Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 983–991 doi: 10.1111/2041-210X.12251

FORUM

Ecologists overestimate the importance of predictor variables in model averaging: a plea for cautious interpretations

Matthias Galipaud[1]*, Mark A. F. Gillingham[1,2,3], Morgan David[4] and François-Xavier Dechaume-Moncharmont[1]

[1]Evolutionary Ecology Group, Université de Bourgogne, UMR CNRS 6282 Biogéosciences, 6 boulevard Gabriel, Dijon 21000, France; [2]Centre de Recherche de la Tour du Valat, Le Sambuc, Arles 13200, France; [3]Department of Evolutionary Genetics, IZW, Alfred-Kowalke-Str. 17, Berlin D-10315 Germany; and [4]University of Antwerp, Department of Biology-Ethology, Drie Eiken Campus, Universiteitsplein 1 Wilrijk 2610, Belgium

Summary

1. Information-theory procedures are powerful tools for multimodel inference and are now standard methods in ecology. When performing model averaging on a given set of models, the importance of a predictor variable is commonly estimated by summing the weights of models where the variable appears, the so-called sum of weights (SW). However, SWs have received little methodological attention and are frequently misinterpreted.

2. We assessed the reliability of SW by performing model selection and averaging on simulated data sets including variables strongly and weakly correlated to the response variable and a variable unrelated to the response. Our aim was to investigate how useful SWs are to inform about the relative importance of predictor variables.

3. SW can take a wide range of possible values, even for predictor variables unrelated to the response. As a consequence, SW with intermediate values cannot be confidently interpreted as denoting importance for the considered predictor variable. Increasing sample size using an alternative information criterion for model selection or using only a subset of candidate models for model averaging did not qualitatively change our results: a variable of a given effect size can take a wide range of SW values.

4. Contrary to what is assumed in many ecological studies, it seems hazardous to define a threshold for SW above which a variable is considered as having a statistical effect on the response and SW is not a measure of effect size. Although we did not consider every possible condition of analysis, it is likely that in most situations, SW is a poor estimate of variable's importance.

Key-words: Akaike Information Criterion, baseline sum of weights, Bayesian information criterion, information theory, model averaging, model selection, multimodel inference, variable importance

Introduction

Information-theoretic (IT) approaches are effective techniques for statistical inference which a growing number of ecologists have now adopted (Johnson & Omland 2004, Stephens *et al.* 2007, Garamszegi *et al.* 2009, Wheeler & Bailer 2009, Burnham, Anderson & Huyvaert 2011, Garamszegi 2011). They allow assessment and comparison of the support of several competing hypotheses whereas standard null-hypothesis testing only assesses whether the data fit to one single null hypothesis or not (Mundry 2011). Ecological questions are fundamentally complex and require many variables and their interactions to be simultaneously taken into account for inference. As a consequence, many different models involving different sets of parameters can be considered as competing hypotheses. Following IT approaches, ecologists first define the set of *R* candidate models which they recognize to be of interest regarding their biological question. These models are then ranked according to their relative support, based on the value of an information criterion. This is called the model selection procedure and is most frequently based on Akaïke Information Criterion (AIC) in ecological studies (Grueber *et al.* 2011, Aho, Derryberry & Peterson 2014, but see Bolker 2008, p. 219). Models with smaller AIC are better supported, and the model with the smallest AIC value is ranked first in the candidate set of models. However, the set of candidate models does not necessarily include a model that comprehensively describes the biological phenomenon under study, that is, it does not necessarily include all the meaningful parameters to explain the data (Link & Barker 2006, Burnham, Anderson & Huyvaert 2011). In fact, in biology, a very large number of parameters may have an effect on the response variable of interest, and

*Correspondence author. E-mail :matthias.galipaud@u-bourgogne.fr

Correction added on 9th October 2014, after first online publication: misalignment of columns in Table 2 corrected experimenters are unlikely to measure every meaningful parameter in their data sets. Also, ecological studies are in some cases very exploratory, intending to detect a statistical link between predictor variables and the response variable. This means that experimenters may unwittingly include spurious variables (i.e. variables unrelated to the response) in their candidate models.

Model selection thus only provides a way to find the model that best approximates the data within the set of models, the so-called best model. The probability that a given model is the best model is given by its model weight (Burnham & Anderson 2004, Symonds & Moussalli 2011, but see Link & Barker 2006). The weight w_i of a model i is computed as the relative likelihood of i divided by the sum of likelihoods of each of the R models: $w_i = \exp(-\Delta_i/2)/\sum_{r=1}^{R}\exp(-\Delta_r/2)$ where Δ_i is calculated as the difference in AIC between a model i and the first-ranked model (Buckland, Burnham & Augustin 1997, Burnham & Anderson 2002 p. 75). It follows that, as soon as the w_i of the first-ranked model is different from 1, there is uncertainty about which model is the best model for inference (Burnham & Anderson 2002, p. 153, Grueber *et al.* 2011). In such cases, it is recommended that interpretations are based on a set of models rather than on the first-ranked model only (Burnham & Anderson 2002, Burnham, Anderson & Huyvaert 2011, Garamszegi *et al.* 2009, Garamszegi 2011). Ecological conclusions are then drawn from the direct comparison of this set of competing models.

However, a clear interpretation of several competing models is sometimes more challenging than interpreting one model only, especially when these models include completely different sets of parameters. To avoid such difficulties, it is recommended that a consensus model including the correct set of variables and their effect size averaged across the set of competing models is built; a procedure called model averaging (Burnham & Anderson 2002, 2004, Grueber *et al.* 2011). Variables averaged effect sizes are thus calculated by summing their estimates in all or a subset of competing models (e.g. with a cumulated weight of 0·95), weighted by the w_i of each considered model (Burnham & Anderson 2004, Lukacs, Burnham & Anderson 2010). Also, the probability that a given predictor variable appears in the best model is estimated by summing the weights w_i of each model where the variable appears (Symonds & Moussalli 2011). A variable's sum of weights (SW) thus varies between 0 and 1. By extension, SW has been more generally used to measure the relative importance of predictor variables (Burnham & Anderson 2002, p. 167–169, Burnham & Anderson 2004).

