 doi: 10.1111/j.2041-210X.2009.00001.x

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur*[1,2], Elena N. Ieno[1,2] and Chris S. Elphick[3]

[1]*Highland Statistics Ltd, Newburgh, UK;* [2]*Oceanlab, University of Aberdeen, Newburgh, UK; and* [3]*Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA*

Summary

1. While teaching statistics to ecologists, the lead authors of this paper have noticed common statistical problems. If a random sample of their work (including scientific papers) produced before doing these courses were selected, half would probably contain violations of the underlying assumptions of the statistical techniques employed.

2. Some violations have little impact on the results or ecological conclusions; yet others increase type I or type II errors, potentially resulting in wrong ecological conclusions. Most of these violations can be avoided by applying better data exploration. These problems are especially troublesome in applied ecology, where management and policy decisions are often at stake.

3. Here, we provide a protocol for data exploration; discuss current tools to detect outliers, heterogeneity of variance, collinearity, dependence of observations, problems with interactions, double zeros in multivariate analysis, zero inflation in generalized linear modelling, and the correct type of relationships between dependent and independent variables; and provide advice on how to address these problems when they arise. We also address misconceptions about normality, and provide advice on data transformations.

4. Data exploration avoids type I and type II errors, among other problems, thereby reducing the chance of making wrong ecological conclusions and poor recommendations. It is therefore essential for good quality management and policy based on statistical analyses.

Key-words: collinearity, data exploration, independence, transformations, type I and II errors, zero inflation

Introduction

The last three decades have seen an enormous expansion of the statistical tools available to applied ecologists. A short list of available techniques includes linear regression, generalized linear (mixed) modelling, generalized additive (mixed) modelling, regression and classification trees, survival analysis, neural networks, multivariate analysis with all its many methods such as principal component analysis (PCA), canonical correspondence analysis (CCA), (non-)metric multidimensional scaling (NMDS), various time series and spatial techniques, etc. Although some of these techniques have been around for some time, the development of fast computers and freely available software such as R (R Development Core Team 2009) makes it possible to routinely apply sophisticated statistical techniques on any type of data. This paper is not about these methods. Instead, it is about the vital step that should, but frequently does not, precede their application.

All statistical techniques have in common the problem of 'rubbish in, rubbish out'. In some methods, for example, a single outlier may determine the final results and conclusions. Heterogeneity (differences in variation) may cause serious trouble in linear regression and analysis of variance models (Fox 2008), and with certain multivariate methods (Huberty 1994).

When the underlying question is to determine which covariates are driving a system, then the most difficult aspect of the analysis is probably how to deal with collinearity (correlation between covariates), which increases type II errors (i.e. failure to reject the null hypothesis when it is untrue). In multivariate analysis applied to data on ecological communities, the presence of double zeros (e.g. two species being jointly absent at various sites) contributes towards similarity in some techniques (e.g. PCA), but not others. Yet other multivariate techniques are sensitive to species with clumped distributions and low abundance (e.g. CCA). In univariate analysis techniques like generalized linear modelling (GLM) for count data, zero inflation of the response variable may cause biased parameter estimates (Cameron & Trivedi 1998). When multivariate techniques use permutation methods to obtain *P*-values, for example in CCA and redundancy analysis (RDA, ter Braak & Verdonschot 1995), or the Mantel test (Legendre & Legendre

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

1998), temporal or spatial correlation between observations can increase type I errors (rejecting the null hypothesis when it is true).

The same holds with regression-type techniques applied on temporally or spatially correlated observations. One of the most used, and misused, techniques is without doubt linear regression. Often, this technique is associated with linear patterns and normality; both concepts are often misunderstood. Linear regression is more than capable of fitting nonlinear relationships, e.g. by using interactions or quadratic terms (Montgomery & Peck 1992). The term 'linear' in linear regression refers to the way parameters are used in the model and not to the type of relationships that are modelled. Knowing whether we have linear or nonlinear patterns between response and explanatory variables is crucial for how we apply linear regression and related techniques. We also need to know whether the data are balanced before including interactions. For example, Zuur, Ieno & Smith (2007) used the covariates sex, location and month to model the gonadosomatic index (the weight of the gonads relative to total body weight) of squid. However, both sexes were not measured at every location in each month due to unbalanced sampling. In fact, the data were so unbalanced that it made more sense to analyse only a subset of the data, and refrain from including certain interactions.

With this wealth of potential pitfalls, ensuring that the scientist does not discover a false covariate effect (type I error), wrongly dismiss a model with a particular covariate (type II error) or produce results determined by only a few influential observations, requires that detailed data exploration be applied before any statistical analysis. The aim of this paper is to provide a protocol for data exploration that identifies potential problems (Fig. 1). In our experience, data exploration can take up to 50% of the time spent on analysis.

Although data exploration is an important part of any analysis, it is important that it be clearly separated from hypothesis testing. Decisions about what models to test should be made *a priori* based on the researcher's biological understanding of the system (Burnham & Anderson 2002). When that understanding is very limited, data exploration can be used as a hypothesis-generating exercise, but this is fundamentally different from the process that we advocate in this paper. Using aspects of a data exploration to search out patterns ('data dredging') can provide guidance for future work, but the results should be viewed very cautiously and inferences about the broader population avoided. Instead, new data should be collected based on the hypotheses generated and independent tests conducted. When data exploration is used in this manner, both the process used and the limitations of any inferences should be clearly stated.

Throughout the paper we focus on the use of graphical tools (Chatfield 1998; Gelman, Pasarica & Dodhia 2002), but in some cases it is also possible to apply tests for normality or homogeneity. The statistical literature, however, warns against certain tests and advocates graphical tools (Montgomery & Peck 1992; Draper & Smith 1998, Quinn & Keough 2002). Läärä (2009) gives seven reasons for not applying preliminary tests for normality, including: most statistical techniques based on normality are robust against violation; for larger data sets the central limit theory implies approximate normality; for small samples the power of the tests is low; and for larger data sets the tests are sensitive to small deviations (contradicting the central limit theory).

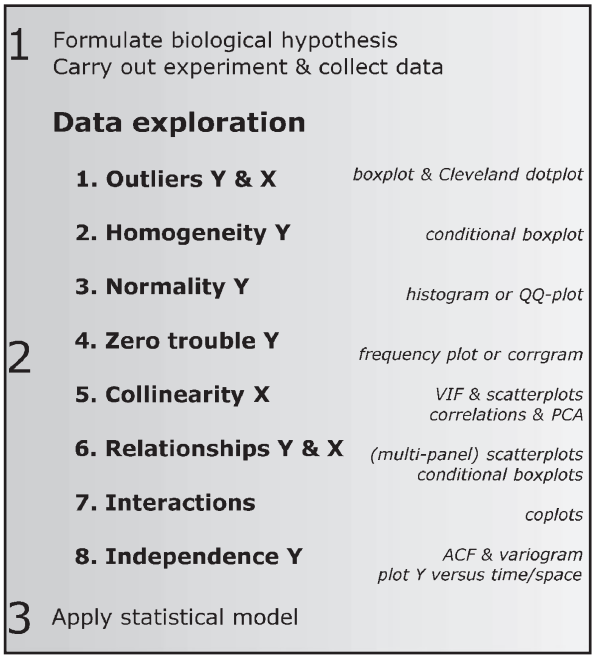

Fig. 1. Protocol for data exploration.

