Received: 10 March 2017 | Accepted: 1 June 2017

DOI: 10.1111/2041-210X.12835

FORUM

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

A farewell to the sum of Akaike weights: The benefits of alternative metrics for variable importance estimations in model selection

Matthias Galipaud[1] | Mark A. F. Gillingham[2] | François-Xavier Dechaume-Moncharmont[3]

[1]Department of Evolutionary Biology, Bielefeld University, Bielefeld, Germany

[2]Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm, Ulm, Germany

[3]Ecology Evolution Team, UMR CNRS 6282 Biogéosciences, University of Bourgogne-Franche-Comté, Dijon, France

Correspondence
Matthias Galipaud
Email: matthias.galipaud@uni-bielefeld.de

Funding information
DFG, Grant/Award Number: DFG Gi 1065/2-1

Handling Editor: Robert Freckleton

Abstract

1. In a previous article, we advocated against using the sum of Akaike weights (SW) as a metric to distinguish between genuine and spurious variables in Information Theoretic (IT) statistical analyses. A recent article (Giam & Olden, *Methods in Ecology and Evolution*, 2016, 7, 388) criticises our finding and instead argues in favour of SW. It points out that (1) we performed a biased data-generation procedure and (2) we erroneously evaluated SW on its capacity to estimate the proportion of variance in the data explained by a variable. We here respond to these points.
2. Giam and Olden's first concern is unfounded. When using the data-generating code they proposed, SW remains very imprecise. To respond to their second concern, we first list the meanings taken by a variable's importance in the context of IT. Although, SW is presented as an estimate of variable relative importance in methodological textbooks (i.e. a variable's rank in importance or its relative contribution to the variance in the data), it is also used as a metric of variable absolute importance (i.e. a variable's absolute effect size or its statistical significance). We then compare SW to alternative metrics on its ability to estimate variable absolute or relative importance.
3. SW values have low repeatability across analyses. As a result, based on SW, it is hard to distinguish between variables with weak and large effects. For estimations of variable absolute importance, experimenters should prefer model-averaged parameter estimates and/or compare nested models based on evidence ratios. Sum of Akaike weights is also a poor metric of variable relative importance. We showed that correct variable ranking in importance was generally more frequent when using model-averaged standardised parameter estimates, than when using SW.
4. To avoid recurrent errors in ecology and evolution, we therefore warn against the use of SW for estimations of variable absolute and relative importance and we propose that experimenters should instead use model-averaged standardised parameter estimates for statistical inferences.

KEYWORDS

Akaike information criterion, effect size, evidence ratio, model-averaging, multi-model inferences, standardised parameter estimates, variable criticality, variable ranking

1 | INTRODUCTION

Information-theoretic (IT) statistical approaches are increasingly used by ecologists. Correspondingly, a number of methodological articles and books have been published to ease their application by empiricists (Burnham & Anderson, 2002; Garamszegi et al., 2009; Grueber, Nakagawa, Laws, & Jamieson, 2011; Behavioural Ecology and Sociobiology special issue "Model selection, multimodel inference and information-theoretic approaches in behavioural ecology" volume 65, 2011). Our own methodological contribution to IT outlines misconceptions found in ecology papers about using the sum of Akaike weights (SW), a commonly proposed estimate of the importance of predictor variables in IT (Galipaud, Gillingham, David, & Dechaume-Moncharmont, 2014). Our article was recently criticised by Giam and Olden (2016). We here address these criticisms. By doing so, we also attempt to clarify the debate over variable relative and absolute importance estimations in IT analyses. We believe that a great deal of misinterpretations of SW originates from a lack of clear definitions in textbooks and methodological articles about what it estimates.

Experimenters can have two different aims when estimating a variable's *absolute* importance. First, it can be related to making dichotomous decisions about a sampled predictor variable's statistical effect on the response variable, in which case it is essentially a problem of variable selection. Experimenters thereby aim at building a parsimonious model for later predictions; that is a model which includes a limited set of variables that optimises predictions by avoiding both under- and overfitting (Burnham & Anderson, 2002). Alternatively or concomitantly, they aim at identifying potential causal relationships in the data, hence forming new hypotheses about the biological role of sampled variables in affecting observed phenomena (Mac Nally, 2000; Stephens, Buskirk, Hayward, & Martínez del Rio, 2005). Second, assessing a variable's absolute importance can be related to estimating the magnitude of its biological effect or the strength of its relationship with the response variable (Nakagawa & Cuthill, 2007). We henceforth refer to it as an estimation of a variable's absolute effect size. The focus of interpretations about variable importance in that context shifts from dichotomous decisions about the presence or absence of a variable's statistical effect to more tempered, biologically sound inferences about how much variables affect a biological phenomenon in the sampled population (Schielzeth, 2010). When estimating a variable's *relative* importance, users are concerned about comparing the importance of variables considered in the set. They thereby aim at testing hypotheses regarding the relative contribution of sampled variables in explaining an observed phenomenon (Johnson & LeBreton, 2004). The information provided by metrics of variable relative importance can be of two natures. First, it can help ranking variables in importance. In such analyses, the variable relative importance metric assigns to each variable a rank number between 1 and *k* (the number of considered variables to rank). Second and more informative are metrics which tell something about how many times more or less important variables are than one another. When they estimate variable's proportionate contribution to the variance in the response, such metrics are referred to as variables *relative weight* or *dispersion importance* (Johnson, 2000; Johnson & LeBreton, 2004). For simplicity, we henceforth generally refer to such metrics as estimates of relative effect size.

Estimations of absolute and relative importance of variables are not independent from one another and they sometimes need to be performed concomitantly for reliable statistical inferences and predictions (Azen, Budescu, & Reiser, 2001). Specifically, the accuracy and precision of an estimator of relative importance directly depends on the number *k* of predictor variables considered in the regression model relative to the sample size *n* (Budescu, 1993; Burnham & Anderson, 2002). As a rule of thumb, when *n* is less than 10 times the value of *k*, chances are that the model over-fits the data, leading to much variation in variables' relative importance estimations (Peduzzi et al., 1996). Conversely, when *k* is very small relative to *n* (e.g. only one or two parameters are considered to model 1,000 observations), the experimenter might have forgotten important predictor variables. The considered model therefore possibly under-fits the data, to the extent that estimates of relative importance are biased. This exemplifies the inherent link between issues of selecting the right set of predictor variables among all possibilities and estimating accurately and precisely the relative importance of selected variables.

