PLOS ONE

RESEARCH ARTICLE

Studentized bootstrap model-averaged tail area intervals

Jiaxu Zeng[1]*, **David Fletcher**[2], **Peter W. Dillingham**[2,3], **Christopher E. Cornwall**[4]

1 Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand, **2** Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand, **3** School of Science and Technology, University of New England, Armidale, Australia, **4** School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand

* jimmy.zeng@otago.ac.nz

Citation: Zeng J, Fletcher D, Dillingham PW, Cornwall CE (2019) Studentized bootstrap model-averaged tail area intervals. PLoS ONE 14(3): e0213715. https://doi.org/10.1371/journal.pone.0213715

Editor: Yinglin Xia, University of Illinois at Chicago College of Medicine, UNITED STATES

Received: November 21, 2018

Accepted: February 27, 2019

Published: March 18, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In many scientific studies, the underlying data-generating process is unknown and multiple statistical models are considered to describe it. For example, in a factorial experiment we might consider models involving just main effects, as well as those that include interactions. Model-averaging is a commonly-used statistical technique to allow for model uncertainty in parameter estimation. In the frequentist setting, the model-averaged estimate of a parameter is a weighted mean of the estimates from the individual models, with the weights typically being based on an information criterion, cross-validation, or bootstrapping. One approach to building a model-averaged confidence interval is to use a Wald interval, based on the model-averaged estimate and its standard error. This has been the default method in many application areas, particularly those in the life sciences. The MA-Wald interval, however, assumes that the studentized model-averaged estimate has a normal distribution, which can be far from true in practice due to the random, data-driven model weights. Recently, the model-averaged tail area Wald interval (MATA-Wald) has been proposed as an alternative to the MA-Wald interval, which only assumes that the studentized estimate from each model has a $N(0, 1)$ or t-distribution, when that model is true. This alternative to the MA-Wald interval has been shown to have better coverage in simulation studies. However, when we have a response variable that is skewed, even these relaxed assumptions may not be valid, and use of these intervals might therefore result in poor coverage. We propose a new interval (MATA-SBoot) which uses a parametric bootstrap approach to *estimate* the distribution of the studentized estimate for each model, when that model is true. This method only requires that the studentized estimate from each model is approximately pivotal, an assumption that will often be true in practice, even for skewed data. We illustrate use of this new interval in the analysis of a three-factor marine global change experiment in which the response variable is assumed to have a lognormal distribution. We also perform a simulation study, based on the example, to compare the lower and upper error rates of this interval with those for existing methods. The results suggest that the MATA-SBoot interval can provide better error rates than existing intervals when we have skewed data, particularly for the upper error rate when the sample size is small.

Introduction

It is well known that calculation of a confidence interval after selection of a best model ignores model uncertainty and can lead to the interval having poor coverage [1–5]. A simple alternative is to use an interval based on the full model. In settings where this model provides a good approximation to the "truth", this will often lead to error rates close to the required levels. Even in these settings, a simpler model may provide a narrower interval with good coverage properties. However, if the data are used to both select a model and to estimate its parameters, the coverage rate can often be much lower than desired. Model-averaging offers a compromise between these two types of intervals, in that we might expect it to lead to a narrower interval than the full model, whilst providing better coverage than an interval based on a single best model [6, 7].

Recently, progress has been made in assessing the theoretical properties of model-averaging, both in terms of optimal weights and construction of confidence intervals. While these results are generally limited to simple settings [8] or rely on asymptotics [9], they provide some insight into the development and understanding of the properties of model-averaged intervals. To complement this work, simulation studies like the one used in this paper are helpful in evaluating the properties of different methods for small sample sizes [4, 7, 10, 11].

Model-averaging is appropriate when interpretation of the parameter of interest, θ, is the same for all models. A common example of such a parameter is the expected value of the response variable for a specified combination of predictor variables. Let M be the number of candidate models, and $\hat{\theta}_m$ be the estimate of θ from model m. In the frequentist setting, the model-averaged estimate of θ is a weighted mean of the estimates from the individual models, given by

$$\bar{\theta} = \sum_{m=1}^{M} w_m \hat{\theta}_m, \tag{1}$$

where w_m is the weight for model m, with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$. There are a number of different methods for selecting the model weights. For the rest of the paper, we consider AIC weights, given by

$$w_m \propto \exp(-\mathrm{AIC}_m/2), \tag{2}$$

where AIC_m is the AIC value for model m. In model selection, AIC tends to select larger models than other information criteria, such as AICc or BIC. It is therefore a natural choice in settings where it is reasonable to assume that the full model is closest to "truth", as in a designed experiment. Previous studies of model-averaged confidence intervals, both theoretical and simulation-based, have also suggested that use of AIC weights is preferable to those based on AICc or BIC [7, 11–13].

Throughout the paper we will refer to the studentized versions of $\bar{\theta}$ and $\hat{\theta}_m$ as T and T_m respectively, i.e.

$$T = \frac{\bar{\theta} - \theta}{\sqrt{\hat{V}(\bar{\theta})}} \quad \text{and} \quad T_m = \frac{\hat{\theta}_m - \theta}{\sqrt{\hat{V}(\hat{\theta}_m)}} \quad (m = 1, \ldots, M). \tag{3}$$

One approach to calculating a model-averaged confidence interval is to use a Wald interval based on $\bar{\theta}$. This involves the assumption that T has a $N(0, 1)$ distribution [6]. This Wald interval has been used in a wide range of application areas [14–24]. We will refer to this interval as the Model-Averaged Wald Interval (MA-Wald).

Recently, [11] proposed a model-averaged tail area (MATA-Wald) interval which involves calculating a weighted average over the models of lower or upper tail areas of the distribution of T_m when model m is true. This involves assuming that T_m in Eq (3) has a $N(0, 1)$ or t-distribution when model m is true. In the context of normal linear regression, the t-distribution version of the MATA-Wald interval has been shown to perform better than the MA-Wald interval [11].

