Title: Why do we still use stepwise modelling in ecology and behaviour?

Mark J. Whittingham[1], Philip A. Stephens[2], Richard B. Bradbury[3] & Robert P. Freckleton[4].

1 Division of Biology, School of Biology and Psychology, Ridley Building, University of Newcastle, Newcastle-Upon-Tyne, NE1 7RU

Corresponding author: e-mail: m.j.whittingham@ncl.ac.uk

2 Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK

3 Royal Society for the Protection of Birds, The Lodge, Sandy, Bedfordshire, SG19 2DL, UK

4 Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

Abstract

1. The biases and shortcomings of stepwise multiple regression are well established within the statistical literature. However an examination of papers published in 2004 by three leading ecological and behavioural journals suggested that the use of this technique remains widespread: of 65 papers in which a multiple regression approach was used, 57% of studies used a stepwise procedure.
2. The principal drawbacks of stepwise multiple regression include bias in parameter estimation, inconsistencies among model selection algorithms, an inherent (but often overlooked) problem of multiple hypothesis testing, and an inappropriate focus or reliance on a single best model. We discuss each of these issue with examples.
3. We use a worked example of data on yellowhammer distribution collected over four years to highlight the pitfalls of stepwise regression. We show that stepwise regression allows models containing significant predictors to be obtained from each year's data. In spite of the significance of the selected models, they vary substantially between years and suggest patterns that are at odds with those determined by analysing the full, four year data set.
4. An Information Theoretic (IT) analysis of the yellowhammer data set illustrates why the varying outcomes of stepwise analyses arise. In particular, the IT approach identifies large numbers of competing models that could describe the data equally well, showing that no one model should be relied upon for inference.

Keywords: multivariate statistical analysis; minimum adequate model (MAM); ecological modelling; statistical bias; habitat selection

Introduction

In the face of complexity, ecologists often strive to identify models that capture the essence of a system, explaining the observed distribution and perhaps ultimately permitting prediction. A first step toward this aim is to collect data on the response of interest, together with data on factors that it is believed might influence that response. Frequently data are observational (i.e. the variance in the dataset has not been generated by experimental manipulation) leading to difficulties in determining which causal factor or factors best explain the observed responses. In these situations, scientific possibility is limited to describing the system and identifying models consistent with the observed phenomenon. One of the most commonly used techniques for this purpose is multiple regression or, more generally, a general linear model with multiple predictors. The statistical theory underlying this methodology is well understood (e.g. Draper & Smith 1981; McCullogh & Nelder 1989), as are the assumptions and limitations of the approach (e.g. Derksen & Keselman 1992; Burnham & Anderson 2002).

Although the scientific primacy of a principle of parsimony is without clear support (Guthery *et al*. 2005), it is usually the case that models with fewer variables also contain fewer nuisance variables and have greater generality (Ginzberg & Jensen 2004). For that reason, research is usually directed towards identifying a relatively parsimonious model that is in general agreement with observed data. A suite of model simplification techniques has been developed, and the notion of a minimum adequate model (MAM) has become commonplace in ecology. A MAM is defined as the model that contains the minimum number of predictors that satisfy some criterion, for example, the model that only contains predictors that are significant at some pre-specified probability level. Finding such a model is not straightforward, and most

statistical packages offer algorithms for model selection in multiple regression. These include algorithms that operate by successive addition or removal of significant or non-significant terms (forward selection and backward elimination, respectively), and those that operate by forwards selection but also check the previous term to see if it can now be eliminated (stepwise regression). Collectively, these algorithms are usually referred to as stepwise multiple regression.

In spite of wide recognition of the limitations of stepwise multiple regression (Grafen & Hails 2002; Hurvich & Tsai 1990; Johnson et al. 2004; Stephens et al. 2005; Steyerberg et al. 1999; Wintle et al. 2003), use of the technique in ecology remains widespread (see further below for a review of applications in major journals). In particular, three problems with the approach are frequently overlooked in ecological analyses, all of which may lead to erroneous conclusions and, potentially, misdirected research. These include bias in parameter estimation, inconsistencies among model selection algorithms, and an inappropriate focus or reliance on a single best model, where data are often inadequate to justify such confidence.

