Cell PRESS

Missing inaction: the dangers of ignoring missing data

Shinichi Nakagawa[1,2] and Robert P. Freckleton[2]

[1] Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand
[2] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

The most common approach to dealing with missing data is to delete cases containing missing observations. However, this approach reduces statistical power and increases estimation bias. A recent study shows how estimates of heritability and selection can be biased when the 'invisible fraction' (missing data due to mortality) is ignored, thus demonstrating the dangers of neglecting missing data in ecology and evolution. We highlight recent advances in the procedures of handling missing data and their relevance and applicability.

> *The best solution to handle missing data is to have none.* – R.A. Fisher

A ubiquitous issue but a neglected topic

Unfortunately, in real-world data sets, missing data are the norm rather than the exception [1–4]. Researchers usually omit cases containing missing data from analyses, concentrating only on sample units for which complete data are available (complete case analysis). At first sight this procedure might seem reasonable, and indeed it might appear that there is no other option available. However, in doing so, researchers often throw away a large part of their data, especially when a data set contains many variables but whole cases are deleted based on only one or two variables not being measured. Even worse, the parameter estimates from such pruned data sets are often incorrect when data are not missing completely at random (MCAR) [5] (Box 1). The illustrations of missing data which are missing at random (MAR) or missing not at random (MNAR) in Box 1 clearly demonstrate potential biases in parameter estimates that can be caused by deleting cases with missing observations when missing data are not MCAR. Given that MCAR is a very strong (and often incorrect) assumption, it is somewhat surprising that the topic of missing data in ecology and evolution has been largely ignored to date. For example, in recent years, *Trends in Ecology and Evolution* has hosted a series of papers discussing major advances in statistical philosophies and reform of statistical practices in ecology and evolution [6–8], but none of these reviews mention the topic of missing data. Possibly a major reason for this is that dealing with missing data is a rather technical issue. However, we believe that recent advances in handling missing data have made it possible to begin to tackle this difficult issue with the aid of techniques that have become well accepted in the statistical literature [2,4].

Here we highlight a recent study that clearly demonstrates the importance of missing data in evolutionary studies, along with some related work showing that ignoring missing data can compromise analyses in general. We then discuss some of the techniques that have been developed to deal with missing data which can be employed by researchers in the field of ecology and evolution.

Visualising the invisible fraction

Although there are a multitude of reasons why data sets contain missing observations, there are often biologically significant reasons for why this might be. For instance, missing data might be particularly important if organisms die before expression of a trait (e.g. secondary sexual traits) or while a trait is still being developed (e.g. weight or height). 'Missing' observations due to premature death in relation to the trait of interest are referred to as the 'invisible fraction' in the evolutionary literature [9]. Although the importance of invisible fractions in trait evolution has long been pointed out [9], researchers have ignored invisible fractions in calculating evolutionary parameters (e.g. heritability and selection gradients and differentials).

A recent paper by Hadfield [10] demonstrates the crucial importance of dealing with the invisible fraction in calculating such evolutionary parameters. The main result of the paper is that when a trait is under viability selection (i.e. a trait relates to survival; e.g. lighter chicks have higher mortality than heavier chicks), missing data due to the invisible fraction are MNAR in most cases (because a sample taken at a certain time after hatching will 'miss' light chicks that have already died). Therefore, estimators of evolutionary parameters such as heritability and selection (e.g. of body weight) are biased if estimated without accounting for the invisible fraction. Although calculating heritabilities and selection gradients is commonplace, very few studies have to date considered this problem. For traits under viability selection, missing observations depend on lifespan. If data sets include lifespan (e.g. if lifespan is the x-variable in Box 1, Figure Ia), missing observations in such traits can be treated as MAR, whereas without information on lifespan, the missing observations remain MNAR. Unfortunately, data on lifespan are rarely measured accurately and are also frequently incomplete (i.e. individuals might die between censoring points, or only after the final censoring point). With such incomplete lifespan data, missing observations of a trait under viability selection (i.e. the invisible fraction) are still MNAR.