Sums of weights are now standard measures among ecologists, because they sometimes see SW as a helpful alternative to *P*-values to assess the significance of predictor variables in IT analyses. Despite popular use among empiricists, little methodological attention has, so far, been paid to SW, especially regarding the exact meaning of predictor variable's relative importance in the context of IT model selection (Burnham & Anderson 2002, p. 167–169, p. 345–347). As a result, SW is subject to an increasing number of erroneous interpretations. We gathered common misconceptions about SW found in recent ecological articles (Table 1). Among those quotations, we identified four kinds of misleading statements as follows: (i) variables with intermediate SW value (e.g. SW = 0·5) are important, (ii) it is possible to define an absolute and universal SW threshold above which variables are considered as important, (iii) SW measures the probability that the variable has an effect, (iv) SW is a measure of effect size. Addressing these issues is both urgent and crucial as they call into question the very validity of many IT-based studies published in ecological journals. Using simple simulations, we illustrate the misleading nature of the four statements listed above and demonstrate that, in many cases, SW utility is limited, if not absent.

A first simulation example

We simulated a data set (sample size n=100) including one response variable y and four predictor variables, x_1, x_2, x_3 and x_4. We controlled for the correlation structure both between the response variables and predictor variables and among predictor variables using a Cholesky decomposition (Genz & Bretz 2009). This method allows one predictor variable with a strong effect to be generated together with other variables with smaller tapering effects, as recommended by Burnham & Anderson (2002 p. 89, 2004). Variable x_1 was strongly correlated to the response variable (Pearson correlation coefficient $r_{y,x_1} = 0{\cdot}70$) whereas x_2 and x_3 were, respectively, moderately ($r_{y,x_2} = 0{\cdot}20$) and weakly ($r_{y,x_3} = 0{\cdot}05$) correlated with the response y. Variable x_4 was uncorrelated with y ($r_{y,x_4} = 0{\cdot}0$). By relying on this correlation structure for the data, we wanted to simulate the condition of analysis commonly met in ecological studies. Even if ecologists have some insights about their biological system, they do not know for sure which predictor variables actually best explain the data. Similarly, they do not know whether they included a spurious variable in the set of candidate models. Our aim was to show that if one takes variables unrelated to the response into account in an analysis, it cannot easily be detected by interpreting SW. Every variable was normally distributed, and there was no evidence for collinearity between predictor variables (Freckleton 2011), as indicated by the maximum value of variance inflation factor of 1·1 (O'Brien 2007).

The majority of ecological studies use AIC or AICc as information criterion for model selection (Grueber *et al.* 2011, Aho, Derryberry & Peterson 2014, see also Table 1). Contrary to AIC, AICc includes a correction for small sample size and is now recommended over AIC (Burnham & Anderson 2002, 2004, Symonds & Moussalli 2011, but see Richards 2005, Turek & Fletcher 2012). In our simulations, we therefore conducted model selection and model averaging based on AICc. Parameters' averaged estimates and SW were calculated across all candidate models. Computations were performed using MuMIn 1.9.13 package (Bartoń 2013) for R 3.0.2 (R Core Team 2013). R code for data set simulations and model averaging procedure are available in the electronic appendix. A summary of the analysis is given in Table 2. As expected, predictor variables x_1 and x_2 occurred among best ranked models and had a SW equal or close to 1. Both variables x_3 and x_4 had intermediate SW values of 0·37. Note that x_4, a variable

Table 1. Example of misleading statements about variable's importance based on sums of weights (SWs). They all come from recent ecological papers. The information criterion (IC) for model ranking was either AIC or AICc. We also reported the sample size N, the number of variables n, and whether interaction terms were taken into account in candidate models. We have anonymized the sentences to focus on errors instead of *ad hominem* criticisms against colleagues

No	Statement	IC	N	n	Interaction
1	All traits had a substantial effect (SW > 0·3) on observed responses	AIC	1317	7	No
2	The predictor variable had an effect on the response variable (SW = 0·38)	AIC	505	5	No
3	The predictor variable obtained some support (SW = 0·37)	AICc	54	4	Yes
4	With SW of 0·45 and 0·37, we can interpret these predictor variables as having around 40% probability that they may indeed play a role in explaining the variability of the response variable	AIC	120	4	No
5	The predictor variable x_1 increased the response variable (SW = 0·53), while x_2 had only a small effect (SW = 0·31)	AIC	12	3	No
6	Five of the ten parameters were important for explaining variation in the response variable, namely x_1 (SW = 0·69), x_2 (SW = 0·68), x_3 (SW = 0·56), x_4 (SW = 0·54) and x_5 (SW = 0·52)	AICc	299	10	Yes
7	A rule of thumb for using these SW was to consider that SW > 0·95, 0·95–0·5 and <0·5 corresponded roughly to the classical p-values <0·01, 0·01–0·5, >0·05. (. . .) We estimated average coefficients for important variables (i.e., SW > 0·5)	AICc	36	10	Yes
8	Strong relationships between a predictor variable and the response variable were indicated by SW from 0·75 to 1, and moderately strong relationships were associated with SW from 0·50 to 0·74; weak relationships were indicated by SW from 0·25 to 0·49	AICc	52	5	No
9	The statistical support of each variable is expressed by SW expressing the probability that the variable affects the response (strong support is indicated by SW 0·6; weaker support is indicated by SW between 0·5 and 0·6	AICc	178	5	Yes
10	SW for x_1, x_2, x_3 and x_4 were 1·0, 0·51, 0·39 and 0·34, respectively, indicating that x_1 followed by x_2 were the two most important variables influencing the response variable	AICc	37	5	Yes
11	We determined the relative importance of each covariate based on SW across the entire model set, with 1 being the most important (present in all models with weight) and 0 the least important. Covariates were considered important if they appeared in top models (ΔAICc < 2·0) and had a relatively high SW (SW > 0·6)	AICc	675	6	Yes

