Chapter 20
Three-Way Nested Data for Age Determination Techniques Applied to Cetaceans

E.N. Ieno, P.L. Luque, G.J. Pierce, A.F. Zuur, M.B. Santos, N.J. Walker, A.A. Saveliev, and G.M. Smith

20.1 Introduction

In the previous case study, we showed how multiple samples from bacteria in honey bees from the same hive gave a nested data structure, and mixed modelling techniques were applied to allow for correlations between observations from the same hive. The bee data provided an example of two-way nested data, and the underlying theory for this was discussed in Chapter 5. In this chapter, we go one step further and use three-way nested data, which extends the two-way approach discussed in Chapter 5. The underlying theory builds on the approach used for two-way data, and we recommend reading Chapter 5 before starting this chapter as we assume familiarity with the theory, model selection, and R code for two-way nested data.

We use a subset of the data analysed in Luque (2008), who compared the results from three staining methods to determine the age of cetaceans stranded in Spain and Scotland. The data are nested in the sense that samples derive from multiple species, and from each species, we have various specimens (individual animals). From each specimen, several teeth were sectioned and tooth sections were stained using three staining methods (the Mayer Haematoxylin, Ehlrich Haematoxylin, and Toluidine Blue methods), giving three age estimates from each tooth. A diagram of the nested structure is given in Fig. 20.1. The three age observations per specimen (obtained by the three staining methods) are likely to be correlated, but we may also expect correlation between age readings within the same species (if, for example, different species have different lifespans and/or different age classes tend to become stranded and thus become the source of samples). The response variable is the estimated age of the animal. Available explanatory variables are sex (male or female), location of stranding (Scotland or Spain), and stain (Mayer Haematoxylin, Ehlrich Haematoxylin, and Toluidine Blue).

In Chapter 4 of West et al. (2006), a three-way nested data set on mathematic scores for students within multiple classes and multiple schools is analysed. From a

E.N. Ieno (✉)
Highland Statistics Ltd., Newburgh, AB41 6FN, United Kingdom

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_20,

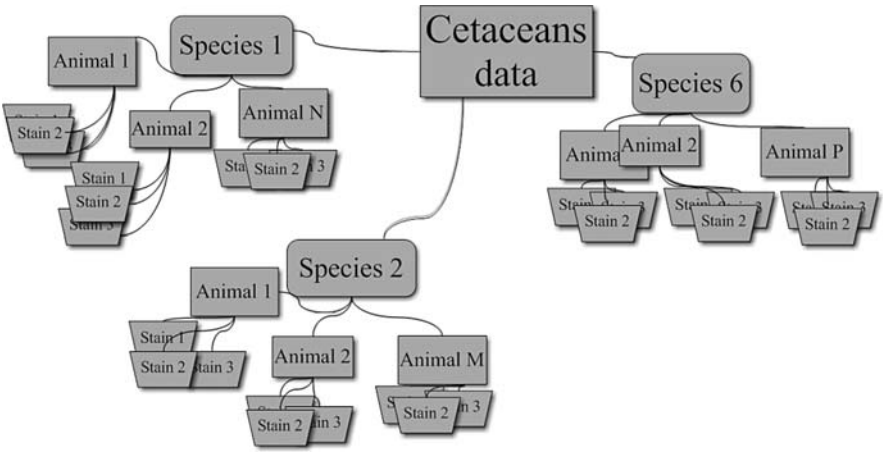

Fig. 20.1 Sketch of the nested structure of the data. Six cetacean species were sampled. These were *Delphinus delphis*, *Lagenorhynchus acutus*, *Phocoena phocoena*, *Stenella coeruleoalba*, *Stenella frontalis*, and *Tursiops truncatus*. For each species, various specimens (animals) were available. The number of specimens per species range between 3 and 25. From each specimen, three estimated age readings were obtained by the three staining methods (labeled as 1, 2, and 3 in the graph)

statistical point of view, there are not many differences between their classroom example and our cetacean data set. In fact, we will closely follow their steps. The only difference is that West et al. (2006) used two different model selection approaches; (i) the step-down approach, which was presented as our protocol with steps 1–10, and (ii) a step-up approach. The classroom data are analysed with the step-up approach. For the cetacean data, we will follow our familiar step-down approach.