In this paper, we give a brief review of the major problems with stepwise multiple regression and we analyse how frequently the technique is used in leading ecological and behavioural journals. We present an example of how focusing on a single model may lead to difficulties of interpretation. Finally, we discuss the problems of analysing and modelling data from complex multivariable ecological datasets.

Problems with multiple regression

Bias in parameter estimation

Stepwise multiple regression requires that model selection (i.e. deciding which

regression variables should be included in the final MAM) is conducted through parameter inference (i.e. testing whether parameters are significantly different from zero) (Chatfield 1995), which can lead to biases in parameters, over-fitting and incorrect significance tests. To see this, consider a simple example, using a single parameter. Consider the linear model which models an observation y_i as a function of parameters α and β, predictor value x_i and some error ε:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (1)$$

which is fitted to data vector **y** and predictor vector **x**. A stepwise approach may be used to decide whether the model in equation (1) is preferable to the simpler model:

$$y_i = \alpha + \varepsilon_i \quad (2)$$

One simple way to do this is to compute the estimate of β (termed b) and then determine whether b is significantly different from zero.

Fig. 1 shows a simple simulation example which illustrates the logical problem in using the test on b to determine which of models (1) and (2) are preferable (see Fig. 1. legend for details). For the simulated data, Fig. 1A shows the sampling distribution of b, a t-distribution. The distribution in Fig. 1A corresponds to the distribution of b when model (1) only is fitted to the data, and no attempt is made at distinguishing between (1) and (2).

Fig. 1B shows the corresponding sampling distribution when model selection based on the significance of b is employed. Accepting model (2) over model (1) is equivalent to accepting a value of $b = 0$ as an estimate of β in model (1). Thus, in Fig. 1B, the distribution of estimates of β has a peak at zero, since most estimates in Fig. 1A are non-significant (i.e. $P > 0.05$). In the right tail an estimate is significant only when it exceeds a critical value. What is clear from Fig. 1B is that the sampling distribution that results from model selection is highly unrepresentative of the

expected distribution of *b* in Fig. 1A. Importantly, any individual estimate *b* in this example is biased: either a value of zero is accepted if the significance test on *b* is non-significant, underestimating β, or values greatly in excess of the true value are accepted if the test on *b* is significant.

This phenomenon is termed model selection bias and will arise in any method of model selection based on the inclusion / exclusion of individual predictors without reference to the suite of other possible models (Chatfield 1995; Burnham & Anderson 1998, 2002). In contrast, in Fig. 1A no individual estimate is biased one way or the other relative to the true value. This bias is important if the model is to be used predictively, and also has implications for other analyses based on the model.

Stepwise algorithms, consistency and interpretation

A second problem with stepwise multiple regression is more widely-recognised and yet appears not to have deterred many ecologists from using the technique. The problem is that the algorithm used (forward selection, backward elimination or stepwise), the order of parameter entry (or deletion), and the number of candidate parameters, can all affect the selected model (e.g. Derksen & Keselman, 1992). This problem is particularly acute where the predictors are correlated (e.g. see Grafen & Hails 2002 for an example). In addition, the number of candidate parameters has a positive effect on the number of nuisance (or noise) variables that are represented in the selected MAM (Derksen & Keselman 1992). Interpreting the quality of the selected model can also be difficult. In particular, it is easy to overlook the fact that a single stepwise regression does not represent one hypothesis test but, rather, involves a large number of tests. This inevitably inflates the probability of Type I errors (false positive results) (Wilkinson 1979). Similarly, searching for a model on the basis of

the data inflates the R^2 value (Cohen & Cohen 1983), overestimating the fit that would be achieved by the same model were more data available. Finally, owing to the selection of variables to include on the basis of the observed data, the distribution of the *F*-statistic is also affected, invalidating tests of the overall statistical significance of the final model (Pope & Webster 1972).

