Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 636–645 doi: 10.1111/2041-210X.12577

SPECIAL FEATURE: 5TH ANNIVERSARY OF *METHODS IN ECOLOGY AND EVOLUTION*

A protocol for conducting and presenting results of regression-type analyses

Alain F. Zuur[1]* and Elena N. Ieno[2]

[1]*Highland Statistics Ltd., 9 St Clair Wynd, Newburgh AB41 6DZ, UK; and* [2]*Highland Statistics Ltd., Av Finlandia 11, Bwg 34, Santa Pola, Spain*

Summary

1. Scientific investigation is of value only insofar as relevant results are obtained and communicated, a task that requires organizing, evaluating, analysing and unambiguously communicating the significance of data. In this context, working with ecological data, reflecting the complexities and interactions of the natural world, can be a challenge. Recent innovations for statistical analysis of multifaceted interrelated data make obtaining more accurate and meaningful results possible, but key decisions of the analyses to use, and which components to present in a scientific paper or report, may be overwhelming.

2. We offer a 10-step protocol to streamline analysis of data that will enhance understanding of the data, the statistical models and the results, and optimize communication with the reader with respect to both the procedure and the outcomes. The protocol takes the investigator from study design and organization of data (formulating relevant questions, visualizing data collection, data exploration, identifying dependency), through conducting analysis (presenting, fitting and validating the model) and presenting output (numerically and visually), to extending the model via simulation. Each step includes procedures to clarify aspects of the data that affect statistical analysis, as well as guidelines for written presentation. Steps are illustrated with examples using data from the literature.

3. Following this protocol will reduce the organization, analysis and presentation of what may be an overwhelming information avalanche into sequential and, more to the point, manageable, steps. It provides guidelines for selecting optimal statistical tools to assess data relevance and significance, for choosing aspects of the analysis to include in a published report and for clearly communicating information.

Key-words: effective communication, protocol, statistical analysis, visualization

Introduction

The statistical analysis of ecological data often presents unique challenges. There may be outliers, correlation between covariates (collinearity), nonlinear relationships among variables, many zero observations, and spatial and temporal dependency structures. The past two decades have seen a surge in development of statistical techniques to analyse such complex data. Here, we present a 10-step protocol (Fig. 1) for the analysis and presentation of ecological data. This protocol will help the author to understand the data and select appropriate statistical models, verify all assumptions, check that the models employed are appropriate, interpret the output and present it in written form.

We focus here on regression-type models such as multiple linear regression, linear mixed modelling, generalized linear (mixed) modelling and generalized additive (mixed) modelling, although similar steps can be followed for analyses based on other statistical approaches. In this paper, we use the free software package R (R Core Team, 2014), but the protocol is not specifically linked to R.

Adherence to this protocol will provide the reader with insight into the process of the data collection and analysis, how the results were obtained and the validity of the modelling. It will facilitate clear communication to the reader and increase the credibility of the analysis and the potential for the work to be accepted for publication.

Step 1: State appropriate questions

In the Introduction to a paper or report, present the underlying biological questions.

Formulating the salient questions prior to gathering data is critical, as it influences the sampling design (Field & Hole 2003) and allows translation of the questions to a statistical model. Presentation of the questions immediately informs the reader of the purpose of the study.

As an example, we use data from Roulin & Bersier (2007), who investigated vocal behaviour of barn owl siblings. One of their primary questions was whether the vocal response of chicks differs depending on the sex of the parent providing

*Correspondence author. E-mail: highstat@highstat.com

Protocol for conducting and presenting results of regression-type analyses

1. **State appropriate questions**
2. **Visualize the experimental design**
3. **Conduct data exploration**
4. **Identify the dependency structure in the data**
5. **Present the statistical model**
6. **Fit the model**
7. **Validate the model**
8. **Interpret and present the numerical output of the model**
9. **Create a visual representation of the model**
10. **Simulate from the model**

Fig. 1. Protocol for statistical analysis of data and presenting results in a scientific paper.

food. They designed a study in which data were collected from 27 barn owl nests. Using microphones, they sampled the number of calls (sibling negotiation) that chicks produced during a short time period. Since food availability may influence how chicks respond via sibling negotiation to the parent bird, chicks in half the nests were provided with extra prey in the morning preceding recording, and prey remains were removed from the other nests (*food treatment*, satiated or deprived). Sampling was carried out between 21·30 and 05·30. The underlying question means that, in the Methods section, we can present a model that uses the three covariates *time*, *food treatment* and *sex of parent* as main terms. We can assess the importance of each of these covariates, taking into account the partial effects of all three.

We can formulate the biological question as 'Does the relationship between sibling negotiation and sex of the parent differ with food treatment, and does the effect of time on sibling negotiation differ with food treatment?' This question means that we need to apply a model that contains all the three main terms and relevant two-way interactions.

It is important to avoid using separate questions such as 'Is there an effect of sex of the parent?' 'Is there a time effect?' and 'Is there a food treatment effect?' This would require that three models be fitted, using a single covariate in each model. A potential problem with this approach is that the residuals of one model may show patterns when compared with the covariates not used in that model, which invalidates the model assumptions.

Step 2: Visualize the experimental design

It is essential that the sampling process be explained in such a way that the reader can immediately comprehend it.

Graphs may be more effective than text in presenting information on variables sampled and how the sampling was conducted (Field & Hole 2003).

Figure 2 illustrates where the barn owl sampling took place. It is apparent that we have data from a relatively large number of nests in Switzerland, all within 10–15 km of one another. It is possible to extend this to a graph showing which nests were sampled on a given date and the sampling time per nest each night (see Data accessibility for R code for such a graph). These data are readily illustrated in multipanel time-series plots.

Create as many graphs that are necessary to allow yourself to clearly visualize the experimental set-up, as this is closely related to Step 4 (dependency) of this protocol. When presenting results, a single graph is sufficient.

Step 3: Conduct data exploration

Data exploration is a crucial step in data analysis.