20.2 Data Exploration

The first question to ask with nested data is how much variation is there between specimens and between species? Figure 20.2 shows a boxplot of age conditional on specimen. Recall that we have three observations per specimen. The graph shows that we have a large between-specimen variation, which means we probably need to use 'animal' as a random effect. The same graph was made for species (Fig. 20.3) and shows there is considerably less between-species variation. This should not be too surprising as each animal has only one true age, but the samples of stranded animals for each species should include the range of age classes present in the populations. Even if one species tends to be longer-lived than another, there will inevitably be a considerable overlap in ages present.

We also made boxplots of age conditional on sex, age conditional on location, and age conditional on staining method. Results are not shown here, but they

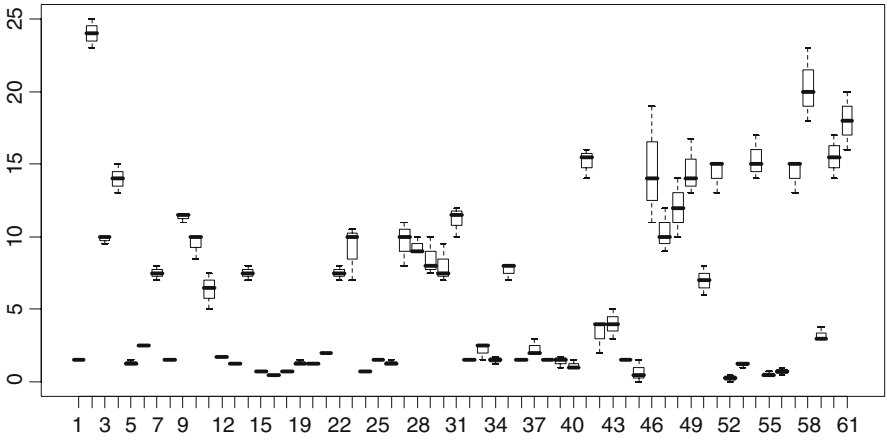

Fig. 20.2 Boxplot of age conditional on animal. Each boxplot consists of three observations from the same animal. Not all numbers are plotted along the horizontal axis due to limited space

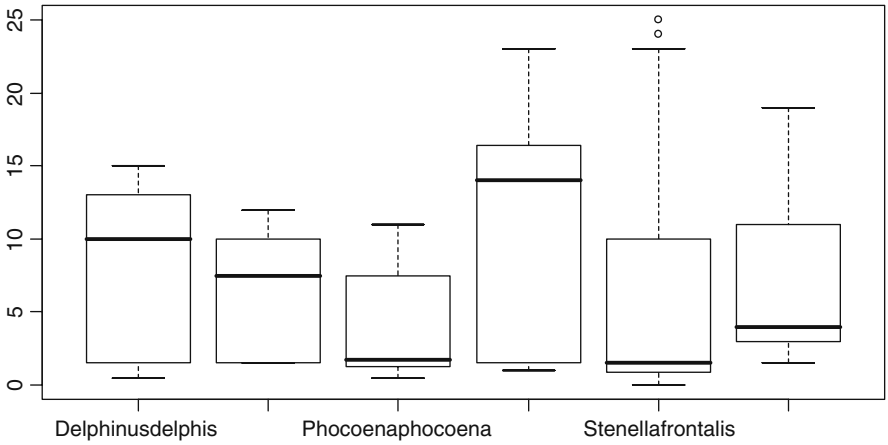

Fig. 20.3 Boxplot of age conditional of species. There is considerably less between-species variation compared to between-animal variation

indicate that the variation in ages recorded in Scotland is considerably less than Spain, indicating we may need to use different variances per country. There are also three observations with undetermined sex, and to allow for interactions between sex, location, and age determination, we removed these observations.