All graphs were produced using the software package R (R Development Core Team 2008). All R code and data used in this paper are available in Appendix S1 (Supporting Information) and from http://www.highstat.com.

Step 1: Are there outliers in Y and X?

In some statistical techniques the results are dominated by outliers; other techniques treat them like any other value. For example, outliers may cause overdispersion in a Poisson GLM or binomial GLM when the outcome is not binary (Hilbe 2007). In contrast, in NMDS using the Jaccard index (Legendre & Legendre 1998), observations are essentially viewed as presences and absences, hence an outlier does not influence the outcome of the analysis in any special way. Consequently, it is important that the researcher understands how a particular technique responds to the presence of outliers. For the moment, we define an outlier as an observation that has a relatively large or small value compared to the majority of observations.

A graphical tool that is typically used for outlier detection is the boxplot. It visualizes the median and the spread of the data. Depending on the software used, the median is typically presented as a horizontal line with the 25% and 75% quartiles forming a box around the median that contains half of the observations. Lines are then drawn from the boxes, and any points beyond these lines are labelled as outliers. Some researchers routinely (but wrongly) remove these observations.

Figure 2a shows an example of such a graph using 1295 observations of a morphometric variable (wing length of the saltmarsh sparrow *Ammodramus caudacutus*; Gjerdrum, Elphick & Rubega 2008). The graph leads one to believe (perhaps wrongly, as we will see in a moment) that there are seven outliers.

Another very useful, but highly neglected, graphical tool to visualize outliers is the Cleveland dotplot (Cleveland 1993). This is a graph in which the row number of an observation is plotted vs. the observation value, thereby providing much more detailed information than a boxplot. Points that stick out on the right-hand side, or on the left-hand side, are observed values that are considerable larger, or smaller, than the majority of the observations, and require further investigation. If such observations exist, it is important to check the raw data for errors and assess whether the observed values are reasonable. Figure 2b shows a Cleveland dotplot for the sparrow wing length data; note that the observations identified by the boxplot are not especially extreme after all. The 'upward' trend in Fig. 2b simply arises because the data in the spreadsheet were sorted by weight. There is one observation of a wing length of about 68 mm that stands out to the left about half way up the graph. This value is not considerably larger than the other values, so we cannot say yet that it is an outlier.

Figure 3 shows a multi-panel Cleveland dotplot for all of the morphometric variables measured; note that some variables have a few relatively large values. Such extreme values could indicate true measurement errors (e.g. some fit the characteristics of 'observer distraction' *sensu* Morgan 2004, whereby the observer's eye is drawn to the wrong number on a measurement scale). Note that one should not try to argue that such large values could have occurred by chance. If they were, then intermediate values should also have been generated by chance, but none were. (A useful exercise is to generate, repeatedly, an equivalent number of random observations from an appropriate distribution, e.g. the Normal distribution, and determine how the number of extreme points compares to the empirical data.) When the most likely explanation is that the extreme observations are measurement (observer) errors, they should be dropped because their presence is likely to dominate the analysis. For example, we applied a discriminant analysis on the full sparrow data set to see whether observations differed among observers, and found that the first two axes were mainly determined by the outliers.

So far, we have loosely defined an 'outlier' as an observation that sticks out from the rest. A more rigorous approach is to consider whether unusual observations exert undue influence on an analysis (e.g. on estimated parameters). We make a distinction between influential observations in the response variable and in the covariates. An example of the latter is when species abundances are modelled as a function of temperature, with nearly all temperature values between 15 and 20 °C, but one of 25 °C. In general, this is not an ideal sampling design because the range 20–25 °C is inadequately sampled. In a field study, however, there may have been only one opportunity to sample the higher temperature. With a large sample size, such observations may be dropped, but with relative small data sets the consequent reduction in sample size may be undesirable, especially if other observations have outliers for other explanatory variables. If omitting such observations is not an option, then consider transforming the explanatory variables.

In regression-type techniques, outliers in the response variables are more complicated to deal with. Transforming the data is an option, but as the response variable is of primary interest, it is better to choose a statistical method that uses a probability distribution that allows greater variation for large mean values (e.g. gamma for continuous data; Poisson or negative binomial for count data) because doing this allows us to

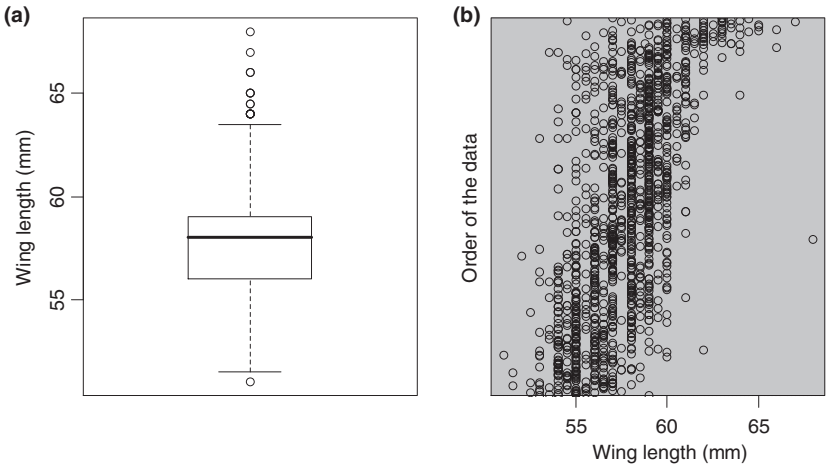

Fig. 2. (a) Boxplot of wing length for 1295 saltmarsh sparrows. The line in the middle of the box represents the median, and the lower and upper ends of the box are the 25% and 75% quartiles respectively. The lines indicate 1.5 times the size of the hinge, which is the 75% minus 25% quartiles. (Note that the interval defined by these lines is not a confidence interval.) Points beyond these lines are (often wrongly) considered to be outliers. In some cases it may be helpful to rotate the boxplot 90° to match the Cleveland dotplot. (b) Cleveland dotplot of the same data. The horizontal axis represents the value of wing length, and the vertical axis corresponds to the order of the data, as imported from the data file (in this case sorted by the bird's weight).