"Best" models and inference

A final source of concern with stepwise regression procedures is their aim of identifying a single "best" MAM as the sole product of analysis. This can suggest a level of confidence in the final model that is not justified by the data, focusing all further analysis and reporting on that single model. Although one model may be selected, other models may have a similarly good fit and it is highly likely that there will be uncertainty surrounding estimates of parameters and even which parameters should be included. Basing inference or conclusions on a single model may be misleading, therefore, because a rather different model may fit the data nearly as well. The selection of a single MAM does not allow such uncertainty to be expressed. We discuss this problem further below.

Current use of stepwise regression

Recognition of all of the problems outlined above is not widespread among ecologists. Recent publications have drawn attention to the problems of bias arising from variable selection on the basis of statistical significance (e.g. Anderson, Burnham & Thompson 2000; Burnham & Anderson 2002) and, as a result, alternative model selection protocols are increasingly used. In particular, use of information theoretic (IT) model selection based on Akaike's Information Criterion (AIC, see further

below) has increased substantially over recent years (Guthery *et al*. 2005; Johnson & Omland 2004; Rushton, Ormerod & Kerby 2004). In spite of this, two of the central messages of Burnham & Anderson (e.g. 2002) have been widely overlooked. These are that models representing different hypotheses should be compared in their entirety, rather than through automated selection procedures, and that further analysis should not be based on a single best model, but should explicitly acknowledge uncertainty among models that are similarly consistent with the data. That these points have been overlooked means that even where authors have used IT model selection, they have often retained the use of stepwise procedures, and based inference on a single best model. Some authors have attempted to overcome some of the limitations of stepwise procedures by checking for consistency between stepwise algorithms (e.g. Post 2005) but this approach is seldom explicit.

In order to assess the prevalence of different stepwise approaches in current literature, MJW reviewed 508 papers published in 2004 in three leading journals: *Journal of Applied Ecology*, *Animal Behaviour* and *Ecology Letters*. In all cases in which a multiple regression approach (excluding ordination techniques) was used, the analytical approach was identified as stepwise or other. Among papers employing stepwise techniques, studies were further subdivided into those that used least squares approaches and those that used IT techniques. Multipredictor regression analyses that did not use stepwise techniques were divided among those that based inference on a global model (i.e. inferences were drawn with all predictors present), and those that used other techniques (typically IT-AIC) to determine a set of well-supported models for inference.

Results of this analysis are presented in Table 1. Overall, 65 papers used a multiple regression approach, of which 57% used a stepwise procedure; however,

there was no statistically significant difference between the proportion of studies using stepwise regression across the three journals ($\chi^2 = 0.145$, $P = 0.98$). Of the studies that used stepwise procedures, six out of 37 (16%) used IT-AIC, whilst the remainder used least squares techniques.

Example

As an empirical example of the problems of using stepwise multiple regression we reanalysed a published data set, collected to determine which factors influence the occurrence of yellowhammers *Emberiza citrinella* L. on lowland farms in the UK (Bradbury *et al*. 2000; see the accompanying electronic supplement for further details of the data and the analytical methods). Previous analyses were conducted using least squares stepwise regression (Bradbury *et al*. 2000). Here we were primarily interested in the limitations of using a single best model for inference, rather than in the limitations of the stepwise approach (which are well-established, see above).

We fitted models to our dataset using least squares procedures (e.g. procedure "lm" in 'R') and compared them using AIC. AIC is a likelihood-based measure of model fit that accounts for the number of parameters estimated in a model (i.e. models with large numbers of parameters are penalised more heavily than those with smaller numbers of parameters), such that the model with the lowest AIC has the 'best' relative fit, given the number of parameters included (Akaike 1974).

The IT methodology developed by Burnham & Anderson (2002) is designed to conduct a comparative model fit analysis for a group of competing models. Specifically, for each model a likelihood weight (for model *i* termed w_i) is calculated. This value has a simple interpretation: it is the probability that of the set of models

considered, model *i* would be the AIC-best model, were the data collected again under identical circumstances. For a set of models the likelihood weights sum to one.

For a dataset in which there is a clear 'best' model, one model would have a very high likelihood weight, and all other models would have very low weights. On the other hand, if all the models are poor, or if most have similar fit, then a number of models will share a similarly low probability. If there is no single model that clearly outperforms all others, the IT methodology may be used to perform model averaging, in which the parameter estimates of all models are combined, the contribution of each model being proportional to its likelihood weight. By contrast, stepwise methodology would identify a single model as pre-eminent, encouraging all further interpretation to be based on that model alone, ignoring the other models with similar fit to the data.

