ASSESSING THE SIGNIFICANCE OF MODEL SELECTION IN ECOLOGY

EDWARD WHEATCROFT

Centre for the Analysis of Time Series, London School of Economics, London, UK, WC2A 2AE
*e-mail corresponding author: e.d.wheatcroft@lse.ac.uk

Abstract.

    Model Selection is a key part of many ecological studies, with Akaike's Information Criterion (AIC) being by far the most commonly used technique for this purpose. Typically, a number of candidate models are defined a priori and ranked according to their expected out-of-sample performance. Model selection, however, only assesses the relative performance of the models and, as pointed out in a recent paper, a large proportion of ecology papers that use model selection do not assess the absolute fit of the 'best' model. In this paper, it is argued that assessing the absolute fit of the 'best' model alone does not go far enough. This is because a model that appears to perform well under model selection is also likely to appear to perform well under measures of absolute fit, even when there is no predictive value. This paper proposes a model selection permutation test that assesses the probability that the model selection statistic of the 'best' model could have occurred by chance alone, whilst taking account of dependencies between the models. It is argued that this test should always be performed as a part of formal model selection. The test is demonstrated on two real population modelling examples of ibex in northern Italy and wild reindeer in Norway. In both cases, the model selection permutation test gives a highly significant result, indicating that the performance of the 'best' model is unlikely to be through chance alone. R code is provided with which to perform the tests.

    Key words: model selection; Akaike's Information Criterion; permutation test; ibex; reindeer.

INTRODUCTION

Model selection is a key element of a large body of research in ecology journals. This is particularly true in population modelling studies in which generalised linear models (GLMs) are typically tested with different combinations of potential predictor variables (Thieme 2018; Jacobson et al. 2004; Imperio et al. 2013). In a large number of cases, Akaike's information criterion (AIC), or its adjusted version for small samples AICc, is used to compare the relative performance of different combinations of variables (henceforth 'models'). The model with the most support according to the information criterion is then usually selected for further use or as a conclusion in itself. There is a noted tendency, however, to neglect to test whether *any* of the models are indeed useful or even 'significant'. After all, the best of a bad bunch of models is still a bad model. MacNally et al. (2017) studied 119 ecology papers that use information criteria to compare the performance of different models and found that only 55 included some measure of the absolute goodness of fit. The authors of that paper suggest both that some measure of absolute perfor-

mance should be shown and that a null model[1] should always be included as a benchmark with which to compare the performance of each of the candidate models.

    Whilst it is essential that some measure of the absolute goodness of fit of the 'best' model is included, it is argued here that doing so does not solve the problem entirely, due to implicit multiple testing that is not taken into account. Suppose that a statistical test measures the significance of the 'best' model, which has been determined by an information criterion or some other method of model selection. Whilst, in this case, only one formal test is actually performed, the model of interest has already been determined as one that appears to perform relatively well under model selection. Those models that achieve relatively good model selection statistics will therefore typically also obtain low p-values from statistical significance tests (e.g. AIC and likelihood ratio tests both reward mod-

---

[1] Null models aim to provide a benchmark that captures the behaviour of a set of observed outcomes under a "null hypothesis" that none of the hypothesised variables or processes of have any predictive value. In the case of generalised linear models, the null model is simply the chosen model structure with no explanatory variables.

els that place high probability on the outcomes, on average). Crucially, this is true both when the model is actually informative. i.e. would perform better than the null model out-of-sample, and when it appears to be informative only by random chance. Therefore, in the latter case, the probability that a statistical test on the 'best' model is wrongly found to be significant is inflated. In statistical terminology, this means that the probability of a type I error is increased. It is argued in this paper that multiple testing needs to be accounted for when assessing the significance of each of the models and a framework is demonstrated with which to do this.

The distinction between assessing the relative and absolute values of a set of candidate models is well known. For example, in the context of ecology, it was pointed out by Symonds and Moussalli (2011) that, in model selection, 'you can have a set of essentially meaningless variables and yet the analysis will still produce a best model'. They therefore suggest that it is 'important to assess the goodness of fit (x^2, R^2) of the model that includes all the predictors under study, arguing that 'if this global model is a good fit, then you can rest assured that the best approximating model will be a good fit also'. This approach seems somewhat ad-hoc since a global model with a large enough number of parameters will always appear to provide a good fit in-sample, regardless of how informative each of the variables are. As an explanation of why testing the significance of the 'best' model is often neglected, Burnham and Anderson (2004) suggest that, historically, it has often been assumed that there is a single 'true' model and that that model is in the candidate set. The Bayesian derivation of the Bayesian Information Criterion (BIC), for example, works under this assumption (Burnham and Anderson (2004)). If the assumption holds, with enough data, one can eventually expect to select the true model, and that model will, by definition, provide a good fit. In practice, few people believe that the 'true' model is ever likely to be a member of the candidate set.

Another suggested approach is to ensure that each variable included in the model is carefully justified *a priori*, such that only variables with a high chance of being informative are included (Burnham and Anderson 2001; Burnham and Anderson 2004). Whilst this is a sensible suggestion, it does not solve the problem since, whilst those variables may be expected to be important, *a priori*, this may not be reflected in the

models once the data have been considered.

The statistical literature on multiple testing is considerable. Perhaps the most well known approach to the problem is the Bonferroni correction which makes a simple adjustment to the significance level according to the number of hypotheses that are tested (Bonferroni (1936)). Other methodologies, such as the Bonferroni-Holm method (Holm (1979)) and Benjamini–Hochberg and Sidak corrections (Benjamini and Hochberg 1995; Šidák 1967), for example, control the order in which tests are applied to limit the number of overall tests, producing a uniformly more powerful approach. These are discussed further in section 1.3.

A weakness of the above approaches is that they assume that each of the hypotheses are independent of each other. If there is dependency between a set of hypotheses, the probability of committing a type I error in at least one of those hypotheses does not generally grow as quickly as when they are independent. Such methods are therefore too conservative in such cases, with the result that the probability of rejecting an informative model is increased, i.e. a type II error is committed (Nakagawa (2004)). To attempt to overcome this problem, the Westfall-Young procedure uses permutation tests to adjust the p-values in multiple correlated hypothesis tests, whilst taking account of the dependency between the hypotheses (Westfall, Young, and Wright (1993)). This provides a test which is far more powerful in such cases.

In ecology, it is common to define candidate models as different combinations of the same set of candidate variables in a generalised linear model (Bolker et al. (2009)). There is therefore a strong degree of dependency between the candidate models and so the Bonferroni correction is unsuitable (along with other similar procedures). In this paper, two permutation tests are proposed, which are based on the Westfall-Young procedure. The first test, referred to as the *single model permutation test*, assesses the significance of individual models on the basis of a model selection statistic. This is then extended to define another test, referred to as the *model selection permutation test*, that measures the significance of the entire model selection procedure, whilst taking into account the dependencies between the candidate models. The result of the first test is an individual p-value for each model whilst the result of the second test is a single p-value relating to the model selection procedure itself. The idea is then that, if the p-val-

ue of the model selection permutation test is smaller than the chosen significance level, the whole model selection procedure can be considered to be 'significant', that is the probability of finding a model selection statistic as good or better than that of the 'best' model by chance is small. Model selection can then go ahead with the reassurance that the performance of the 'best' models is unlikely to have occurred simply due to random chance.