Ecologists are accustomed to dichotomous choices in traditional null-hypothesis testing, and they have thus often used SW for estimations of absolute variable importance in IT analyses (Table 1 in Galipaud et al., 2014). We herein expose drawbacks of SW as both a metric of absolute and relative importance and review alternative methods for statistical inferences. We thereby respond to Giam & Olden's comments and reiterate our plea against the use of SW. This article is an additional contribution to the debate initiated in the forum section of *Methods in Ecology and Evolution* (Galipaud et al., 2014; Giam & Olden, 2016). We therefore only briefly re-introduce IT model selection procedure, to concentrate on the specific issue of estimating variable importance.

2 | INFORMATION-THEORETIC MODEL SELECTION

In ecology and evolution, IT approaches are mainly used as a model selection procedure when several predictor variables are considered to model the response (Burnham & Anderson, 2004; Lukacs, Burnham, & Anderson, 2010; Stephens et al., 2005; Whittingham, Stephens, Bradbury, & Freckleton, 2006). Classically, there can be two, quite different, aims to model selection. (1) It can be used for predictions: it helps identifying a parsimonious set of parameter estimates that models the response, limiting both variance and bias in estimation and allowing reliable predictions of unknown outcomes (Stephens, Buskirk, & Martínez del Rio, 2007). (2) Instead of (or alongside) predictions, model selection can be used for statistical inferences: it helps discriminating between important and unimportant variables in the sample, hence inferring major (potentially causal) influences of a set of variables on the response (Mac Nally, 2000; Stephens et al., 2005). In both cases, rather than estimating variables relative importance, users need to select important variables to include in a best model (or a set of good models) to use for interpretations. Such variable selection is often performed using so-called stepwise selection and *p*-values as a metric of variables absolute importance. The drawbacks of this procedure are now well-exposed and methodological studies strongly suggest using IT model selection instead (Whittingham et al., 2006).

In IT model selection, variable importance is commonly estimated using SW (Burnham & Anderson, 2002; Burnham, Anderson, & Huyvaert, 2011). To understand how this metric is calculated, let us briefly summarise the IT model selection procedure. Given that full ecological reality is, at best, only approximated using statistical modelling, IT approaches compute, for each model considered, the amount of information loss between the model and reality. This information loss, termed the relative Kullback-Leibler information (henceforth RKLI), can be estimated using a bias-corrected maximum likelihood estimation for parameters in each considered model: this estimate is the Akaike information criterion (henceforth AIC, Akaike, 1973; see Burnham & Anderson, 2002, pp. 58–64 for a more complete description of AIC mathematical derivation). According to AIC, the estimated best model for inferences therefore minimises information difference to full ecological reality. Because full reality is unknown, an AIC value in itself carries no information about how close is any particular model to reality independent of others. However, by estimating differences in models RKLI, AIC provides a way to compare models in terms of their relative strength of evidence. AIC is an estimation, and the model with the smallest AIC can substantially vary across samples. A major benefit of IT over alternative approaches for model selection is that it accounts for such uncertainty (Burnham & Anderson, 2002; Burnham et al., 2011). After ranking models according to AIC (or its derivatives AICc, QAIC etc.), users calculate for each model its Akaike weight as the weight of evidence that it is the RKLI best model (Burnham & Anderson, 2002; p. 75). If several models have similar weights, IT model selection therefore allows inferences and predictions to rely on a set of alternative models of comparable support rather than on one-first ranked model. Because basing interpretations on multiple models simultaneously is not straightforward, it is often recommended to average variable parameter estimates over models in the set; a method called model-averaging (Burnham & Anderson, 2002). Two different types of model-averaging are used in IT model selection. Under the full model-averaging technique, parameter estimates are multiplied by the Akaike weight of their corresponding model (Lukacs et al., 2010). In models where the variable is absent, its estimate is set to 0. As a result, full model-averaged estimates shrink towards 0 (Symonds & Moussalli, 2011). On the contrary, the natural model-averaging technique involves averaging each variable's parameter estimates only over the models where the variable appears (Symonds & Moussalli, 2011). For this calculation, Akaike weights of models where the variable appears are first rescaled to sum up to 1 before being multiplied by the variable parameter estimates. To facilitate multi-model interpretations, a variable's SW (also called parameter weight) can also be calculated as the sum of Akaike weights of models where the variable appears. In textbooks and methodological papers, SWs are described as estimates of variables' relative importance (Burnham & Anderson, 2002; pp. 168, 281-282). However, that SW is an estimate of variable relative rather than absolute importance remains ambiguous in the ecological literature. Very influential methodological articles and textbooks have presented Akaike weights as the probability that a given model is the best RKLI model (e.g. Burnham & Anderson's, 2002 textbook is cited more than 36,000 times and the articles of Symonds & Moussalli, 2011; Burnham et al., 2011 published specifically to target an ecologists audience cumulate above 1,600 citations). As a consequence, SW is also largely defined as the probability that the variable of interest is included in this best model, independent of other variables in the set (Symonds & Moussalli, 2011). Such estimates of variable importance are also referred to as measures of variable *criticality*; variables with large SW values are considered critical to parsimonious predictions (Azen et al., 2001; Johnson & LeBreton, 2004). Under that definition, SWs appear primarily useful for variable selection rather than relative importance estimations, as they supposedly provide means of deciding whether or not to consider any given variable (and their corresponding parameter estimates) in a RKLI best model for statistical inferences and predictions. This ambiguity has certainly led ecologists accustomed to using *p*-values to detect significant effects in stepwise model selection to interpret SW as an estimate of variable importance in absolute rather than relative terms. Specifically, variables with a high probability to appear in the RKLI best model are often interpreted as having a statistically significant effect on the response (see Table 1 in Galipaud et al., 2014).

3 | THE CURRENT DEBATE OVER SW RELIABILITY

Using simulated datasets, we showed that SW of variables with given effect sizes vary widely among simulations, even for large sample sizes, under different model parametrisations and using different model ranking procedures, making them imprecise metrics for statistical inferences about variables importance (Galipaud et al., 2014). In their article, Giam and Olden (2016) make two main criticisms to our study. (1) They point out that the way we generated data may have biased our results, and (2) they claim that we did not use the appropriate benchmark of SW to evaluate its ability to estimate variable's relative importance. We address these two concerns below. Our response to the first point is rather technical. The reader mainly interested in variable importance estimations in IT analyses can therefore confidently skip the following paragraph without impairing her/his understanding of this article.