Data exploration provides insight into the data and helps familiarize the author with all aspects of his/her own data and, importantly, its limitations (Chatfield 1995; Quinn & Keough 2002; Zuur, Ieno & Elphick 2010; Ieno & Zuur 2015). The limitations have consequences for the choice of models to be applied. Zuur, Ieno & Elphick (2010) described a 10-step protocol for data exploration consisting of investigation of outliers, homogeneity, normality, zero trouble, collinearity, relationships, interactions and independence.

As an example, we use an oystercatcher feeding data set (Ieno & Zuur 2015). The aim of the study was to investigate how the relationship between the length of clams preyed upon by oystercatchers and feeding type (hammering or stabbing) differs with month (December and January) and feeding plot (three locations), which led to creation of a model containing a three-way interaction term. However, detailed data exploration showed that, for stabbers in location A in December, there were only two observations, both showing the same value (Fig. 3). If we had ignored this situation and applied a multiple linear regression model with all main terms and interaction terms, the significance of the three-way interaction term, and therefore the principal biological conclusion, would have been driven only by these two observations.

Expect to spend a considerable amount of time on data exploration. The Methods section can include the statement that 'Data exploration was carried out following the protocol described in Zuur, Ieno & Elphick (2010)'. The results of the data exploration can be summarized in the Results section in two or three lines of text. A graph, such as a multipanel scatterplot, depicting the data may be helpful.

Step 4: Identify the dependency structure in the data

It is rare to encounter a data set in which the observations of the response variable are independent.

Hurlbert (1984) discussed the lack of statistical independence of replicates (pseudoreplication) in ecological field experiments. Pseudoreplication in regression models results in

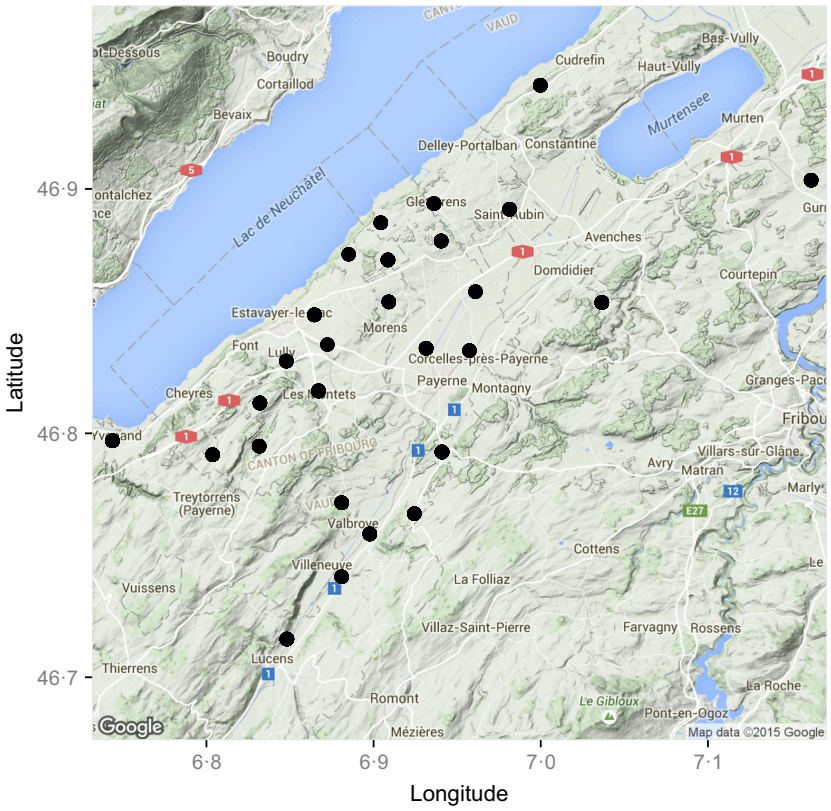

Fig. 2. Locations of the 27 nests for the owl data.

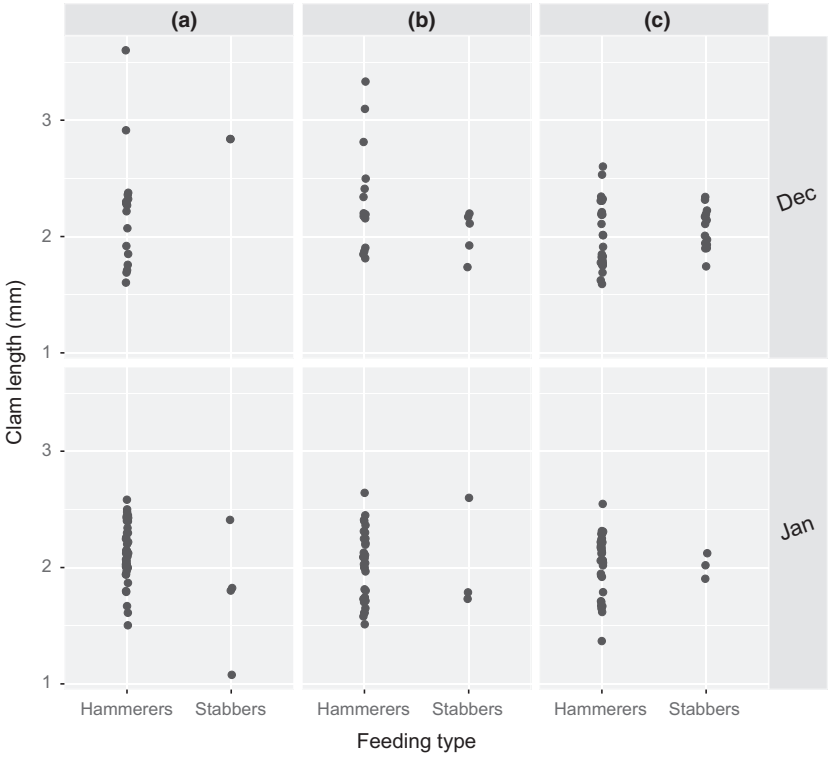

Fig. 3. Length of clams preyed upon by oystercatchers vs. month, feeding type and feeding plot. Note the two points with the same value in the top left panel. A small amount of random noise was added to each point to differentiate observations with the same value.

biased parameter estimates and increased type I errors. Pseudoreplication may be dealt with by applying a generalized linear mixed-effects model (GLMM) (Pinheiro & Bates 2000; Bolker 2008; Zuur *et al.* 2009; Zuur, Saveliev & Ieno 2012; Zuur, Hilbe & Ieno 2013). The GLMM entered the ecological data analysis toolbox in the past 10 years, and its execution is not routinely taught in many university statistics courses for ecologists.