When data are MNAR, it is necessary to make some assumptions about how the data are missing. With

Corresponding author: Nakagawa, S. (shinichi.nakagawa@otago.ac.nz).

Box 1. Problems of missing data and their bewildering classification

The mechanisms (distribution patterns) of missing data are traditionally divided into three classes: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [5]. If missing data are not MCAR, then there are potential problems in analysing data as though they were, but the precise outcome depends on the way in which they are missing, specifically whether data are MAR or MNAR. Figure I shows simple examples of this in the case of a bivariate regression.

In Figure I, assuming complete observations in the *x*-variable (red and blue points), missing data (red points) in (a) are MAR, missing data in (b) are MNAR and missing data in (c) are MCAR. This is because in (a), missing data in the *y*-variable (termed missingness) depend on the *x*-variable ($x < 0$) whereas in (b), missing data in the *y*-variable (missingness) depend on the *y*-variable itself ($y < 0$). For example, (a) represents a situation where the lifespan (*x*; possibly log scale) of chicks is correlated with weight at 13 days after hatching (*y*); the weight information of chicks that die before the 13th day is not available, although information on lifespan is available for all individuals. For (b), imagine a slightly different situation where lifespan (*y*; log scale) and hatching weight (*x*) are correlated; lifespan is only measured after a certain point (e.g. 13 days after hatching) whereas the information on hatching weight for all individuals is available. It is important to note, although confusing, that if the *x*-variable (e.g. lifespan) is not among those measured in a situation such as (a), missing data in (a) should be classified as MNAR. Thus, MNAR comes in two forms: (i) missingness depends on the missing value itself or (ii) missingness depends on an unobserved variable (see Refs [1–4,10,18] for more technical and precise definitions for missing mechanisms, along with definitions for 'ignorability' of missing data; MCAR and MAR are referred to as 'ignorable' whereas MNAR is 'non-ignorable').

In addition to obvious biases in the means and variances of *x* and *y* due to missing data in (a) and (b) but not in (c), the key point in Figure I is that the estimates of the slope might be biased if missing data are not MCAR, depending on the nature of the missing data. In (a), the expected slope is unbiased as missingness depends on *x*, with the result that covariance between *x* and *y* is reduced by the same amount as the reduction in standard deviation of *x*. However, in (b), this is not true and the slope is biased as the missingness is determined by *y*. In (c), the slope is unbiased. Moreover, in (c), the expected R^2 is the same as for the original data, whereas in the case of (a) and (b), the R^2 is reduced. Clearly, even in this simple example, the consequences of missing data are not straightforward to predict. The situation will become much more complex if a multivariate data set is considered. Moreover, in reality, a data set might contain variables with missing observations which are MCAR, MAR or MNAR.

Notably, this classification (i.e. MCAR, MAR and MNAR) has been criticised because of the confusing nature of its terminology (e.g. MAR does not mean that missing data are distributed at 'random'). Furthermore, MNAR can be difficult to distinguish from MAR owing to the very fact that we have no information regarding missing values when MNAR (but see Box 2) [4]. Therefore, importantly, the most practical assumption is MAR, which is a basis of recent advances of handling missing data (e.g. Box 2) [1–4,18].