In Galipaud et al. (2014), we simulated a response variable *y*, three predictor variables more or less correlated with *y* (see Appendix S1 in supporting information for correlation values) and one spurious variable uncorrelated with *y*. The empirical correlation between the response and each predictor variable was constrained and did not deviate from the expected correlation value by more than ±0.01, regardless of the sample size. Instead of controlling for the empirical correlation, Giam and Olden (2016) fixed correlations in the population they sampled from; the empirical correlation between the response and predictor variables in these datasets corresponded to the expected (population) value on average but with a sample variance decreasing with increasing sample sizes. This is indeed a convenient method to simulate the process of experimental sampling and its associated uncertainty. In our previous article, our point was to emphasise the large variance inherent to SW calculation. Specifically, because spurious variables may appear in top ranked models (even for very large sample size) their SW may remain variable although sampling variance is reduced to its minimum. In order to disentangle the confounding effect of the sampling variance from the genuine variability of SW, we thus decided to control for the correlation structure of our dataset. Although Giam and Olden (2016) did not explain precisely in which direction our procedure could have biased our results, they argued that it was inappropriate to study SW reliability as an estimate of variable relative importance. We, on the contrary, believe that it was a conservative procedure in the sense that exposing SW large variability under restricted sampling variance conditions would guarantee that it is an imprecise metric under more realistic settings. However, we understand Giam & Olden's concern, and agree that our initial procedure traded-off empirical realism for smaller sampling variance. Accordingly, we repeated our initial simulations using Giam and Olden's (2016) alternative data-generating code (R simulation codes are accessible in supporting information). The main results of this new analysis are shown in Figure 1. They only differ very marginally from our initial results (Figure 1). With both data-generating procedures, SWs for genuine and spurious variables exhibited a wide range of variation. Incidentally, the theoretical range taken by SW is even wider under the Giam & Olden's data-generating procedure, which is not surprising given that it accounts for more realistic sampling variances (Figure 1). As a consequence, Giam and Olden's (2016) concerns about the relevance of our conclusions on SW variability were not justified.

In their second comment, Giam and Olden (2016) argued that we based our criticisms on the erroneous assumption that SW for each

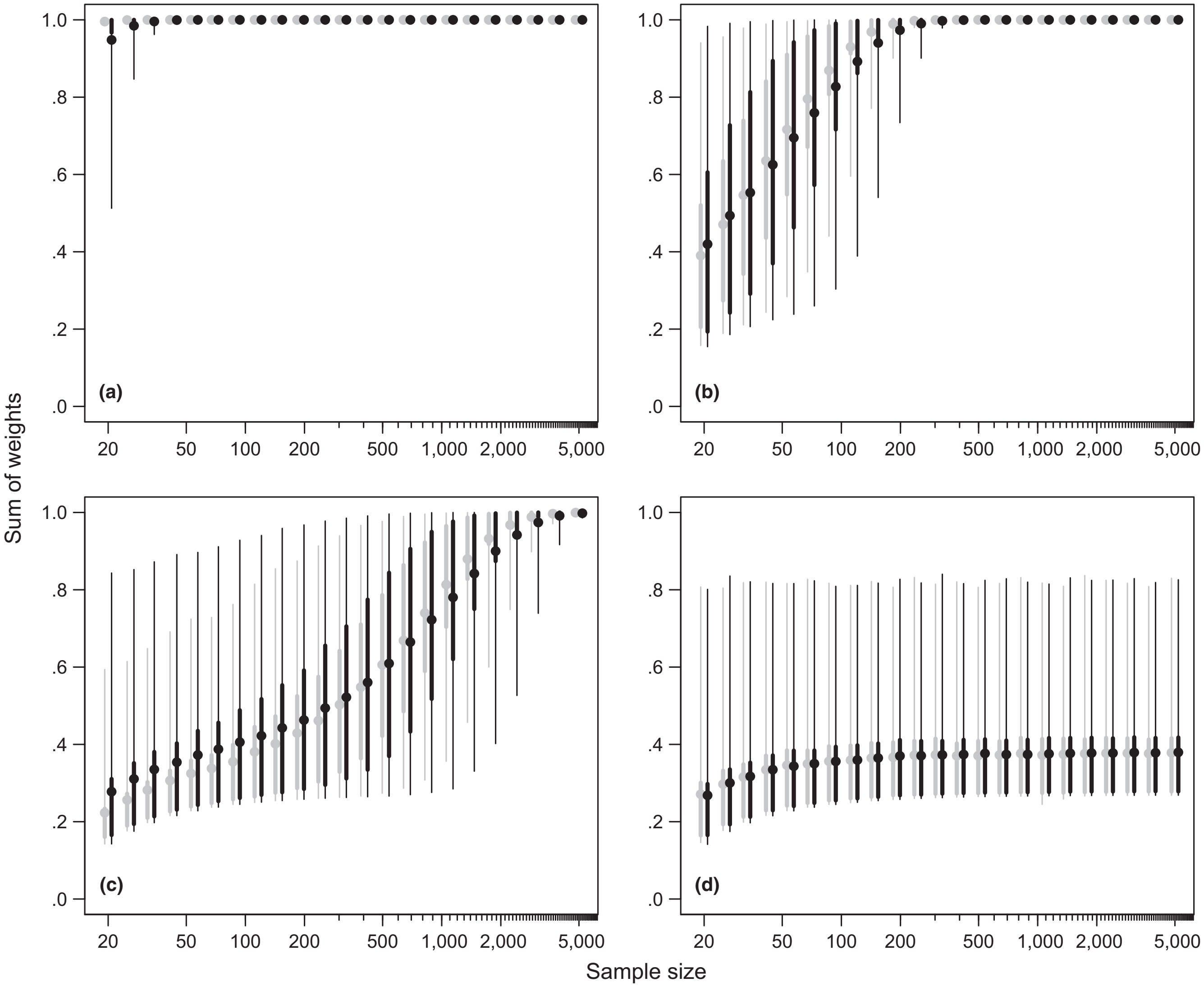

FIGURE 1 Effect of sample size on the expected distribution of AICc-based SW for each four predictor variables: (a) *x*1, (b) *x*2, (c) *x*3, (d) *x*4. We simulated data using Giam and Olden's (2016) generating code (black dots and lines) and using our initial code in Galipaud et al., 2014 (grey dots and lines). For each sample size, the mean SW (dot), interquartile interval (thick line) and 95% interval (thin line) were calculated from 10,000 simulated independent datasets

predictor variable should correspond to its squared bivariate correlation coefficient with the response variable (r^2). This is not what we stated or implied in our study. Sum of Akaike weights is derived from Akaike weights which are essentially relative measures and therefore has nothing to do with the goodness-of-fit of models considered in the set. Sum of Akaike weights should therefore not be mistaken for a variable's estimate of the proportion of variance it explains in the response (Galipaud et al., 2014). In the following sections, we expand the discussion initiated in our first article, which exposed the imprecision of SW, to further discuss the relative reliability of SW. We review the costs and benefits of SW compared to alternative methods for estimations of absolute and relative variable importance in IT.