The model selection permutation test proposed in this paper has been utilised in another recent paper entitled 'Effects of weather and hunting on wild reindeer population dynamics in Hardangervidda National Park' (Bargmann et al. 2020) on which the author of this paper is also named. In that paper, the test is referred to as a 'sanity check' test and a reference to this paper is provided. As such, some of the analysis from that paper is reproduced here with the primary focus here being the application of the proposed tests. Additionally, in this paper, an example is used in the form of a population modelling analysis of ibex populations in the northern part of Italy which was taken from an existing paper published in 2004 (Jacobson et al. (2004)).

Approaches to Model Selection

Model selection is a key part of many studies in a wide range of disciplines, including ecology (Johnson and Omland (2004)). The standard approach is to define a set of candidate models *a priori* and to attempt to rank them according to how well they would generalise out-of-sample. The basis of model selection techniques is that a fair comparison is needed between models with different numbers of parameters. If an extra parameter is added, the fit of the model will necessarily improve in-sample but may be 'overfitted' and will not improve out-of-sample. Model selection techniques therefore attempt to account for this issue.

Model selection techniques typically fall into two different categories. Information criteria weigh up the in-sample fit of the model with the number of parameters to be selected such that extra parameters are penalised. Cross-validation, on the other hand, divides the dataset such that parameter selection is always performed on data that are distinct from those on which the performance of the model is tested.

By far the most commonly used information criterion in ecology is Akaike's Information Criterion (AIC) and its corrected version for small samples

AICc (Akaike (1974); Wagenmakers and Farrell (2004)). AIC is given by $AIC = -2log(\hat{L}) + 2K$ where $\hat{L}$ is the maximised likelihood and K is the number of parameters selected. In each case, the model with the lowest AIC is considered to be the most appropriate when applied out-of-sample. For small sample sizes, AIC is slightly biased and thus a corrected, unbiased, version is often used.

The exact nature of AICc depends on the setting but, under the assumption that the model is univariate, linear and has Gaussian residuals (Claeskens and Hjort (2008)), it is defined by

$$AICc = -2log(\hat{L}) + \frac{2K(K + 1)}{n - K - 1}.$$

Cross-validation takes a different approach to model selection. Here, the data are divided into two sets: a training set, over which the parameters are selected, and a test set, on which the model is tested with those parameters. The process is then repeated with different subsets of the data set used as the training and test sets. Under leave-one-out cross-validation, the test set consists of a single point and the training set consists of each of the remaining points. This process is repeated such that each data point forms the test set exactly once. Leave-one-out cross-validation can be used alongside any method of forecast evaluation and, in this paper, is performed alongside the ignorance score (see appendix A for details) such that the forecasts can be evaluated probabilistically. In fact, this approach can be shown to be asymptotically equivalent to AIC but will usually be expected to give a different ordering of models for finite sample sizes (Stone (1974)). For a discussion on the relative merits of point forecasting and probabilistic forecasting, see appendix C.

Permutation Tests

A permutation test is a nonparametric statistical test in which the significance of a test statistic is obtained by calculating its distribution under all different permutations of the set of observed outcomes. For example, a permutation test for the slope parameter of a simple linear regression would be performed by permuting the positions of the *y* values (the dependent variable), keeping the *x* values (the predictor variables) in their original positions and calculating the slope parameter under all possible combinations

of y. The position of the slope parameter that has been calculated from the data in their original positions would then be compared to this distribution to calculate a p-value. In practice, it is often computationally prohibitive to consider all possible permutations and thus permutations are randomly chosen a fixed number of times. Such tests are called *randomised permutation tests*. Permutation tests have a number of advantages over standard parametric tests. Unlike the latter, no assumptions about the distribution of the test statistic under the null hypothesis are required since the method draws from the exact distribution (though it is assumed that the data points are exchangeable). Permutation tests thus give an exact test and, as such, randomised permutation tests are asymptotically exact. The general nature of permutation tests allows them to be applied in a wide range of settings without knowing the underlying sampling distribution. In this paper, two types of permutation test are demonstrated. The first assesses the significance of a single model without taking into consideration the other models in the model selection procedure whilst the second assesses the significance of the entire model selection procedure and thus takes into account multiple testing.

Multiple Testing

The problem of multiple testing is well-known and has been widely studied. Remedies to the problem typically involve adjustments to the p-value of each hypothesis test to reflect the number that are tested. Much of the literature on multiple testing focuses on controlling the *familywise error rate* (FWER) α_f, defined as the probability of wrongly rejecting at least one of the hypotheses. Whilst, under standard hypothesis testing, α_f usually grows with the number of hypotheses, the aim here is usually to limit the FWER to α_f. Perhaps the most common approach to the problem is the Bonferroni correction which adjusts the required significance level for each test to $\alpha_{f\,m}$ where m is the number of hypotheses tested. A major weakness of the Bonferroni correction, however, is that it assumes that each of the significance tests are independent of each other. When this is not the case, the test is too conservative and the true FWER is less than α_f, resulting in a loss of power. Several modifications to the Bonferroni correction, such as the Bonferroni-Holm (Holm (1979)), Benjamini-Hochberg (Benjamini and Hochberg (1995)) and Sidak (Šidák (1967)) corrections, have been proposed that aim to increase the power by adjusting the

order in which hypotheses are considered. None of these approaches take into account dependency between hypotheses, however.

An alternative approach to multiple testing was proposed by Westfall and Young in 1993 and aims to account for dependency between tests (Westfall, Young, and Wright (1993)). The approach makes use of permutation tests by randomly permuting the outcomes and calculating adjusted p-values for each hypothesis. An adjusted p-value for the *ith* hypothesis is given by...

$$\tilde{p}_i = Pr\left(\min_{1\leq j\leq m} P_j < p_i \vee H_0^C\right)$$

where P_i denotes the observed p-value for the *ith* test, H_0^C is the 'complete' null hypothesis that all null hypotheses are true and P_i is the p-value of the *ith* hypothesis under a given permutation of the outcomes. The adjusted p-value of the *ith* hypothesis corresponds to the probability of obtaining a p-value as small or smaller from at least one of the m hypotheses that are tested simultaneously.

Materials and Methods

In this paper, an additional step in the process of model selection is proposed. The 'model selection permutation test' assesses the probability that a statistic at least as favourable as that of the 'best' model could have occurred by chance, given the candidate models. If this probability is deemed to be below some pre-defined threshold, normal model selection should then take place and the best model(s) chosen. By taking this additional step, confidence can be had that the information contained in the models is indeed informative. First, however, a permutation for individual models is proposed. R code is provided for both tests in Appendix D and is also available at https://github.com/edjw87/Assessing-the-Significance-of-Model-Selection-in-Ecology.