The following R code was used to access the data, make the two boxplots, and remove the three observations where sex was undetermined. By now, you should be familiar with this code. The object `Cetaceans2` contains the male and female data (the class unknown was dropped).

```
> library(AED); data(Cetaceans)
> Cetaceans$fSpecies <- factor(Cetaceans$Species)
> Cetaceans$fDolphinID <- factor(Cetaceans$DolphinID)
> boxplot(Age ~ fSpecies, data = Cetaceans)
> boxplot(Age ~ fDolphinD), data = Cetaceans)
> I <- Cetaceans$Sex==0
> Cetaceans2 <- Cetaceans[!I,]
```

20.3 Data Analysis

The starting point for the analysis is a model of the form

$$\text{Age}_{ijk} = \text{fixed part}_{ijk} + \text{random part}_{ijk} \tag{20.1}$$

The variable Age_{ijk} is the age of observation i in animal j of species k. The index k runs from 1 to 6 and i from 1 to 3. The number of animals per species differs. We start discussing the fixed part of the model and then the random part. Recall from Chapters 4 and 5 that the protocol dictates that we start with a model that contains as many fixed explanatory variables as possible. In this case, we have three nominal explanatory variables. We therefore start with a model containing sex, stain, and location as main terms, all two-way interactions, and the three-way interaction. Hence, the fixed part consists of

$$\begin{aligned}&\text{Sex}_{ijk} + \text{Stain}_{ijk} + \text{Location}_{ijk} + \text{Sex}_{ijk} \times \text{Stain}_{ijk} + \text{Sex}_{ijk} \times \text{Location}_{ijk} + \\ &\text{Stain}_{ijk} \times \text{Location}_{ijk} + \text{Sex}_{ijk} \times \text{Stain}_{ijk} \times \text{Location}_{ijk}\end{aligned}$$

This model is fitted with the `gls` function to serve as a reference model. The following code was used for this.

```
> library(nlme)
> Cetaceans2$fSex <- factor(Cetaceans2$Sex)
> Cetaceans2$fLocation <- factor(Cetaceans2$Location)
> Cetaceans2$fStain <- factor(Cetaceans2$Stain)
> f1 <- formula(Age ~ fSex * fStain * fLocation)
> M1 <- gls(f1, method = "REML", data = Cetaceans2)
```

We can now go to step 2 of the analysis. The random effect 'animal' is nested within the random effect 'species'. Just as West et al. (2006), we argue that if the random effect 'animal' is included in the model, then the random effect 'species' should also be included in the model. Making our starting point for the random part,

$$a_k + a_{j|k} + \varepsilon_{ijk}$$

The term ε_{ijk} is the unexplained error and represents the within-animal variation. It is assumed to be normally distributed with mean 0 and variance σ^2. However, the data exploration indicated that there may be different spread per location and we should be prepared at some stage to test whether multiple variances are needed per location. But to avoid too many steps at once, we will wait until we reach steps 3–5 of the analysis before considering this in any detail.

Recall that the index k refers to species k. We assume that a_k is normally distributed with mean 0 and variance ${\sigma_{\text{species}}}^2$. The term a_k is a random intercept and allows for variation between the species. The amount of variation is determined by ${\sigma_{\text{species}}}^2$. The term $a_{j|k}$ looks intimidating, but represents the variation between animals (index j) of the same species (index k). We assume it is normally distributed with mean 0 and variance ${\sigma_{\text{animal}}}^2$. Summarising, a_k allows for variation between the species and $a_{j|k}$ for the variation between animals within the same species.

Therefore, our starting model contains a sex, location, and stain effect as well as all their interactions, and we also use random intercepts that model between-species variation and between-animal variation within the species.