Although use of the MA-Wald or the MATA-Wald interval will often be preferable to one based on the full model or on a best model [4, 7], they will clearly not perform well if each T_m is skewed. This might occur when we have a response variable that is skewed, for several reasons. First, the distribution of each $\hat{\theta}_m$ may be non-normal. Second, each $\hat{\theta}_m$ and its estimated standard error may be positively correlated [25]. Finally, the estimated standard error of each $\hat{\theta}_m$ may be more variable than assumed. If the response variable is positively skewed these effects can lead to both T and T_m being negatively skewed, which will cause the upper confidence limit to be too low and the upper error rate to be too high.

To overcome these problems, a studentized-bootstrap approach can be used to *estimate* the distribution of T_m when model m is true. This involves the less-stringent requirement that each T_m is approximately pivotal when model m is true. This will often be a reasonable assumption, even for skewed data. We therefore extend the MATA-Wald interval using a parametric studentized bootstrap, and refer to this as the studentized-bootstrap model-averaged tail area (MATA-SBoot) interval.

The use of bootstrapping in model-averaging was discussed by [6], who considered use of a model-averaged parametric percentile bootstrap (PB) interval. This involves generating bootstrap samples from a fitted model, typically the full model. For each bootstrap sample, the best model is selected and this provides an estimate, $\hat{\theta}^*$. The PB interval is then given by the appropriate percentiles of $\hat{\theta}^*$. Use of this interval on real data was considered by [6] and [26], but its coverage properties are not well known. In the single-model setting, the percentile bootstrap is known to be first-order accurate, whereas the studentized bootstrap is second-order accurate [27]. We would therefore expect the MATA-SBoot interval to perform better than the PB interval.

The outline of the paper is as follows. First, we describe the MA-Wald, MATA-Wald, and PB intervals, and introduce the MATA-SBoot interval. We then demonstrate use of the MATA-SBoot interval in a real-life setting that involves using a lognormal model to analyse a three-factor experiment designed to assess the effects of global change on a coralline algae. We use a simulation study based on this example to compare the new MATA-SBoot interval to existing methods including the Wald interval from the full model, which we refer to as Full-Wald, and finish with a discussion and suggestions for further research.

Methods

As in the single-model setting, we might transform the parameter of interest before model averaging, in order to better satisfy the assumptions associated with a particular method. For example, in the context of logistic regression we might calculate a model-averaged confidence interval for a probability π by back-transforming the corresponding interval for logit (π). We return to this point when we use a log-transformation in both the example and the simulation study.

MA-Wald interval

The MA-Wald interval was proposed by [3], and is given by

$$\bar{\theta} \pm z_{1-\alpha}\sqrt{\hat{V}(\bar{\theta})}, \tag{4}$$

where 100 $(1 - 2\alpha)$% is the nominal coverage,

$$\hat{V}(\bar{\theta}) = \sum_{m=1}^{M} w_m \left\{ \left(\frac{t_{v_m,1-\alpha}}{z_{1-\alpha}} \right)^2 \hat{V}\left(\hat{\theta}_m\right) + \left(\hat{\theta}_m - \bar{\theta}\right)^2 \right\}, \quad (5)$$

$V(\hat{\theta}_m)$ is the variance of $\hat{\theta}_m$ *conditional* upon model m being true (estimated in the usual way after fitting model m), $t_{v_m,1-\alpha}$ is the 100 $(1 - \alpha)$th percentile of the t-distribution with v_m degrees of freedom, v_m is the residual degrees of freedom associated with model m, and $z_{1-\alpha}$ is the 100 $(1 - \alpha)$th percentile of the N (0, 1) distribution [3]. Use of the ratio $t_{v_m,1-\alpha}/z_{1-\alpha}$ in Eq (5) is motivated by a desire to allow for differences between models in the uncertainty associated with $\hat{V}(\hat{\theta}_m)$.

This interval is based on the assumption that the sampling distribution of $\bar{\theta}$ is approximately normal [6]. This assumption is unlikely to be satisfied due to the randomness of the weights, and reliable estimation of the standard error of $\bar{\theta}$ is also difficult [9]. One motivation for the estimate in Eq (5) is that it can be regarded as a frequentist analogue of the variance of a model-averaged posterior distribution for θ [3]. As mentioned in the Introduction, T will often be negatively skewed when the response variable is positively skewed, leading to this interval having poor coverage.

An alternative Wald interval was proposed by [9]; as this does not have any advantages over the Wald interval from the full model [8, 28], we do not consider it further.

MATA-Wald interval

The MATA-Wald interval is based on a Wald interval obtained from each model [11]. The t-version of the 100 $(1 - 2\alpha)$% MATA-Wald interval $[\theta_L, \theta_U]$ is obtained by solving the equations

$$\sum_{m=1}^{M} w_m Pr(T_{v_m} \leq t_{L,m}) = \sum_{m=1}^{M} w_m Pr(T_{v_m} \geq t_{U,m}) = \alpha, \quad (6)$$

where T_{v_m} has a t-distribution with v_m degrees of freedom,

$$t_{L,m} = \frac{\hat{\theta}_m - \theta_U}{\sqrt{\hat{V}(\hat{\theta}_m)}}, \quad \text{and} \quad t_{U,m} = \frac{\hat{\theta}_m - \theta_L}{\sqrt{\hat{V}(\hat{\theta}_m)}}.$$

Use of this interval is based on the assumption that T_m in Eq (3) has a t-distribution with v_m degrees of freedom when model m is true. This assumption will be exact when we are averaging over a set of normal linear models, and may be a reasonable approximation in other settings. In general, for likelihood-based models it is common practice to assume that the sample size is large enough for T_m to have a N (0, 1) distribution when model m is true. This leads to the z-version of the MATA-Wald interval, in which each T_{v_m} in Eq (6) is replaced by $Z \sim N$ (0, 1). Unlike the t-version, this makes no allowance for the uncertainty associated with $\hat{V}(\hat{\theta}_m)$, which will clearly be undesirable if the sample size is small. The t-version of the MATA-Wald interval is therefore likely to be generally more reliable, and will always have a higher coverage rate than the z-version.