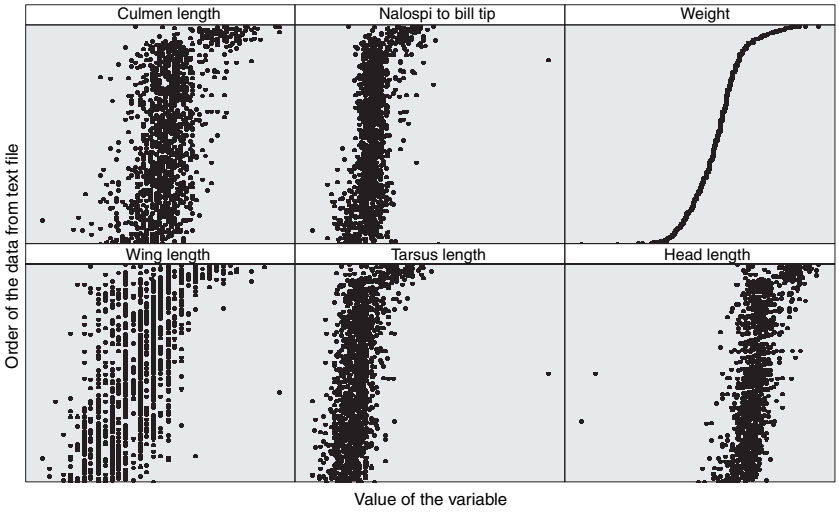

Fig. 3. Multi-panel Cleveland dotplot for six morphometric variables taken from the sparrow data, after sorting the observations from heaviest to lightest (hence the shape of the weight graph). Axis labels were suppressed to improve visual presentation. Note that some variables have a few unusually small or large values. Observations also can be plotted, or mean values superimposed, by subgroup (e.g. observer or sex) to see whether there are differences among subsets of the data.

work with the original data. For multivariate analyses, this approach is not an option because these methods are not based on probability distributions. Instead, we can use a different measure of association. For example, the Euclidean distance is rather sensitive to large values because it is based on Pythagoras' theorem, whereas the Chord distance down-weights large values (Legendre & Legendre 1998).

Some statistical packages come with a whole series of diagnostic tools to identify influential observations. For example, the Cook statistic in linear regression (Fox 2008) gives information on the change in regression parameters as each observation is sequentially, and individually, omitted. The problem with such tools is that when there are multiple 'outliers' with similar values, they will not be detected. Hence, one should investigate the presence of such observations using the graphical tools discussed in this paper, before applying a statistical analysis.

Ultimately, it is up to the ecologist to decide what to do with outliers. Outliers in a covariate may arise due to poor experimental design, in which case dropping the observation or transforming the covariate are sensible options. Observer and measurement errors are a valid justification for dropping observations. But outliers in the response variable may require a more refined approach, especially when they represent genuine variation in the variable being measured. Taking detailed field or experiment notes can be especially helpful for documenting when unusual events occur, and thus providing objective information with which to re-examine outliers. Regardless of how the issue is addressed, it is important to know whether there are outliers and to report how they were handled; data exploration allows this to be done.

Step 2: Do we have homogeneity of variance?

Homogeneity of variance is an important assumption in analysis of variance (ANOVA), other regression-related models and in multivariate techniques like discriminant analysis. Figure 4 shows conditional boxplots of the food intake rates of Hudsonian godwits (*Limosa haemastica*), a long-distance migrant shorebird, on a mudflat in Argentina (E. Ieno, unpublished data). To apply an ANOVA on these data to test whether mean intake rates differ by sex, time period or a combination of these two variables (i.e. an interaction), we have to assume that (i) variation in the observations from the sexes is similar; (ii) variation in observations from the three time periods is similar; and (iii) variation between the three time periods within the sexes is similar. In this case, there seems to be slightly less variation in the winter data for males and more variation in the male data from the summer. However, such small differences in variation are not something to worry about. More serious examples of violation can be found in Zuur *et al.* (2009a). Fox (2008) shows that for a simplistic linear regression model heterogeneity seriously degrades the least-square estimators when the ratio between the largest and smallest variance is 4 (conservative) or more.

In regression-type models, verification of homogeneity should be done using the residuals of the model; i.e. by plotting residuals vs. fitted values, and making a similar set of conditional boxplots for the residuals. In all these graphs the residual variation should be similar. The solution to heterogeneity of variance is either a transformation of the response variable to stabilize the variance, or applying statistical techniques that do not require homogeneity (e.g. generalized least squares; Pinheiro & Bates 2000; Zuur *et al.* 2009a).

Step 3: Are the data normally distributed?

Various statistical techniques assume normality, and this has led many of our postgraduate course participants to produce histogram after histogram of their data (e.g. Fig. 5a). It is important, however, to know whether the statistical technique to be used does assume normality, and what *exactly* is assumed to be normally distributed? For example, a PCA does not require normality (Jolliffe 2002). Linear regression does assume normality, but is reasonably robust against violation of the assumption (Fitzmaurice, Laird & Ware 2004). If you want to apply a statistical test to determine whether there is sig-

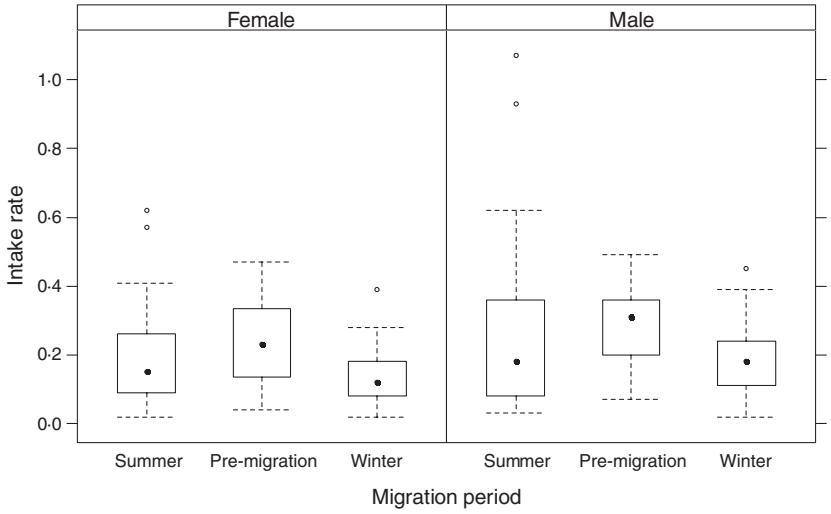

Fig. 4. Multi-panel conditional boxplots for the godwit foraging data. The three boxplots in each panel correspond to three time periods. We are interested in whether the mean values change between sexes and time periods, but need to assume that variation is similar in each group.

nificant group separation in a discriminant analysis, however, normality of observations of a particular variable *within each group* is important (Huberty 1994). Simple *t*-tests also assume that the observations in each group are normally distributed; hence histograms for the raw data of every group should be examined.

In linear regression, we actually assume normality of all the replicate observations at a particular covariate value (Fig. 6; Montgomery & Peck 1992), an assumption that cannot be verified unless one has many replicates at each sampled covariate value. However, normality of the raw data implies normality of the residuals. Therefore, we can make histograms of residuals to get some impression of normality (Quinn & Keough 2002; Zuur *et al*. 2007), even though we cannot fully test the assumption.

Even when the normality assumption is apparently violated, the situation may be more complicated than it seems. The shape of the histogram in Fig. 5a, for example, indicates skewness, which may suggest to one that data transformation is needed. Figure 5b shows a multi-panel histogram for the same variable except that the data are plotted by month; this lets us see that the skewness of the original histogram is probably caused by sparrow weight changes over time. Under these circumstances, it would not be advisable to transform the data as differences among months may be made smaller, and more difficult to detect.