Table 2. Summary of the model selection procedure applied to the first simulated data set. For each of the 16 models, we reported parameter estimates, total number of estimable parameters (k), the log-likelihood ($\log(\mathcal{L})$), AICc criterion, Δ_i=AICc$_i$–minAICc, Akaike weight (w_i) and adjusted R^2. Models are ordered in terms of Δ_i for AICc. At the bottom of the table, we reported, for the four variables (x_1, x_2, x_3, x_4) and the intercept (int), model-averaged estimates $\hat{\beta}$ with their 95% confidence interval (95%CI) and their sum of weights (SW)

	int	x_1	x_2	x_3	x_4	k	$\log(\mathcal{L})$	*AICc*	Δ_i	w_i	R^2
	−3·79	0·676	0·199			4	−105·97	220·4	0·00	0·353	0·559
	−4·60	0·678	0·201	0·076		5	−105·26	221·2	0·79	0·238	0·566
	−4·86	0·684	0·204		0·094	5	− 105·27	221·2	0·81	0·236	0·566
	−5·25	0·684	0·204	0·060	0·073	6	− 104·86	222·6	2·26	0·114	0·570
	−1·83	0·678				3	− 109·72	225·7	5·32	0·025	0·520
	−2·56	0·680		0·071		4	− 109·15	226·7	6·35	0·015	0·526
	−2·71	0·686			0·081	4	− 109·23	226·9	6·53	0·014	0·525
	−3·08	0·685		0·057	0·061	5	− 108·90	228·4	8·06	0·006	0·528
	2·73		0·208			3	− 141·32	288·9	68·53	0	0·430
	2·10		0·210	0·061		4	− 141·10	290·6	70·25	0	0·475
	4·82					2	−143·39	290·9	70·54	0	0
	2·62		0·209		0·011	4	− 141·42	291·1	70·69	0	0·341
	4·26			0·055		3	− 143·21	292·7	72·31	0	0·038
	2·19		0·210	0·064	−0·011	5	− 141·09	292·8	72·46	0	0·476
	4·84				−0·002	3	−143·39	293·0	72·67	0	0
	4·44			0·061	−0·024	4	−143·19	294·8	74·44	0	0·042
SW	1·00	1·00	0·94	0·37	0·37						
$\hat{\beta}$	−4·31	0·679	0·201	0·070	0·086						
95%CI	[–7·0;–1·6]	[0·54;0·81]	[0·06; 0·34]	[−0·06;0·20]	[−0·08;0·25]						

unrelated to the response, also appeared among best ranked models and, contrary to what is assumed in the literature (e.g. Garamszegi *et al.* 2009, Table 1, quotation 11), had a SW different from 0.

SW range of variation

We further investigated SW values by repeating the analysis described above over 10 000 independent simulations. We estimated the distributions of SW for the strongly correlated variable x_1 (Fig. 1a), the two more weakly correlated variables x_2 (Fig. 1b) and x_3 (Fig. 1c) and for x_4 which was unrelated to y (Fig. 1d). As expected, variable x_1 had a SW consistently equal to 1. However, the distribution of SW values for other variables was very large. Mean SW for x_2 was 0·9, with 95% of SW ranging from 0·49 to 1. For x_3 and x_4, mean SW were, respectively, 0·37 (95% of SW ranging from 0·25 to 0·79) and 0·36 (95% of SW ranging from 0·25 to 0·83). This means that predictor variables with different effects on the response could have similar SW values, so that SW alone is not necessarily informative about the relative importance of different variables. Besides, with x_4 having a SW consistently >0 and often reaching values above 0·5, it seems misleading to set up a threshold for SW above which a predictor variable is considered as important (e.g. Table 1, quotations 7 and 9). Below, we investigate whether larger sample sizes, alternative information criteria or smaller subsets of models used for model averaging could improve SW performances.

EFFECT OF SAMPLE SIZE

With increasing sample size, mean SW for genuine variables x_1, x_2 and x_3 increased up to 1 (Fig. 2a,b,c). However, SW for x_4 did not reach 0 and its distribution did not narrow, so that x_4 could still have SW values as important as 0·8 even for $n = 5000$ (Fig. 2d). Therefore, increasing sample size is not likely to lower the risk of overestimating the support of predictor variables when interpreting intermediate values of AIC-based SW. This is not surprising considering that AIC is an asymptotically efficient criterion, in the sense that it seeks to optimize predictive accuracy (Burnham & Anderson 2004, Aho, Derryberry & Peterson 2014). As a consequence, best ranked models tend to consist of a growing number of variables as sample size increases; more parsimonious models include more parameters as sample size increases (Johnson & Omland 2004, Link & Barker 2006, Bolker 2008). That is why SW for genuine variables reached 1. That is also why x_4 has a non-null SW: it is sometimes mistakenly considered as improving predictive accuracy. For this reason, AIC is not asymptotically optimal when the goal is to confirm or reject hypotheses (Aho, Derryberry & Peterson 2014). It follows that AIC-based SW does not seem to provide an accurate means of estimating the presence or absence of an effect of predictor variables on the response.