Percentile bootstrap interval

This method involves generating B bootstrap samples from one of the fitted models, and for each sample selecting the best model according to some criterion. When applying this method

in the example and the simulation study, we use AIC to select the best model. The best model for each bootstrap sample provides an estimate $\hat{\theta}^*$. The 100 $(1 - 2\alpha)$% PB interval is then given by the $100\alpha^{\text{th}}$ and 100 $(1 - \alpha)^{\text{th}}$ percentiles of the distribution of $\hat{\theta}^*$ over all bootstrap samples.

If the models are nested, as in a factorial experiment, it is natural to use the fitted full model to generate the bootstrap samples, as we expect this to provide a good approximation to the "truth". In our simulation study, we therefore generate bootstrap samples in this manner. In related work, [29] recommended that bootstrapping should generally be from the full model.

MATA-SBoot interval

In the single-model setting, a parametric studentized bootstrap interval is given by

$$\left[\hat{\theta} - t_U^*\sqrt{\hat{V}(\hat{\theta})},\ \hat{\theta} - t_L^*\sqrt{\hat{V}(\hat{\theta})}\right],$$

where t_L^* and t_U^* are the $100\alpha^{\text{th}}$ and $100(1 - \alpha)^{\text{th}}$ percentiles of the distribution of

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{\hat{V}(\hat{\theta}^*)}}, \tag{7}$$

and $\hat{\theta}^*$ is the estimate of θ obtained from a bootstrap sample generated from the fitted model. Suppose $[\theta_L, \theta_U]$ denotes the resulting interval. The limits of this interval satisfy the equations

$$Pr(T^* \leq t_L^*) = Pr(T^* \geq t_U^*) = \alpha,$$

where T^* is given by Eq (7),

$$t_L^* = \frac{\hat{\theta} - \theta_U}{\sqrt{\hat{V}(\hat{\theta})}} \quad \text{and} \quad t_U^* = \frac{\hat{\theta} - \theta_L}{\sqrt{\hat{V}(\hat{\theta})}}.$$

By analogy, in the multi-model setting the MATA-SBoot interval $[\theta_L, \theta_U]$ is obtained by solving the equations

$$\sum_{m=1}^{M} w_m Pr(T_m^* \leq t_{L,m}^*) = \sum_{m=1}^{M} w_m Pr(T_m^* \geq t_{U,m}^*) = \alpha, \tag{8}$$

where

$$T_m^* = \frac{\hat{\theta}_m^* - \hat{\theta}_m}{\sqrt{\hat{V}(\hat{\theta}_m^*)}}, \quad t_{L,m}^* = \frac{\hat{\theta}_m - \theta_U}{\sqrt{\hat{V}(\hat{\theta}_m)}}, \quad t_{U,m}^* = \frac{\hat{\theta}_m - \theta_L}{\sqrt{\hat{V}(\hat{\theta}_m)}},$$

and $\hat{\theta}_m^*$ is the estimate of θ obtained from fitting model m to that bootstrap sample. The probabilities in Eq (8) are estimated from the bootstrap distribution of T_m^*, based on B bootstrap samples generated from the fitted version of model m.

Use of the bootstrap in this way avoids the need to assume a parametric distribution for T_m. We need only require that T_m be approximately pivotal when model m is true, an assumption that will be reasonable in many settings [30, 31].

When T_m has a N (0, 1) or t-distribution, the MATA-SBoot interval will be identical to the corresponding MATA-Wald interval, as long as B is chosen to be sufficiently large. We would therefore expect the MATA-SBoot interval to perform at least as well as the two versions of the MATA-Wald interval.

A factorial design example

Ocean acidification is the process of increasing absorption of anthropogenically-derived CO_2 by surface seawater [32, 33]. This has negative repercussions for calcareous species, altering growth and calcification rates [34]. Metabolic processes have the potential to modulate the effects of ocean acidification, e.g. photosynthetic uptake of CO_2 by macroalgae could increase pH back to current levels in large macroalgal forests [35], or at the surface of the macroalga [36, 37]. This has been shown to alleviate the negative effects of ocean acidification for species capable of raising seawater pH [38].

In multi-stressor global change experiments the importance or existence of interactions is generally unknown, so it is not always clear which statistical model should be used to make predictions about physiological responses. While numerous studies have attempted to answer this question (e.g. the meta-analysis in [39]), testing for interactions and then using the selected model to make predictions is precisely the setting that is known to result in poor error rates. Recently, [40] used the MA-Wald interval in Eq (4) to make predictions in a global change experiment, the first example we know of model-averaging being used in this important research area. It is therefore of interest to assess whether there is a better choice of interval.

In this example we use data originally presented in [36] to illustrate the use of model averaging in an investigation of the effect of assemblages of upright and crustose coralline algae to modify their local environment within and immediately above their canopies. Several response variables were measured; our choice of surface hydronium ion concentration ($[H_3O^+]$, standardized by bulk concentration) is purely for illustration. In a unidirectional flume, bulk seawater *pH* (ambient pH 8.00, and simulated ocean acidification pH 7.65), *irradiance* (darkness and photosynthetically saturating light), and the effect of water *velocity* (0.015 and 0.040 m s^{-1}) were tested on hydronium ion gradients using a 2^3 factorial design with five replicates.

We focus on estimation of the mean hydronium ion concentration for each of the eight combinations of the factor levels, which we denote as θ_{ijk} ($i, j, k = 1, 2$). Thus $\theta_{ijk} \equiv \mathrm{E}(Y_{ijkl})$, where Y_{ijkl} is the hydronium ion concentration for replicate l associated with treatment combination (i, j, k) ($l = 1, \ldots, r$). In the example we have $r = 5$, while in the simulation study we consider a range of values for r. We assumed the following lognormal model for Y_{ijkl}

$$\log(Y_{ijkl}) = \mu_{ijk} + \varepsilon_{ijkl}, \tag{9}$$

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}, \tag{10}$$

where μ is the overall effect, $\{\alpha_i, \beta_j, \gamma_k\}$ are the main effects, $\{\alpha\beta_{ij}, \alpha\gamma_{jk}, \beta\gamma_{jk}\}$ are the two-way interactions, $\alpha\beta\gamma_{ijk}$ is the three-way interaction, and $\varepsilon_{ijkl} \sim N(0, \sigma^2)$. In the context of this study, μ_{ijk} is proportional to the mean surface pH of the algae for treatment combination (i, j, k).