A Single-model Permutation Test

A permutation test is now described with which to test the significance of individual models in a model selection procedure. The test is performed by comparing the observed model selection statistic with J draws from the distribution of that statistic under the null hypothesis that the outcomes are independent of the model predictions. An approximate p-value is calculated by counting the proportion of permutations in which the model selection statistic is smaller

(assuming a negatively oriented statistic) than the observed statistic. The single-model permutation test is formally described below:

1. Calculate the model selection statistic M.
2. Set $j = 1$.
3. Randomly permute the outcomes.
4. Calculate the model selection statistic M_j under the new ordering.
5. Set $j = j + 1$.
6. Repeat steps two to four until $j = J$.
7. Calculate the approximate p-value

$$p = \frac{1}{J} \sum_{j=1}^{J} I\left(\tilde{M}_j < M\right)$$

where I is the indicator function.

In fact, the test outlined above is a standard permutation test and, as such, is somewhat similar to that of the Westfall-Young permutation test and is asymptotically exact. However, whilst that test uses individual p-values to calculate adjusted p-values for each hypothesis (or model), the above test uses model selection statistics which do not necessarily naturally have p-values associated with them.

The single model permutation test provides a simple basis with which to assess the significance of a single model. Note that, for a single model, when the chosen model selection statistic is an information criterion, the penalty for the number of parameters is always constant and therefore the test is equivalent to performing a permutation test on the log-likelihood. However, in the next section, the test is extended to multiple models with different numbers of parameters and it is here in which the value of permutation tests for model selection statistics becomes apparent.

A Model-selection Permutation Test

A permutation test for a model selection procedure is now defined. The aim here is to estimate the probability that the 'best' model selection statistic could have occurred solely by chance, that is under the hypothesis that none of the models are informative. We call this test the *model-selection permutation test* which, again, is asymptotically exact.

Under the model selection permutation test, the outcomes are randomly permuted as they are for the single model permutation test. Here, the null hypothesis is that the model predictions are independent of the outcomes for all tested models. For a given permutation of the outcomes, the chosen model selection statistic is calculated for each model. The compari-

son of interest is between the observed 'best' model selection statistic and the statistic of the 'best' model under J permutations. The p-value is estimated by counting the proportion of permutations in which the model selection statistic of the 'best' model is more favourable than that of the 'best' model under the true ordering of the outcomes. Formally, the procedure is performed as follows:

1. Calculate the model selection statistic for each model $\tilde{M}_{j,1}, .., \tilde{M}_{j,m}$.
2. Set $j = 1$
3. Randomly permute the outcomes.
4. Calculate the model selection statistic for each model $\tilde{M}_{j,1}, .., \tilde{M}_{j,m}$.
5. Set $j = j + 1$
6. Repeat steps two to four until $j = J$.
7. Calculate an estimated p-value

$$p = \frac{1}{J} \sum_{j=1}^{J} I\left(min(M_{j,i}, .., M_{j,m}) < min(\tilde{M}_1, .., \tilde{M}_j)\right)$$
$$< min(M_1, .., M_j)\right)$$

where I is the indicator function.

Experiment One: Demonstration of Type I Error Inflation

The aim of experiment one is to demonstrate the use of the model selection permutation test in a case in which, by construction, none of the models are informative. It is shown that, as expected, the probability that the 'best' model is significant according to the single model permutation test increases with the number of candidate models. This represents, by definition, inflation in the probability of a type I error. It is then demonstrated that, for the model selection permutation test, the probability of a type I error is consistent with the prescribed significance level and is not affected by the number of candidate models. Two cases are used for illustration:

The experiment is conducted as follows: Let $y = y_{1,...}, y_{20}$ be a set of outcomes, each of which are independent and identically distributed *iid* draws from a standard Gaussian distribution $N(0,1)$. Let $X_1, \ldots, X_k$ be k predictor variables for y where each one is defined as a set of 20 random iid draws $X_i = x_{1,i}, .., x_{20,i}$ from $N(0,1)$ which are independent of y and all other predictor variables. Define a model to be some combination of predictor variables in a multiple linear regression with y as the dependent

variable. As such, none of the models have any predictive value out-of-sample and thus the null model is, by design, the optimal choice.

For each model, AICc is obtained and a p-value is then calculated for the 'best' model using the single model permutation test. In addition, a p-value for the model selection permutation test is obtained. In both cases, the number of random permutations is set to $J = 512$. Each test is deemed to be significant if its p-value is less than 0.05 (i.e. both tests are carried out at the 5 percent level).

The whole procedure is repeated 256 times, each with newly generated predictor and dependent variables. The proportion of those 256 repeats in which the 'best' model is found to be significant and the proportion in which the model selection permutation test is found to be significant is then calculated.

The following cases are considered:

1. independent models - each model is a linear regression with exactly one of $X_1, \ldots, X_n$ as a single predictor variable. There are thus n candidate models.

2. dependent models - k candidate variables $X_1, \ldots, X_k$ are defined and n distinct combinations are randomly selected, without replacement, as candidate models.

In the former case, by construction, each model is independent whilst, in the latter, since some candidate models have shared predictor variables, there is a dependency structure between models. For each value of k, there are $2^k - 1$ possible combinations of variables (excluding the null model) and n random combinations are chosen. Only values of n up to $2^k - 1$ are therefore possible. Therefore, for $n = 2^k - 1$, all of the possible combinations of variables are tested and only a random subset are tested for $n \leq 2^k - 1$.

Population Modelling Examples

Two real population modelling examples from ecology are used to demonstrate both the single model and model selection permutation tests. Both examples are published in existing papers and are presented here with the minimal details required to effectively demonstrate the methodology presented in this paper. Further details can be found in the papers themselves.

Experiment Two: Ibex

The first population modelling example was published in *Ecology* in 2004 in 'Climate forcing and density dependence in a mountain ungulate population' (Jacobson et al. (2004)). In that paper, the authors fit 20 different population models to attempt to explain changes in the ibex population of Gran Paradiso National Park in Northwestern Italy between the years of 1956 and 2000, using combinations of the following predictor variables:

- Current population.
- Snow cover.
- Interaction between snow cover and current population.

Five different combinations of the three variables were fitted with both the Modified Stochastic Ricker and Modified Stochastic Gompertz models (defined in the appendix) such that a total of ten models were assessed. In addition, the same ten combinations of variables were fitted with a threshold such that different coefficients were applied to the predictor variables when the snow density exceeds and falls below 154cm. The total number of models tested is therefore 20.

The *relative population change* in year i is defined as

$$R_i = log\left(\frac{n_{i+1}}{n_i}\right)$$

where n_i and n_{i+1} are the population counts in years i and $i + 1$ respectively. The Modified Stochastic Ricker and Modified Stochastic Gompertz models are generalised linear models such that the relative population change is modelled as a linear function of the chosen predictor variables. The Stochastic Ricker and Gompertz models differ only in the way they treat the current population size as a predictor variable.

AIC was calculated for each model based on its performance in predicting the relative population change (rather than the actual population size). Although, in that paper, AIC was the only model selection statistic considered, here, for illustration, the models are also compared using leave-one-out-cross-validation with the mean ignorance score as the evaluation method. The model selection statistics for each model are presented relative to that of the null model, i.e. with the statistic of the null model subtracted, such that a negative value indicates more support for the model than for the null model.

To demonstrate the two tests defined in this paper, the single model permutation test is performed for each model and an estimated p-value is calculated. In addition, results from the model selection per-

mutation test are shown to assess the credibility of the overall model selection procedure.