Identifying the dependency structure of an ecological study is not always simple, as the following example shows. Sick *et al.* (2014a, b) investigated social strategies throughout the course of the day in chacma baboons *Papio ursinus*. Data were

collected from 60 baboons from two troops in the Tsaobis Leopard Park in Namibia. An individual baboon was followed for 1 h (*focal hour*), during which its grooming and dominance interactions were recorded, including the identity of the baboon being groomed (*receiver*). Multiple observations of the same baboon in a focal hour was the first source of dependency. During the 6-month sampling period, each baboon was repeatedly sampled, which is another source of dependency. To increase complexity, the receiver represents another level of dependency. Dealing with the pseudoreplication requires a mixed-effects model with a two-way nested and crossed random effect. The response variable was the difference in rank within the troop of the groomer and the receiver, represented as a value ranging from 0 to 1.

A flow chart may help clarify the dependency structure in the experimental design (Fig. 4). Readers may not be familiar with the purpose of, or need for, complex models; thus, the Methods section can include a short explanation of the applied statistical models and justify their use. State clearly which observations of the response variable are dependent, as they account for the components of the model that deal with dependent data. You can include a statement such as 'This data set consists of multiple observations of rank differences of a given baboon and receivers within a focal hour, along with multiple observations of a given receiver. We therefore applied a mixed-effects model with the random effect *focal hour* nested within the random effect *baboon* and a crossed random effect *receiver*'. You can present a figure describing the dependency structure (Fig. 4) or text alone.

A similar statement for the owl data could be 'We sampled each nest multiple times and therefore applied a GLMM in which *nest* is used as random intercept, as this models a dependency structure among sibling negotiation observations of the same nest'.

Step 5: Present the statistical model

Presenting a statistical model as mathematical notation facilitates comprehension of how the data fit into the analysis. Presenting it in a paper clarifies the process for the reader.

In order to ensure that results are replicable, it is crucial to clearly specify the covariates and how they were used (categorical or continuous), interactions, random effects, distribution of the response variable, etc. Recording the statistical models in mathematical terms during the analysis may avoid fitting ill-formulated models and, consequently, software warnings and error messages.

The equation can be presented in mathematical notation containing all regression parameters and indices or in words, as in eqn (1). The latter approach is easier on the eye when the covariates contain categorical covariates of more than two levels.

For the owl data, we can state 'To model the number of sibling calls as a function of the covariates, a Poisson GLMM with a log link function was used [Equation (1)]. The log link function ensures positive fitted values, and the Poisson distribution is typically used for count data. Fixed covariates are *sex of the parent* (categorical with two levels), *arrival time* (continuous) and *food treatment* (categorical with two levels). The interaction terms are *food treatment* × *sex of parent* and *food treatment* × *arrival time*. To incorporate the dependency among observations of the same nest, we used *nest* as random intercept.

$$\begin{aligned} \mathrm{NCalls}_{ij} &\sim \mathrm{Poisson}(\mu_{ij}) \\ \mathrm{E}(\mathrm{NCalls}_{ij}) &= \mu_{ij} \\ \log(\mu_{ij}) &= \mathrm{SexParent}_{ij} + \mathrm{FoodTreatment}_{ij} \\ &\quad + \mathrm{ArrivalTime}_{ij} + \mathrm{SexParent}_{ij} \times \mathrm{FoodTreatment}_{ij} \\ &\quad + \mathrm{SexParent}_{ij} \times \mathrm{ArrivalTime}_{ij} + \mathrm{Nest}_i \\ \mathrm{Nest}_i &\sim N(0, \sigma^2) \end{aligned} \quad \text{(eqn 1)}$$

where NCalls_{ij} is the jth observation in nest i, and $i = 1,\ldots, 27$, and Nest_i is the random intercept, which is assumed to be normally distributed with mean 0 and variance σ^2.

The equation, along with the introductory paragraph, clearly shows the distribution used for the response variable, the covariates included, whether they are categorical or continuous and why an advanced model like a GLMM is being used to incorporate a dependency structure via the use of random intercepts.

The goal of the baboon study was to learn how social relationships change over short time-scales, and one specific question was whether subordinate groomers are more likely to

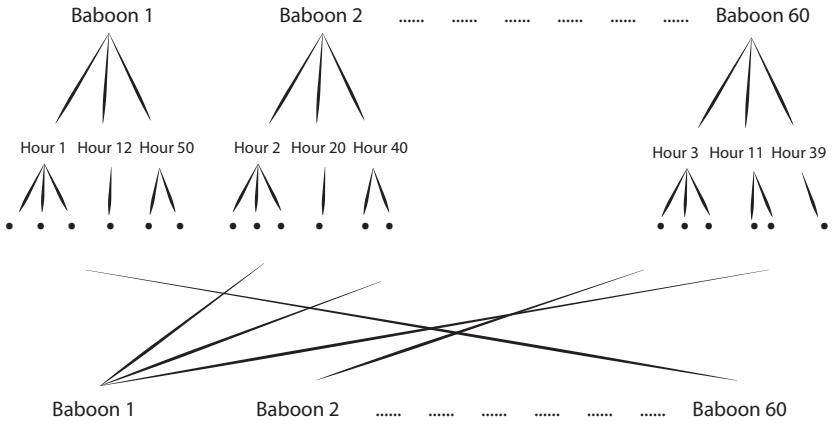

Fig. 4. Nested structure for the baboon data. Top row: A baboon is randomly selected. During a focal hour (second row), multiple observations are made. Dots represent the observations. The bottom row shows that receivers are repeatedly sampled.

groom dominate animals earlier in the day. This behaviour might be affected by the *relatedness* of groomer and receiver as well as *group size*. The Methods section can include 'We modelled the *rank difference* of groomer and receiver as a function of *relatedness* of groomer and receiver, *group size* (small versus large), *time* (of day), and the interactions *time* × *relatedness* and *group size* × *relatedness*. The response variable (rank difference) was coded as a value from 0 to 1; therefore, a beta distribution with a logistic link function was used'.