We obtained a confidence interval for θ_{ijk} by back-transformation of the interval for $\eta_{ijk} \equiv \log(\theta_{ijk}) = \mu_{ijk} + \sigma^2/2$. The estimate of η_{ijk} from model m is given by

$$\hat{\eta}_{ijk,m} = \hat{\mu}_{ijk,m} + \frac{\hat{\sigma}^2_m}{2}, \tag{11}$$

where $\hat{\mu}_{ijk,m}$ is the mean for combination (i, j, k) on the log-scale, and $\hat{\sigma}^2_m$ is the residual mean square from an analysis of variance on this scale. An unbiased estimate of the variance of $\hat{\eta}_{ijk,m}$

is given by

$$\hat{V}\left(\hat{\eta}_{ijk,m}\right) = \frac{\hat{\sigma}_m^2}{r} + \frac{(\hat{\sigma}_m^2)^2}{2(v_m + 2)}. \quad (12)$$

The expressions for $\hat{\eta}_{ijk,m}$ and $\hat{V}(\hat{\eta}_{ijk,m})$ both involve $\hat{\sigma}_m^2$. When model m is true, these two estimates will therefore be positively correlated. In addition, $\hat{\eta}_{ijk,m}$ will have a non-normal distribution, and $v_m \hat{V}(\hat{\eta}_{ijk,m})/V(\hat{\eta}_{ijk,m})$ will not have a $\chi^2_{v_m}$ distribution. These effects will mean that the assumptions underlying the MA-Wald interval and both versions of the MATA-Wald interval are invalid. The distribution of both T and T_m will then be negatively skewed, leading to the corresponding interval having an upper limit that is too low, and consequently an upper error rate that is too high. All three of the above effects will be more noticeable for smaller values of r, and for larger values of σ^2. For the MA-Wald interval there is the additional issue that the model weights are estimated, rather than fixed, and $\bar{\eta}_{ijk} = \sum_{m=1}^{M} w_m \hat{\eta}_{ijk,m}$ may then have a non-normal distribution even if each $\hat{\eta}_{ijk,m}$ is close to normal (as they would be if each v_m were large).

Model-averaging was performed using the set of all possible models. As usual, interaction terms were included only if lower-order terms were also in the model. The AIC weights showed non-negligible support for several models (Table 1). For each θ_{ijk}, we calculated the MA-Wald interval, both versions of the MATA-Wald interval, the percentile bootstrap interval, and the MATA-SBoot interval. We also calculated a Wald interval from the full model (Full-Wald), which is equivalent to using the MA-Wald interval, or the t-version of the MATA-Wald interval, with all the weight given to the full model.

Table 1. AIC model weights obtained when modelling hydronium ion concentrations. The main effects of pH, irradiance and water velocity are denoted by P, I and V, respectively; PI, IV and PV denote the corresponding two-way interactions and PIV is the three-way interaction.

Model	AIC weight
Null	0.000
P	0.000
I	0.000
V	0.000
P+I	0.028
P+V	0.000
I+V	0.000
P+I+V	0.013
P+I +PI	0.325
P+V+PV	0.000
I+V+IV	0.000
P+I+V+PI	0.150
P+I+V+PV	0.005
P+I+V+IV	0.017
P+I+V+PI+PV	0.057
P+I+V+PI+IV	0.263
P+I+V+PV+IV	0.006
P+I+V+PI+PV+IV	0.100
P+I+V+PI+PV+IV+PIV	0.038