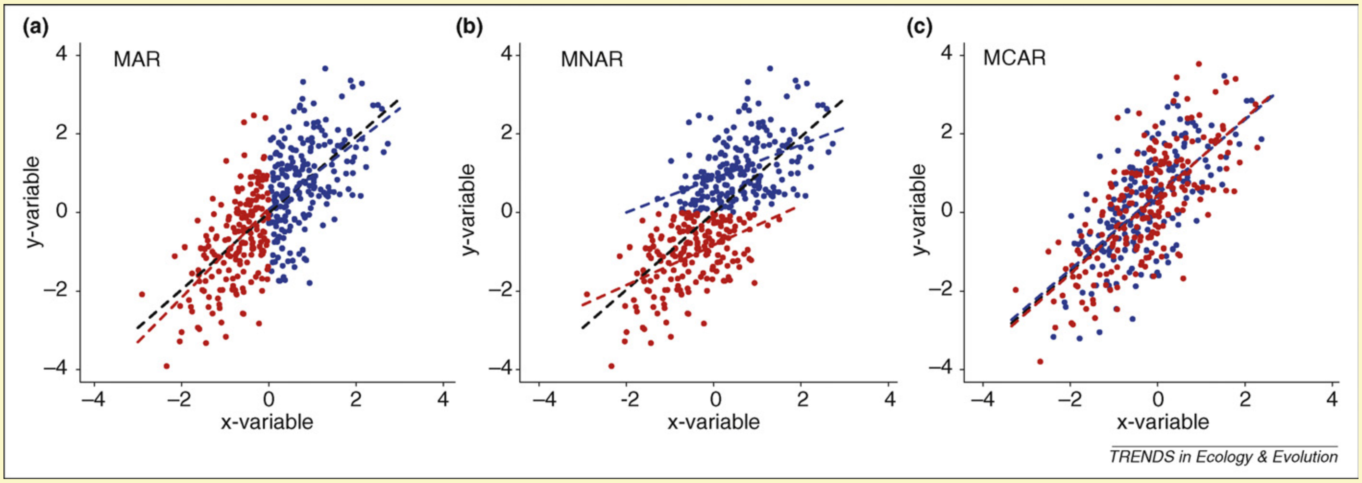

Figure I. Illustrations of the classification for the mechanism of missing data. Blue points are observations whereas red points are missing observations in the *y*-variable; statistics for complete data (blue and red combined) are slope (*b*) = 1, standard error (*se*) = 0.05 and $R^2 = 0.5$. Assuming observations in the *x*-variable are complete, **(a)** represents missing at random (MAR), **(b)** represents missing not at random (MNAR) and **(c)** represents missing completely at random (MCAR). For the observed data (blue points), the estimated slope, *se* and R^2, are (a) $b = 0.86$, $se = 0.11$, $R^2 = 0.29$, (b) $b = 0.432$, $se = 0.06$, $R^2 = 0.23$ and (c) $b = 0.957$, $se = 0.07$, $R^2 = 0.49$.

censored survival data, it is generally assumed that the model of survival between censoring points is consistent with what is observed at the censoring points. For example, imagine that the population mean body size increases systematically between three censoring points (because small individuals are the first to die). Then, it is reasonable to assume that among those individuals who died in the first interval, the smaller ones died shortly after the first censoring point, and the larger ones died shortly before the second censoring point. However, this assumption might not be justified in certain situations (e.g. involving traits with less predictable expression or development, such as secondary sexual characters).

In the quantitative genetic framework, as Hadfield [10] points out, this assumption on the invisible fraction (e.g. relationship between lifespan and a trait) can be verified and adjusted with pedigree information. This is because some information on trait values of individuals that died prematurely can be obtained from observed trait values of their relatives that exhibit similar trait values (e.g. the lifespan of an individual that died without a weight measurement can be compared with the weights of its relatives to verify the relationship between lifespan and weight). Thus, if missing observations are modelled accurately with a pedigree, the bias in heritability and selection estimates can be reduced.

One of the major conclusions of the paper by Hadfield was surprise regarding the neglect of missing data in evolutionary biology in general, and a call for more attention to this problem [10]. Indeed, missing observations in ecology and evolution might often not be MCAR. For example, in addition to the situation described by Hadfield

Box 2. Multiple imputation: the highly praised yet underused method

Multiple imputation (MI) procedures generally include three steps: data imputation, routine analysis and pooling results for parameter estimation (Figure Ia; Figure I also includes conceptual representations of single imputation [b] and data augmentation [c]).