4 | ESTIMATING VARIABLE ABSOLUTE IMPORTANCE IN IT

A metric of variable absolute importance must possess two principal characteristics to be practical. First, if the aim is variable selection, there must be a threshold above or below which corresponding variables are considered important. This is for instance usually the case of interpretations based on *p*-values, where claims of statistical significance, and therefore variable importance, are traditionally made when the metric falls below the α threshold of 0.05. If experimenters rather aim at discussing the magnitude of variables' effect, interpretations can, but need not, be based on ranges of values taken by an importance metric for which, based on knowledge of the studied biological system, the variable is considered to have great, moderate or little importance (Nakagawa & Cuthill, 2007; Raftery, 1995). Second, to be practical, the metric must be precise, that is, repeatable across analyses. Because SW is defined as an estimate of relative importance, textbooks logically do not discuss potential SW thresholds above which variables could be considered as having a significant effect. To address this gap, some ecologists have haphazardly proposed their own arbitrary SW thresholds (Table 1 in Galipaud et al., 2014). Such an uninformed procedure is often misleading. Specifically, it usually ignores the fact that the lowest theoretical value taken by SW is not 0; mean SWs for spurious variables (i.e. variables with an effect size of 0) are above 0 in regression analyses (Figure 1, Burnham & Anderson, 2002, p. 345). In fact, in AIC analyses ranking all nested models from the most highly parametrised model without interactions between variables (hereafter full set of candidate models), it can be shown that the lowest SW bound for continuous variables is $1/(1 + e^1) \approx 0.27$ (Burnham, 2015; unpublished manuscript). It can also be shown that spurious variables have on average SWs corresponding to $1/(1 +\sqrt{e^1}) \approx 0.38$, which is consistent with our simulations (Figure 1d, Burnham, 2015, unpubl. manuscript). Any attempt to set thresholds for dichotomous or more continuous interpretations about variable absolute importance would not only have to consider this basal SW value but it would also need to consider the theoretical variability of SW taken by variables with different effect sizes across analyses. If SWs are imprecise, important variables can have low SWs, unimportant variables can have high SWs and setting SW thresholds for variable importance is hazardous at best.

4.1 | Model-averaging for estimations of variable absolute importance

Intuitively, a more straightforward approach than SW to estimate variables absolute importance would be to interpret variables' effect sizes. Sampled variables with parameter estimates close or equal to zero would not be considered as being important to explain the variance in the response. Conversely, variables with non-zero estimates could be interpreted as having little, moderate or great importance based on the experimenter's expertise on the studied phenomenon. In IT, when absolute importance estimations must account for uncertainty in model ranking, such interpretations can be made based on model-averaged parameter estimates. We performed simulations following the data-generating procedure proposed by Giam & Olden in order to generate distributions of model-averaged parameter estimates for each variable considered in the set over a wide range of possible sample sizes (see our response to Giam & Olden's first comment for a description of the data-generating procedure and Appendix S1 for a description of the simulation procedure). Results are shown in Figure 2. As expected, estimated averaged parameters are generally consistent with the fact that: *x1* has a large effect on the response, *x2* has a moderate effect on the response and *x3* and *x4* have little or no effect on the response. Like SW, model-averaged parameter estimates are variable across analyses (Figure 2). However, they overall are more repeatable than SW over a wide range of sample sizes (Figure 3, Appendix S1). Unlike SW, when sample size increases, model-averaged parameter estimates become perfectly consistent across analyses (i.e. they reach a repeatability of 1, Figure 3). This result is well-illustrated when comparing the variability of SW in Figure 1d with the perfect consistency of model-averaged effect sizes in Figure 2d for large sample sizes.

Of course, using model-averaged parameter estimates for absolute importance also have drawbacks. Specifically, as a result of the shrinkage, full model-averaged parameter estimates are biased downward. This bias mostly happens under low sample sizes and for variables with weak effects (Figure 2). It has also been shown to be less severe for full model-averaged metrics than when considering variable importance metrics in the first ranked model only (Burnham & Anderson, 2002; pp. 151–153, but see Richards, Whittingham, & Stephens, 2011). Nevertheless, users must be cautious when interpreting full model-averaged metrics and should not take them as accurate averaged estimates of population effect size. When the performed analysis aims at estimating variables' absolute effect size, natural model-averaging should thus be preferred. On the contrary, when the goal is rather to select the correct set of variables in a best model for parsimonious predictions and inferences, full model-averaging should be preferred. This is because consistently biasing parameter estimates towards 0 avoids giving too much importance to variables with weak effects when sample size is small, hence avoiding over-fitting (Lukacs et al., 2010). For a given sample size, full model-averaging therefore allows assessing variables' absolute importance either directly by considering variables with averaged effect sizes close or equal to zero as being unimportant or indirectly by consistently