https://doi.org/10.1371/journal.pone.0213715.t001

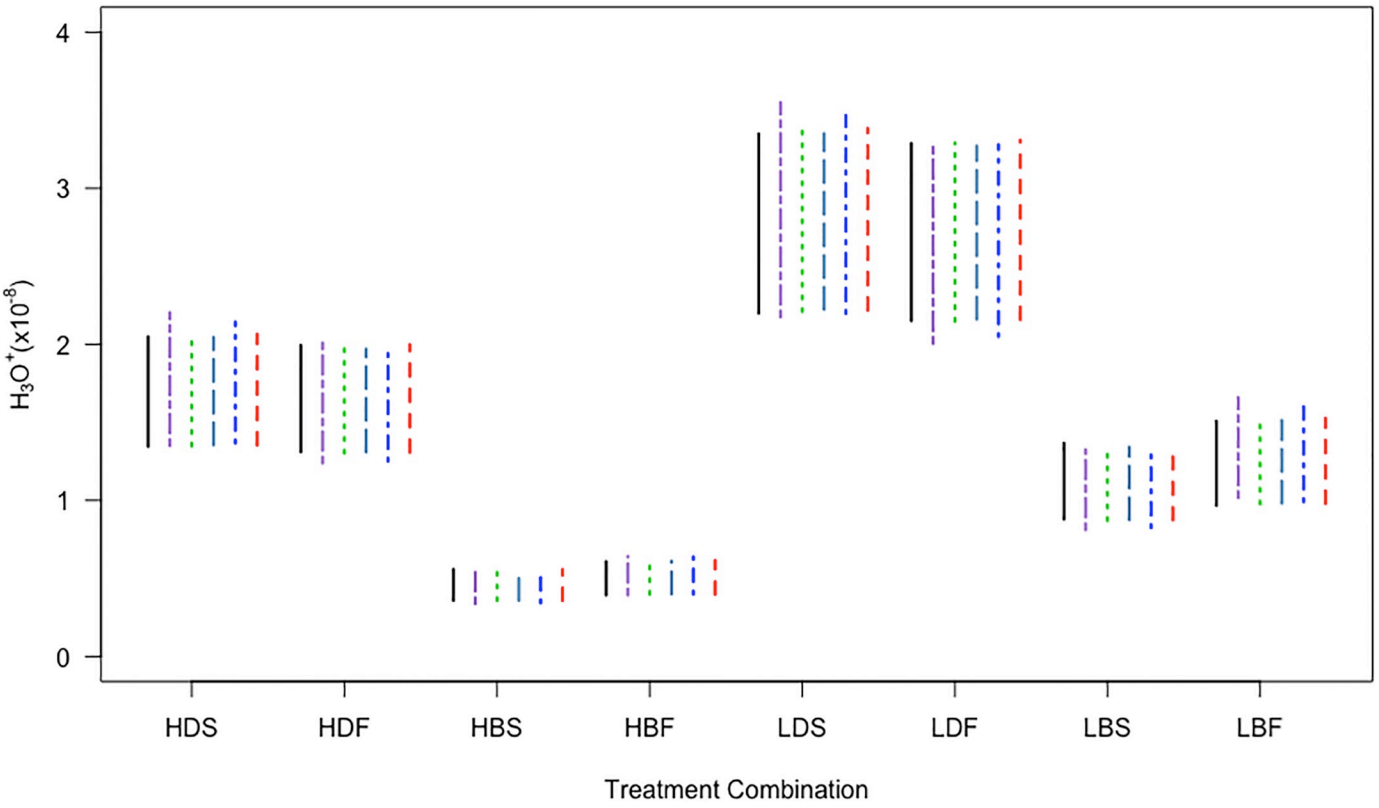

Fig 1. Model-averaged confidence intervals for mean hydronium ion concentration. The intervals are labelled as follows: Full-Wald (long dash-dotted purple line), MA-Wald (solid black line), *t*-version of MATA-Wald (dotted green line), *z*-version of MATA-Wald (long dashed steelblue line), percentile bootstrap (dash-dotted blue line), and MATA-SBoot (dashed red line). The treatment combination denotes the level of each of three factors: pH (H = 8.00, L = 7.65), irradiance (D = darkness, B = photosynthetically saturating light), and velocity (S = 0.015 m s^{-1}, F = 0.040 m s^{-1}).

https://doi.org/10.1371/journal.pone.0213715.g001

For each of the eight combinations of factor levels, the six intervals were broadly similar, the main difference being that the Full-Wald and the percentile interval were generally wider (Fig 1). Although the intervals in this example are similar, we would expect the MATA-Wald and MATA-SBoot intervals to perform quite differently when the sample size is small, as T_m will then be more skewed. We therefore consider a range of sample sizes in the simulation study.

Simulations

We carried out a simulation study in order to compare the six types of interval. We considered the same setting as the example in Section 3, namely a 2^3 factorial experiment. The data were generated using the lognormal model in Eqs (9) and (10). As the performance of an interval will not be influenced by the value of μ, we set $\mu = 0$. We set $\sigma^2 = 1$ as this corresponds to a log-normal distribution that is clearly skewed, with a skewness coefficient of 6.2. We return to the choice of σ^2 when discussing the results.

In order to broaden the conclusions of the study, a different set of parameter values was generated for each simulation run, as in [7]. Thus each main effect and interaction was specified as having a "magnitude" of 2 (High), 1 (Medium) or 0.1 (Low). The corresponding parameter value was then selected from a normal distribution with mean zero and standard deviation equal to this magnitude. The three magnitudes were chosen to be greater than, the same as, or less than σ^2. As we usually expect main effects to be at least as large as two-way

interactions, which in turn will often be at least as large as three-way interactions, we chose the following ten scenarios: LLL, MLL, HLL, MML, HML, MMM, HMM, HHL, HHM, HHH, where, for example, HML is a scenario in which the main effects, two-way interactions and three-way interaction have high, medium and low magnitudes respectively.

To assess the performance of each interval for various levels of replication, we considered $r = 2$, 5 and 50. The choice $r = 2$ represents the lowest possible sample size for this type of study, corresponding to the greatest skewness of T and T_m (for fixed σ^2). The choice $r = 50$ is unlikely to be used in practice, and was included solely to check for asymptotic convergence of the methods. We used 10^5 simulations for each of the ten scenarios, as this allowed us to achieve binomial standard errors for the lower and upper error rates of approximately 0.3%. For the bootstrap-based intervals, we used $B = 9999$. As for the real data, we first calculated a confidence interval for η_{ijk} and back-transformed it to obtain an interval for $\theta_{ijk} \equiv \exp(\eta_{ijk})$, which we denote as $[\theta^L_{ijk}, \theta^U_{ijk}]$. Model-averaging was performed over all 19 possible models.

The performance of each interval was summarised by its mean, over the eight combinations of factor levels, of the lower and upper error rate. We also calculated the mean lower and upper relative half-widths for each treatment combination, and averaged these over the eight combinations, where the relative lower and upper half-width are defined as $(\theta^L_{ijk} - \theta_{ijk})/\theta_{ijk}$ and $(\theta^U_{ijk} - \theta_{ijk})/\theta_{ijk}$ respectively. All calculations were implemented in R Version 3.4.2 [41], and the solutions to Eqs (6) and (8) were found using the *uniroot* function. We also include example code in Supplementary Information (S1 File) demonstrating calculation of the MATA-SBoot interval for a single dataset, and note that functions to calculate the MATA-Wald interval are available in the MATA library of R [11].

Results

The clearest difference between the methods are for the upper confidence limit, with the MATA-SBoot interval generally having an upper error rate that is closest to the nominal level (Figs 2 to 4). This improvement in the upper error rate is most marked for $r = 2$, as we expected. The MATA-SBoot interval also provided a lower error rate that was close to the nominal level. Because the MATA-SBoot increases its width to account for skewness, it was always wider than the MA-Wald and MATA-Wald intervals and usually wider than the PB interval. Interestingly, for the LLL, MLL, and HLL scenarios, the improvement was achieved with little increase in the upper half-width relative to other model-averaging techniques, while substantially outperforming the Wald-Full interval. However, for scenarios with higher magnitude effects and few replicates (Fig 2), the increased width from skewness led to substantially wider upper half-widths for MATA-SBoot intervals relative to the others.