Experiment three: Reindeer

This example comes from a study of the population of wild reindeer in Hardangervidda National Park in Southern Norway (Bargmann et al., 2019). The aim of the study was to attempt to understand the factors that cause the population to change over time. This was done using a Modified Stochastic Ricker population model (defined in the appendix) with various combinations of factors as inputs. The following climatic factors were considered as potential predictors of the population:

(a) Mean temperature over January and February.

(b) Days in February/March in which the temperature exceed $0°C$.

(c) The number of summer growing degree days from June to September (days above 5 degrees Celsius).

(d) The current size of the population (density dependence).

(e) The proportion of the population hunted and killed.

(f) Interaction between proportion killed and chosen weather variable.

(g) Interaction between population size and chosen weather variable.

The winter of 2010 was significantly colder than each of the other years in the data set and was found to be an influential observation (according to Cook's distance). Given this, the analysis was performed twice: with and without that year included. The corrected version of Akaike's Information Criterion (AICc) was used to rank the performance of the models.

In this paper, the analysis is repeated and, for the purpose of demonstration, the models are also compared using the cross-validated mean ignorance score, as an alternative model selection technique. The analysis is performed with the year 2010 removed (see above). Following the original paper, a slightly different approach is taken to that of the ibex example. Whilst, in the ibex case, the performance of the models is assessed in terms of prediction of the relative population change, in this case, forecasts of the actual population counts are produced. To do this, Monte-Carlo simulation is used with a large sample and forecast distributions are produced using kernel density estimation.

Both the single model permutation test and the model selection permutation test are performed alongside both the cross-validated mean ignorance and the AICc for the original set of candidate models. The experiment is then repeated with a subset of the models to demonstrate a case in which the model selection permutation test is found not to be significant.

RESULTS

Experiment One: Demonstration of Type I Error Inflation

The results of experiment one are now presented. In figure 1, the dashed lines show, as a function of n, the proportion of repeats in which the 'best' model, as selected by AICc, is found to be significant under the single model permutation test and the solid lines show the proportion of repeats in which the model selection permutation test is found to be significant. The black line corresponds to the case in which all of the models are independent of each other and the red, green and magenta lines show the case in which there a total of 2, 3, and 4 candidate variables respectively. Both tests are performed at the 5 percent level. The grey area denotes the interval in which the proportions would fall with 95 percent probability if the
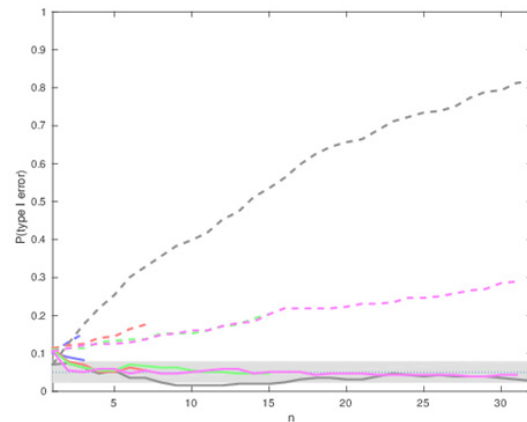

Figure 1. Estimated probability of a type I error at the 5 percent level as a function of the number of models (n) for (i) the single model permutation test (dashed lines) and (ii) the model selection permutation test (solid lines). Black lines show the results for the independent case and the blue, red, green and magenta lines show the dependent case for k (the number of independent predictor variables) equal to 1,2,3 and 4 respectively. The grey area denotes the interval in which the proportions would fall with 95 percent probability if the underlying probability of a significant result (a type I error) were truly 5 percent.

underlying probability of a wrongly significant result (a type I error) were truly 5 percent. If the proportion falls outside of this range, there is significant evidence that the probability of wrongly rejecting, and therefore committing a type I error, is different to the prescribed significance level.

As expected, as the number of candidate models is increased, the probability of a type I error for the 'best' model is inflated beyond the prescribed significance level. Note, however, how dependency between models decreases this probability. This is because fewer predictor variables are considered and

therefore the probability of finding one that happens to be 'significantly' correlated with the outcomes by chance is reduced. This demonstrates the impact dependency between models can have and makes clear the importance of taking this into account.

Since the solid lines generally fall within the grey area, the proportion of cases in which a significant result is wrongly found for the model selection permutation test is generally consistent with the significance level of 5 percent both when the models are independent of each other (black line) and when there are dependencies between them (coloured

Table 1: AIC and cross-validated mean ignorance scores expressed relative to the null model for each ibex model. Columns headed by b, c and e indicate whether density dependence, snow cover and the interaction between the two have been included in the model respectively. The column headed by "Threshold" indicates whether a threshold model has been used. Also shown are the total number of free parameters and whether the Ricker or Gompertz model has been used (R or G). The p-value of each model from the single model permutation test when using the AIC and the cross-validated mean ignorance are shown. The models are listed in order of their AIC values.

| Model	b	c	e	Threshold	Pars	DD	AIC	p-value	Mean Ign	p-value
M11	*	*	*	Yes	7	R	-51.2	$0/(2^{16})$	-0.98	$0/(2^{16})$
M12	*	*	*	Yes	7	G	-50.6	$0/(2^{16})$	-0.94	$0/(2^{16})$
M13		*	*	Yes	5	R	-46.1	$0/(2^{16})$	-0.84	$0/(2^{16})$
M14		*	*	Yes	5	G	-45.8	$0/(2^{16})$	-0.82	$0/(2^{16})$
M15	*		*	Yes	5	R	-37.2	$0/(2^{16})$	-0.66	$0/(2^{16})$
M4		*	*	No	3	G	-36.8	$0/(2^{16})$	-0.67	$0/(2^{16})$
M2	*	*	*	No	4	G	-34.8	$0/(2^{16})$	-0.62	$0/(2^{16})$
M16	*		*	Yes	5	G	-34.7	$0/(2^{16})$	-0.58	$0/(2^{16})$
M3		*	*	No	3	R	-33.9	$0/(2^{16})$	-0.57	$0/(2^{16})$
M18	*	*		Yes	5	G	-32.4	$1/(2^{16})$	-0.52	$1/(2^{16})$
M1	*	*	*	No	4	R	-39.1	$0/(2^{16})$	-0.50	$1/(2^{16})$
M5	*		*	No	3	R	-30.9	$0/(2^{16})$	-0.56	$0/(2^{16})$
M6	*		*	No	3	G	-30.6	$0/(2^{16})$	-0.55	$0/(2^{16})$
M8	*	*		No	3	G	-30.0	$0/(2^{16})$	-0.54	$0/(2^{16})$
M7	*	*		No	3	R	-27.1	$0/(2^{16})$	-0.47	$0/(2^{16})$
M17	*	*		Yes	5	R	-24.8	$0/(2^{16})$	-0.27	$44/(2^{16})$
M9			*	No	2	R	-22.1	$0/(2^{16})$	-0.43	$0/(2^{16})$
M19			*	Yes	3	R	-20.8	$0/(2^{16})$	-0.37	$0/(2^{16})$
M10			*	No	2	G	-10.4	$34/(2^{16})$	-0.16	$368/(2^{16})$
M20			*	Yes	3	G	-10.1	$85/(2^{16})$	-0.12	$445/(2^{16})$
M0				No	1		0	-	0	-
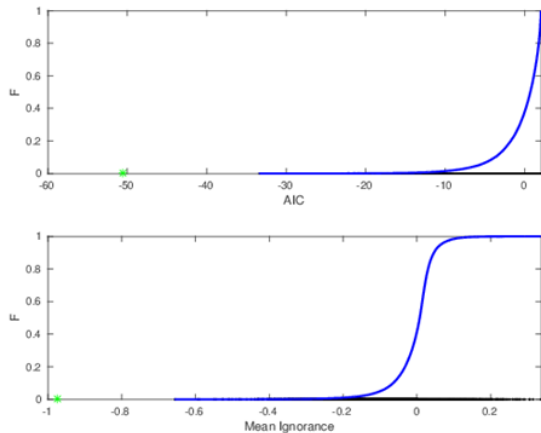
Figure 2: Top: Smallest AIC values from resampled outcomes (black dots), their CDF and the smallest AIC from the observed data (green star) for the ibex example. Bottom: the same for cross-validated mean ignorance.