For the yellowhammer dataset, there were nine predictors, and we fitted all possible subsets of these parameters. For each model we generated a likelihood weight, and we ranked all models from best fitting to worst fitting on the basis of AIC values. We plotted summed likelihood weights against model rank (Fig. 2). These plots are effectively cumulative probability plots, with the summed probability measuring the probability that the cumulative set of models would include the AIC-best model were the data re-collected. At a given cumulative probability level (e.g. 95%) this is sometimes termed a confidence set.

The yellowhammer dataset was collected over four years. We analysed the data separately for each year, and for all years combined. The data from the four years analysed separately failed to yield a model that, in terms of likelihood weights, was clearly better than the alternative models (Fig. 2*a, b*). For instance, in Fig. 2*a* the four years of study required 77, 114, 172 and 159 models to yield a summed probability of 0.95. The implication is therefore, that any one of a large number of models could

have been selected as the best fitting model in each year. The best-fitting model is, in a sense, a random draw from this set of similarly well supported models. This interpretation is backed up by Table 2 which shows the minimum adequate models selected for the four separate years. The models selected are highly variable from year to year, with no variable selected in all four years.

The analysis of the combined dataset yielded a smaller set of credible models, with only 42 models required to reach a probability of 0.95. However this is still too large a number to be able to base all inference and conclusions on one model with any confidence. The MAM for this dataset includes most of the variables found to be significant in the analysis of the single years. However, the likelihood weight for this model was only 0.028; it was not the AIC-best model, which itself had an AIC weight of only 0.048. Either of these models would be a poor one on which to base inference.

Discussion

Biases and shortcomings of stepwise multiple regression are well established. Surprisingly, however, we found that of recent papers in three leading ecological and behavioural journals, approximately half of those that employed multiple regression did so using a stepwise procedure (Table 1). Our example, using detailed data on yellowhammer habitat selection highlights the dangers of this approach. In particular, although the yellowhammer field study was conducted on a large scale, a single year's data was clearly insufficient to identify a single best model to explain yellowhammer territory occupancy, or even a small number of similarly well-supported models for that purpose. Even with four years' data, representing a comprehensive autecological study, as many as 42 models provided similarly good explanations of the observed data. To select a single MAM from this set without acknowledging the considerable

uncertainty that remains, would be entirely misleading. A full model approach (i.e. including all predictors and all four years' data) gives, in this case, a very similar result to one derived using the IT methodology (see Table 2). This re-inforces the point that conclusions based on data collected in any one year may be erroneous.

Multiple regression is a widely used statistical method within ecology with 13% of the papers we reviewed using this method. It was notable that within two of the journals sampled (*Animal Behaviour* and *Ecology Letters*) only between 8-9% of studies used a multiple regression approach whereas in *Journal of Applied Ecology* 26% (23/88) used such an approach. Therefore the problems we report may very likely be more widespread within landscape studies (which tend to collect large numbers of potentially explanatory factors) than in studies with more restricted experimental designs (e.g. laboratory experiments which are common within behavioural science).

As with our example, it is likely that many studies employing stepwise procedures conceal much uncertainty when selecting a single MAM. Most ecological datasets usually include a set of predictors with a tapered distribution of effect sizes (Burnham & Anderson 2002) and almost all analyses will therefore contain equivocal variables close to statistical significance. Estimated effects are likely to be strong, intermediate and weak, or zero. For predictors with zero or weak effects, MAMs are likely to yield biased estimates of parameters (e.g. Fig. 1) and a high Type I error rate. Furthermore, when correlations exist between the predictors, different combinations of predictors may yield models with similar explanatory power (e.g. Grafen & Hails 2002). The methodology underlying MAMs is generally not designed to analyse marginal effects.