The z-version of the MATA-Wald interval performed worst of all the model-averaging methods on the upper error rate, but very well on the lower error rate, presumably because any skewness to the left still allowed the right-hand tail T_m to be similar to that for a $N(0, 1)$ distribution. The generally superior performance of the t-version over the z-version of the MATA-Wald interval is due to it making some allowance for the uncertainty associated with $\hat{V}(\hat{\theta}_m)$, the difference being largest when r is small. The MA-Wald interval incorporates the ratio $t_{\nu_m,1-\alpha}/z_{1-\alpha}$ as a means of allowing for this uncertainty, which provides another reason for its performance being similar to that for the t-version of the MATA-Wald interval.

Unlike the model-averaged intervals, the interval widths for Full-Wald do not vary across the different simulation scenarios. Particularly, they do not take advantage of the smaller widths possible through model-averaging when weight is placed on reduced models. Consequently, the intervals are wider or equal to the intervals from MA-Wald, MATA-Wald, or PB

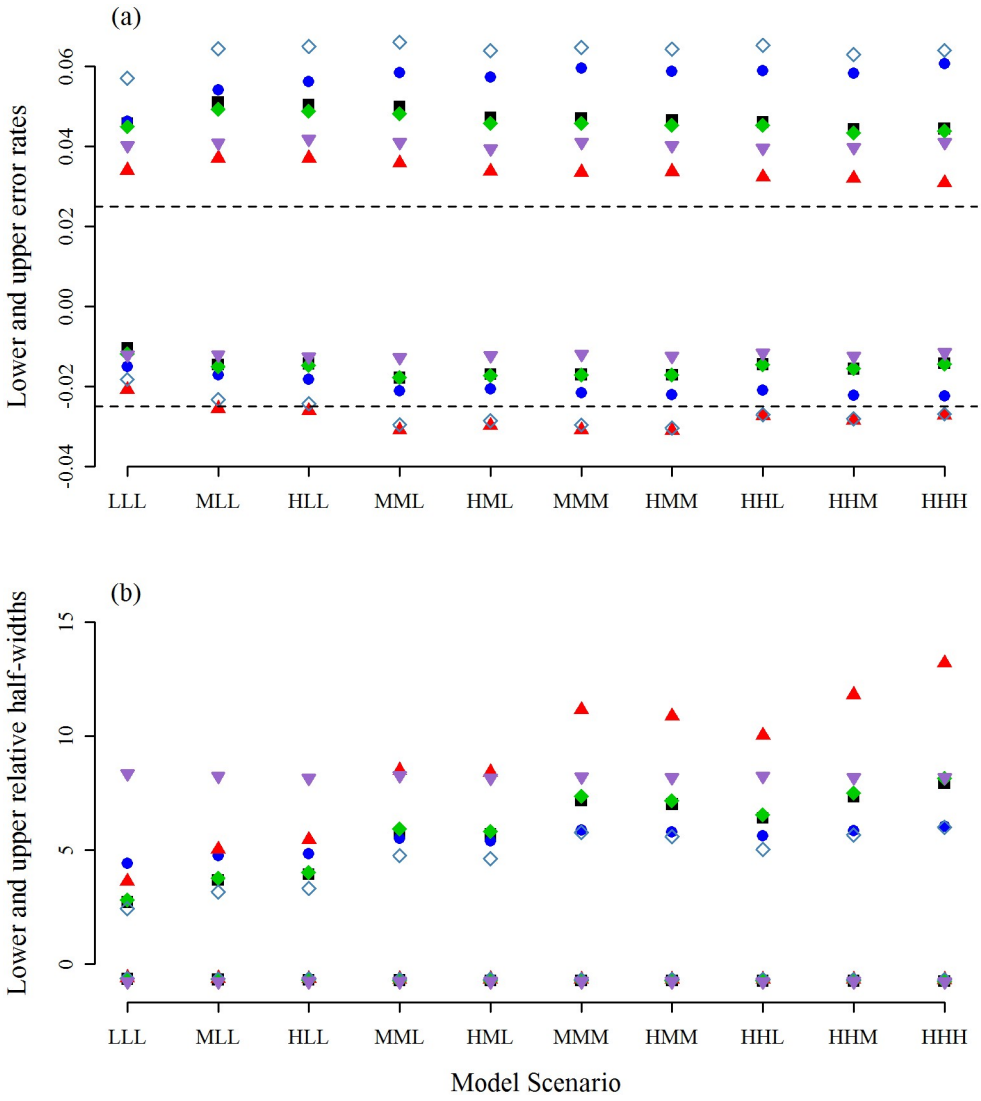

Fig 2. Error rates and relative half-widths when $r = 2$. The results obtained using the Full-Wald (down-pointing purple triangle), the MA-Wald (black square), the *t*-version of the MATA-Wald interval (green diamond), the *z*-version of the MATA-Wald interval (unfilled steelblue diamond), the percentile interval (blue circle), and the MATA-SBoot interval (up-pointing red triangle). For simplicity, the lower error rates are plotted on the negative axis. The nominal rate is shown as a dashed line.

https://doi.org/10.1371/journal.pone.0213715.g002

intervals. However, the increased upper error rates of the Full-Wald interval also produce narrower upper half-widths than MATA-SBoot for scenarios with large magnitude effects (Fig 2).

All methods except the PB interval had approximately the same error rates for $r = 50$ (Fig 4). The PB interval is the only method that does not involve studentization, and suffers from being unnecessarily wide when $r = 50$, with the upper and lower error rates both being less than required, especially for the LLL, MLL, and HLL scenarios.