For the benefit of readers unfamiliar with beta distribution, include a sentence stating that 'To ensure that the fitted values range from 0 to 1, we cannot use a Gaussian linear mixed-effects model and instead apply a model with a beta distribution, which can be used if the response variable is a continuous variable ranging from x_1 and x_2, with a logistic link function'. The model specification will be of value to readers not familiar with this type of model. The initial model is of the form

$$\begin{aligned}
\mathrm{RD}_{ijk} &\sim \mathrm{Beta}(\pi_{ijk}) \\
\mathrm{E}(\mathrm{RD}_{ijk}) &= \pi_{ijk} \\
\mathrm{var}(\mathrm{RD})_{ijk} &= \pi_{ijk} \times (1-\pi_{ijk})/(1+\theta) \\
\mathrm{logit}(\pi_{ijk}) &= \mathrm{Time}_{ijk} + \mathrm{Relatedness}_{ijk} + \mathrm{GroupSize}_{ijk} \\
&\quad + \mathrm{Time}_{ijk} \times \mathrm{Relatedness}_{ijk} \\
&\quad + \mathrm{Relatedness}_{ijk} \times \mathrm{GroupSize}_{ijk} + \mathrm{Groomer}_i \text{'} \\
&\quad + \mathrm{Hour}_j + \mathrm{Receiver}_l \\
\mathrm{Groomer}_i &\sim N(0, \sigma^2_{\mathrm{Groomer}}) \\
\mathrm{Hour}_j &\sim N(0, \sigma^2_{\mathrm{Hour}}) \\
\mathrm{Receiver}_l &\sim N(0, \sigma^2_{\mathrm{Receiver}})
\end{aligned} \qquad \text{(eqn 2)}$$

where RD_{ijk} is the kth observed rank difference between groomer i and receiver l in hour j, and θ is an unknown parameter controlling the variance.

In eqn (2), we can see which terms are used as random intercepts, which covariates are used as main terms and the interactions included. Writing this equation out during the analysis clarifies the function of each component in the model.

Step 6: Fit the model

Fitting the model should be the straightforward part of the analysis.

One of the most popular statistical software packages available is R. R software code for a wide variety of statistical techniques can be found in Venables & Ripley (2002), Dalgaard (2002), Crawley (2012) and in textbooks containing detailed R code for more specific topics, for instance Pinheiro & Bates (2000), Bolker (2008), and Zuur *et al.* (2009), Zuur, Hilbe & Ieno (2013) for mixed modelling; and Wood (2006) and Zuur (2012), Zuur, Saveliev & Ieno 2014) for GAM and generalized additive mixed-effects model (GAMM).

In the Methods section, mention the software and R packages that you used. The *citation( )* command in R shows how to do this. For all R packages used, include a sentence such as 'the package lme4 (Bates *et al.* 2014) in the software R (R Core Team 2014) was used to fit the model in Equation (1)'.

Recent years have seen a rise in applied statistics textbooks that use Bayesian analysis (McCarthy 2007; Kéry 2010; Lunn *et al.* 2012; Zuur, Hilbe & Ieno 2013; Zuur, Saveliev & Ieno 2014; Korner-Nievergelt *et al.* 2015), and, while this technique will become widespread in future ecology studies, it may be some time before readers are familiar with it. The use of Bayesian methods such as Markov chain Monte Carlo (MCMC) techniques may require a short one-line explanation in a paper. 'A Bayesian analysis framework with MCMC was adopted to fit the model in Equation (2). MCMC is essentially a simulation technique to obtain the distribution of each parameter in a model', may clarify the approach.

When reporting MCMC analysis, include a statement in the Methods section such as 'To fit the model in Equation (2), MCMC was applied using JAGS (Plummer 2003) via the package R2jags (Su & Yajima 2012) in R (R Core Team 2014). We used a burn-in of 50 000 iterations, three chains, a thinning rate of 10 and 15 000 iterations for each posterior distribution. Diffuse normal priors were used for the regression parameters and diffuse uniform priors for the standard deviation parameters'. In the Results section, state 'Mixing of the chains was good'. Reviewers and readers may not be familiar with Bayesian concepts such as mixing, chains. You may need to explain in a few sentences the workings of MCMC and the meaning of thinning, mixing, chains and burn-in.

Step 7: Validate the model

Model validation confirms that the model complies with underlying assumptions.

All regression-type techniques are based on a series of assumptions, those of prime importance being independence and absence of residual patterns. Violation of these assumptions may result in biased parameter estimates and type I errors (Quinn & Keough 2002). To validate the fitted models, standardized or Pearson residuals must be plotted against fitted values, against each covariate in the model, against each covariate not in the model and against time and space coordinates, if relevant (Zuur *et al.* 2009). If the data include temporal or spatial aspects, use autocorrelation functions and/or variograms to assess independence of residuals (Schabenberger & Pierce 2002).

The Methods section can include 'Model assumptions were verified by plotting residuals versus fitted values, versus each covariate in the model and versus each covariate not in the model. We assessed the residuals for temporal and spatial dependency'. In the Results section, state 'Model validation indicated no problems', assuming that this was the case. It may be an option to include one or two graphs in an online Supporting information. Be prepared to provide these graphs if a reviewer asks for them.