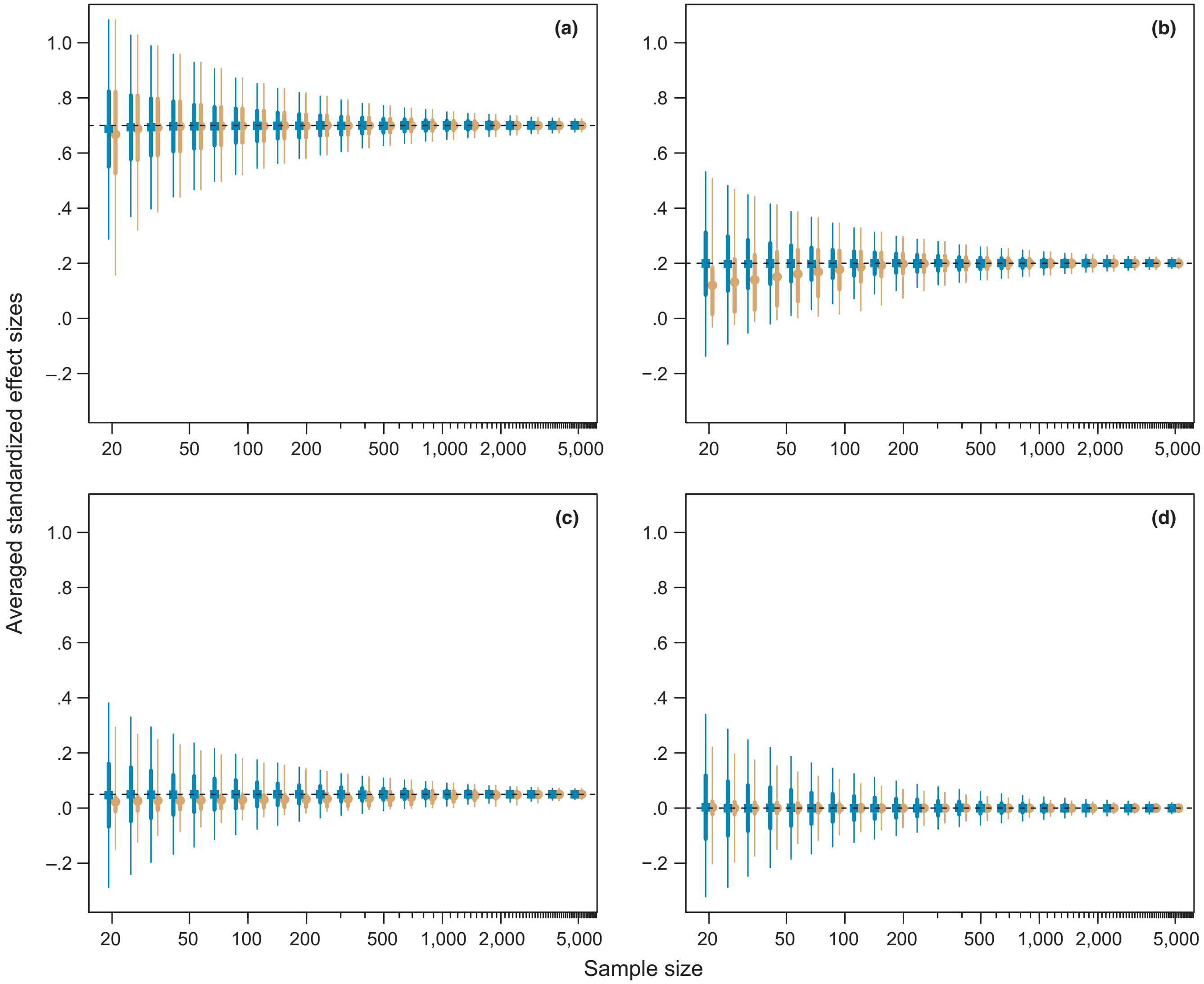

FIGURE 2 Effect of sample size on the expected distribution of model-averaged standardised effect sizes for each four predictor variables using the full (orange dots) or the natural (blue squares) model-averaging technique: (a) *x*1, (b) *x*2, (c) *x*3, (d) *x*4. For each sample size, the mean model-averaged effect size (dot or squares), interquartile interval (thick line) and 95% interval (thin line) were calculated from 10,000 independent datasets simulated using Giam & Olden's data generating procedure

attributing low regression coefficients to unimportant variables in models for statistical predictions.

4.2 | Absolute importance estimations based on nested model comparison

One common but incorrect practice consists in considering all models that fall within ΔAIC < 2 of the top ranked model as being truly competitive (Arnold, 2010; Burnham et al., 2011). In cases where these models are simply more complex forms of the top ranked model, including for instance one additional continuous predictor variable, it is indeed somewhat dubious that this extra variable has a statistical effect on the response, as it does not sufficiently increase the model fit to overcome the 2 unit penalty in AIC (Arnold, 2010; Burnham & Anderson, 2002, p.131, Arnold, 2010; Richards, 2005; Stephens et al., 2007). More formally, users can compare the likelihood of two nested models. The test statistic *D* for a likelihood-ratio test between a more complex model *j* and its simpler version *i* can be expressed in terms of their AIC difference ΔAIC_{ji}: $D = -2\log(L_i) + 2\log(L_j) = 2(k_j - k_i) - \Delta\text{AIC}_{ji}$ (Burnham & Anderson, 1998, p. 61), where *L* is the maximum likelihood estimator of each model and *k* is its number of estimated parameters (note that this equation is different for alternative information criteria such as AICc or QAIC). This test statistic follows a Chi-squared distribution with $k_j - k_i$ degrees of freedom. It follows for instance that, a model including one extra continuous variable assumed to have an effect size of zero (i.e. the null hypothesis) has a probability of .843 (as calculated using the Chi-squared distribution with one degree of freedom) of having a greater AIC score than a simpler model excluding the variable (i.e. $0 < \Delta\text{AIC}_{ji} < 2$). This probability, equivalent to a *p*-value, falls below the traditional .05 threshold when the more complex model ranks higher than the simpler one and has a smaller AIC score by at least 1.84 (i.e. $\Delta\text{AIC}_{ji} < -1.84$). In

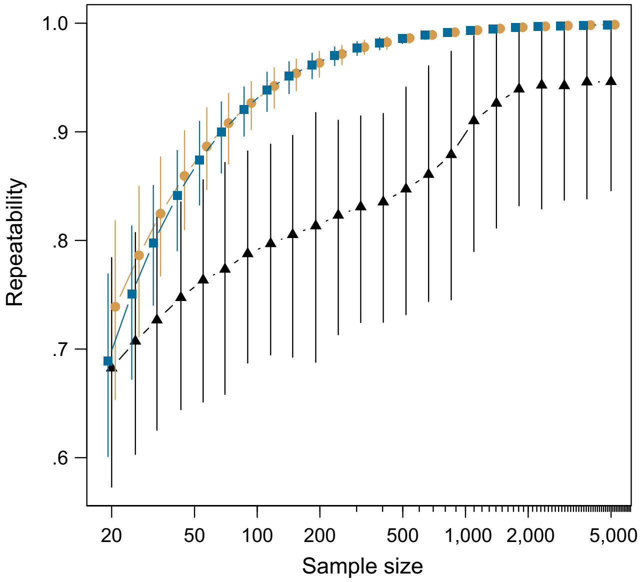

FIGURE 3 Mean repeatability of SW (black triangles) and full (orange dots) and natural (blue squares) model-averaged standardised effect sizes as a function of sample size. Bars represent the 95% confidence interval range for each metric's repeatability at each sample size. A small jitter was added to each point to make them distinguishable from one another at each sample size. See Appendix S1 for a description of the simulation procedure

such a situation, the null-hypothesis of no effect of the extra variable is rejected at the level α = 0.05 and the continuous variable could be considered statistically "significant." It is, however, important to remind the reader that such an analysis only has a heuristic value when assessing the importance of predictor variables in IT model selection (Lukacs et al., 2007). A *p*-value here only informs about the probability of observing a ΔAIC between two nested models given that the extra variable included in the more complex model is spurious. It does not inform about the probability that this variable is truly spurious given the data at hand (Lukacs et al., 2007; Raftery, 1995). This well-known limitation of *p*-values has led some authors to warn against the use of null-hypothesis testing in IT statistical analyses (Lukacs et al., 2007; but see Stephens et al., 2005).