ALTERNATIVE INFORMATION CRITERION

Some authors discussed the relative performance of model selection based on Akaike vs. Bayesian information criteria (BIC), also known as Schwarz criterion (e.g. George 2000, Burnham & Anderson 2004, Link & Barker 2006, Bolker 2008, p. 219–220). Although BIC is much less frequently used in ecological papers (Grueber *et al.* 2011, Aho, Derryberry & Peterson 2014), it penalizes overly complex models (Link & Barker 2006, Bolker 2008) and is therefore susceptible to exclude spurious variables from better ranked models. We repeated the simulations described above using BIC in model selection and model averaging. As for AICc, SW for genuine variables in BIC-based analyses reached 1 for large sample sizes (Fig. 3a–c). However, BIC-based SW have better asymptotic properties with regard to x_4. The values of SW for this variable uncorrelated with y tend towards 0 with increasing sample size (Fig. 3d), thus increasing the experimenter's confidence about the absence of effect of x_4 on the response. This result is in

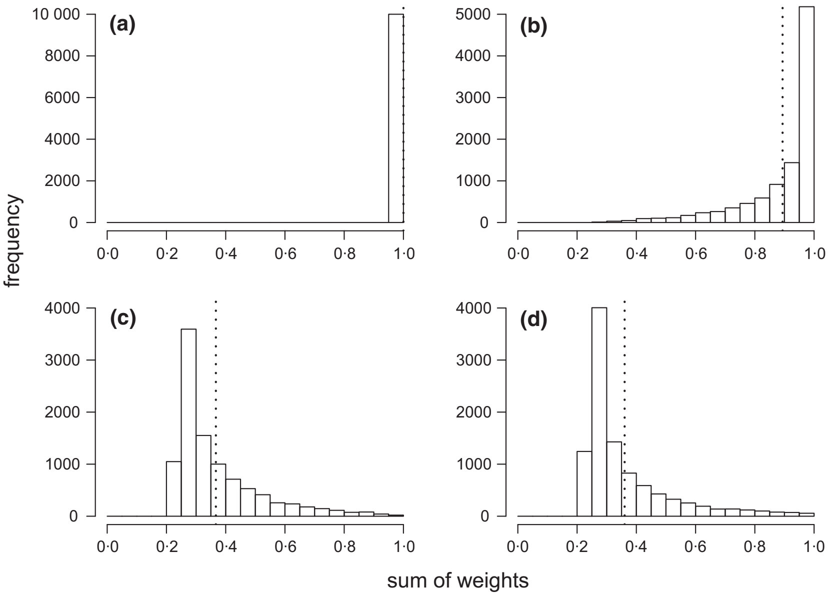

Fig. 1. Expected distribution of AICc-based SW for each predictor variable: (a) x_1, (b) x_2, (c) x_3, (d) x_4. Their distribution was estimated from 10 000 simulated independent data sets. For each data set, the sample size was $n = 100$. Predictor variables x_1, x_2 and x_3 were randomly generated to be, respectively, strongly (mean $r = 0·7$, range in sample from 0·69 to 0·71), moderately (mean $r = 0·2$, range in sample from 0·19 to 0·21) and weakly (mean $r = 0·05$, range in sample from 0·04 to 0·06) correlated with the response variable y, whereas x_4 was uncorrelated with y (mean $r = 0·0$, 95% range in sample from $-0·2$ to 0·2). The vertical dotted line shows the mean of each distribution.

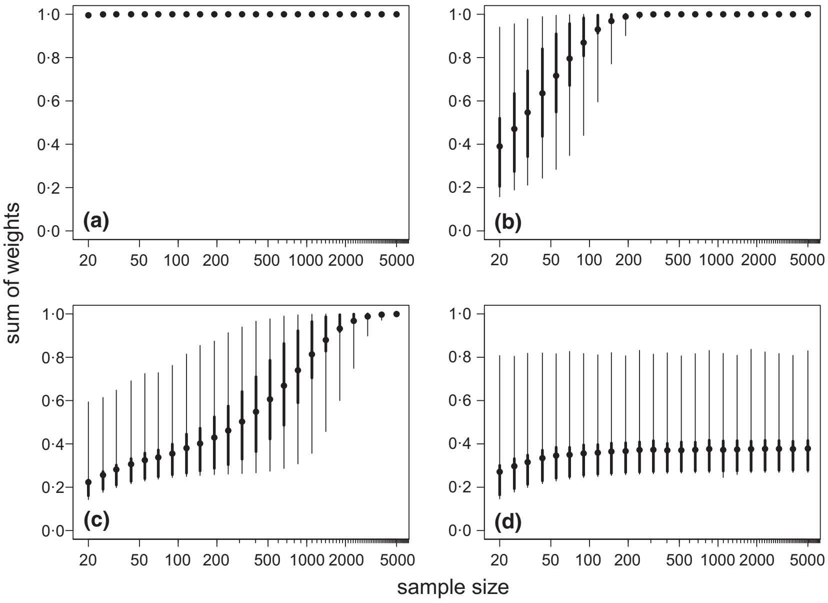

Fig. 2. Effect of sample size on the expected distribution of AICc-based SW for each predictor variable: (a) x_1, (b) x_2, (c) x_3, (d) x_4. Predictor variables x_1, x_2 and x_3 were randomly generated to be, respectively, strongly (mean $r = 0{\cdot}7$, range in sample from 0·69 to 0·71), moderately (mean $r = 0{\cdot}2$, range in sample from 0·19 to 0·21), and weakly (mean $r = 0{\cdot}05$, range in sample from 0·04 to 0·06) correlated with the response variable y, whereas x_4 was uncorrelated with y (mean $r = 0{\cdot}0$). For each sample size, we computed the mean SW (dot), interquartile interval (thick line) and 95% interval (thin line) from 10 000 simulated independent data sets.