The imputation step is a 'filling-in' process by assuming MAR (several methods are available for imputation, although for a data set containing both categorical and continuous variables, imputation can be a difficult task; reviewed in Ref. [2]). This step is repeated *m* times (*m* between 3 and 10 is usually sufficient [24]); each set of imputed missing values is unique. The second step involves each 'complete' data set being analysed. The third step aggregates parameter estimates from results of each analysis. This pooling process is fairly straightforward using the equation [4,18,24,25]

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j,$$

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j,$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{Q}_j - \overline{Q})^2,$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B,$$

where $\bar{Q}$ is the mean of $\hat{Q}_j$ which is a parameter estimated from the *j*th data set ($j = 1, 2, \ldots, m$), $\bar{U}$ is the within-imputation variance derived from $\hat{U}_j$ which is the standard error associated with $\hat{Q}_j$, B is the between-imputation variance estimates and T is the total variance (the overall standard error is the square root of T). Additionally, missing information in the original data set can be quantified using the rate of missing information or γ (a unique and important property of MI ranging from 0 to 1) [4,18,24,25]:

$$\gamma = \frac{r + 2/(d\,f + 3)}{r + 1},$$

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

$$d\,f = (m - 1)\left(1 + \frac{m\bar{U}}{(m + 1)B}\right)^2,$$

where *r* represents the relative increase in variance (or uncertainty of estimates) due to missing values and *df* is the degree of freedom associated with the *t* distribution ($t = \bar{Q}/\sqrt{T}$; used for calculating confidence intervals and/or statistical significance of parameter estimates, e.g. slopes and intercepts). Notably, γ is small when missing data are MAR (Box 1), but will be substantial when missing data are MNAR (for more details, see Refs [4,24]).

It is important to clarify that data augmentation is superior to MI in a number of respects: (i) the number of replacements of missing observations in a data set can be infinite (not $m = 3$–10) and (ii) feedback between missing data replacements and parameter estimates exists (Figure I). A useful guide for implementing MI using software is found at http://www.multiple-imputation.com, and more detailed information regarding MI is found in the form of Frequently Asked Questions at http://www.stat.psu.edu/~jls/mifaq.html, along with useful references.