As part of this analysis (step 2), the first model comparison is between the model with the two random effects a_k and $a_{j|k}$ and a model without them. Recall that these random effects are nested. If the between-animal variation is important, then we should use both random effects. So, we will not test whether the random effect a_k on its own is important. The following code applies the model with both random effects and compares the model with and without the random effects using the `anova` command.

```
> M2 <- lme(f1, random =~1 | fSpecies / fDolphinID,
            data = Cetaceans2, method = "REML")
> anova(M1, M2)
```

It is important that you define the variables `fSpecies` and `fDolphinID` as factors before the `lme` command or R will give an error message. The output of the `anova` command is as follows:

```
   Model df       AIC       BIC    logLik   Test  L.Ratio p-value
M1     1 13 1101.4488 1141.8261 -537.7244
M2     2 15  740.3277  786.9168 -355.1638 1 vs 2 365.1212  <.0001
```

The AIC indicates that the model with the two random effects is considerably better. The likelihood ratio statistic is $L = 365.12$, and the cited p-value indicates that we can reject the null hypothesis H_0: ${\sigma_{\text{animal}}} = 0$ in favour of the alternative H_1: ${\sigma_{\text{animal}}} > 0$. However, note that we are testing on the boundary, and therefore, the cited critical p-value should be multiplied with 0.5; see also Chapter 5 or Chapter 4 in West et al. (2006). Even after applying this correction, we still come to the same conclusion that we need the two random intercepts.

We now have two options. We can either apply a model validation, check for homogeneity (especially plotting residuals versus location), or extend the model by allowing for multiple variance based on location and see whether it improves the model. The motivation for the last approach is because the data exploration showed a clear difference in spread per location. Recall from Chapter 4 that adding such a variance structure extends the model to

$$\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma_s^2)$$

The index s refers to the two locations, allowing the residuals from the two locations to have a different spread. Based on the data exploration, we decided to include the multiple variance structure and see whether it improved the model. Some might argue that this variance structure should have been used in the starting model, but there are two reasons for not doing this; firstly because we prefer to start as simple as possible and secondly the explanatory variables could have explained the differences in spread. The following R code was used to extend the model.

```
> M3 <- lme(f1, random =~ 1 | fSpecies / fDolphinID,
          weights = varIdent(form =~ 1 | fLocation),
          data = Cetaceans2)
```

The only new bit of code is the weights option with the `varIdent` variance structure (see also Chapter 4). The `anova` command shows that the AIC of this model is 733.01 and the likelihood ratio statistic is $L = 9.30$ ($df = 1$, $p = 0.002$), making this the best model so far. We can now proceed to steps 7–9 of the analysis to find the optimal fixed structure for our selected random structure.

To find the optimal model in terms of the fixed explanatory variables, we can either use the t-statistics, sequential F-tests, or likelihood ratio tests. In this instance, as we have factors with more than two levels (stain), we decided to use the third option. This part of the analysis was described earlier in Chapter 5, and we assume the reader is familiar with the tedious repetitive process of fitting a full model dropping all allowable terms in turn, applying likelihood ratio tests of nested models dropping the least significant term, and repeating the whole process until all terms are significant.

We first fitted a model with all terms (main terms: all two-way interactions and the three-way interaction) and then a model without the three-way interaction. Both models were fitted with maximum likelihood estimation (ML). The likelihood ratio test indicated that we could drop the three-way interaction ($L = 5.05$, $df = 2$, $p = 0.07$). The process then continued by dropping each of the three two-way interaction terms in turn and identified the least significant with the likelihood ratio test. The first interaction term to go out was the sex $\times$ location term ($L = 0.35$, $df = 1$, $p = 0.55$), followed by the sex $\times$ stain interaction ($L = 1.5$, $df = 2$, $p = 0.46$), and finally, sex as a main term was dropped ($L = 0.68$, $df = 1$, $p = 0.40$) as it was not included in the remaining two-way interaction term. At this point, the fixed part of the model contained stain, location, and the stain $\times$ location interaction. Dropping