Alternatively, the relative likelihood of a more complex model (i.e. hypothesis) *j* and a simpler model *i* can be calculated as $l_{ji} = L_j/L_i = \exp(\Delta AIC_{ji}/2)$. It can also be easily calculated as the ratio of the Akaike weights of the two models: w_j/w_i (Burnham et al., 2011; Lukacs et al., 2007). If *j* and *i* have the same a priori probability to be true, which is likely the case in exploratory analyses, l_{ji} corresponds to their evidence ratio (Burnham & Anderson, 2004). When it is large, for instance l_{ji} = 50, it means that the empirical support for model *j* is fifty times that of model *i*. The extra parameter(s) that *j* includes is (are) therefore likely significantly related to the response.

Instead of dichotomous decisions based on *p*-values, the comparison of nested models based on l_{ji} seems to be a reliable method to investigate the strength of evidence for variables' effect on the response. It can be performed simply by interpreting the differences in models' information criterion score. For instance, a difference in AIC of 2 between the two nested models corresponds to l_{ji} = 2.7, whereas a difference in AIC of 10 corresponds to $l_{ji} \approx 150$ and provides much more support for the effect of the extra variable(s) included in model *j*. Of course, using such a method supposes that the experimenter chooses a priori which hypotheses are to be compared to respond to her/his biological question. This emphasises the importance of careful specification of a sensible model set before analyses instead of considering the full set of candidate models. Specifically, the aim of the statistical analysis here is hypothesis testing rather than brute force variable selection and parameter estimations. When the set of models a priori selected is small (sometimes down to only 2), this however comes at the cost of potential biases or inaccuracy in parameter estimations. Such a procedure also seems like an unsatisfactory solution for ecologists accustomed to stepwise model selection, where the goal is to discard potentially spurious variables among the large set of measured variables. In such cases, the choice of variables to include in the basal model to which alternative nested models are compared is likely arbitrary. It is yet unclear whether or not a reliable method exists for such an a priori uninformed analysis (Burnham et al., 2011).

5 | ESTIMATING VARIABLE RELATIVE IMPORTANCE IN IT

Giam and Olden (2016) suggest that the *ranking* of a set of predictor variables according to SW corresponds to their ranking according to each variable's r^2 (squared bivariate coefficient of correlation between the response variable and each predictor variable). Users interpreting relative SW values are presumably able to conclude about variables' relative rank in explaining the variance in the response, making SW a reliable estimate of variable relative importance. We think otherwise, and explain why below.

At first glance, the fact that SW performs relatively poorly as an estimate of variable absolute importance does not necessarily disqualify it as an estimate of variable relative importance. However, as explained in the introduction, the reliability of relative importance metrics is inherently linked to the experimenter's ability to select a parsimonious set of variables to compare, would that be using statistical methods (i.e. variable selection) or a priori, based on her/his expertise on the studied phenomenon. For a given sample size, selecting too many variables might lead to over-fitting and imprecise estimations of their relative rank in importance or their relative effect sizes. On the contrary, selecting too few variables might lead to under-fitting and biases in relative importance estimations. In addition, relative importance estimations are only meaningful when considered variables are already known to have a biologically significant effect on the response. Surely, comparing the relative importance of a set of variables with negligible effects on the response would contribute little to the understanding of the studied biological phenomenon. Sum of Akaike weights conveys in itself no information about the magnitude of the effect of considered variables. This results from the fact that model Akaike weights are essentially a relative metric and sum up to 1, regardless of the goodness-of-fit of models considered in the set.

Now what if every considered variable has an important biological effect on the response? Can SW be then used as an estimate of variable relative importance? Statistical textbooks and methodological articles indeed strongly warn against the inclusion of spurious variables when performing IT model selection, arguing that IT should be used for confirmatory rather than explorative analyses (Anderson, Burnham, Gould, & Cherry, 2001; Burnham & Anderson, 2002, 2004; Burnham et al., 2011). Prior to analyses, users are advised to carefully decide which models to include in the set, only considering those for which they have the conviction, or at least strong suspicions, that they play a role in explaining the biological phenomenon under study. Such an a priori model specification is principally done based on background evidence and expertise on the biological system. Specifically, it is a good practice to avoid data dredging; one should not blindly consider all possible models out of a set of predictor variables and, instead, exclude variables that are a priori poor from a biological point of view (Burnham & Anderson, 2002). One important (albeit frequently overlooked) condition required for the calculation of SW, is that variables must appear in balanced number in the model set (Burnham & Anderson, 2002). If a predictor variable appears in all of the models considered a priori in the set whereas another is only included in one model, the former is mechanically more likely to have a greater SW than the latter, irrespective of their true population relative importance. Overcoming this bias is most easily achieved by considering all possible models (i.e. not specifying a careful a priori model set), hence unfortunately also favouring the inclusion of unimportant models in the set and possibly spurious variables. Finally, even if users rely on a priori knowledge of the system to build a biologically wise model set, refrain from resorting to data dredging and consider variables appearing in equal number in the model set, the presence of spurious variables is still possible. Previous studies reporting the effect of a predictor variable on a biological phenomenon do not guarantee that this variable influences the response in the user's studied species or under the particular settings of his or her current experiment or field study (Dochtermann & Jenkins, 2011; Stephens et al., 2005). This is particularly true in fields where researchers rarely attempt to faithfully replicate a previous study and rather test somewhat similar hypotheses on different species (such so-called quasireplications are for instance prevalent in ecology and evolution, Kelly, 2006; Nakagawa & Parker, 2015).

5.1 | SW performs relatively poorly at ranking variables in importance

Even when controlling for the potential inclusion of spurious variables in candidate models, the low repeatability of SW revealed previously also makes it an impractical metric to estimate variables' relative importance. For low to moderate sample sizes, SW theoretical distributions largely overlap for different variables, making it difficult to distinguish them in terms of importance (Figure 1). Incidentally, in simulations performed by Giam and Olden (2016), SW only poorly estimates the ranking of variables in their relative contribution to the variance in the response. For instance when the sample size was $n = 100$, SW correctly estimated variables ranking in only 18.2% of the simulations (Giam & Olden, 2016). We repeated those simulations, also assessing the rank of variables based on their relative model-averaged standardised parameter estimates (see Appendix S1 for a description of the simulation procedure). Our results show that, over the whole range of considered sample sizes, the proportion of retrieved correct ranking was higher when based on model-averaged standardised estimates compared to when it was based on SW (Figure 4). As sample size increases, assessment of variable relative ranking importance becomes more and more accurate when using model-averaged standardised parameter estimates, to the point that it retrieves correct ranking in 100% of the simulations. On the contrary, when using SW, relative ranking importance estimations become impossible at large sample sizes because every SW for non-spurious variables necessarily tends towards and rapidly reaches 1 (Figures 1a–c, 4). Note that this is very much in accordance with SW being an estimate of variable criticality, albeit imprecise. For large sample sizes, SW for genuine variables reaches 1 (regardless of their effect size), which indicates that they are all critical for prediction and inference (Azen et al., 2001). This property of SW also disqualifies it as an estimate of variable's relative effect size; the expected ratio of SWs for two non-spurious variables is not constant across samples sizes because it necessarily tends towards 1 when sample size increases; the ratio of SWs tends towards 1/1=1.