Step 4: Are there lots of zeros in the data?

Elphick & Oring (1998, 2003) investigated the effects of straw management on waterbird abundance in flooded rice fields. One possible statistical analysis is to model the number of birds as a function of time, water depth, farm, field management method, temperature, etc. Because this analysis involves modelling a count, GLM is the appropriate analysis. Figure 7 shows a frequency plot illustrating how often each value for total waterbird abundance occurred. The extremely high number of zeros tells us that we should not apply an ordinary Poisson or negative binomial GLM as these would produce biased parameter estimates and standard errors. Instead one should consider zero inflated GLMs (Cameron & Trivedi 1998; Zuur *et al*. 2009a).

One can also analyse data for multiple species simultaneously using multivariate techniques. For such analyses, we need to consider what it means when two species are jointly absent. This result could say something important about the ecological characteristics of a site, for example that it contains conditions that are unfavourable to both species. By extension,

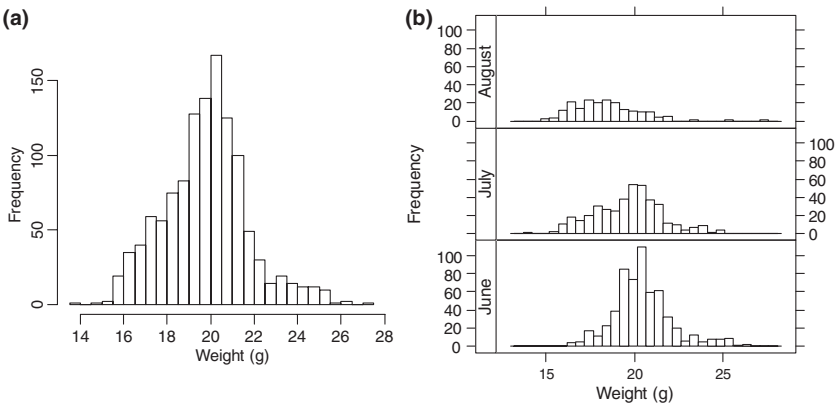

Fig. 5. (a) Histogram of the weight of 1193 sparrows (only the June, July and August data were used). Note that the distribution is skewed. (b) Histograms for the weight of the sparrows, broken down by month. Note that the centre of the distribution is shifting, and this is causing the skewed distributed for the aggregated data shown in (a).

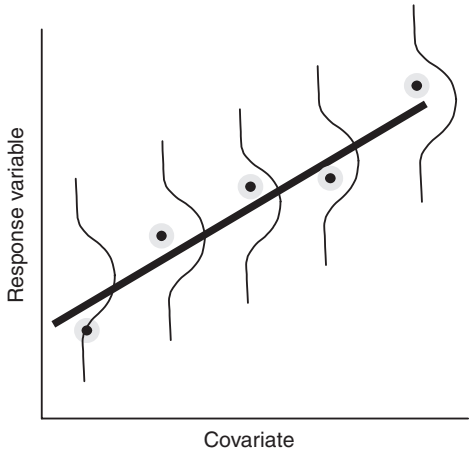

Fig. 6. Visualization of two underlying assumptions in linear regression: normality and homogeneity. The dots represent observed values and a regression line is added. At each covariate value, we assume that observations are normally distributed with the same spread (homogeneity). Normality and homogeneity at each covariate value cannot be verified unless many (> 25) replicates per covariate value are taken, which is seldom the case in ecological studies. In practice, a histogram of pooled residuals should be made, but this does not provide conclusive evidence for normality. The same limitations holds if residuals are plotted vs. fitted values to verify homogeneity.

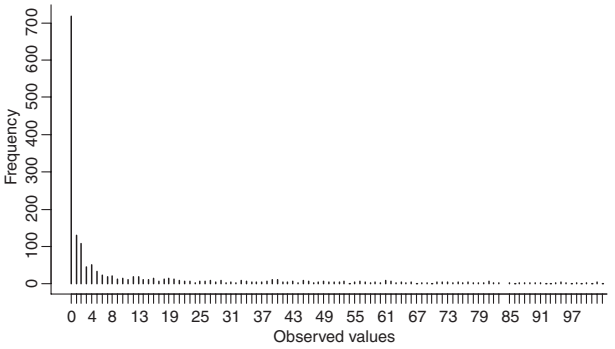

Fig. 7. Frequency plot showing the number of observations with a certain number of waterbirds for the rice field data; 718 of 2035 observations equal zero. Plotting data for individual species would result in even higher frequencies of zeros.

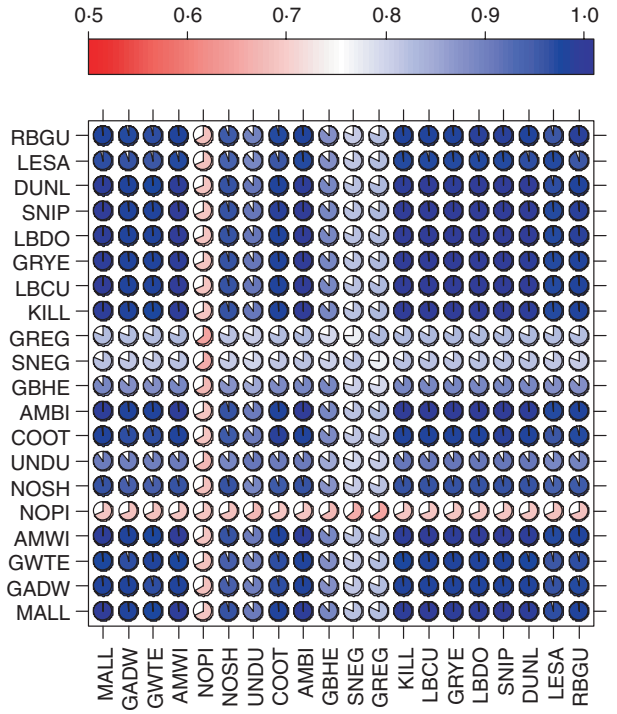

Fig. 8. A corrgram showing the frequency with which pairs of waterbird species both have zero abundance. The colour and the amount that a circle has been filled correspond to the proportion of observations with double zeros. The diagonal running from bottom left to top right represents the percentage of observations of a variable equal to zero. Four-letter acronyms represent different waterbird species. The top bar relates the colours in the graph to the proportion of zeros.

when two sites both have the same joint absences, this might mean that the sites are ecologically similar. On the other hand, if a species has a highly clumped distribution, or is simply rare, then joint absences might arise through chance and say nothing about the suitability of a given site for a species, the similarity among the habitat needs of species or the ecological similarity of sites. A high frequency of zeros, thus, can greatly complicate interpretation of such analyses. Irrespective of our attitude to joint absences, we need to know whether there are double zeros in the data. This means that for each species-pair, we need to calculate how often both had zero abundance for the same observation (e.g. site). We can either present this information in a table, or use advanced graphical tools like a corrgram (Fig. 8; Sarkar 2008). In our waterbird example, the frequency of double zeros is very high. All the blue circles correspond to species that have more than 80% of their observations jointly zero. This result is consistent with the biology of the species studied, most of which form large flocks and have highly clumped distributions. A PCA would label such species as similar, although their ecological use of habitats is often quite different (e.g. Elphick & Oring 1998). Alternative multivariate analyses that ignore double zeros are discussed in Legendre & Legendre (1998) and Zuur *et al.* (2007).