We present an example of how model validation detected problems in the Poisson GLMM applied to the owl data. When conducting model validation for the Poisson GLMM, we calculated a dispersion statistic as 5·43, indicating overdispersion. It is important to find the source of the overdispersion and adjust the model accordingly (Hilbe 2011), or the

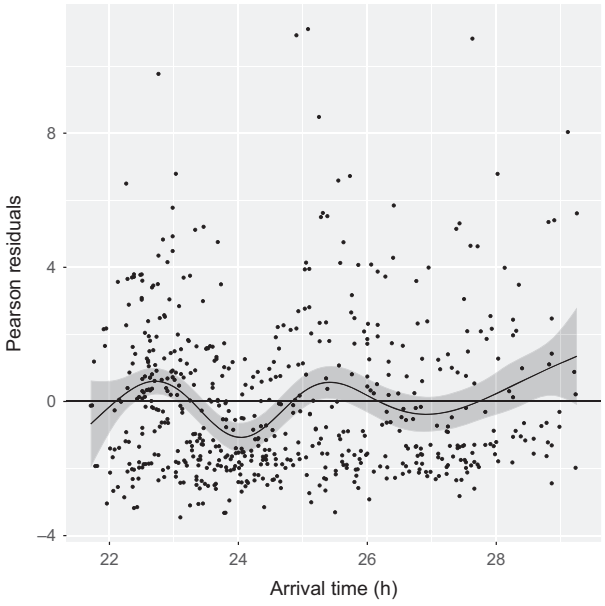

Fig. 5. Pearson residuals vs. *arrival time* for the Poisson GLMM applied to the owl data. A thin plate regression spline smoother with 95% confidence intervals was added to aid visual interpretation. The smoother was fitted using the mgcv package (Wood 2006) and explains 7% of the variation in the Pearson residuals.

estimated parameters may be biased and standard errors too small. To identify the cause of overdispersion, we plotted Pearson residuals vs. *arrival time* (Fig. 5). The graph showed a clear pattern, indicating that we may need to allow for a nonlinear *arrival time* effect. Other factors contributing to the overdispersion may have been a missing covariate, the relatively large number of zero observations (25%) or dependency within (or between) nests.

Step 8: Interpret and present the numerical output of the model

Grasping the biological relevance of the numerical output of a statistical model may be a challenge for readers.

R software programs tend to produce an avalanche of information when fitting a model (Crawley 2012). Identifying important factors and translating output of a statistical model into meaningful biological information can be a complex task, especially if the model contains interaction terms, in which case intercepts and slopes change depending on the values of the interaction terms (Quinn & Keough 2002). It is important to be aware that corrections to intercepts and slopes are a source of confusion for some readers and may differ with software and software settings.

Your first task is to determine which numerical information to present in a paper. For multiple linear regression models, provide a table with the estimated parameters, standard errors, *t*-values, R^2 and the estimated variance. The same can be done with mixed-effects models; however, you must include multiple variances. If possible, calculate an intraclass correlation (Zuur *et al.* 2009), for which you will need to determine R^2 for mixed models (Nakagawa & Schielzeth 2012). Similar information can be presented for GLM and GLMMs. For GAM and GAMMs, you must include the effective degrees of freedom of the smoothers and indicate the type of smoothers used (Wood 2006). The specific format of tables and figures will depend on the guidelines of the journal to which you are submitting your paper.

There is an ongoing debate in the literature (Halsey *et al.* 2015) over whether *P*-values should be included. Their interpretation is prone to abuse, and, for most of the frequentist techniques mentioned, *P*-values are approximate at best. They must be interpreted with care, and this should be emphasized in the paper. An alternative is to present 95% confidence intervals for the regression parameters and effect size estimates and their precision (Halsey *et al.* 2015).

Next, we discuss interpreting and communicating the numerical output. Ignoring for the moment its model validation problems, the numerical output of the Poisson GLMM applied on the owl data is given in Table 1. There are two categorical covariates, each of two levels; therefore, we have four equations. It may help to write them out:

$$\log(\mu_{ij}) = \begin{cases} 5{\cdot}169 - 0{\cdot}129 \times \text{ArrivalTime}_{ij} & \text{deprived and female} \\ 5{\cdot}160 - 0{\cdot}129 \times \text{ArrivalTime}_{ij} & \text{deprived and male} \\ 4{\cdot}515 - 0{\cdot}129 \times \text{ArrivalTime}_{ij} & \text{satiated and female} \\ 4{\cdot}635 - 0{\cdot}130 \times \text{ArrivalTime}_{ij} & \text{satiated and male} \end{cases} \quad \text{(eqn 3)}$$

The numbers in the equations are obtained by modifying the intercept and the slope based on the values of the categorical covariates *food treatment* and *sex of parent* (Quinn & Keough 2002). For example, the value 5·160 was obtained by subtracting 0·009 from the intercept (Table 1). This is the correction of the intercept for an observation of deprived treatment and male parent. These equations can be presented in the Results section as text or superimposed on a graph that illustrates the modelling results (Step 9). However, the overdispersion and nonlinear pattern in the Pearson residuals (Fig. 5) invalidate the results in Table 1.

The MCMC output for the baboon model is presented in eqn (2). Table 2 shows the posterior mean values, posterior standard errors and 95% credible intervals. Provide a conceptual explanation of these terms. The fitted model for a large group size is

$$\text{logit}(\pi_{ijk}) = 0{\cdot}289 - 0{\cdot}052 \times \text{Time}_{ijk} - 0{\cdot}036 \times \text{Relatedness}_{ijk} + 0{\cdot}039 \times \text{Time}_{ijk} \times \text{Relatedness}_{ijk} \quad \text{(eqn 4)}$$

For a small group, the intercept and one slope needs correction, resulting in

$$\text{logit}(\pi_{ijk}) = 0{\cdot}021 - 0{\cdot}052 \times \text{Time}_{ijk} + 0{\cdot}214 \times \text{Relatedness}_{ijk} + 0{\cdot}039 \times \text{Time}_{ijk} \times \text{Relatedness}_{ijk} \quad \text{(eqn 5)}$$

Even for statisticians, interpreting this model is a brain-teaser, and for that reason, we introduce Step 9.