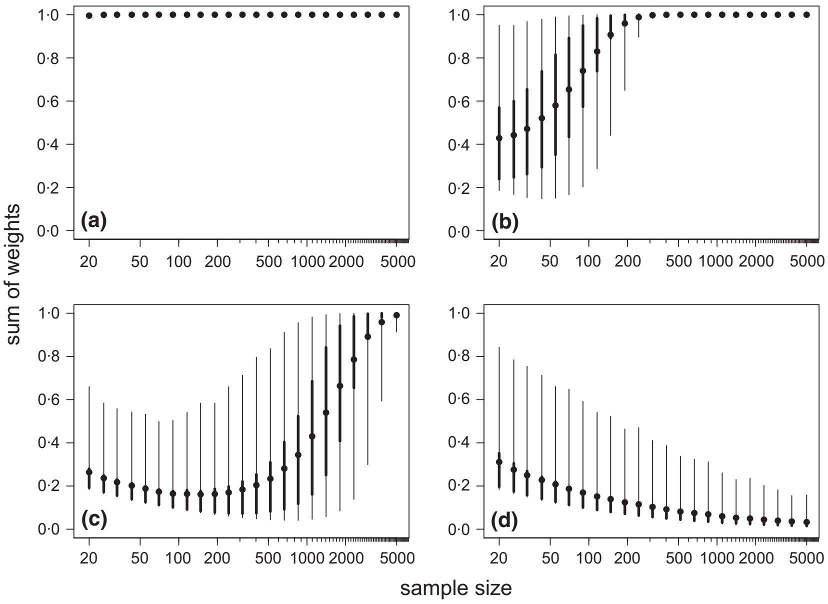

Fig. 3. Effect of sample size on the distribution of BIC-based SW for each predictor variable: (a) x_1, (b) x_2, (c) x_3, (d) x_4. Predictor variables x_1, x_2 and x_3 were randomly generated to be, respectively, strongly (mean $r = 0{\cdot}7$, range in sample from 0·69 to 0·71), moderately (mean $r = 0{\cdot}2$, range in sample from 0·19 to 0·21) and weakly (mean $r = 0{\cdot}05$, range in sample from 0·04 to 0·06) correlated with the response variable y, whereas x_4 was uncorrelated with y (mean $r = 0{\cdot}0$). For each sample size, we computed the mean SW (dot), interquartile interval (thick line) and 95% interval (thin line) from 10 000 simulated independent data sets.

accordance with the fact that BIC is asymptotically consistent, that is it selects the correct number of variables (Aho, Derryberry & Peterson 2014, Bolker 2008).

Nonetheless, for small-to-moderate sample sizes (up to $n = 500$), the distribution of both AIC- and BIC-based SW for x_4 – which was unrelated to the response – largely overlapped the distribution of SW for genuine variables (e.g. x_3). This limits the accuracy of interpretation about variables' importance for analyses based on both criteria. In addition, we generally observed a good qualitative agreement between AIC-based and BIC-based SW in our simulations. For a given sample size, these two measures appeared to be linked by a monotonic relationship (Fig. 4). For low sample sizes, there was good quantitative accordance between them. For medium-to-large sample size, the absolute values of SW differed but the ranks were preserved: the lowest AIC-based SW corresponded to the lowest BIC-based SW.

To conclude, although BIC-based SWs seem to provide more accurate measures of predictor variables' importance than AIC-based SWs in very large sample sizes, both criteria result in almost non-informative SW under sample sizes more commonly found in ecological studies (Table 1).

SMALLER SUBSET OF MODELS

Many papers addressed the issue of whether the model averaging procedure should be based on every ranked models (e.g. Grueber *et al.* 2011, Richards, Whittingham & Stephens 2011). To avoid giving too much importance to spurious variables, they argue that only a subset of best ranked models should be used for the analysis: for example, a confidence set of models based on 95% cumulated model weights, (Burnham & Anderson 2002), the subset of models with Δ_i lower than 2, 4, 6 and 10 (Richards 2008, Bolker *et al.* 2009), or a confidence

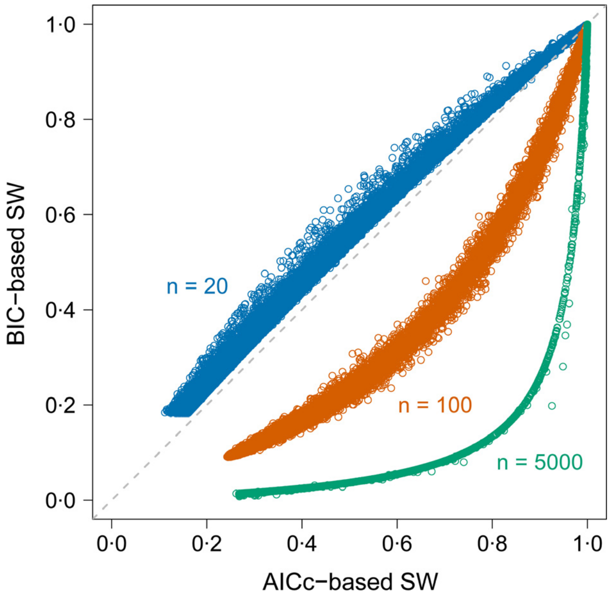

Fig. 4. Relationship between AICc-based and BIC-based SW, computed from 10,000 independent simulated data sets of sample size n = 20 (blue), n = 100 (orange) or n = 5000 (green). We represented every SW computed for each data set. The grey dashed line corresponds to the line of perfect concordance between the two measures of SW. For the three sample sizes, the Spearman coefficient of correlation was larger than 0·99.