Step 5: Is there collinearity among the covariates?

If the underlying question in a study is which covariates are driving the response variable(s), then the biggest problem to overcome is often collinearity. Collinearity is the existence of correlation between covariates. Common examples are covariates like weight and length, or water depth and distance to the shoreline. If collinearity is ignored, one is likely to end up with a confusing statistical analysis in which nothing is significant, but where dropping one covariate can make the others significant, or even change the sign of estimated parameters. The effect of collinearity is illustrated in the context of multiple linear regression, but similar problems exist in analysis of variance, mixed effects models, RDA, CCA, GLMs or GAMs. Table 1 gives the results of a multiple linear regression in which

the number of saltmarsh sparrows captured in a study plot is modelled as a function of covariates that describe the relative abundance of various plant species (for details, see Gjerdrum, Elphick & Rubega 2005; Gjerdrum *et al.* 2008). The second column of the table gives the estimated *P*-values of the *t*-statistics for each regression parameter when all covariates are included in the model. Note that only one covariate, that for the per cent cover of the rush *Juncus gerardii*, is weakly significant at the 5% level.

In linear regression, an expression for the variances of the parameters β_j is given by (Draper & Smith 1998; Fox 2008):

$$\text{Variance}(\beta_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)S_j^2}$$

The term S_j depends on covariate values, n is the sample size and σ^2 is the variance of the residuals, but these terms are not relevant to the current discussion (and therefore their mathematical formulation is not given here). It is the first expression that is important. The term R_j^2 is the R^2 from a linear regression model in which covariate X_j is used as a response variable, and all other covariates as explanatory variables. A high R^2 in such a model means that most of the variation in covariate X_j is explained by all other covariates, which means that there is collinearity. The price one pays for this situation is that the standard errors of the parameters are inflated with the square root of $1/(1 - R_j^2)$, also called the variance inflation factor (VIF), which means that the *P*-values get larger making it more difficult to detect an effect. This phenomenon is illustrated in Table 1; the third column of the table gives the VIF values for all covariates and shows that there is a high level of collinearity. One strategy for addressing this problem is to sequentially drop the covariate with the highest VIF, recalculate the VIFs and repeat this process until all VIFs are smaller than a preselected threshold. Montgomery & Peck (1992) used a value of 10, but a more stringent approach is to use values as low as 3 as we did here. High, or even moderate, collinearity is especially problematic when ecological signals are weak. In that case, even a VIF of 2 may cause nonsignificant parameter estimates, compared to the situation without collinearity. Following this process caused three variables to be dropped from our analysis: the tall *Spartina alterniflora*, and those for plant height and stem density. With the collinearity problem removed, the *Juncus* variable is shown to be highly significant (Table 1). Sequentially dropping further nonsignificant terms one at a time gives a model with only the *Juncus* and Shrub variables, but with little further change in *P*-values, showing how dropping collinear variables can have a bigger impact on *P*-values than dropping nonsignificant covariates.

Other ways to detect collinearity include pairwise scatterplots comparing covariates, correlation coefficients or a PCA biplot (Jolliffe 2002) applied on all covariates. Collinearity can also be expected if temporal (e.g. month, year) or spatial variables (e.g. latitude, longitude) are used together with covariates like temperature, rainfall, etc. Therefore, one should always plot all covariates against temporal and spatial covariates. The easiest way to solve collinearity is by dropping collinear covariates. The choice of which covariates to drop can be based on the VIFs, or perhaps better, on common sense or biological knowledge. An alternative consideration, especially when future work on the topic will be done, is how easy alternative covariates are to measure in terms of effort and cost. Whenever two covariates X and Z are collinear, and Z is used in the statistical analysis, then the biological discussion in which the effect of Z is explained should include mention of the collinearity, and recognize that it might well be X that is driving the system (cf. Gjerdrum *et al.* 2008). For a discussion of collinearity in combination with measurement errors on the covariates, see Carroll *et al.* (2006).

Table 1. *P*-values of the *t*-statistic for three linear regression models and variance inflation factor (VIF) values for the full model. In the full model, the number of banded sparrows, which is a measure of how many birds were present, is modelled as a function of the covariates listed in the first column. In the second and third columns, the *P*-values and VIF values for the full model are presented (note that no variables have been removed yet). In the fourth column *P*-values are presented for the model after collinearity has been removed by sequentially deleting each variable for which the VIF value was highest until all remaining VIFs were below 3. In the last column, only variables with significant *P*-values remain, giving the most parsimonious explanation for the number of sparrows in a plot

Covariate	*P*-value (full model)	VIF	*P*-value (collinearity removed)	*P*-value (reduced model)
% *Juncus gerardii*	0·0203	44·9953	0·0001	0·00004
% Shrub	0·9600	2·7818	0·0568	0·0727
Height of thatch	0·9989	1·6712	0·8263	
% *Spartina patens*	0·0640	159·3506	0·3312	
% *Distichlis spicata*	0·0527	53·7545	0·2538	
% Bare ground	0·0666	12·0586	0·8908	
% Other vegetation	0·0730	5·8170	0·9462	
% *Phragmites australis*	0·0715	3·7490	0·2734	
% Tall sedge	0·2160	4·4093	0·4313	
% Water	0·0568	17·0677	0·6942	
% *Spartina alterniflora* (short)	0·0549	121·4637	0·2949	
% *Spartina alterniflora* (tall)	0·0960	159·3828		
Maximum vegetation height	0·2432	6·1200		
Vegetation stem density	0·7219	3·2064		

Step 6: What are the relationships between Y and X variables?

Another essential part of data exploration, especially in univariate analysis, is plotting the response variable vs. each covariate (Fig. 9). Note that the variable for the per cent of tall sedge in a plot (%Tall sedge) should be dropped from any analysis as it has only one non-zero value. This result shows that the boxplots and Cleveland dotplots should not only be applied on the response variable but also on covariates (i.e. we should not have calculated the VIFs with %Tall sedge included in the previous section). There are no clear patterns in Fig. 9 between the response and explanatory variables, except perhaps for the amount of *Juncus* (see also Table 1). Note that the absence of clear patterns does not mean that there are no relationships; it just means that there are no clear two-way relationships. A model with multiple explanatory variables may still provide a good fit.

Besides visualizing relationships between variables, scatterplots are also useful to detect observations that do not comply with the general pattern between two variables. Figure 10 shows a multi-panel scatterplot (also called a pair plot) for the 1295 saltmarsh sparrows for which we have morphological data. Any observation that sticks out from the black cloud needs further investigation; these may be different species, measurement errors, typing mistakes or they may be correct values after all. Note that the large wing length observation that we picked up with the Cleveland dotplot in Fig. 2b has average values for all other variables, suggesting that it is indeed something that should be checked. The lower panels in Fig. 10 contain Pearson correlation coefficients, which can be influenced by outliers meaning that outliers can even contribute to collinearity.