lines). Unlike the approach in which the 'best' model is tested, the model selection permutation test is not impacted by these dependencies and thus we can be confident that the probability of a type I error is kept at the prescribed significance level.

Experiment Two: Ibex

The results of the model selection procedure for the ibex example are shown in table 1. Consistent with the original paper, columns headed by *b*, *c* and *e* indicate whether density dependence, snow cover and the interaction between the two, respectively, have been included in the model. It is found that almost all of the models outperform the null model under both model selection methods. Estimated p-values calculated using the single model permutation test are shown for each model based on the two model selection techniques considered. In all cases, the estimated p-values are found to be extremely small.

Although it seems unlikely that the performance of the models could be explained simply through random chance, it is rigorous to use the model selection permutation test to assess the overall significance of the model selection procedure. For both model selection techniques considered, out of $2^{16} = 16384$ permutations tested, none were found in which the 'best' model outperformed that for the observed outcomes and thus the estimated p-value is zero. A CDF of the AIC of the 'best' model (relative to the null model) under each permutation of the outcomes is shown in the top panel of figure 2 along with the

AIC of the best model from the observed data set. The equivalent, but with the cross-validated mean ignorance, is shown in the lower panel. From, these results, it is clear that it is extremely unlikely that the 'best model' in the model selection procedure occurred purely by chance. Given its strong significance, confidence can be had that the results indicate genuine predictive skill.

Experiment three: Reindeer

The results of the model selection procedure for the reindeer case are shown in table 2. Here, those variables that are included in the model are indicated with a star. The letters correspond to the variables listed in the Materials and Methods section. The AICc and mean ignorance (both shown relative to that of the null model) are shown for each model along with estimated p-values obtained from the single model permutation test.

Here, whilst a number of the models are found to be strongly significant, the p-values of those models are typically larger than for the best ibex population models in experiment two. It is therefore prudent to apply the model selection permutation test to assess the probability that the model selection statistics of the 'best' model could have occurred by chance. Here, we obtain estimated p-values of 0.0083 and 0.0016 for the AICc and cross-validated mean ignorance respectively. The fact that these p-values are small means confidence can be had that the 'best' model is indeed informative relative to the null model and did not simply occur by chance.

The reindeer example is now used to demonstrate a case in which, whilst one or more of the models is found to be significant, the probability that this occurred by chance is found to be high. Consider a model selection procedure in which the best six models according to the AICc in table 2 were not used as candidate models and therefore the selection is between the ten remaining models. At least one of the candidate models is significant at the 5 percent level for both model selection techniques. However, given the number of candidate models, caution is advised. Applying the model selection permutation test, p-values of 0.0659 and 0.1902 are obtained for the AICc and cross-validated mean ignorance respectively. Here, in both cases, the test is insignificant at the 5 percent level and there is a relatively high probability that the significance of the individual models

Table 2: AICc and cross-validated mean ignorance scores expressed relative to the null model. A star in each column headed by the letters a to h indicates whether the variables listed in the Materials and Methods section have been included in the model. Estimated p-values calculated using the single model permutation test are given for each model for each model selection technique (under the heading p(IGN) and p(AICc) for the mean ignorance and AIC respectively).

Model	a	b	c	d	e	f	g	h	Δ AICc	Mean ign.	p (IGN)	p (AICc)
M3	*				*				−15.30	−0.5864	0.0002	0.0002
M5	*				*		*		−14.04	−0.4871	0.0002	0.0005
M10		*			*		*		−12.48	−0.4224	0.0007	0.0024
M1	*								−10.89	−0.3920	0.0015	0.0010
M8		*			*				−8.54	−0.2443	0.0081	0.0171
M2	*			*					−8.35	−0.2887	0.0042	0.0042
M4	*			*		*			−6.87	−0.0184	0.0132	0.1665
M6		*							−4.85	−0.1462	0.0276	0.0305
M7		*		*					−3.94	−0.1189	0.0649	0.0449
M9		*		*		*			−2.59	−0.1226	0.1123	0.0381
M14			*	*		*			−1.47	+0.2154	0.2065	0.6255
M13			*		*				−1.21	+0.0342	0.2942	0.6755
M16				*	*			*	−0.74	−0.0215	0.5430	0.2366
M11			*						−0.03	+0.1014	0.2527	0.7715
M0									0	0	-	-
M12			*	*					+0.41	+0.1617	0.4324	0.8652
M15			*		*		*		+1.33	+0.1113	0.4573	0.6875

simply occurred by chance.

DISCUSSION

There is a clear and obvious need in ecology for authors to assess the absolute value of the 'best' model in a model selection procedure. Currently, this step is all too often completely absent. The single model permutation test defined in this paper provides a generalised approach with which to assess the significance of a model. However, by selecting the 'best' model via model selection and proceeding to evaluate its significance, the probability of a type I error can be inflated far beyond the significance level. This is because the model with the best model selection statistic has already been determined as one that per-

forms well relative to the other models, perhaps by chance.

One can imagine that, if each of the models were independent, intuition could be used to assess the impact of multiple testing. Caution would be advised if one out of a total of twenty models were significant at the five percent level, for example. The Bonferroni correction works on this basis. Commonly, in model selection in ecology, the same variables are present in multiple models. Given this dependency, this intuition is lost and therefore more formal methods are required. The model selection permutation test has been proposed for situations such as these. The test estimates the probability that the 'best' model could have occurred by chance, whilst taking the depen-

dency structure between the models into account. As such, the test gives a clear and intuitive approach to the problem of significance in model selection by assessing the entire model selection procedure.

The tests described in this paper can be used to assess whether a set of variables can provide better predictions than the null model in a population modelling procedure. Although the focus here is on ecology and, in particular, population modelling, the methodology is highly applicable to other fields in which model selection is applied. For example, in sports forecasting, one may want to determine which combination of factors most impact the probability of scoring a goal or the outcome of a game.

Whilst the tests described can help provide confidence that the best candidate variables are more informative than the null model in terms of making predictions, it should be highly stressed that, even if a model can be shown to significantly outperform the null model, it is not necessarily the case that the model is fit for a particular purpose. Before using the model, further evidence regarding the suitability of the model in a particular setting should be gathered. Nonetheless, the tests described in this paper provide a key step towards rigorous model selection in ecology which, in turn, allows for better modelling and hence a better understanding of the factors that impact animal populations.