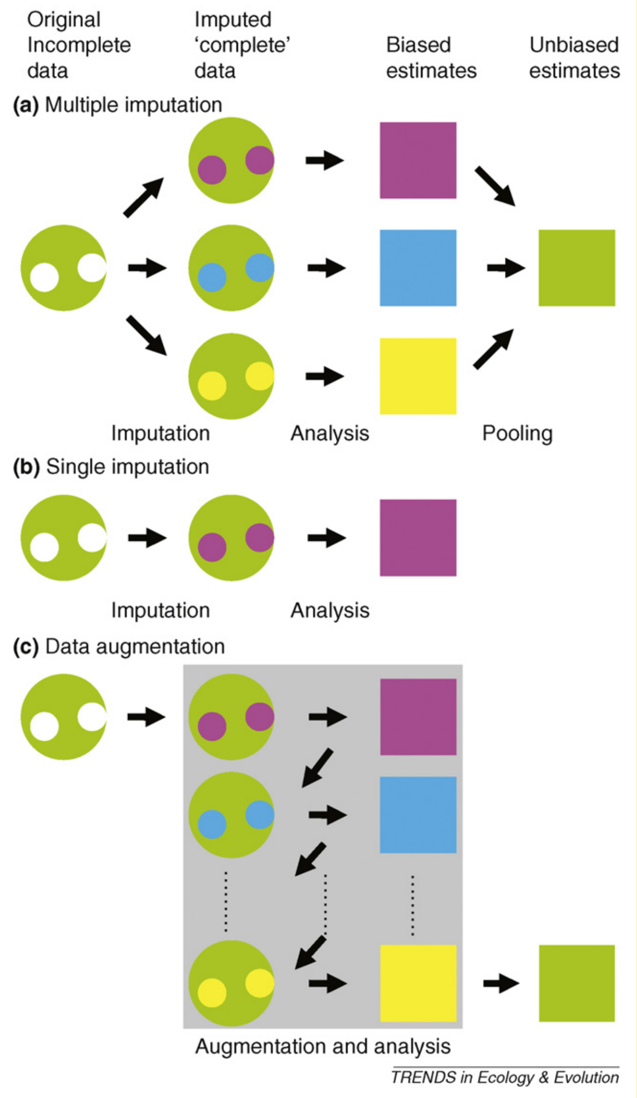

Figure I. Diagram outlining **(a)** multiple imputation ($m = 3$; modified from http://www.multiple-imputation.com), **(b)** single imputation and **(c)** data augmentation; a grey box represents implicit data replacements integrated in data analysis, and there are feedback processes between augmentation and analysis. Note that all three procedures can use the same probability models (e.g. using maximum likelihood or Markov chain Monte Carlo) for imputation or augmentation of missing data.

[10], older individuals might be more difficult to catch or, more generally, different life-history strategies within a species might cause differential capture/observation rates of individuals [11]. In comparative analysis, data on rare species are more likely to be missing than those for common species [12]. Also, missing observations due to extinction should be treated as MAR rather than MCAR because the occurrence of extinction events is probably not completely at random [13]. For example, a study which investigated intrinsic and extrinsic factors to predict extinction risk in Australian marsupials suggested that missing observations due to extinction were not MCAR [14]. This was indirectly demonstrated by the fact that the results from the data set in which the missing observations were deleted were different from the results obtained by using a data set in which the missing observations were handled properly. Therefore, missingness itself often tells something interesting about biological phenomena.

These arguments suggest that researchers should at least start to investigate whether missing data in their data sets are MCAR. Fortunately, diagnostic procedures for distinguishing whether or not missing data are MCAR are relatively straightforward, although distinguishing MNAR from MAR is usually more difficult (see Ref. [4] for a review of numerical and graphical methods; see also Box 2). Graphical procedures can be easily employed for the former; for example, replacing missing observations with 'outlier' values in a multivariate data set and scatterplotting each bivariate combination in the data set will easily detect whether or not missing data are MCAR (i.e. clusters of the outlier values will be found if not MCAR). Some statistical packages such as SPSS have special graphical output which shows missing patterns in a data set (e.g. the data matrix plot; any add-on routines for microarray analysis should have this type of plot function) [4].

Advances in handling missing data

Once a data set has been found to contain missing data, there are three broad categories of methods for handling such data: deletion, augmentation and imputation. Data deletion, as mentioned above, is the most commonly used. Its advantage is simplicity. However, data deletion can bias parameter estimates and reduce statistical power (e.g. see Box 1).

The other two procedures, data augmentation and imputation (Box 2, Figure I), can be very similar, but the main difference is that data augmentation (Box 2, Figure Ic) does not explicitly replace missing values, whereas data imputation (Box 2, Figure Ia,b) will substitute missing observations with imputed values which we can actually see [4]. In data augmentation procedures (Box 2, Figure Ic), parameter estimation, based on observed data, is augmented by the extra information gained from assuming certain underlying distributions and probability models (note that such assumptions are also usually made in data imputation) [4]. Data augmentation procedures included maximum likelihood (ML), expectation maximisation (EM) [15], Markov chain Monte Carlo (MCMC) [16], weighting, dummy code adjustment and more (the first three are referred to as model-based procedures; for more details of these procedures, see Refs [1–4,17]). For example, an ML procedure assumes a multivariate normal distribution among both response and predictor variables in a data set (under MAR) and estimates parameters which maximise the probability of observed data, taking missing observations into account (i.e. estimating and averaging over probabilities of all possible missing values and their combinations) [1–4,18].

Data imputation can be divided into two types: single-imputation and multiple-imputation procedures (Box 2, Figure Ia,b). Single imputation (Box 2, Figure Ib) such as replacing missing values with the mean (i.e. mean substitution) is often worse than data deletion in terms of parameter estimation and especially the estimation of parameter uncertainty (i.e. standard errors), because single imputation ignores any uncertainty of imputed values [1,3,4]. For example, in evolutionary biology, this approach has been used in comparative analyses where missing values are substituted with mean trait values for genera or close relatives, but the approach has been criticised [19,20]. Even when some model-based procedures such as ML, EM and MCMC are used, single imputation suffers from inaccuracy in estimates of parameter uncertainty.