Instead of using stepwise procedures, two analyses are arguably valid: a full model including all effects, or the analysis using IT-AIC methods (the approach that we demonstrated here). The full model tests a single set of hypotheses on a single model. The expected parameter estimates are unbiased (e.g. Fig. 1), and the statistical properties of the generalised linear model are well understood (e.g. McCullough & Nelder 1989). If the main aim of the study in question were to analyse whether each of the predictors affected the distribution of birds, and whether the effects were consistent between years, this analysis should be entirely justifiable.

The downsides of using the full model for analysis and inference are that (i) the model may not be the 'best' model for the data in question, as other models may fit the data equally as well; (ii) if we wished to use the model predictively, it includes variables that are non-significant; (iii) the analysis would rely on null-hypothesis testing. The first argument is not relevant to comparisons of the effects of different predictors. The reason why this model may not be the best model is precisely that it includes predictors that are non-significant. The analysis is designed to reveal those predictors that are significant, and those that are not. Hence we would not expect this model to be the best model.

The second problem is that a full model will contain estimates for all parameters, irrespective of whether they are statistically significant or not. This can generate an excess of noise, resulting in a model that is unsatisfactory for prediction. By contrast, techniques exist for multi-model parameter estimation, particularly within the IT framework (e.g. Burnham & Anderson 2002). This approach allows model uncertainty to be measured at the same time as parameter uncertainty to assess the likely bias in parameters resulting from selection. The advantage of using this approach for prediction, rather than the full model, is that the contribution of each

predictor (in making predictions) is determined by its performance across the whole suite of models.

The third problem with basing inference on the global model, is where tests of individual parameters (designed to determine how important they are) are conducted using null hypothesis testing (NHT). NHT has been the focus of much criticism in recent decades (e.g. Carver 1978, Cohen 1994, Johnson 1999, Anderson *et al*. 2000). In particular, two problems of NHT apply directly to the issue of parameter testing within the global model. First, NHT is essentially binary in nature; either the tested parameter is (statistically) 'significant' or it is not. Wherever the threshold for significance is drawn, this can lead to dramatic differences in inference arising from very small differences in the dataset. For example, consider a threshold for significance drawn at $P = 0.05$. Imagine that our estimate for a parameter coefficient, β, was 2.5, with a 95% confidence interval between $-0.1 < \beta < 5.1$. Here, we would reject the estimate of β and assume that $\beta = 0$ was a more reliable estimate. However, if the estimate of β was the same but with a confidence interval $0.1 < \beta < 4.9$, then we would accept that $\beta = 2.5$. The second problem of NHT that applies to analyses of the global model is that, assuming we have reason to include the variable of interest in the model, then a null hypothesis of "no effect" (representing a coefficient estimate of $\beta = 0$) is a "silly null". Indeed, in the previous example, an estimate of $\beta = 5.0$ is as plausible as an estimate of $\beta = 0.0$, and is arguable more plausible, given that we had a priori reasons to believe that the tested parameter should be important.

The full model is appropriate if the data are taken from an experiment (Burnham & Anderson 2002). This is because an experiment will be designed in order to examine all main effects as well as, potentially some of the interactions. In this case

the parameter estimates for one variable should be unaffected by the inclusion (or otherwise) of other factors.

Stepwise regression is most likely to lead to problems when it is used for data mining exercises. For example, it is common within landscape ecology studies for large numbers of predictors to be collected that are potentially associated with a particular organism or group of organisms. This is often the case when the underlying ecology of an organism is poorly known. Such studies sometimes use MAMs to reduce the list of predictors down to a manageable number. As we have shown the MAM approach will lead to errors for such datasets.

In our IT analysis we considered all possible subsets of models including these. This might be considered a large number of competing models to consider. The key issue with the dataset we explored here (and another discussed elsewhere by Whittingham *et al.* 2005) is that the variables included in the analysis represent a small proportion of the possible variables that could have been included. This subset was selected on the basis of a priori considerations (i.e. with reference to the known ecology of yellowhammers and similar farmland birds). Consequently, the analysis is not a 'shot-gun' attempt to find significant variables, but is more precisely testing the relative effects of a realistic set of candidate predictors (a form of magnitude of effects estimation, *sensu* Guthery *et al*. 2005). That this set is large is a typical problem in ecological analyses.