Table 3. Comparison of the mean AICc-based SW (and their 95% range between brackets) for different subsets of models used for the model averaging. See text for explanations. These values were computed from 10 000 simulated data sets

	x_1	x_2	x_3	x_4
All models	1	0·89	0·37	0·36
	[1; 1]	[0·49; 1]	[0·25; 0·79]	[0·25; 0·83]
95% of the	1	0·93	0·35	0·33
cumulated weights	[1; 1]	[0·53; 1]	[0·17; 0·85]	[0·16; 0·90]
Δ < 2	1	0·95	0·42	0·41
	[1; 1]	[0·48; 1]	[0·17; 1]	[0·17; 1]
Δ < 4	1	0·93	0·37	0·36
	[1; 1]	[0·48; 1]	[0·19; 0·86]	[0·18; 0·88]
Δ < 6	1	0·91	0·38	0·36
	[1; 1]	[0·48; 1]	[0·24; 0·83]	[0·25; 0·84]
Δ < 10	1	0·90	0·38	0·36
	[1; 1]	[0·48; 1]	[0·25; 0·83]	[0·25; 0·84]
Without	1	0·89	0·26	0·27
nested models	[1; 1]	[0·34; 1]	[0·04; 0·86]	[0·04; 0·89]

set of models built by removing nested models complicating the first-ranked model within the subset of models with $\Delta_i < 6$ (Richards 2008, Richards, Whittingham & Stephens 2011). None of these alternative procedures qualitatively changed our findings. The range of variation of SW was still very large for predictor variables which were moderately, weakly or uncorrelated to the response (Table 3).

ADDING INTERACTION TERMS IN MODELS

Sums of weights are likely to strongly vary between studies and analyses undertaken. In our simulations, adding interactions between variables in candidate models (as commonly done in ecological studies) generally increased the SW values of predictor variables. Figure 5 shows the result of 10 000 independent repetitions of model averaging procedures applied to the four variables considered earlier and all their two-way interactions. For x_4, which was uncorrelated to the response, 95% of SW ranged from 0·49 to 0·98, Fig. 5d). Adding interactions between predictor variables increased the number of models where variables appeared, therefore artificially inflating their SW values (Dochtermann & Jenkins 2011, Grueber *et al.* 2011). Incidentally, the presence of interaction terms in candidate models severely impairs any interpretation about variable importance based on SW.

NON-INFORMATIVE SW

Due to all the reasons listed above, there are many situations where intermediate to large SW values cannot be confidently interpreted as indicating importance. Although particularly common in ecological studies, SW-based analyses combining low-to-intermediate sample sizes and interaction terms must be avoided or, at least, be interpreted with the greatest caution. We do not aim to provide the reader with all possible theoretical ranges of SW variation for given predictor variables. Many factors may influence SW theoretical distribution. A non-exhaustive list could for example include variable type (categorical or continuous), link function and distribution of the residuals (binomial, Poisson, ...) or parametrization of the models (number and effect sizes of parameters, Burnham & Anderson 2004). Instead, we believe that it is much more valuable to provide ecologists with a method allowing to assess the reliability of SW in the context of their analyses.

It has been proposed that, to assess the importance of a predictor variable, one should first estimate the SW distribution expected if this variable had no predictive value (Burnham & Anderson 2002, p. 345–347). Such a baseline SW distribution can be calculated by randomization methods. This can be done by permuting the response variable y, thus reorganizing the data as if all predictor variables considered for the analysis were uncorrelated to the response (see the R code in Data S1 in the Supporting Information for an example). Applying model averaging over a great number of permuted data sets allows estimation of a baseline SW distribution for all considered predictor variables simultaneously. Another possibility could involve permuting only one predictor variable of interest within the data set to estimate its baseline SW distribution alone while keeping all the other variables unchanged. Running 10 000 independent permutations of y on the simulated data set used in Table 2, we estimated a baseline SW distribution for each predictor variable (i.e. x_1, x_2, x_3 and x_4). This distribution is the same for the four predictor variables with 95% of baseline SW values ranging from 0·25 to 0·82. Such a large variation in baseline SW renders any interpretation about the variable's importance very difficult. Of course, a strong assumption behind the use of AIC in model selection is that