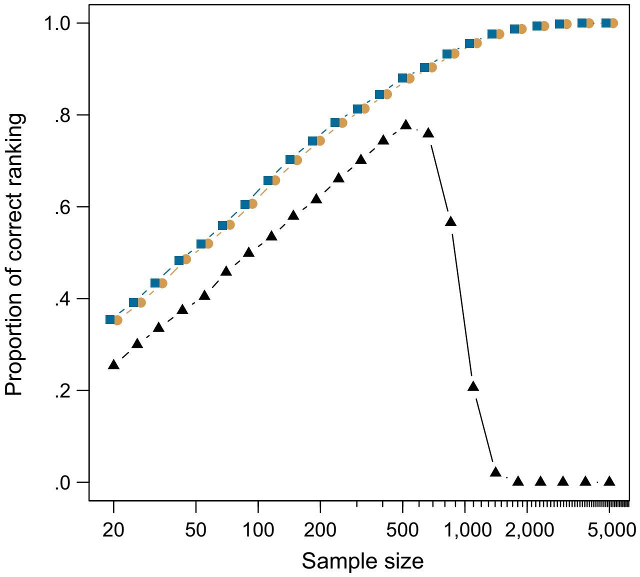

FIGURE 4 Effect of sample size on the proportion of correct variables ranking in importance over 10,000 repetitions of the simulation and for each variable importance metric: SW (black triangles), full model-averaged standardised effect sizes (orange dots), natural model-averaged standardised effect sizes (blue squares). See Appendix S1 for a description of the simulation procedure

5.2 | Inferences about variable relative effect sizes based on model-averaged standardised parameter estimates

The limitations of SW cited above are most easily overcome when the used relative importance metric also provides information about variable's absolute importance. In that sense, a non-exhaustive list of better candidates for variable relative importance inferences includes: model-averaged variables standardised parameter estimates, variables model-averaged absolute value of the *t*-statistics (Cade, 2015) and I_{weighted}, the metric proposed by Giam and Olden (2016). Note that these metrics all provide estimates of effect sizes on one common scale for all variables in the set. This is indeed a necessary condition for assessments of variable relative importance. That is why, contrary to assessments of absolute importance, predictor variables need to be standardised for accurate assessments of their relative effect sizes based on model-averaged parameter estimates (Cade, 2015; Nakagawa & Cuthill, 2007; Schielzeth, 2010). For each variable, standardisation is usually achieved by subtracting its mean $\bar{x}$ from each value x_i and dividing the resulting vector by the variable's standard deviation (Schielzeth, 2010). When variables are not correlated within models, their standardised parameter estimates correspond to their bivariate coefficient of correlation with the response (Cade, 2015; Schielzeth, 2010; see also Appendix S2 in supporting information for a brief discussion about variable importance estimations under multicollinearity). A relatively straightforward measure of variable relative effect size can be calculated as the ratio between variables model-averaged metrics. In our simulations, the proportion of correct ratio between any two variables averaged standardised parameter estimates tended towards 1 with increasing sample size (Figure 5). This results from the fact that the uncertainty around estimations decreases and virtually reaches 0 as sample size increases (Figure 2). It is worth noting that, because of their biases, full model-averaged estimates perform relatively poorly in retrieving correct ratios in importance compared to natural model-averaged estimates (Figure 5, ratios *x2:x1*, *x3:x1*, *x3:x2*). If, however, full-averaged parameter estimates are not biased, as it is the case for *x4* for instance, full model-averaging produces more precise estimates than natural model averaging (Figure 2d, Figure 3) and therefore performs better at retrieving correct ratios in importance (Figure 5, ratios *x4:x1*, *x4:x2*, *x4:x3*).

6 | CONCLUSIONS

An increasing number of methodological papers advocate for the use of IT model selection over unreliable alternatives. This trend must be accompanied by methodological solutions for ecologists to accommodate their need of assessing variables' importance. The current debate over SW reliability and misconceptions found in the ecology and evolution literature most probably originate from a lack of clear