References

Akaike, H. 1974. "A New Look at the Statistical Model Identification." *Automatic Control, IEEE Transactions on* 19 (6) (Dec): 716–723.

Bargmann, T., Wheatcroft, E., Imperio, S. and Vetaas, O.R., 2020. Effects of weather and hunting on wild reindeer population dynamics in Hardangervidda National Park. *Population Ecology*.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Bernardo, José M. 1979. "Expected Information as Expected Utility." *the Annals of Statistics*: 686–690.

Bolker, Benjamin M, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. 2009. "Generalized Linear Mixed Models: a Practical Guide for Ecology and Evolution." *Trends in Ecology & Evolution* 24 (3): 127–135.

Bonferroni, Carlo E. 1936. "Teoria Statistica Delle Classi E Calcolo Delle Probabilita: Libreria Internazionale Seeber."

Burnham, Kenneth P, and David R Anderson. 2001. "Kullback-Leibler Information as a Basis for Strong Inference in Ecological Studies." *Wildlife Research* 28 (2): 111–119.

Burnham, K. P. & Anderson, D. R. (2004), 'Multimodel inference: understanding aic and bic in model selection', Sociological methods & research. 669 33(2), 261–304.

Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. doi:10.1017/CBO9780511790485.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Good, I. J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society: Series B* 14: 107–114.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*: 65–70.

Imperio, Simona, Radames Bionda, Ramona Viterbi, and Antonello Provenzale. 2013. "Climate Change and Human Disturbance Can Lead to Local Extinction of Alpine Rock Ptarmigan: New Insight from the Western Italian Alps." *PloS One* 8 (11): e81598.

J.Bröcker, and L. Smith. 2007. "Scoring Probabilistic Forecasts: the Importance of Being Proper." *Tellus A* 22 (2).

Jacobson, Andrew R., Antonello Provenzale, Achaz von Hardenberg, Bruno Bassano, and Marco Festa-Bianchet. 2004. "CLIMATE FORCING AND DENSITY DEPENDENCE IN a MOUNTAIN UNGULATE POPULATION." *Ecology* 85 (6): 1598–1610. doi:10.1890/02-0753. http://

dx.doi.org/10.1890/02-0753.

Johnson, Jerald B, and Kristian S Omland. 2004. "Model Selection in Ecology and Evolution." *Trends in Ecology & Evolution* 19 (2): 101–108.

Mac Nally, Ralph, Richard P. Duncan, James R. Thomson, and Jian D. L. Yen. 2017. "Model Selection Using Information Criteria, but Is the 'Best' Model Any Good?" *Journal of Applied Ecology*: n/a–n/a. doi:10.1111/1365-2664.13060. http://dx.doi.org/10.1111/1365-2664.13060.

Murphy, Allan H, and Edward S Epstein. 1989. "Skill Scores and Correlation Coefficients in Model Verification." *Monthly Weather Review* 117 (3): 572–582.

Nakagawa, Shinichi. 2004. "A Farewell to Bonferroni: the Problems of Low Statistical Power and Publication Bias." *Behavioral Ecology* 15 (6): 1044–1045.

Roulston, Mark S., and Leonard A. Smith. 2002. "Evaluating Probabilistic Forecasts Using Information Theory." *Monthly Weather Review* 130: 1653–1660.

Šidák, Zbyněk. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62 (318): 626–633.

Stone, Mervyn. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society: Series B (Methodological)* 36 (2): 111–133.

Symonds, Matthew RE, and Adnan Moussalli. 2011. "A Brief Guide to Model Selection, Multimodel Inference and Model Averaging in Behavioural Ecology Using Akaike's Information Criterion." *Behavioral Ecology and Sociobiology* 65 (1): 13–21.

Thieme, Horst R. 2018. *Mathematics in Population Biology*. Princeton University Press.

Wagenmakers, Eric-Jan., and Simon. Farrell. 2004. "AIC Model Selection Using Akaike Weights." *Psychonomic Bulletin and Review* 11 (1): 192–196.

Westfall, Peter H, S Stanley Young, and S Paul Wright. 1993. "On Adjusting P-Values for Multiplicity." *Biometrics* 49 (3): 941–945.

Wheatcroft, Edward. 2015. "Improving Predictability of the Future by Grasping Probability Less Tightly." PhD thesis, The London School of Economics; Political Science (LSE).

APPENDIX

Population Modelling

The permutation tests described in this paper are demonstrated using two population modelling examples taken from existing papers. Background methodology relevant to both papers is described here. Each of the two examples make use of population models. The Modified Stochastic Gompertz Model is defined

$$R_i = a + b log(n_{i-1}) + \sum_{i=1}^{V} V_i + \varepsilon$$

and the Modified Stochastic Ricker model is defined by

$$R_i = a + bn_{i-1} + \sum_{i=1}^{V} V_i + \varepsilon$$

where n_i is the $R_i = log\left(\frac{n_{i+1}}{n_i}\right)$t in the *ith* year,

is called the *relative population change*, V_i is the *ith* explanatory variable, and ε is a random draw from a Gaussian distribution with mean zero and variance σ^2. The two models are very similar and only differ in how the current population is used as an explanatory variable (i.e. which form of so called 'density dependence' is considered). The parameters *a*, *b* and $c_1,..,c_V$ are to be selected using least-squares estimation. The Stochastic Gompertz and Ricker Models automatically give probabilistic forecasts of the relative population change in the form of of a Gaussian distribution $N(\hat{R}_i, \sigma^2)$. The forecast distribution of the relative population change can be used to estimate a forecast distribution of the actual population. In this paper, where applicable (for the reindeer case), this is done using Monte-Carlo simulation with 10,000 samples.

A 'null' model distribution naturally arises from the Modified Stochastic Ricker or Gompertz Model with all parameters except for the intercept and the variance set to zero. The null model therefore takes the form $R_i \sim N(a, \sigma^2)$ where a and σ are parameters to be selected.

Evaluating Probabilistic Forecasts

Probabilistic forecasts are usually evaluated using functions of the forecast and the outcome called

scoring rules. A wide range of scoring rules have been proposed and there is still some debate over which are the most appropriate (Gneiting and Raftery (2007)). A property of scoring rules generally considered to be of high importance is called *propriety*. A score is *proper* if it is optimised in expectation when the distribution from which the outcome is drawn is issued as the forecast (Bröcker and Smith (2007)). Propriety would therefore discourage a forecaster in possession of that forecast distribution from issuing a different one to achieve a better score. It is worth noting that no similar property exists for measures of the performance of point forecasts. For example, R^2 needn't favour a forecast based on the true distribution of the outcome.

An example of a proper scoring rule is the *ignorance score* (Good 1952; Roulston and Smith 2002) defined by

$$IGN = -log_2(p(Y))$$

where $p(Y)$ is the probability density placed on the outcome. The ignorance score is negatively oriented and hence smaller values indicate better forecast skill. The score is also *local* because it only takes the probability at the outcome into consideration (Gneiting and Raftery (2007)) and, in fact, can be shown to be the only scoring rule that is both proper and local (Bernardo (1979)). An advantage of the ignorance score is in its interpretation. The difference in the mean ignorance between two sets of forecasts can be interpreted as the base 2 logarithm of the ratio of the density placed on the outcome by each, measured in bits. For example, if the mean ignorance of one set of forecasts is 3 bits smaller than another, it places 2^3 times more probability density on the outcome, on average. The ignorance score is used in the results section of this paper alongside leave-one-out cross-validation.