By contrast, multiple imputation (MI) (Box 2, Figure Ia) is probably the most praised and fastest-growing method in handling missing data, and is indeed becoming the standard method in social and medical sciences [2,4,18,21]. As far as we are aware, the study regarding extinction risk mentioned above [14] is among the few which employs MI (or even considers the problem of missing data) in the field of evolution and ecology [6,10]. The critical difference between MI and the other methods of handling missing data are the capability of MI to provide information (i.e. statistical parameters) regarding the impact of missing data on parameter estimation (see Box 2). In summary, model-based data augmentation and MI are the methods recommended by statisticians to deal with missing data [1,2,18].

Concluding thoughts

More and more researchers in ecology and evolution use criterion-based model selection procedures such as the information-theoretic approach (e.g. Akaike's information criterion; AIC) and the Bayesian approach (e.g. Bayesian or deviance information criterion; BIC and DIC) [6–8]. Any model selection procedure requires complete cases (at least in predictor variables). However, as we have discussed, the deletion of missing data which are not MCAR will bias parameter estimates from models. Furthermore, model ranking according to statistical criteria such as AIC and the associated statistics (e.g. Akaike weights [22]) will be similarly biased as a consequence. This is not a trivial issue, and yet many researchers are probably unaware of it. Model-based data augmentation and/or MI should be incorporated before or as a part of model selection procedures.

Given the current accessibility of software packages devoted to handling missing data for data augmentation and MI (both stand-alone packages and add-on routines for statistical packages [2,4]), researchers should be able to appropriately deal with missing data in their data sets in most cases. We believe that MI can be immediately employed by many researchers because researchers can use familiar statistical software (Box 2), whereas data augmentation is more demanding (e.g. implementing a data augmentation process within statistical analysis [10,23]). Under some circumstances, model-based data augmentation and MI procedures can be difficult to implement, particularly if there is a large amount of missing data, or if a data set includes variables with different distributions and/or variables which have nested or repeated structures [2,4,17]. Nonetheless, all these procedures remain much more feasible solutions for missing data than Fisher's blithe advice not to have any missing values.

Acknowledgements

We thank Hamish Spencer, Sheena Townsend, Tom Webb and three anonymous reviewers for comments which greatly improved this paper. R.P.F. is funded by a Royal Society University Research Fellowship. S.N. is grateful to Terry Burke, who provided financial support.

References

1 Allison, P.D. (2002) *Missing Data,* Sage
2 Horton, N.J. and Kleinman, K.P. (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.* 61, 79–90
3 Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data,* John Wiley & Sons
4 McKnight, P.E. *et al.* (2007) *Missing Data: A Gentle Introduction,* Guilford Press