Table 1. Estimated regression parameters, standard errors, *z*-values and *P*-values for the Poisson GLMM presented in eqn (1). The estimated value for σ_{Nest} is 0·484.

	Estimate	Std. error	*z* value	*P*-value
Intercept	5·169	0·292	17·665	<0·05
FoodTreatmentSatiated	−0·654	0·468	−1·395	0·162
ArrivalTime	−0·129	0·011	−11·472	<0·05
SexParentMale	−0·009	0·045	−0·208	0·834
FoodTreatmentSatiated : SexParentMale	0·129	0·070	1·842	0·065
FoodTreatmentSatiated : ArrivalTime	−0·000	0·019	−0·026	0·979

Table 2. Posterior mean values, standard errors and 95% credible intervals for the parameters in the beta model presented in eqn (2).

	Mean	Std. error	2·5%	97·5%
Intercept	0·289	0·138	0·014	0·560
Time.std	−0·052	0·023	−0·098	−0·007
Relatedness.std	−0·036	0·040	−0·115	0·042
GroupSizesmall	−0·267	0·213	−0·685	0·157
Time.std : Relatedness.std	0·039	0·021	−0·002	0·081
Relatedness.std : GroupSizesmall	0·250	0·053	0·147	0·354
sigma.gr	0·554	0·069	0·433	0·704
sigma.fh	0·328	0·043	0·238	0·408
sigma.rc	0·538	0·061	0·431	0·671
theta	13·289	0·877	11·630	15·069

Step 9: Create a visual representation of the model

To aid the interpretation of statistical results, create a graphic depiction of the fitted model.

In the previous step, we showed that interpreting the biological implications of the model for the baboon data is a major challenge. Sketching the fitted values will aid in comprehension; graphs may be more effective at imparting information than are numbers and equations. For models with multiple covariates, graphing requires some R coding skills (Sarkar 2008; Wickham 2009). Strive to avoid 'chartjunk', a phrase introduced by Tufte (1983) referring to redundant information in a graph. See Rougier, Droettboom & Bourne (2014) for a detailed discussion on improving graphs and Gelman, Pasarica & Dodhia (2002), who make a strong case for transforming tables into graphs.

The Poisson GLMM model for the owl data was overdispersed, and the residuals contained nonlinear patterns. The model can be improved by application of a GAMM allowing for a nonlinear *arrival time* effect, fitting a zero-inflated model or employing a model with a more complex dependency structure. To choose the appropriate approach, plot the model fit of the Poisson GLMM (Fig. 6). This graph, together with the model validation graphs, can provide direction on how to proceed with the analysis. If the Poisson GLMM were the optimal and final model, a graph such as Fig. 6 could be included in the paper, but, in this example, the GAMM is required.

The model for the baboon data contains an interaction between a continuous and a categorical covariate, as well as between two continuous covariates. Despite writing the equations for small and large groups, it is difficult to interpret the model and to determine the contribution of each component. Graphing the model fit will aid in comprehension. Figure 7 shows two planes, one representing the small group and for the other representing the large group. The small group exhibited a positive effect of *relatedness* on rank differences early in the day and a negative effect later in the day, as well as a negative *time* effect for low *relatedness* values and a positive *time* effect for higher *relatedness* values. The large group shows a negative *relatedness* effect early in the day.

Step 10: Simulate from the model

When simulating data from the model, the simulated data should be comparable to the original data. If not, the model needs improvement.

After you have fitted a model, you can simulate a large number of data sets from the model to verify that the model complies with the data (Gelman *et al.* 2014), to obtain predictions under specific covariate conditions for environmental management purposes (Zuur, Hilbe & Ieno 2013) and to better recognize the modelled data's implication for real-world situations.

We simulated 10 000 data sets from the Poisson GLMM for the owl data. For each simulated data set, we calculated the percentage of zeros. The majority of simulated data sets contained 3–4% zeros, and several had 8% (Fig. 8). The percentage of zeros in the observed data was 26%, indicating that the Poisson GLMM in eqn (1) cannot cope with the relatively large number of zeros. Similar simulation steps can be carried out for the dispersion statistic, minimum value, maximum value, etc. You can also calculate a frequency table for each simulated data set and use these to create an average frequency table to compare with the frequency table of the observed data.

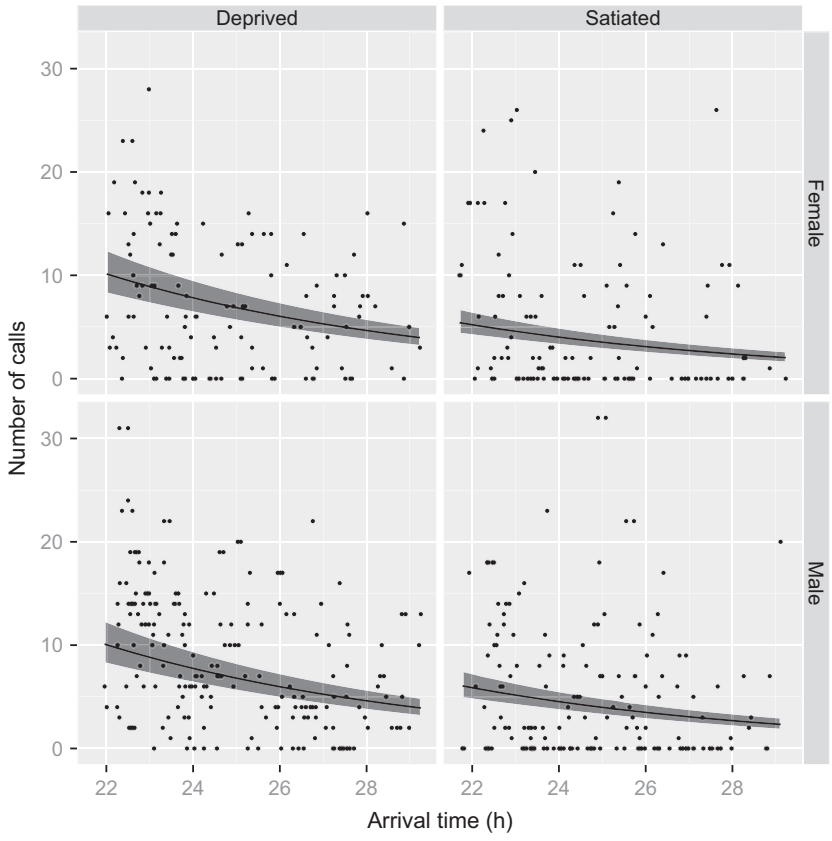

Fig. 6. Fit of the Poisson GLMM in eqn (1) for the owl data.