Step 7: Should we consider interactions?

Staying with the sparrow morphometric data, suppose that one asks whether the relationship between wing length and weight changes over the months and differs between sexes. A common approach to this analysis is to apply a linear regression model in which weight is the response variable and wing length (continuous), sex (categorical) and month (categorical)

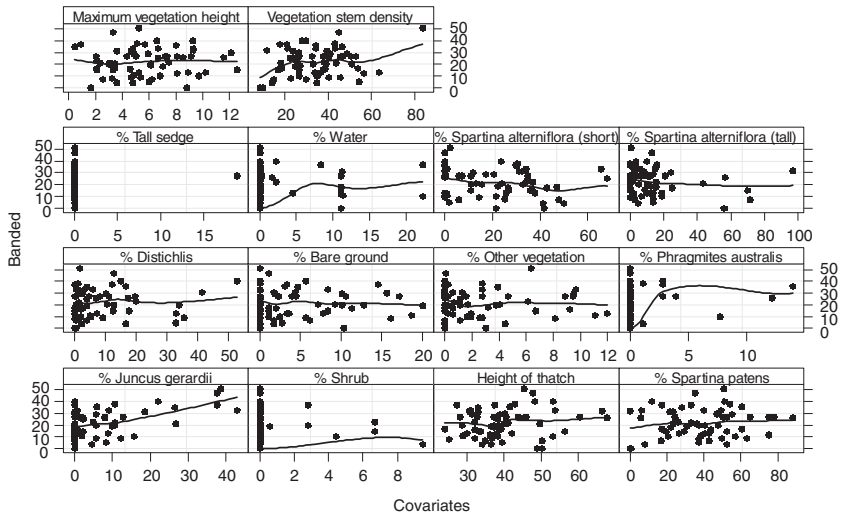

Fig. 9. Multi-panel scatterplots between the number of banded sparrows and each covariate. A LOESS smoother was added to aid visual interpretation.

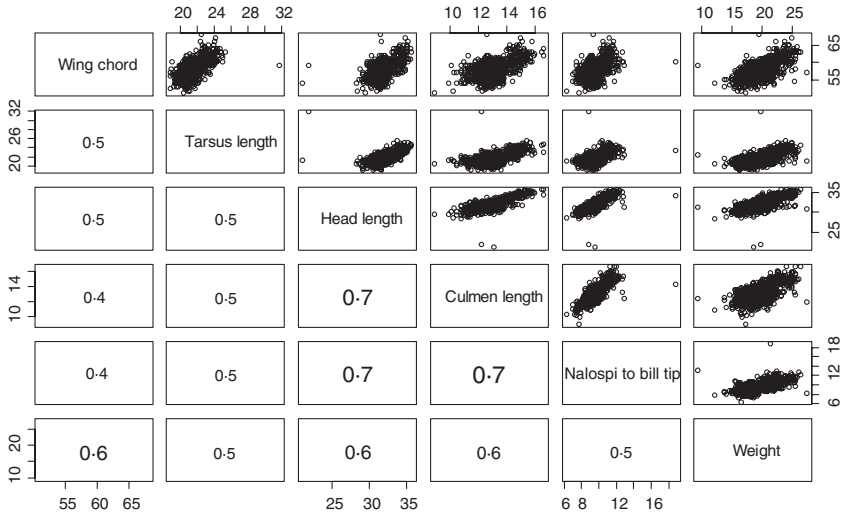

Fig. 10. Multi-panel scatterplot of morphometric data for the 1295 saltmarsh sparrows. The upper/right panels show pairwise scatterplots between each variable, and the lower/left panels contain Pearson correlation coefficients. The font size of the correlation coefficient is proportional to its value. Note that there are various outliers.

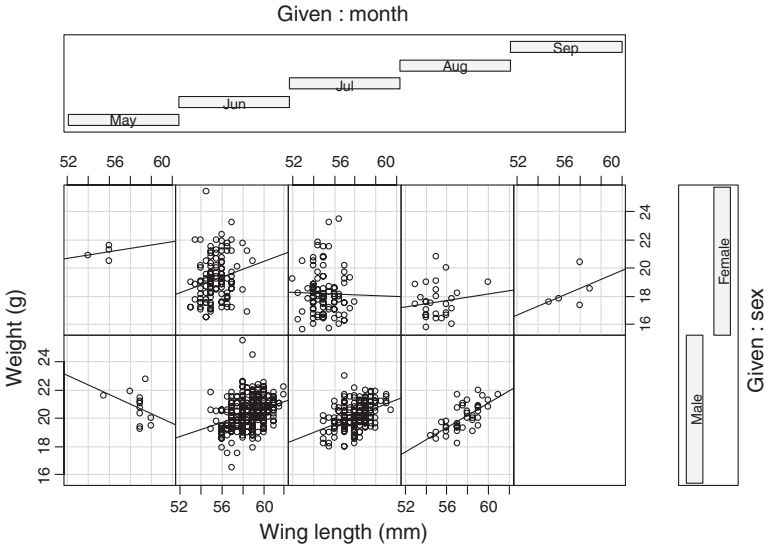

Fig. 11. Coplot for the sparrow data. The lower left panel shows a scatterplot between wing length and weight for males in May, and the upper right panel for females in September. On each panel, a bivariate linear regression model was fitted to aid visual interpretation.

are covariates. Results showed that the three-way interaction is significant, indicating that the relationship between weight and wing length is indeed changing over the months and between sexes. However, there is a problem with this analysis. Figure 11 shows the data in a coplot, which is an excellent graphical tool to visualize the potential presence of interactions. The graph contains multiple scatterplots of wing length and weight; one for each month and sex combination. A bivariate linear regression line is added to each scatterplot; if all lines are parallel, then there is probably no significant interaction (although only the regression analysis can tell us whether this is indeed the case). In our example, lines have different slopes, indicating the potential presence of interactions. In some months, however, the number of observations is very small, and there are no data at all from males in September. A sensible approach would be to repeat the analysis for only the June–August period.

Step 8: Are observations of the response variable independent?

A crucial assumption of most statistical techniques is that observations are independent of one another (Hurlbert 1984), meaning that information from any one observation should not provide information on another after the effects of other variables have been accounted for. This concept is best explained with examples.

The observations from the sparrow abundance data set were taken at multiple locations. If birds at locations close to each other have characteristics that are more similar to each other than to birds from locations separated by larger distances, then we would violate the independence assumption. Another example is when multiple individuals of the same family (e.g. all of the young from one nest) are sampled; these individuals might be more similar to each other than random individuals in the population, because they share a similar genetic make-up and similar parental provisioning history.