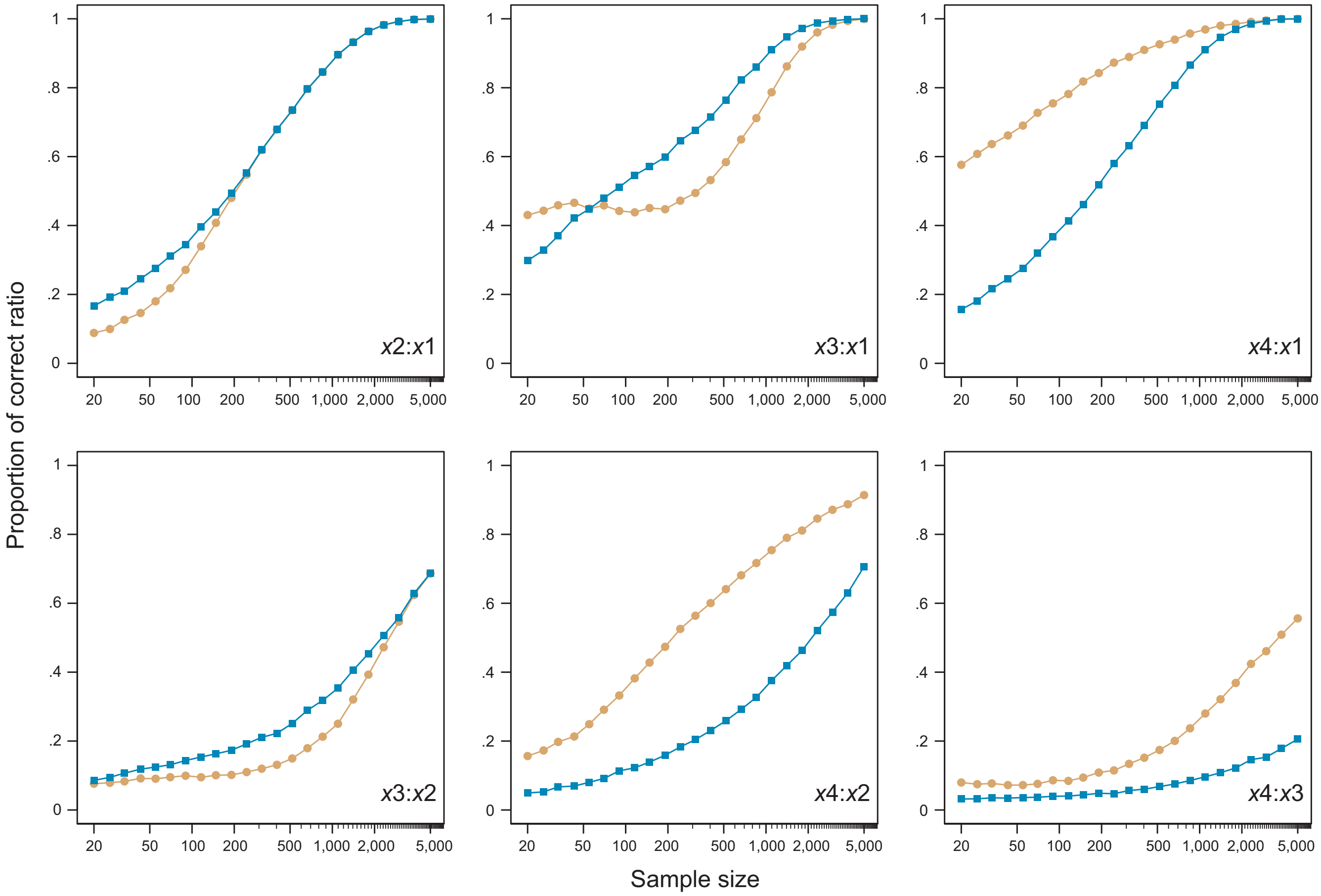

FIGURE 5 Effect of sample size on the proportion of correct estimations of variables relative effects over 10,000 repetitions of the simulation using full (orange dots) and natural model-averaged standardised effect sizes (blue squares) as estimates of variable importance. See Appendix S1 for a description of the simulation procedure

and unambiguous definitions about SW and about the type of statistical inferences it allows. The imprecision of SW revealed in previous studies (Cade, 2015; Galipaud et al., 2014; Giam & Olden, 2016; Murray & Conner, 2009; Smith, Koper, Francis, & Fahrig, 2009), and its relatively poor reliability demonstrated in this article now argues against SW use for statistical inferences. It is timely to consider alternative metrics or techniques in order to avoid recurrent errors in IT model selection. Sum of Akaike weights performs poorly as both an estimate of variable absolute and relative importance compared to model-averaged parameter estimates. Of course, model-averaged metrics are also subject to limitations, especially under multicollinearity or when considering more complex model structures (Appendix S2, see also Cade, 2015 for problems and guidance concerning model-averaging under multicollinearity and when fitting generalised linear models). Nonetheless, our suggestion to ecologists is, at the very least, to be cautious in SW interpretation, and at best, to avoid using it and prefer aforementioned alternative methods to interpret variable importance.

ACKNOWLEDGMENT

We thank Holger Schielzeth, Tim Schmoll, Steven Ramm, Alexandre Courtiol, Brian Cade, two anonymous reviewers and the members of the *Stats Club* of Bielefeld University's Evolutionary Biology and Animal Behaviour departments for very helpful discussions and comments on the manuscript. M.A.F.G. was supported by a DFG grant (DFG Gi 1065/2-1). Calculations were performed using HPC resources from DSI-CCUB (Université de Bourgogne). None of the authors declare any conflicts of interest.

AUTHORS' CONTRIBUTIONS

M.G. and F.X.D.M. performed the simulations presented in the manuscript; M.G., M.A.F.G. and F.X.D.M. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

This article does not contain collected data. Data used to draw figures were simulated and can be retrieved by running the R simulation codes we provide in supporting information.

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akaemiai Kiado.

Anderson, D. R., Burnham, K. P., Gould, W. R., & Cherry, S. (2001). Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, *29*, 311–316.

Arnold, T. W. (2010). Uninformative parameters and model selection using Akaike's Information Criterion. *Journal of Wildlife Management*, *74*, 1175–1178.

Azen, R., Budescu, D. V., & Reiser, B. (2001). Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology*, *54*, 201–225.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542–551.

Burnham, K. (2015). Multimodel inference: Understanding AIC relative variable importance values. Retrieved from https://sites.warnercnr.colostate.edu/kenburnham/wpcontent/uploads/sites/25/2016/08/VARIMP.pdf.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and multimodel inference: A practical information-theoretic approach*, 1st ed. New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd ed. New York, NY: Springer.

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.

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, *96*, 2370–2382.

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

Galipaud, M., Gillingham, M. A. F., David, M., & Dechaume-Moncharmont, F. X. (2014). Ecologists overestimate the importance of predictor variables in model averaging: A plea for cautious interpretations. *Methods in Ecology and Evolution*, *5*, 983–991.

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

Giam, X., & Olden, J. D. (2016). Quantifying variable importance in a multimodel inference framework. *Methods in Ecology and Evolution*, *7*, 388–397.

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. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*, 1–19.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, *7*, 238–257.

Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: How and why it should be done but rarely ever is. *The Quarterly Review of Biology*, *81*, 221–236.

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.

Lukacs, P. M., Thompson, W. L., Kendall, W. L., Gould, W. R., Doherty, P. F., Burnham, K. P., & Anderson, D. R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *Journal of Applied Ecology*, *44*, 456–460.

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between – And reconciliation of – predictive and explanatory models. *Biodiversity and Conservation*, *9*, 655–671.

Murray, K., & Conner, M. M. (2009). Methods to quantify variable importance: Implications for the analysis of noisy ecological data. *Ecology*, *90*, 348–355.

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

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*, 88+.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*, 1373–1379.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

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

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 Sociobiology*, *65*, 77–89.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*, 103–113.

Smith, A., Koper, N., Francis, C. M., & Fahrig, L. (2009). Confronting collinearity: Comparing methods for disentangling the effects of habitat loss and fragmentation. *Landscape Ecology*, *24*, 1271–1285.

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

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.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modeling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Galipaud M, Gillingham MAF, Dechaume-Moncharmont F-X. A farewell to the sum of Akaike weights: The benefits of alternative metrics for variable importance estimations in model selection. *Methods Ecol Evol*. 2017;8:1668–1678. https://doi.org/10.1111/2041-210X.12835