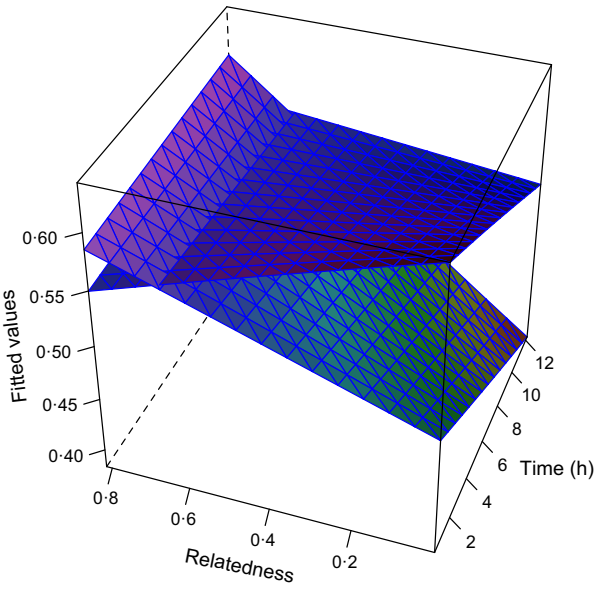

Fig. 7. Fit of the beta model in eqn (2) for the baboon data. The separate planes represent the small (lower right) and large group (top right).

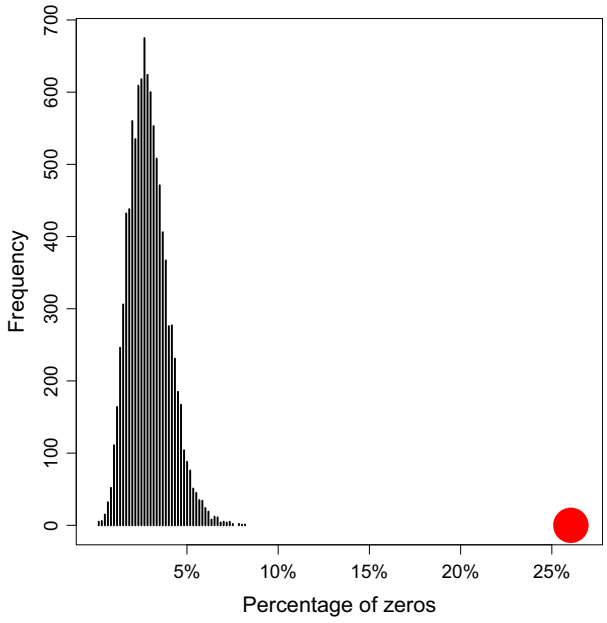

Fig. 8. 10 000 data sets were simulated from the Poisson GLMM. The number of zeros was counted for each simulated data set. The graph shows frequency of simulated data sets with 0, 1, 2,. . ., zeros. The majority of simulated data sets contained 2–4% zeros. The red dot represents observed data, which contained 26% zeros.

This should indicate whether the model complies with the data.

Instead of simulating data from the model, you can use 90% of the data for fitting the model and the remaining 10% to assess how well the model performs for prediction. This is called cross-validation, and a wide range of procedures for its execution is available. Harrel (2001) discusses various options for cross-validation as part of the model validation process. In MCMC literature, simulating data from a model is well described; see Lawson (2013) or Gelman *et al.* (2014).

It is also possible to use simulation techniques to investigate 'what if' scenarios. Zuur, Hilbe & Ieno (2013) analysed zero-inflated and spatially correlated common scoter *Melanitta nigra* data. The purpose of the research was to investigate whether measures taken to compensate for the loss of sea habitat resulting from the creation of a large port extension in Rotterdam were effective in safeguarding populations of common scoters in the south-west coastal zone of the Netherlands. One such measure was to create sanctuary zones free of disturbance, primarily caused by sea vessels, along the studied coastal area. Different scenarios of vessel disturbance were fed into the fitted model. Results of such simulations provide guidelines for policy-makers and will be the core of published results.

Simulation of data should be part of the model validation process. In a paper, mention the results in one or two lines.

Discussion

Ecology is a developing scientific field that creates increasingly complex data sets obtained through expanding sampling techniques, the analysis of which requires sophisticated statistical approaches. Whereas 20 years ago one could use descriptive techniques, such as non-metric multidimensional scaling, to present information, methods providing more complete and usable information are currently available, but may be challenging to employ. We have presented a common sense protocol that comprises a series of essential steps in analysing data and presenting results in written form, to facilitate clarity in communication of results.

The idea of using a step-by-step approach when reporting the results of a statistical analysis is not new; see for example Fukuda & Ohashi (1997) and references therein. A wide array of information and recommendations for presenting the numerical output of statistical models is available online. A scientific paper reflects a process that proceeds from formulating basic questions, sampling and analysing data, to visualizing results and drawing conclusions. The 10-step protocol presented here can be used as a guide for data analysis and for presenting the statistical aspects of a study. More general information how to write a paper, design tables, use appropriate font sizes for legends and labels, references, titles, abstract content, etc., can be found in, for example, Gustavii (2003).

When presenting the data in the Results section, limit the amount of non-essential information. A barrage of facts such as the mean, median, standard deviation and other summary statistics of the response variable and covariates is often difficult to interpret, as well as superfluous when the results of more sophisticated techniques to estimate mean values and covariate effects are presented.