Discussion

We have focussed on using a bootstrap-based method to construct a model-averaged confidence interval. Bootstrapping can also be used to select model weights [26, 42–44]. For

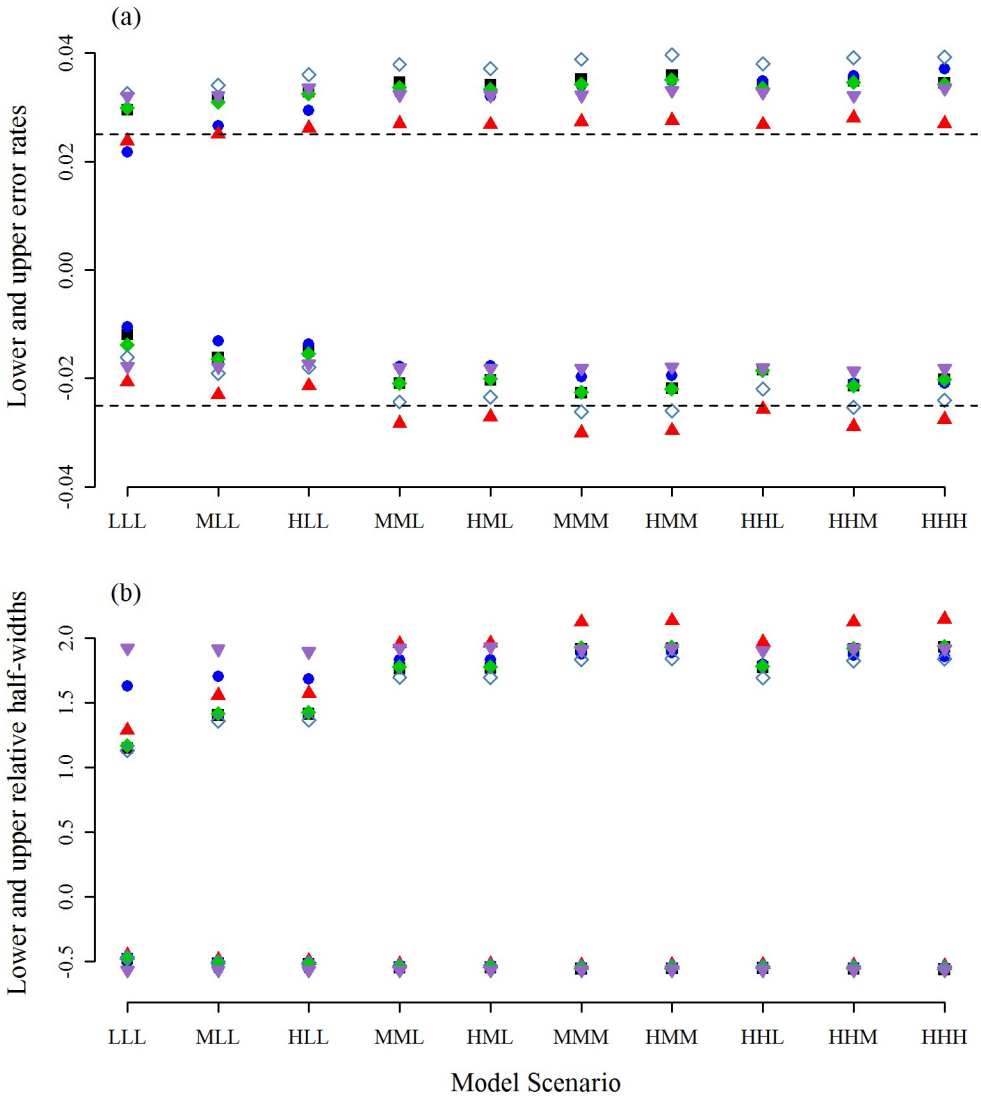

Fig 3. Error rates and relative half-widths when $r = 5$. The results obtained using the Full-Wald (down-pointing purple triangle), the MA-Wald (black square), the *t*-version of the MATA-Wald interval (green diamond), the *z*-version of the MATA-Wald interval (unfilled steelblue diamond), the percentile interval (blue circle), and the MATA-SBoot interval (up-pointing red triangle). For simplicity, the lower error rates are plotted on the negative axis. The nominal rate is shown as a dashed line.

https://doi.org/10.1371/journal.pone.0213715.g003

example, we might choose w_m to be the proportion of times over all bootstrap samples that model m is selected as the best model. This type of weight is implicit in calculation of the PB interval, as well as in the use of bagging to calculate a model-averaged point estimate, a technique that has been used widely in machine learning [45, 46]. In our simulation study, when calculating the PB interval, we found that using AIC to select the best model led to this weight being very similar to the AIC weight in Eq (2), in agreement with the results of [6].

In our simulation setting, the MATA-SBoot interval provided a consistent improvement over existing methods when the sample size was small enough for T and each T_m to be skewed. Our results suggest that this interval has a better error rate than the other methods for small r, while maintaining good error rates and small relative half-widths for large r. This difference is most marked for the upper error rate, as T and T_m are both negatively skewed.