Point and Probabilistic Forecasting

It was noted by MacNally et al. (2017) that authors commonly neglect to include a measure of the absolute performance of the 'best' model alongside a model selection procedure. Of those papers that do include such a measure, the vast majority were found to use R^2, adjusted R^2 or related measures. Although the low number of cases in which no absolute measure of fit is provided is concerning, those measures that *are* commonly used for this purpose can be problematic themselves. AIC and its corrected version are founded in information theory and ap-proximate the expected difference in information loss from approximating the underlying system with different candidate models. Since 'information' in this case relates to the probability or probability density assigned to the outcome, it is therefore a measure of probabilistic performance. It can be noted that generalised linear models, as commonly used in ecology, naturally provide a set of probabilistic forecasts. R^2 and similar related metrics, however, are measures of deterministic performance, i.e. they only consider a forecast to be a single number. This means that, whilst models are selected according to the performance of the resulting *probabilistic* forecasts, they are evaluated as *point* forecasts. This seems like an inconsistent approach to the forecasting problem as a whole.

In fact, probabilistic forecasts can contain a great deal of information that cannot be communicated in point forecasts. In the case of a Gaussian forecast distribution, for example, the variance can be of great value in understanding the uncertainty in the point estimate defined by the mean. For more complex forecast distributions, a single number such as the mean may be entirely inadequate. Consider, for example, a herd of terrestrial animals that, according to a probabilistic forecast distribution, is equally likely to be on either side of a large lake that runs from north to south. It is difficult, in this case, to define a single number from the distribution that represents a useful point forecast. After all, it makes little sense to predict the mean of that distribution since it would fall within the lake, an area in which there is little or no chance of the herd residing. Equally, it would make little sense to forecast that the herd will be on a particular side of the lake since each are deemed equally likely. In summary, to use a point rather than a probabilistic forecast, information must be discarded.

In addition to the issues described above, measures of the predictive performance of point forecasts tend to be fraught with problems. For example R^2, which appears to be the most commonly used measure in ecology papers, is widely known to be a poor measure of forecast performance (Wheatcroft (2015)). Firstly, the correlation between two variables does not depend on the scale on which they are given. This means that, if a set of predictions and outcomes are correlated, it doesn't necessarily mean that those predictions are 'close' to the outcomes. For example, a set of temperature forecasts measured in Fahrenheit when the outcomes are measured in Cel-

sius may still have a high R^2 value. This has been widely acknowledged and, for example, Murphy describes R^2 as a measure of *potential* rather than absolute skill (Murphy and Epstein (1989)). Secondly, other well known problems with using correlation coefficients apply. For example, influential observations can greatly increase the correlation between two variables without much, or any, actual improvement in predictive performance (Wheatcroft (2015)).

R Code

The code at the end of this section can be used to apply the two permutation tests described in this paper. It is demonstrated using the ibex example but it is straightforward to apply to any user-defined model. The code and instructions can also be found at https://github.com/edjw87/Assessing-the-Significance-of-Model-Selection-in-Ecology.

Running the code for the Ibex example

The following code can be used to apply the single model and model selection permutation tests in R. The code consists of the following two functions and a script demonstrating their usage:

1) **model_sel_permutations.R** (function to perform the single model and model selection permutation tests)

2) **Ibex.R** (function to use within model_sel_permutations.R to perform tests for the ibex example)

3) **main.R** (script demonstrating how to use the two functions to perform the tests)

The **model_sel_permutations.R** function performs the tests with a given, user defined, function. This is demonstrated using the Ibex.R function but it is straightforward to use an alternative function (instructions for doing this are described below). Further details on each of the functions is given below and their use is demonstrated in **main.R**

1) **Ibex.R**

Let n be the number of yearly observations (in this case of the relative population change), k the number of potential predictor variables and m the number of models (each one representing a different combination of predictor variables).

The function **Ibex.R** takes as arguments:

(i) *outcome* - a vector of outcomes of length n

(ii) *vars* - an nxk matrix of predictor variables

(iii) *combinations* - an mxk matrix of 1s and 0s which determines the predictor variables that are to be included in each model. Each row corresponds to a different combination of variables (model) and the number in each column determines whether that variable (as defined in the columns of "vars") is included or not (1 or 0).

(iv) *threshold_value* - a vector of length m defining a threshold value for each model on which the regression is performed. Set this to "NA" if threshold model is not used.

(v) *threshold_variable* - Variable corresponding to column in "combinations" on which the threshold is to be applied.

The function fits each model and returns a vector of AIC values with each one corresponding to a different combination of variables and thresholds (as determined by the *combinations* argument to the function).

2) model_sel_permutations.R

This is the function that performs the two permutation tests. It takes as arguments :

(i) *repeats* - a single number defining the number of permutations of the outcomes used to calculate the estimated p-values of the two tests.

(ii) *FUN* - The name of the model function (in this case "Ibex")

(iii) *pars* - a list consisting of the inputs to the model function in the order they appear. Each element of the list should be given the name of the argument. E.g., for the ibex example, we define

>**pars=list("outcome"=*Y*,"vars"=*vars*,"combinations"=*var_combinations*,"threshold_value"=*threshold_value*,"threshold_variable"=*threshold_variable*)**

For the ibex example, the tests are then run with the command:

>**output=model_sel_permutations(repeats,Ibex,pars)**

and the outputs form a list, the elements of which can be accessed by typing

>**print('Single model p-values:') >print(output$single_model_p_value)**

>**print('Model selection p-value:') >print(output$model_selection_p_value)**

Running the code with a user-defined model

The **Ibex.R** function provided is suitable for other population modelling examples using the Stochastic Ricker and Gompertz models as well as other linear generalised linear models (GLMs). However, it is also straightforward to write a function for other models by replacing Ibex.R with a user-defined function. The only requirement of the function is that

the first argument must be a vector of outcomes. The tests can then be performed by running

>**output=model_sel_permutations(*repeats*,"-function_name",*pars*)**

where "function_name" is replaced with the name of the function and *pars* is a list consisting of the inputs to the user-defined function (see the ibex example for syntax).

