespie DOS, Hatchwell BJ, Burke T (2007a) Predictable males and unpredictable females: sex difference in repeatability of parental care in a wild bird population. J Evol Biol 20:1674–1681

Nakagawa S, Ockendon N, Gillespie DOS, Hatchwell BJ, Burke T (2007b) Does the badge of status influence parental care and investment in house sparrows? An experimental test. Oecologia 153:749–760

Ockendon N, Griffith SC, Burke T (2009) Extrapair paternity in an insular population of house sparrows after the experimental introduction of individuals from the mainland. Behav Ecol 20:305–312

R Development Core Team (2009) R: a language and environment for statistical computing, 282nd edn. R Foundation for Statistical Computing, Vienna

Richards SA, Whittingham MJ, Stephens PA (2010) Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1035-8

Rubin DB (1976) Inference and missing data. Biometrika 63:581–590

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489

Rushton SP, Ormerod SJ, Kerby G (2004) New paradigms for modelling species distributions? J Appl Ecol 41:193–200

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Schafer JL (1999) Multiple imputation: a primer. Stat Meth Med Res 8:3–15

Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Meth 7:147–177

Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2005) Information theory and hypothesis testing: a call for pluralism. J Appl Ecol 42:4–12

Stephens PA, Buskirk SW, del Rio CM (2007) Inference in ecology and evolution. Trends Ecol Evol 22:192–197

Still AW (1992) On the number of subjects used in animal behaviour experiments. Anim Behav 30:873–880

Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond B Biol Sci 269:137–142

Symonds MRE, Moussalli A (2010) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol. doi:0.1007/s00265-010-1037-6

van Buuren S, Groothuis-Oudshoorn K (2009) mice: Multivariate imputation by chained equations. In: R package version 1.21. http://www.stefvanbuuren.nl

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182–1189