We have dealt in this paper with problems in formal model selection. However, a great deal of selection occurs informally in exploratory data analysis. For example, researchers may conduct preliminary analyses to reduce the set of predictors examined and reported in publications, or may use statistical tests in the exploratory phase to guide them towards the final model. This part of the analytical process is

generally not reported; however it is clear that a great deal of selection may occur prior to the final output. Such an approach (termed ‘data-dredging’ by Burnham & Anderson 2002) may suffer from all of the limitations we have outlined above, although is less straightforward to recognise or correct. It cannot be stressed enough how important it is to either specify hypotheses a priori, or to describe in detail how the final reported analysis was determined.

In summary we have demonstrated that use of stepwise multiple regression is widespread within ecology and some areas of behavioural science. We have outlined the three main weaknesses of this technique (namely: bias in parameter estimation, inconsistencies among model selection algorithms, and an inappropriate focus or reliance on a single best model) and shown how erroneous conclusions can be drawn with a worked example. We suggest that use of stepwise multiple regression is bad practice. Ecologists and behavioural scientists should make use of alternative (e.g. IT) methods or, where appropriate, should fit a full model (i.e. one containing all predictors). Full (or global) models are unlikely to be well-suited for prediction, however, and we recommend multi-model averaging techniques where prediction is the desired end.

Acknowledgements

We thank the landowners on the nine farms on which yellowhammer data was collected. We also thank BBSRC for funding the project on yellowhammers between 1994-7. MJW was supported by a BBSRC David Phillips Fellowship, and RPF by a Royal Society University Research Fellowship. Two anonymous referees provided helpful criticism of an earlier draft.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

Anderson, D.R., Burnham, K.P. & Thompson, W.L. (2000) Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, **64**, 912-923.

Bradbury, R.B., Kyrkos, A., Morris, A.J., Clark, S.C., Perkins, A.J. & Wilson, J.D. (2000) Habitat associations and breeding success of yellowhammers on lowland farmland. *Journal of Applied Ecology*, **37**, 789-805.

Burnham, K.P. & Anderson, D.R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practice information-theoretic approach*. Springer Verlag, New York.

Carver, R.P. (1978) The case against statistical significance testing. *Harvard Educational Review* 48, 378-399

Chatfield, C. (1995) *Problem solving: a statistician's guide*. Chapman & Hall, London.

Cohen, J. (1994) The earth is round (P < .05). *American Psychologist*, **49**, 997-1003.

Cohen, J. & Cohen, P. (1983) *Applied multiple regression/correlation analysis for the behavioural sciences*. Erlbaum.

Derksen, S. & Keselman, H.J. (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical & Statistical Psychology*, **45**, 265-282.

Draper, N. & Smith, H. (1981) *Applied regression analysis – 2nd edition*. John Wiley and Sons, Somerset, UK.

Ginzburg, L.R. & Jensen, C.X.J. (2004) Rules of thumb for judging ecological theories. *Trends in Ecology & Evolution*, **19**, 121-126.

Grafen, A. & Hails, R. (2002) *Modern Statistics for the Life Sciences*. Oxford University Press, Oxford.

Guthery, F.S. Brennan, L.A. Peterson, M.J. & Lusk, J.J. Information theory in wildlife science: critique and viewpoint. *Journal of Wildlife Management*, **69**, 457-465.

Hurvich, C.M. & Tsai, C.L. (1990) The impact of model selection on inference in linear regression. *American Statistician*, 44, 214-217.

Johnson, D.H. (1999) The insignificance of statistical significance testing. *Journal of*

Wildlife Management 63, 763-772.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101-108.

Johnson, C.J., Seip, D.R., & Boyce, M.S. (2004) A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, 41, 238-251.

McCullogh, P. & Nelder, J.A. (1989) *Generalized Linear Models* – 2nd Edition. Chapman & Hall, London.

Pope, P.T. & Webster, J.T. (1972) Use of an F-statistic in stepwise regression procedures. *Technometrics*, **14**, 327-340.

Post, E. (2005) Large-scale spatial gradients in herbivore population dynamics. *Ecology*, **86**, 2320-2328.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193-200.