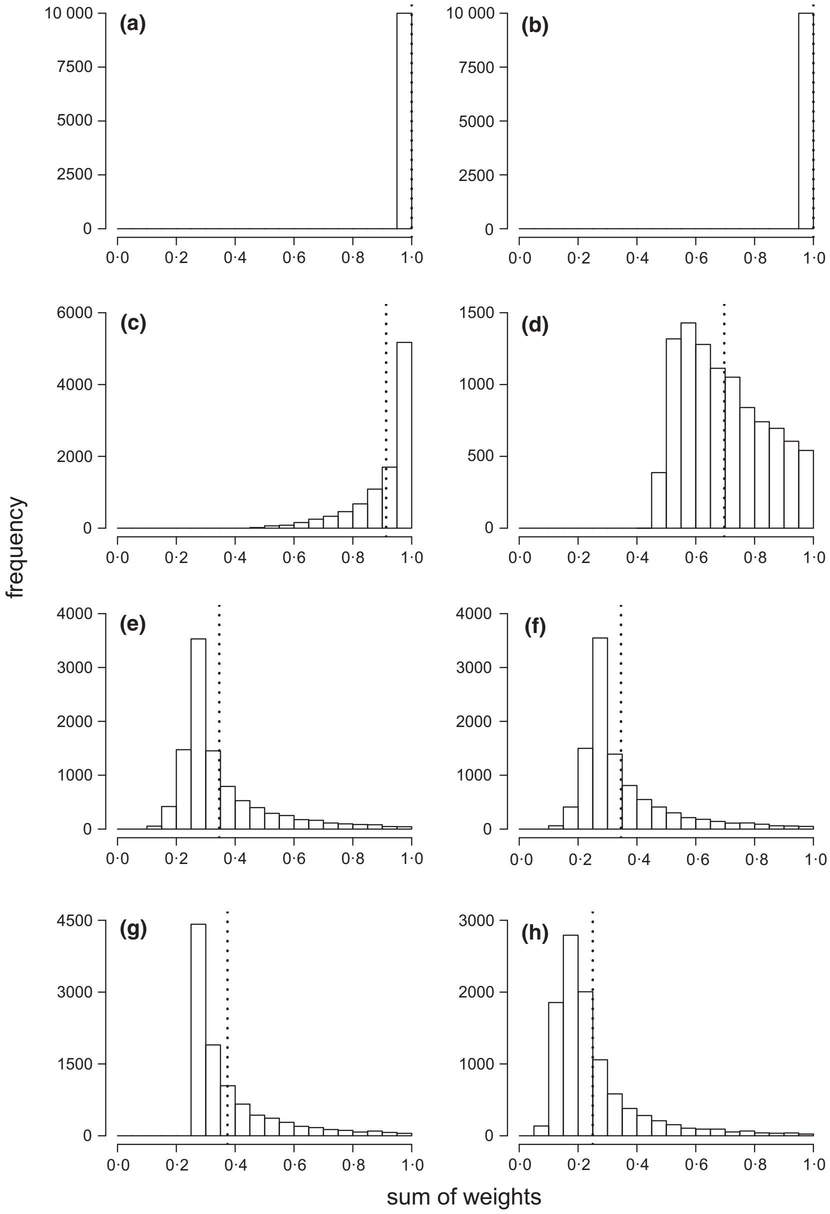

Fig. 5. Expected distribution of AICc-based SW, computed based on 10 000 simulated data sets, for the four variables and some of their interactions: (a) x_1, (b) x_2, (c) x_3, (d) x_4, (e) interaction between x_1 and x_3, (f) interaction between x_2 and x_3, (g) interaction between x_1 and x_2, (h) interaction between x_3 and x_4. For each data set, the sample size was n=1000. Predictor variables x_1, x_2 and x_3 were randomly generated to be, respectively, strongly (mean r = 0·7, range in sample from 0·69 to 0·71), moderately (mean r = 0·2, range in sample from 0·19 to 0·21) and weakly (mean r = 0·05, range in sample from 0·04 to 0·06) correlated with the response variable y, whereas x_4 was uncorrelated with y (mean r = 0·0, 95% range in sample from −0·2 to 0·2). Candidate models only include two-way interactions. The vertical dotted line shows the mean of each distribution.

every parameters considered in candidate models have an effect on the response, no matter how small it is (Burnham & Anderson 2002, Symonds & Moussalli 2011). But under real conditions of data collection and analysis, experimenters could include a spurious variable with initial conviction that it has an effect on the response (Symonds & Moussalli 2011). Acknowledging this possibility, baseline SWs, together with Δ_i, w_i and averaged parameter estimates, help to detect the conditions of analysis under which interpretations about variable's importance based on SW must be avoided, that is when baseline SWs can take a wide range of possible values (up to 1).

Discussion

The SW is one of the most common estimates of predictor variable's importance in ecological papers relying on model averaging analyses. However, experimental studies frequently fail to correctly interpret SW. In the following paragraphs, we will explain why the four statements presented in the introduction are misleading in the light of our findings.

MISCONCEPTION 1

Many authors performing AIC-based multimodel inferences consider predictor variables with SW between 0·3 and 0·7 as important (e.g. Table 1, quotations 1–6, 11). However, we showed that variables uncorrelated to the response do not have SW = 0 and, more importantly, often reach SW values similar to those of genuine variables, up to 1. Consequently, ecologists cannot interpret the importance of variables based on their SW alone.

MISCONCEPTION 2

A natural temptation is to try to define an absolute SW threshold for significant variable effect on the response (e.g. Table 1, quotations 7, 9, 11). According to some authors (e.g. Calcagno & de Mazancourt 2010), SW larger than a threshold (for instance SW > 0·8) would always denote importance, whatever the data set and candidate models considered. We believe that this is incorrect. The distributions of SW are likely to be highly variable between studies, and it is dubious that such an absolute threshold can be defined in every situation. Ecologists are trained to use 95% confidence intervals as a criterion for variable support. However, we strongly warn the reader against blindly using SW = 0·8 as a threshold just because it corresponds to the upper limit of 95% range of SW distribution in many of our simulations. Following the same reasoning, it is also hopeless to define absolute ranges of SW values denoting weak, moderate or strong support for the corresponding variable (e.g. Table 1, quotations 7–9). We must insist on the fact that our aim here is not to introduce binary thinking into IT. The upper limit of the 95% range of SW distribution should not be used as a threshold to dichotomize between significant and non-significant variables. Such a binary distinction is in conflict with the IT philosophy.

MISCONCEPTION 3

For several authors, SW is a measure of the probability for the corresponding variable to play a role in explaining the data (e.g. Table 1, quotations 4, 11). This formulation is misleading and should be avoided. In its pure mathematical sense, SW is indeed a probability for the corresponding variable to appear in the best model (Burnham, Anderson & Huyvaert 2011, Symonds & Moussalli 2011). However, SW should not be taken as a probability that the variable has a statistical effect. If it was so, the absence of effect would be associated with SW = 0. This was hardly ever the case in our simulations, except when using BIC as the criterion for model selection and averaging with very large sample sizes.