When such dependence arises, the statistical model used to analyse the data needs to account for it. For example, by modelling any spatial or temporal relationships, or by nesting data in a hierarchical structure (e.g. nestlings could be nested within nests). Testing for independence, however, is not always easy. In Zuur *et al.* (2009a) a large number of data sets were analysed in which dependence among observations played a role. Examples include the amount of bioluminescence at sites along an oceanic depth gradient, nitrogen isotope ratios in whale teeth as a function of age, pH values in Irish rivers, the number of amphibians killed by cars at various locations along a road, feeding behaviour of different godwits on a beach, the number of disease-causing spores affecting larval honey bees from multiple hives and the number of calls from owl chicks upon arrival of a parent. Another commonly encountered situation where non-independence must be addressed is when there is phylogenetic structure (i.e. dependence due to shared ancestry) within a data set.

There are many ways to include a temporal or spatial dependence structure in a model for analysis. These include using lagged response variables as covariates (Brockwell & Davis 2002), mixed effects modelling (Pinheiro & Bates 2000), imposing a residual correlation structure using generalized least squares (Zuur *et al.* 2009a) or allowing regression parameters to change over time (Harvey 1989). It is also possible to fit a model with and without a correlation structure, and compare the models using a selection criterion or hypothesis test (Pinheiro & Bates 2000). The presence of a dependence structure in the raw data may be modelled with a covariate such as month or temperature, or the inclusion of a smoothing function of time or a two-dimensional smoother of spatial coordinates (Wood 2006). Regardless of the method used, the model residuals should not contain any dependence structure. Quite often a residual correlation structure is caused by an important covariate that was not measured. If this is the case, it may not be possible to resolve the problem.

When using regression techniques, the independence assumption is rather important and violation may increase the type I error. For example, Ostrom (1990) showed that ignoring auto-correlation may give *P*-values that are 400% inflated.

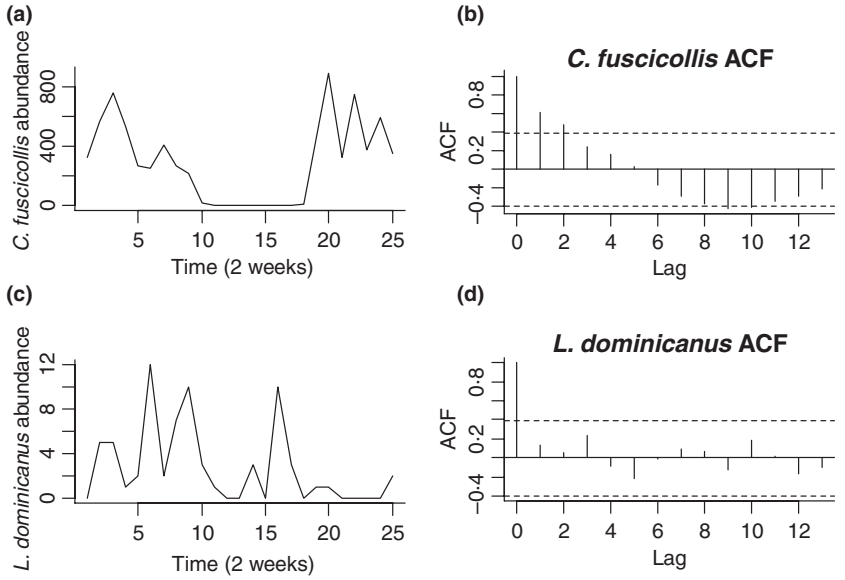

Fig. 12: (a) Number of *Calidris fuscicollis* plotted vs. time (1 unit = 2 weeks). (b) Auto-correlation function for the *C. fuscicollis* time series showing a significant correlation at time lags of 2 and 4 weeks (1 time lag = 2 weeks). (c) Number of *Larus dominicanus* vs. time. (d) Auto-correlation function for *L. dominicanus* showing no significant correlation. Dotted lines in panels (b) and (d) are *c.* 95% confidence bands. The auto-correlation with time lag 0 is, by definition, equal to 1.

Hence, it is important to check whether there is dependence in the raw data before doing the analysis, and also the residuals afterwards. These checks can be made by plotting the response variable vs. time or spatial coordinates. Any clear pattern is a sign of dependence. This approach is more difficult if there is no clear sequence to the observations (e.g. multiple observations on the same object), but in this case one can include a dependence structure using random effects (Pinheiro & Bates 2000; Fitzmaurice *et al.* 2004; Brown & Prescott 2006; Zuur *et al.* 2009a). Figure 12a,c shows a short time series illustrating the observed abundance of two bird species on a mudflat in Argentina over a 52 week period (E. Ieno, unpublished data). The first time series shows high numbers of white-rumped sandpipers *Calidris fuscicollis* during the first 20 weeks, followed by zeros (because the species migrates), and then an abundance increase again after 38 weeks. The second time series does not show a clear pattern in the abundance of kelp gulls (*Larus dominicanus*).

A more formal way to assess the presence of temporal dependence is to plot auto-correlation functions (ACF) for regularly spaced time series, or variograms for irregularly spaced time series and spatial data (Schabenberger & Pierce 2002). An ACF calculates the Pearson correlation between a time series and the same time series shifted by k time units. Figures 12b,d show the auto-correlation of the time series in panels (a) and (c). Panel (b) shows a significant correlation with a time lag of $k = 1$ and $k = 2$. This means that abundances at time t depend on abundances at time $t - 1$ and $t - 2$, and any of the methods mentioned above could be applied. For the *L. dominicanus* time series, there is no significant auto-correlation.

Discussion

All of the problems described in this paper, and the strategies to address them, apply throughout ecological research, but they are particularly relevant when results are to be used to guide management decisions or public policy because of the repercussions of making a mistake. Increasing attention has been paid in recent years to the body of data supporting particular management practices (Roberts, Stewart & Pullin 2006; Pullin & Knight 2009), and applied ecologists have become increasingly sophisticated in the statistical methods that they use (e.g. Ellison 2004; Stephens *et al.* 2005; Robinson & Hamann 2008; Koper & Manseau 2009; Law *et al.* 2009; Sonderegger *et al.* 2009). But more fundamental questions about the appropriateness of the underlying data for a given analysis can be just as important to ensuring that the best policies are derived from ecological studies.

In this paper, we have discussed a series of pitfalls that can seriously influence the results of an analysis. Some of these problems are well known, some less so, but even the well-known assumptions continue to be violated frequently in the ecological literature. In all cases, the problems can lead to statistical models that are wrong. Such problems can be avoided only by applying a systematic data exploration before embarking on the analysis (Fig. 1).

Although we have presented our protocol as a linear sequence, it should be used flexibly. Not every data set requires each step. For example, some statistical techniques do not require normality (e.g. PCA), and therefore there is no point in making histograms. The best order to apply the steps may also depend on the specific data set. And for some techniques, assumptions can be verified only by applying data explorations steps after the analysis has been performed. For example, in linear regression, normality and homogeneity should be verified using the residuals produced by the model. Rather than simplistically following through the protocol, ticking off each point in order, we would encourage users to treat it as a series of questions to be asked of the data. Once satisfied that each issue has been adequately addressed in a way that makes biological sense, the data set should be ready for the main analysis.