the interaction gave $L = 19.14$ ($df = 2$, $p < 0.001$), giving the optimal model in terms of fixed terms. In words, it is given by

$$\text{Age}_{ijk} = \text{Stain}_{ijk} + \text{Location}_{ijk} + \text{Stain}_{ijk} \times \text{Location}_{ijk} + a_k + a_{j|k} + \varepsilon_{ijk}$$

We refitted the model with REML and applied a model validation. There are no problems with homogeneity. The numerical output of the model is obtained with the `summary` command:

```
> options(digits=4)
> summary(M3)

Linear mixed-effects model fit by REML
 Data: Cetaceans2
    AIC   BIC logLik
  734.7 766.1 -357.4

Random effects:
 Formula: ~1 | fSpecies
        (Intercept)
StdDev:       1.285

 Formula: ~1 | fDolphinID %in% fSpecies
        (Intercept) Residual
StdDev:       5.503   0.6496

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | fLocation
 Parameter estimates:
Scotland    Spain
   1.000    1.596
Fixed effects: list(f1)
                              Value Std.Error  DF t-value p-value
(Intercept)                   4.050    1.3502 114   3.000  0.0033
fStainMayer                   0.398    0.1624 114   2.454  0.0157
fStainToluidine               0.227    0.1624 114   1.395  0.1657
fLocationSpain                3.928    1.8085  52   2.172  0.0345
fStainMayer:fLocationSpain    1.481    0.3255 114   4.551  0.0000
fStainToluidine:fLocationSpain 0.672   0.3255 114   2.063  0.0414
 Correlation:
                              (Intr) fStnMy fStnTl fLctnS fSM:LS
fStainMayer                   -0.060
fStainToluidine               -0.060  0.500
fLocationSpain                -0.718  0.045  0.045
fStainMayer:fLocationSpain     0.030 -0.499 -0.249 -0.090
fStainToluidine:fLocationSpain 0.030 -0.249 -0.499 -0.090  0.500

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-2.97802 -0.31902 -0.04765  0.30647  3.33243
```

```
Number of Observations: 177
Number of Groups:
                fSpecies fDolphinID %in% fSpecies
                       6                       59
```

The Mayer staining method when applied to samples from Spain is significantly different from the Ehlrich (baseline) method when applied to sample from Scotland (baseline). The Toluidine method is also significantly different from the Ehlrich method, but a p-value of 0.04 is not really that impressive. It may be an option to change the baseline and see whether the Mayer and Toluidine methods differ from each other. Note that the main term location makes a major contribution for Spain to the fitted values.

The `summary` command also gives information on the random terms. The estimated values for σ, σ_{animal}, and σ_{species} are 0.64, 5.50, and 1.18, respectively. The multiplication factors for the different standard deviations per stratum for location are 1 for Scotland and 1.59 for Spain. The residual spread in Spain is therefore considerably larger than it is in Scotland. This means that more of the age variation is explained by the available explanatory variables in Scotland than in Spain.

20.3.1 Intraclass Correlations

We now discuss the interpretation of the random intercepts. The output above shows that the random effect a_k, representing the between species variation, is N(0, 1.28^2), the random effect $a_{j|k}$, representing the between animal variation in the same species, is N(0, 5.50^2) for observations from Scotland, the random noise ε_{ijk} is N(0, 0.64^2), and for observations from Spain, the random noise ε_{ijk} is N(0, 1.02^2). The value of 1.02 is obtained by multiplying 0.64 and 1.59. The values can be used to calculate the intraclass correlation at the species level ($\text{ICC}_{\text{species}}$) and at the animal level ($\text{ICC}_{\text{animal}}$). The formulae were taken from West et al. (2006) and are as follows:

$$ICC_{\text{species}} = \frac{\sigma^2_{\text{species}}}{\sigma^2_{\text{species}} + \sigma^2_{\text{animal}} + \sigma^2}$$

$$ICC_{\text{animal}} = \frac{\sigma^2_{\text{species}} + \sigma^2_{\text{animal}}}{\sigma^2_{\text{species}} + \sigma^2_{\text{animal}} + \sigma^2}$$