MISCONCEPTION 4

Some statements about variable importance are ambiguous when implying that SW could represent a measure of effect size (e.g. Table 1, quotations 1, 5, 8). In the same way that a *p*-value is not a measure of effect size (Nakagawa & Cuthill 2007, Mundry 2011), large SW should certainly not be interpreted as denoting large effects. Accordingly, referring to SW as a measure of variable *relative importance* (as it is the case in the ecological literature, see also Table 1, quotation 11) is somewhat confusing. For instance, SW for x_2 and x_3 strongly increased with increasing sample size even though we simulated these variables as having constant effect sizes. Increasing SW values only means that we are indeed much more certain about the presence of x_2 and x_3 in the best approximating model with increasing sample size. In contrast to SW, unstandardized measures of effect size are provided by averaged parameter estimates (Nakagawa & Cuthill 2007). Although some authors pointed out possible biases in relying on model-averaged estimates for inference (Richards 2005, Richards, Whittingham & Stephens 2011), others believe that, for sound interpretations, they should always be preferred to parameter estimates of the first-ranked model (Burnham & Anderson 2002, Garamszegi *et al.* 2009, Burnham, Anderson & Huyvaert 2011, Symonds & Moussalli 2011). In addition, unbiased calculations of 95% confidence intervals around averaged estimates (Fletcher & Dillingham 2011, Turek & Fletcher 2012) help to interpret the uncertainty in parameters estimation and thus to understand the uncertainty in variable's importance.

Beyond model averaging, interpretations based on SW should only be carried out if candidate models in the selected set fit the data. A model weight does not give, in itself, an absolute evaluation of how informative the model is with regard to the observed variation in the data. We noted that papers relying on IT approaches seldom report deviance or variance explained by candidate models, although it allows assessment of models fits and provides absolute measures of their explanatory power. For example, R^2 is a metric that ecologists are already trained to report and interpret. A model selection may lead to models with large weights and predictor variables with SW values close or equal to 1 even if those models have very low R^2. In that case, because the fit of the models under consideration is poor, little can be learned despite large model weights and variable SW (Burnham, Anderson & Huyvaert 2011, Symonds & Moussalli 2011). Even if one cannot base the model selection procedure on the analysis of R^2 (Burnham & Anderson 2002, p .95), authors should systematically report R^2 values alongside their models.

To conclude, there is an inherent need among ecologists to assess the importance of predictor variables. In IT, this role is often played by SW. However, interpreting SW should be carried out with particular caution because a given variable can take a wide range of SW values under sample sizes generally found in ecological studies. This implies that spurious variables unwittingly included in the analysis may not easily be excluded from an averaged best model for inference or may not be detected as having no effect on the response.

Acknowledgements

We thank François Rousset and Jacques Labonne for early feedback on the research undertaken as well as Bob O'Hara and two anonymous referees for constructive comments on the manuscript. Calculations were performed using HPC resources from DSI-CCUB (Université de Bourgogne).

Data accessibility

R script for simulations is provided as Supporting Information.

References

Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95**, 631–636.

Bartoń K. (2013) *MuMIn: Multi-Model Inference, R Package Version 1.9.13*. CRAN http://CRAN.Rproject. org/package=MuMIn.

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

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Buckland, S.T., Burnham, K.P. & Augustin, N.H. (1997) Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.

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

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, **33**, 261–304.

Burnham, K.P., Anderson, D.R. & Huyvaert, K.P. (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, **65**, 23–35.

Calcagno, V. & de Mazancourt, C. (2010) glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, **34**, 1–29.

Dochtermann, N.A. & Jenkins, S.H. (2011) Developing multiple hypotheses in behavioral ecology. *Behavioral Ecology and Sociobiology*, **65**, 37–45.

Fletcher, D. & Dillingham, P.W. (2011) Model-averaged confidence intervals for factorial experiments. *Computational Statistics and Data Analysis*, **55**, 3041–3048.

Freckleton, R.P. (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, **65**, 91–101.

Garamszegi, L.Z. (2011) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. *Behavioral Ecology and Sociobiology*, **65**, 1–11.

Garamszegi, L.Z., Calhim, S., Dochtermann, N., Hegyi, G., Hurd, P.L., Jargensen, C., *et al*. (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. *Behavioral Ecology*, **20**, 1363–1375.

Genz, A. & Bretz, F. (2009) *Computation of multivariate normal and t probabilities*. Lecture Notes in Statistics (vol. 195). Springer, New York.

George, E.I. (2000) The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Grueber, C.E., Nakagawa, S., Laws, R.J. & Jamieson, I.G. (2011) Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*, **24**, 699–711.

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

Link, W.A. & Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87**, 2626–2635.

Lukacs, P.M., Burnham, K.P. & Anderson, D.R. (2010) Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, **62**, 117–125.

Mundry, R. (2011) Issues in information theory-based statistical inference - a commentary from a frequentist's perspective. *Behavioral Ecology and Sociobiology*, **65**, 57–68.

Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605.

O'Brien, R.M. (2007) A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, **41**, 673–690.

R Core Team (2013) *R: A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards, S.A. (2005) Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology*, **86**, 2805–2814.

Richards, S.A. (2008) Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, **45**, 218–227.

Richards, S.A., Whittingham, M.J. & Stephens, P.A. (2011) Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioral Ecology and Sociobiology*, **65**, 77–89.

Stephens, P.A., Buskirk, S.W. & Martínez del Rio, C. (2007) Inference in ecology and evolution. *Trends in Ecology and Evolution*, **22**, 192–197.

Symonds, M.R.E. & Moussalli, A. (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**, 13–21.

Turek, D. & Fletcher, D. (2012) Model-averaged Wald confidence intervals. *Computational Statistics and Data Analysis*, **56**, 2809–2815.

Wheeler, M.W. & Bailer, A.J. (2009) Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*, **16**, 37–51.

Received 17 June 2014; accepted 24 July 2014
Handling Editor: Robert B. O'Hara

Supporting Information

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

Data S1. R script for simulations.