5 Rubin, D.B. (1976) Inference and missing data. *Biometrika* 63, 581–590
6 Clark, J.S. and Gelfand, A.E. (2006) A future for models and data in environmental science. *Trends Ecol. Evol.* 21, 375–380
7 Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.* 19, 101–108
8 Stephens, P.A. *et al.* (2007) Inference in ecology and evolution. *Trends Ecol. Evol.* 22, 192–197
9 Grafen, A. (1988) On the uses of data on lifetime reproductive success. In *Reproductive Success* (Clutton-Brock, T.H., ed.), pp. 454–471, University of Chicago Press
10 Hadfield, J.D. (2008) Estimating evolutionary parameters when viability selection is operating. *Proc. R. Soc. Lond. B Biol. Sci.* 275, 723–734
11 Lebreton, J.D. *et al.* (1992) Modeling survival and testing biological hypotheses using marked animals – a unified approach with case-studies. *Ecol. Monogr.* 62, 67–118
12 Kunin, W.E. and Gaston, K.J. (1997) *The Biology of Rarity: Causes and Consequences of Rare-Common Differences,* Chapman & Hall
13 Maddison, W.P. *et al.* (2007) Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* 56, 701–710
14 Fisher, D.O. *et al.* (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. *Proc. R. Soc. Lond. B Biol. Sci.* 270, 1801–1808
15 Dempster, A.P. *et al.* (1977) Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. B* 39, 1–38
16 Tanner, M.A. and Wing, H.W. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–540
17 Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data,* Chapman & Hall
18 Schafer, J.L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177
19 Freckleton, R.P. *et al.* (2003) Bergmann's rule and body size in mammals. *Am. Nat.* 161, 821–825
20 Smith, R.J. and Jungers, W.L. (1997) Body mass in comparative primatology. *J. Hum. Evol.* 32, 523–559
21 Raghunathan, T.E. (2004) What do we do with missing data? Some options for analysis of incomplete data. *Annu. Rev. Public Health* 25, 99–117
22 Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* Springer-Verlag
23 Clark, J.S. and Bjornstad, C.N. (2004) Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology* 85, 3140–3150
24 Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons
25 Schafer, J.L. (1999) Multiple imputation: a primer. *Stat. Methods Med. Res.* 8, 3–15

0169-5347/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tree.2008.06.014 Available online 25 September 2008

Research Focus

Testate amoebae and nutrient cycling: peering into the black box of soil ecology

David M. Wilkinson

School of Natural Sciences and Psychology, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

In some areas of ecology and evolution, such as the behavioural ecology of many well-studied bird species, it is increasingly difficult to make surprising new discoveries. However, this is not the case in many areas of soil and/or microbial ecology. Two recent studies suggest that the testate amoebae, a microbial group unfamiliar to most biologists, might play a much larger role in soil nutrient cycling than has hitherto been suspected.

The importance of soil

The soil is largely 'out of sight' to an ecologist without a spade and, for much of the 20th century, this meant that it was also 'out of mind' to most ecologists. At best, it tended to be treated as a black box, with the behaviour of its inhabitants lumped together under simple labels such as decomposers or nitrogen fixers [1]. Slowly, things are changing: indeed, it has been noticeable that since I started attending major ecology meetings (in the mid-1980 s), the number of papers and sessions on topics such as soil ecology or mycorrhizae has been increasing. One reason for this increase in interest might be the realisation that studying changes in soil respiration is crucial to predicting the future of the soil as a carbon sink [2]. This could be vital for understanding the effects of global warming.

The silica cycle

When ecology textbooks describe soil microbiology, it is often in the context of nutrient cycling. One of the cycles

Box 1. Testate amoebae

Testate amoebae (also known as testate rhizopods or thecamoebians) are protozoa in which the single cell is enclosed within a shell usually referred to as a test, with a size range of 5–300 μm [11]. The tests are usually composed of either self-secreted material – which can be siliceous or proteinaceous – or 'agglutinated' tests, which incorporate material from the environment (such as sand grains, diatoms or the scales of smaller siliceous testates which have been consumed as prey) [12].

Like many microbes, testate amoebae have a relatively modest fossil record – for example, occasionally being preserved in amber. However, recently fossils very similar to modern testates have been described from rocks of around 740 million years old [13]. Although polyphyletic (traditionally placed in the phylum Rhizopoda), testates appear to form a reasonably uniform ecological grouping, occurring around the world in a range of terrestrial and freshwater habitats. They are especially common in habitats with high organic matter content, such as organic-rich soils, peats and mosses [12]. Many, but not all, of the identified morphospecies are cosmopolitan in their distribution [14,15].

The presence of tests means that taxa of testate amoebae can be identified by morphology and their populations can be enumerated by direct counting. Testate amoebae thus represent a microbial group whose ecology can be studied by approaches very similar to those used in the study of macroscopic organisms. There is also a long history of studies of testate amoebae autecology, dating back to 19th century microscopists (Figure I).

Corresponding author: Wilkinson, D.M. (d.m.wilkinson@ljmu.ac.uk).