The only difference is that we need to calculate these ICCs for both Scotland and for Spain as we have two σs. The actual calculations are just a matter of filling in the values and give the ICCs for Scotland: $\text{ICC}_{\text{species}} = 0.05$ and $\text{ICC}_{\text{animal}} = 0.98$ and for Spain: $\text{ICC}_{\text{species}} = 0.05$ and $\text{ICC}_{\text{animal}} = 0.96$. Hence, there is massive correlation between the three observations of the same animal for data of both countries, as we would of course expect for this dataset. The correlation between animals of the same species is low. Just as West et al. (2006), we can show the implications of these ICC

values. Suppose we have three animals from the same species. The model we fitted implies the following marginal correlation structure for the Spanish data.

	1	2	3	4	5	6	7	8	9
1	1	0.96	0.96	0.05	0.05	0.05	0.05	0.05	0.05
2		1	0.96	0.05	0.05	0.05	0.05	0.05	0.05
3			1	0.05	0.05	0.05	0.05	0.05	0.05
4				1	0.96	0.96	0.05	0.05	0.05
5					1	0.96	0.05	0.05	0.05
6						1	0.05	0.05	0.05
7							1	0.96	0.96
8								1	0.96
9									1

The values 1–3 are teeth from the same animal. The correlation between these three observations is very high (0.96). The same holds for observations 4–6; these are also from the same, albeit a different, animal. These are also highly correlated. And the same holds for observations 7–9, which are all from a third animal. However, the correlation between two age observations from different animals, say 1 and 4, is low (0.05). The lower part of the correlation matrix is identical to the upper part.

20.4 Discussion

Some of the results displayed above are obvious from the nature of the data. We expect the three staining methods to give similar results on age for the same animal, and we would expect a fairly wide overlap in the ages of animals available for different species. However the overlap between age ranges for the different species is not complete, as for example, common dolphins live longer than harbour porpoises.

It is less obvious what the 'country' effect means (and why ages should be more variable in one country than another) as all the teeth were prepared and assessed by the same team. There was a different range of species among strandings in the two countries and although we restricted the analysis to the three most common species, their relative abundance differs between countries. So there could be some confounding of country and species effects. There may also have been differences in the effectiveness of staining due to different storage procedures for teeth used by the local sampling programmes (in Scotland, teeth are normally stored in alcohol, whereas the Spanish samples had been stored frozen).

However, one important effect we have ignored is that the variation in age readings probably depends on age: teeth of older animals are more difficult to interpret because the later incremental growth layers are closer together. Spain had a higher proportion of common dolphins in the sample (as compared to dominance of porpoises in the Scottish sample) meaning the Spanish sample was biased towards

older, larger animals. Thus, if we had included length (highly correlated with age but independent of the age measurements) as an explanatory variable, the country effect may have disappeared.

The objective of the original study was to compare the efficacy of several staining methods to prepare dolphin teeth used to determine age. The heterogeneity of the available teeth samples presented challenges that could not be easily overcome without the availability of mixed effects modelling. The availability of mixed effects modelling should not be considered a replacement for good sampling design, but it does offer a solution to problems created by opportunistic sampling, where these are the only data available.

20.5 What to Write in a Paper

We can be very short: The same as we suggested in Chapter 19. Present boxplots to emphasise the large between-animal variation, discuss the need for mixed effects modelling, explain the model selection approach and present the results, discuss the model validation, and explain what it means in terms of biology.