R functions and scripts

main.R
```
### read in necessary functions

source('Ibex.R')
source('model_sel_permutations.R')

## read ibex data file

data=read.csv('GP_data.csv')

### These are the candidate predictor
variables

postmp=which(is.na(data[,3])==FALSE)
###locations in which there are no
missing snow density values
n=data[postmp,2]  ###current popula-
tion count
x=data[postmp,3]  ###log population
count
sd=data[postmp,5]       ###snow densi-
ty
n_sd=n*sd ### interaction between snow
density and population count
x_sd=x*sd ### interaction between snow
density and log population count

### This is the dependent variable

Y=data[postmp,4] ### Vector of "rel-
ative population change".  These are
the outcomes that we want to predict.

vars=cbind(n,x,sd,n_sd,x_sd) ### cre-
ate a matrix of predictor variables
for use in Ibex.R

### define included variables in each
model

var_combinations=matrix-
(rep(0,20*5,1),20,5) ### initialise
matrix of selected variables
```

```
### Define combinations of variables
(models). One row for each combina-
tion.

var_combinations[1,]=c(1,0,1,1,0)  #m1
var_combinations[2,]=c(0,1,1,0,1)  #m2
var_combinations[3,]=c(0,0,1,1,0)  #m3
var_combinations[4,]=c(0,0,1,0,1)  #m4
var_combinations[5,]=c(1,0,0,1,0)  #m5
var_combinations[6,]=c(0,1,0,0,1)  #m6
var_combinations[7,]=c(1,0,1,0,0)  #m7
var_combinations[8,]=c(0,1,1,0,0)  #m8
var_combinations[9,]=c(0,0,0,1,0)  #m9
var_combinations[10,]=c(0,0,0,0,1)
#m10
var_combinations[11,]=c(1,0,1,1,0)
#m11
var_combinations[12,]=c(0,1,1,0,1)
#m12
var_combinations[13,]=c(0,0,1,1,0)
#m13
var_combinations[14,]=c(0,0,1,0,1)
#m14
var_combinations[15,]=c(1,0,0,1,0)
#m15
var_combinations[16,]=c(0,1,0,0,1)
#m16
var_combinations[17,]=c(1,0,1,0,0)
#m17
var_combinations[18,]=c(0,1,1,0,0)
#m18
var_combinations[19,]=c(0,0,0,1,0)
#m19
var_combinations[20,]=c(0,0,0,0,1)
#m20

threshold_val-
ue=c(rep(NA,10),rep(154,10))  ###
define threshold values for each mod-
el (vector with entry for each model).
Here models 11-20 are threshold models
and so a threshold is defined
threshold_variable=3 ### select vari-
able column in "vars" on which to add
threshold (scalar).  Here, we set the
threshold on the third variable (snow
density).

repeats=4096 #### define number of per-
mutations in permutation test
pars=list("outcome"=Y,"vars"=vars,"-
combinations"=var_combinations,"-
threshold_value"=threshold_val-
ue,"threshold_variable"=thresh-
old_variable) ### make a list for the
```

parameters in Ibex.R. The first argument must be the outcomes that are to be permuted

```
output=model_sel_permutations(repeats,Ibex,pars) ### run the permutation test and save them in "output" which is a list.

print('Single model p-values') ### print the results
print(output$single_model_p_value) ### single model p-values are given in the first element of the list.
print('Model selection p-value')
print(output$model_selection_p_value) ### model selection p-value is given in the second element of the list
```

Ibex.R
```
Ibex <- function (outcome,vars,combinations,threshold_value,threshold_variable) {
### function to produce the AIC of different combinations of predictor variables in a multiple linear regression
### The arguments are as follows:
### outcome - a vector of outcomes of length n
### vars - an nxk matrix of predictor variables
### combinations - an mxk matrix consisting of 1s and 0s which determines the predictor variables that are to be included in the model. Each row corresponds to a different combination of variables (model) and each column determines whether that variable is included or not (1 or 0).
### threshold_value - threshold value for selected variable
### threshold_variable - variable (the column in "vars") on which threshold is applied

output=numeric() ### initialise output vector
for (i in 1:dim(combinations)[1]){
### loop through each combination of variables (model)
    includedvars=which(combinations[i,]==1)
    subvars=vars[,includedvars] ### find which variables are to be included
```

```
    if (is.na(threshold_value[i])){
### if there is no threshold, perform regression on selected variables
        if (length(subvars)==0)
{ ### if no variables, perform regression with only a constant term and calculate AIC
            output[i]=-2*logLik(lm(outcome~1))+2*1
        } else{ ### if variables defined, perform regression and calculate AIC
            output[i]=-2*logLik(lm(outcome~subvars))+(sum(combinations[i,])+2)*2 ### Perform regression and calculate AIC
        }
    } else{ ### if threshold value is supplied, find separate parameters on each side of the threshold
        if (length(subvars)==0){
            output[i]=-2*logLik(lm(outcome~1))+2*1 ### if no variables, perform regression with only a constant term and calculate AIC
        } else{
            if (length(includedvars)>1){
                vars1=vars[,includedvars] #### pull out selected variables into a new matrix
                vars2=vars[,includedvars] #### pull out selected variables into another new matrix
            } else{
                vars1=matrix(vars[,includedvars]) #### if only one variable is selected, convert vector into a matrix
                vars2=matrix(vars[,includedvars]) ## as above
            }
            pos1=which(vars[,threshold_variable]<threshold_value[i]) ### find points below threshold
            vars1[pos1,]=0
##### put zeros into the matrix when selected variable is less than threshold
            pos2=which(vars[,threshold_variable]>=threshold_value[i]) ### find points above or equal to threshold
            vars2[pos2,]=0
```

```
##### put zeros into the matrix when
selected variable is greater than or
equal threshold
               varsnew=cbind(-
vars1,vars2) ### bind the 2 matrices
into 1 big one.
               output[i]=-2*log-
Lik(lm(outcome~varsnew))+(2*sum(combi-
nations[i,])+2)*2 ### Perform regres-
sion and calculate AIC for threshold
model
          }
     }
}

return(output)  #### return AIC values
for each combination of variables
}
```

model_sel_permutations.R
```
model_sel_permutations <- function
(repeats,FUN,pars) {
##### function to calculate p-values
from single model and model selection
permutation tests

### calculate model selection statis-
tics from the outcomes in their origi-
nal order.
STATtrue=do.call(FUN, pars)

#### Calculate model selection statis-
tics under each permutation
STATperm=numeric() ### initialise ma-
trix to store statistics from each
permutation
for (i in 1:repeats)  ### use a random
permutation of the outcomes and calcu-
late the model selection statistic
{
     pars[[1]]=sample(as.numeric(un-
list(pars[1]))) ### permute the out-
comes and
     tmp=as.numeric(do.call(FUN,
pars))+rnorm(1)*0.000001 ### apply the
user function with the permuted out-
comes (rnorm is added to break poten-
tial ties)
     STATperm=rbind(STATperm,tmp) ###
Record model selection statistics from
particular permutation of the outcomes
}

#### Calculate p-values of each model
under single model permutation test
single_model_p_value=numeric() ###
```

```
initialise vector for p-values

for (j in 1:dim(STATperm)[2]) ### loop
through each combination of variables
{
     single_model_p_value[-
j]=(which(sort(c(STATtrue[j],STATper-
m[,j]))==STATtrue[j])-1)/repeats ###
p-value for each combination of vari-
ables
}


#### Calculate p-value for model se-
lection permutation test

tmp=numeric()  ##### find minimum mod-
el selection statistic over all models
for each repeat and store in "tmp".
for (k in 1:dim(STATperm)[1]){
     tmp[k]=min(STATperm[k,])
}

model_selection_p_val-
ue=(which(sort(c(min(STATtrue),t-
mp))==min(STATtrue))-1)/repeats ###
Calculate model selection p-value

output=list("single_model_p_value"=s-
ingle_model_p_value,"model_selec-
tion_p_value"=model_selection_p_value)
### define outputs

return(output)
}
```