Stephens, P.A., Buskirk, S.W., Hayward, G.D., & Martinez del Rio, C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, 42, 4-12.

Steyerberg, E.W., Eijkemans, M.J.C., & Habbema, J.D.F. (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52, 935-942.

Whittingham, M.J., Swetnam, R.D., Wilson, J.D., Chamberlain, D.E. & Freckleton, R.P. (2005) Habitat selection by yellowhammers *Emberiza citrinella* on lowland farmland at two spatial scales: implications for conservation management. *Journal of Applied Ecology*, **42**, 270-280.

Wilkinson, L. (1979) Tests of significance in stepwise regression. *Psychological Bulletin*, **86**, 168-174.

Wintle, B.A., McCarthy, M.A., Volinsky, C.T., & Kavanagh, R.P. (2003) The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17, 1579-1590.

Table 1. Proportion of studies from a range of primary ecological journals (all issues in 2004 included in this analysis) that used stepwise multiple regression for at least one component of their study. Studies using two-way ANOVA (or similar) for replicated experiments are not included as they are not really multivariate analyses that would require this approach (see discussion). Note: (1) in some cases it was not possible to determine exactly how the statistical analysis was performed, these cases are omitted from this Table. (2) *The number of studies in which it was possible to use stepwise methods is indicated in the denominator, e.g. 23 in this case, and the number that did so as the numerator, e.g. in this case 12, the remaining studies used alternative methods which are listed in the final column.

	% of studies using stepwise regression	Number of papers published by journal in 2004	Ratio of predictors to sample size for analyses using stepwise regression (no. of cases in which based given in parentheses)	Alternative approaches
Journal of Applied Ecology	52% (12/23)*	88	24 (8)	7 studies fitted full model, 1 used heirarchical partitioning and 3 used an IT approach.
Ecology Letters	58% (7/12)	139	66 (3)	4 studies fitted full model, 1 used an IT approach.
Animal Behaviour	60% (18/30)	281	9 (6)	All 12 studies fitted full model.

Table 2. Minimum adequate models constructed to explain the distribution of yellowhammers in four separate years. Data were collected from a variable number of farms in each year and these are indicated in brackets after each year. Note: (1) boundary length and a code for farm forced into all models, therefore number of predictors entered into all models was 11. * P<0.05, ** P<0.01, *** P<0.001; (2) For comparison with the results of the full model we calculated selection probabilities using IT methodology (see Whittingham *et al*. 2005). + - the model selection probability is the probability that a given predictor will appear in the AIC-best model, and is derived from the IT-AIC analysis.

	1994 (5)	1995 (5)	1996 (8)	1997 (9)	1994 - 1995	IT Selection probability+
Hedge presence	*	**			P = 0.058	0.73
Tree-line presence			*	*	***	0.67
Ditch presence	**	*		*	***	1.00
Road adjacent	*				*	0.61
Width of margin	***	*	***		***	1.00
Pasture adjacent	**		*	***	***	1.00
Silage ley adjacent						0.48
Winter rape						0.64
Beans adjacent		*				0.37
n	185	185	347	387	1103	
Ratio of sample size to predictors	21	21	32	35	123	

Figure Legends

Figure 1. Model selection bias in a simple simulation. Data were generated according to the model $y = 1 + 0.5\,x + e$, where e was an error term with zero mean and standard deviation = 1. Datasets of sample size $n = 10$ were drawn, and a linear model fitted. Fig. 1A shows the distribution of estimates of the slope parameter. The slope parameter was tested against a slope of zero, and the linear model (main text, equation 1) rejected in favour of the simpler model (main text, equation 2) if the test was non-significant (i.e. a slope of zero was accepted for $P < 0.05$). Fig. 1B shows the resultant sampling distribution based on this model selection method.

Figure 2. Cumulative probability curves for the models fitted to the data on yellowhammer distributions. The curves show the summed probabilities for the models ranked from lowest to highest AIC score. (a) Models fitted separately to the data from the four years separately (each line represents a different year). (b) Models fitted to the combined dataset. The horizontal lines show a probability of 0.95, i.e. encompassing the set of models which, under repeated sampling, would be expected to contain the AIC-best model with a probability of 0.95.