Ecological field data tend to be noisy, field conditions unpredictable and prior knowledge often limited. In the applied realm, changes in funding, policy, and research priorities further complicate matters. This situation is especially so for long-term studies, where the initial goals often change with circumstances (e.g. the use of many data sets to examine species responses to climate change). For all these reasons, the idealized situation whereby an ecologist carefully designs their analysis *a priori* and then collects data may be compromised or irrelevant. Having the analytical flexibility to adjust one's analyses to such circumstance is an important skill for an applied ecologist, but it requires a thorough understanding of the constraining assumptions imposed by a given data set.

When problems arise, the best solutions vary. Frequently, however, ecologists simply transform data to avoid assumption violations. There are three main reasons for a transformation; to reduce the effect of outliers (especially in covariates), to stabilize the variance and to linearize relationships. However, using more advanced techniques like GLS and GAMs, heterogeneity and nonlinearity problems can be solved, making transformation less important. Zuur *et al.* (2009a) showed how the use of a data transformation resulted in different conclusions about long-term trends compared to an appropriate analysis using untransformed data; hence it may be best to avoid transforming response variables. If a transformation is used, automatic selection tools such as Mosteller and Tukey's bulging rule (Mosteller & Tukey 1977) should be used with great caution because these methods ignore the effects of covariates. Another argument against transformations is the need to subsequently back-transform values to make predictions; it may not always be clear how to do this and still be able to interpret results on the original scale of the response variable. It is also important to ensure that the transformation actually solves the problem at hand; even commonly recommended transformations do not always work. The bottom line is that the choice of a specific transformation is a matter of trial and error.

It is a given fact that data exploration should not be used to define the questions that a study sets out to test. Every step of the exploration should be reported, and any outlier removed should be justified and mentioned. Reasons for data transformations need to be justified based on the exploratory analysis (e.g. evidence that model assumptions were violated and that the transformation rectified the situation).

Applying data exploration (e.g. scatterplots to visualize relationships between response and explanatory variables) to create hypotheses and then using the same data to test these hypotheses should be avoided. If one has limited *a priori* knowledge, then a valid approach is to create two data sets; apply data exploration on the first data set to create hypotheses and use the second data set to test the hypotheses. Such a process, however, is only practical for larger data sets. Regardless of the specific situation, the routine use and transparent reporting of systematic data exploration would improve the quality of ecological research and any applied recommendations that it produces.

Acknowledgements

We thank Anatoly Saveliev, and two anonymous reviewers for comments on an earlier draft.

References

Brockwell, P.J. & Davis, R.A. (2002) *Introduction to Time Series and Forecasting*, 2nd edn. Springer-Verlag, New York.

Brown, H. & Prescott, R. (2006) *Applied Mixed Models in Medicine*, 2nd edn. John Wiley and Sons, New York.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference. A Practical Information–Theoretic Approach*, 2nd edn. Springer, New York.

Cameron, A.C. & Trivedi, P.K. (1998) *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edn. Chapman & Hall, Boca Raton, FL.

Chatfield, C. (1998) *Problem Solving: A Statistician's Guide*. Chapman & Hall, Boca Raton, FL.

Cleveland, W.S. (1993) *Visualizing Data*. Hobart Press, Summit, NJ.

Draper, N.R. & Smith, H. (1998) *Applied Regression Analysis*, 3rd edn. John Wiley and Sons, New York.

Ellison, A.M. (2004) Bayesian inference in ecology. *Ecology Letters*, **7**, 509–520.

Elphick, C.S. & Oring, L.W. (1998) Winter management of Californian rice fields for waterbirds. *Journal of Applied Ecology*, **35**, 95–108.

Elphick, C.S. & Oring, L.W. (2003) Conservation implications of flooding rice fields on winter waterbird communities. *Agriculture, Ecosystems and Environment*, **94**, 17–29.

Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. (2004) *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, NJ.

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, 2nd edn. Sage Publications, CA.

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

Gjerdrum, C., Elphick, C.S. & Rubega, M. (2005) What determines nest site selection and nesting success in saltmarsh breeding sparrows? *Condor*, **107**, 849–862.

Gjerdrum, C., Elphick, C.S. & Rubega, M.A. (2008) How well can we model numbers and productivity of saltmarsh sharp-tailed sparrows (*Ammodramus caudacutus*) using habitat features? *Auk*, **125**, 608–617.

Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, UK.

Hilbe, J.M. (2007) *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK.

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

Jolliffe, I.T. (2002) *Principal Component Analysis*, 2nd edn. Springer, New York.

Koper, N. & Manseau, M. (2009) Generalized estimating equations and generalized linear mixed-effects models for modelling resources selection. *Journal of Applied Ecology*, **46**, 590–599.

Läärä, E. (2009) Statistics: reasoning on uncertainty, and the insignificance of testing null. *Annales Zoologici Fennici*, **46**, 138–157.

Law, R., Illian, J., Burslem, D.F.R.P., Gratzer, G., Gunatilleke, C.V.S. & Gunatilleke, I.A.U.N. (2009) Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, **97**, 616–628.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Second English Edition. Elsevier, Amsterdam.

Montgomery, D.C. & Peck, E.A. (1992) *Introduction to Linear Regression Analysis*. Wiley, New York.

Morgan, J.H. (2004) Remarks on the taking and recording of biometric measurements in bird ringing. *The Ring*, **26**, 71–78.

Mosteller, F. & Tukey, J.W. (1977) *Data Analysis and Regression: A Second Course in Statistics*. Addison Wesley, Reading, MA.

Ostrom, C.W. (1990) *Time Series Analysis: Regression Techniques*, 2nd edn. Sage Publications Inc, Thousand Oaks/Newbury Park, CA.

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

Pullin, A.S. & Knight, T.M. (2009) Doing more good than harm – building an evidence-based for conservation and environmental management. *Biological Conservation*, **142**, 931–934.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK.
R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. URL http://www.R-project.org.
Roberts, P.D., Stewart, G.B. & Pullin, A.S. (2006) Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. *Biological Conservation*, **132**, 409–423.
Robinson, A.P. & Hamann, J.D. (2008) Correcting for spatial autocorrelation in sequential sampling. *Journal of Applied Ecology*, **45**, 1221–1227.
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.
Sonderegger, D.L., Wang, H., Clements, W.H. & Noon, B.R. (2009) Using SiZer to detect thresholds in ecological data. *Frontiers in Ecology and the Environment*, **7**, 190–195.
Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martínez del Rio, C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, **42**, 4–12.
ter Braak, C.J.F. & Verdonschot, P.F.M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Science*, **57**, 225–289.
Wood, S.N. (2006) *Generalized Additive Models. An Introdcution with R. Chapman Hall/CRC*, Boca Raton, FL. Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G. (2009a) *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.
Zuur, A.F., Ieno, E.N. & Smith, G.M. (2007) *Analysing Ecological Data*. Springer, New York.
Zuur, A.F., Ieno, E.N. & Meesters, E.H.W.G. (2009b) *A Beginner's Guide to R*. Springer, New York.

Received 13 August 2009; accepted 8 October 2009
Handling Editor: Robert P. Frecklenton

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Data sets and R code used for analysis.

As a service to our authors and readers, this journal provides support ing information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.