Finally, take care to ensure that the biological conclusions are consistent with the results of the statistical analyses. Use appropriate citations. Make the raw data and R code used freely available via websites like DRYAD (http://www.datadryad.org/). This allows other researchers to verify, replicate, reuse and extend your analyses.

Acknowledgements

We thank Dr Aaron MacNeil and an anonymous reviewer for comments on an earlier draft. We thank Alexandre Roulin for providing the owl data and Peter van Horssen for providing R code to convert the Swiss spatial coordinates to WGS84 format. We thank The Lucidus Consultancy for English language editing and comments.

Data accessibility

The baboon data are available from Dryad (http://dx.doi.org/10.5061/dryad.n4k6p). The owl data, oystercatcher data, baboon data (with modified variable names) and the R code are deposited in Dryad repository: http://datadryad.org/resource/doi:10.5061/dryad.v4t42 (Zuur & Ieno 2016).

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014) *lme4: Linear Mixed-Effects Models using Eigen and S4*. R package version 1.1-7. http://CRAN.R-project.org/package = lme4.

Bolker, M.B. (2008) *Ecological Models and Data in R*. Princeton University Press, Woodstock.

Chatfield, C. (1995) *Problem Solving: A Statistician's Guide*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Crawley, M.J. (2012) *The R Book*. John Wiley & Son, West Sussex.

Dalgaard, P. (2002) *Introductory Statistics with R*. Springer, New York.

Field, A. & Hole, G.J. (2003) *How to Design and Report Experiments*. Sage publications Ltd., London.

Fukuda, H. & Ohashi, Y. (1997) A guideline for reporting results of statistical analysis in Japanese Journal of Clinical Oncology. *Japanese Journal of Clinical Oncology*, **27**, 121–127.

Gelman, A., Pasarica, C. & Dodhia, R. (2002) Let's practice what we preach: turning tables into graphs. *The American Statistician*, **56**, 121–130.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2014) *Bayesian Data Analysis*, 3rd edn. CRC Press, Boca Raton, FL.

Gustavii, B. (2003) *How to Write and Illustrate a Scientific Paper*. Cambridge University Press, Cambridge.

Halsey, L.G., Curran-Everett, D., Vowler, S.L. & Drummond, G.B. (2015) The fickle P value generates irreproducible results. *Nature Methods*, **12**, 179–185.

Harrel, F.E. (2001) *Regression Modelling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

Hilbe, J.M. (2011) *Negative Binomial Regression*. Cambridge University Press, Cambridge.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

Ieno, E.N. & Zuur, A.F. (2015) *A Beginner's Guide to Data Exploration and Visualisation with R*. Highland Statistics, Newburgh.

Kéry, M. (2010) *Introduction to WinBUGS for Ecologists: Bayesian Approach to Regression, ANOVA, Mixed Models and Related Analyses*. Elsevier, San Diego.

Korner-Nievergelt, F., Roth, F., von Felten, S., Guélat, J., Almasi, B. & Korner-Nievergelt, P. (2015) *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan: Including Comparisons to Frequentist Statistics*. Elsevier, Amsterdam.

Lawson, A.B. (2013) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, 2nd edn. CRC, Boca Raton, FL.

Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2012) *The BUGS Book: A Practical Introduction to Bayesian Analysis*. CRC Press, Boca Raton, FL.

McCarthy, M.A. (2007) *Bayesian Methods for Ecology*. Cambridge University Press, Cambridge.

Nakagawa, S. & Schielzeth, H. (2012) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142.

Pinheiro, J. & Bates, D. (2000) *Mixed Effects Models in S and S-Plus*. Springer, New York.

Plummer, M. (2003) JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.
R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Rougier, N.P., Droettboom, M. & Bourne, P.E. (2014) Ten simple rules for better figures. *PLOS Computational Biology*, **10**, 1–7.
Roulin, A. & Bersier, L. (2007) Nestling barn owls beg more intensely in the presence of their mother than in the presence of their father. *Animal Behaviour*, **74**, 1099–1106.
Sarkar, D. (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York.
Schabenberger, O. & Pierce, F.J. (2002) *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press, Boca Raton, FL.
Sick, C., Carter, A.J., Marshall, H.H., Knapp, L.A., Dabelsteen, T. & Cowlishaw, G. (2014a) Evidence for varying social strategies across the day in chacma baboons. *Biology Letters*, **10**, 0140249.
Sick, C., Carter, A.J., Marshall, H.H., Knapp, L.A., Dabelsteen, T. & Cowlishaw, G. (2014b) Data from: Evidence for varying social strategies across the day in chacma baboons. *Dryad Digital Repository*. http://dx.doi.org/10.5061/dryad.n4k6p.
Su, Y.S. & Yajima, M. (2012) *R2jags: A Package for Running Jags from R*. R package version 0.03-08. http://CRAN.R-project.org/package = R2jags.
Tufte, E.G. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.
Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York.
Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.
Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, London.
Zuur, A.F. (2012) *A Beginner's Guide to Generalized Additive Models with R*. Highland Statistics Ltd, Newburgh.
Zuur, A.F., Hilbe, J.M. & Ieno, E.N. (2013) *Beginner's Guide to GLM and GLMM with R*. Highland Statistics Ltd, Newburgh.
Zuur, A.F. & Ieno, E.N. (2016) Data from: A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*. http://dx.doi.org/10.5061/dryad.v4t42.
Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods Ecology and Evolution*, **1**, 3–14.
Zuur, A.F., Saveliev, A.A. & Ieno, E.N. (2012) *Zero Inflated Models and Generalized Linear Mixed Models with R*. Highland Statistics Ltd, Newburgh.
Zuur, A.F., Saveliev, A.A. & Ieno, E.N. (2014) *A Beginner's Guide to Generalized Additive Mixed Models with R*. Highland Statistics Ltd, Newburgh.
Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology*. Springer, New York.

Received 31 January 2016; accepted 29 March 2016
Handling Editor: Robert Freckleton