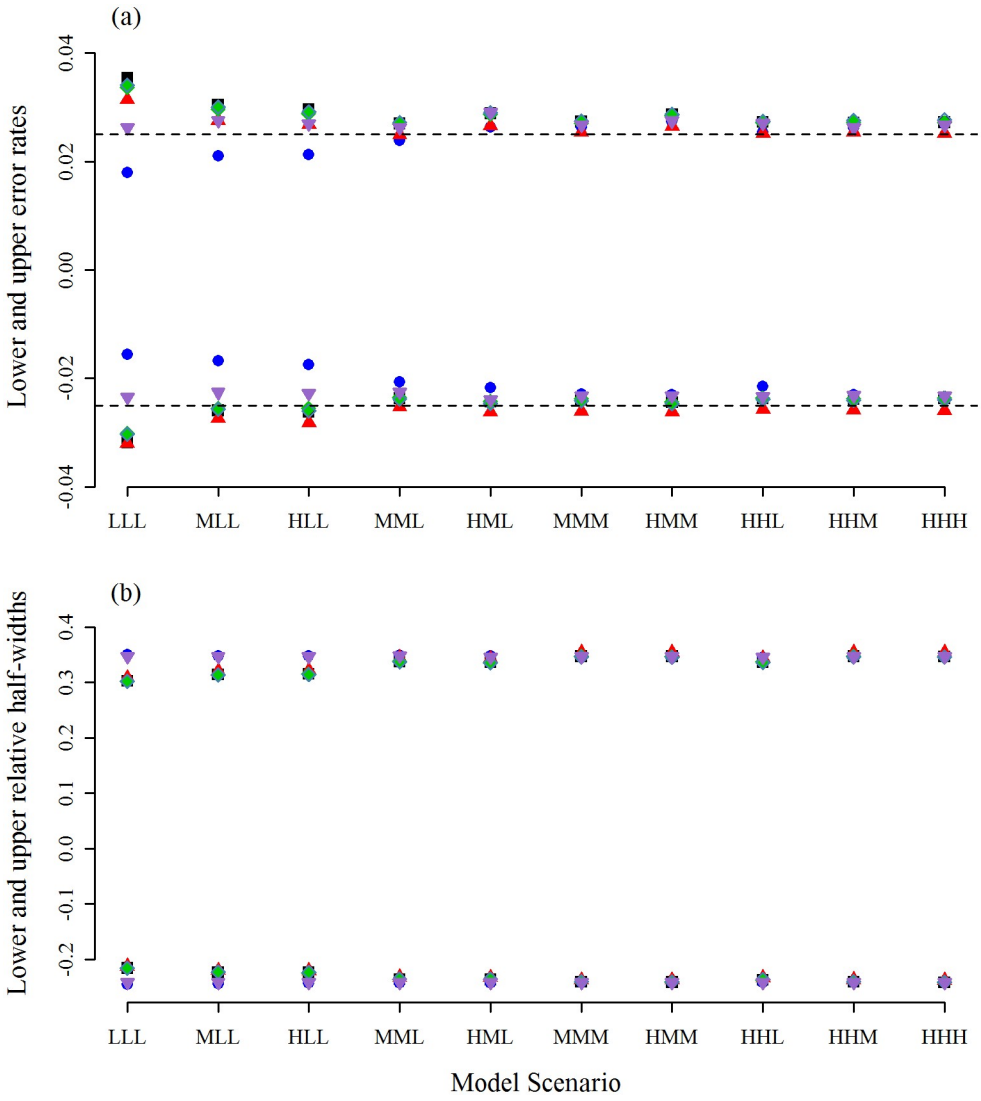

Fig 4. Error rates and relative half-widths when $r = 50$. The results obtained using the Full-Wald (down-pointing purple triangle), the MA-Wald (black square), the *t*-version of the MATA-Wald interval (green diamond), the *z*-version of the MATA-Wald interval (unfilled steelblue diamond), the percentile interval (blue circle), and the MATA-SBoot interval (up-pointing red triangle). For simplicity, the lower error rates are plotted on the negative axis. The nominal rate is shown as a dashed line.

https://doi.org/10.1371/journal.pone.0213715.g004

When it is reasonable to assume that T_m has a $N(0, 1)$ or t-distribution, the MATA-SBoot interval is equivalent to the relevant MATA-Wald interval, as long as B is large enough. The MATA-Wald interval has the advantage of being computational quicker, which might be important when the number of models is large or some of the models are complex.

We chose to set $\sigma^2 = 1$ in the simulations, which corresponds to a skewness coefficient of 6.2. If we had used $\sigma^2 < 1$ there would have been less skewness and the results for the MATA-Wald and MATA-SBoot intervals would be more similar. Conversely, if had set $\sigma^2 > 1$, there would have been more skewness and the benefits of using the MATA-SBoot interval would be even clearer.

The MATA-SBoot interval will obviously not perform well if the studentized bootstrap itself is prone to problems, such as when the standard error of $\hat{\theta}_m$ is poorly estimated and/or T_m is

clearly not pivotal. This caveat is similar to that given by [11] for the MATA-Wald interval [8]. Likewise, in general the MATA approach to constructing a model-averaged confidence interval does not guarantee that the coverage will be exactly as desired, even if the distributional assumptions underlying its use are met [8, 12, 13].

In the simulations the response variable was known to have a lognormal distribution. This allowed us to use an unbiased estimate of the standard error of each $\hat{\theta}_m$ in both versions of the MATA-Wald interval and in the MATA-SBoot interval. In some settings, the true distribution of the response variable may differ from what we assume. We would expect the MATA-SBoot interval to be more robust than the other methods to such misspecification, as long as T_m is approximately pivotal.

The parameter of interest in the example and simulation study was the population mean (for each treatment combination). With skewed data, we might consider estimation of the population median instead. This would amount to removing the second terms on the right-hand side of both Eqs (11) and (12), leading to $\hat{\eta}_{ijk,m}$ and $\hat{V}(\hat{\eta}_{ijk,m})$ being uncorrelated. The studentized version of $\hat{\eta}_{ijk,m}$ would then have a t-distribution with ν_m degrees of freedom. In this case, the t-version of the MATA-Wald interval and the MATA-SBoot interval would be identical, as long as B were chosen to be large enough.

It is difficult to establish theoretical results about the performance of model-averaged confidence intervals, due to the randomness of the model weights. Moreover, model uncertainty will usually arise when the sample size is relatively small, suggesting that asymptotic theory may not be that relevant. We have therefore used simulation to assess the potential benefits of our proposed method. We have considered a range of model-scenarios and used a random-effects generating model in order to make our conclusions more robust. However, as with any simulation study, the conclusions are strictly limited to a particular setting. It would be helpful to assess the performance of the MATA-SBoot interval in other settings, in order to broaden our conclusions.

It is important to note that there exists no unique way to assess the performance of the methods for simulation studies. In our simulation study, we focus on assessing the overall performance of the methods across a number of scenarios. An alternative approach is to compare the minimum coverage of the methods across all possible parameter values, as suggested in [47]. It would be helpful to explore this option in future to further assess the performance of the methods.

For global change studies such as our example, a primary goal is to make predictions under future climate scenarios with an appropriate measure of uncertainty. Use of the MATA-SBoot method can sometimes lead to a narrower confidence interval than one from the full model, whilst maintaining error rates that are generally close to the nominal level, while in other cases the interval must be larger to account for the underlying skewness. Compared to other model-averaging techniques, interval widths are increased due to skewness but, again, this is to maintain approximately the correct level of uncertainty.

Supporting information

S1 File. R code and data for obtaining MATA-SBoot intervals for the hydronium ion example.
(R)

Acknowledgments

The authors thank Associate Professor Catriona Hurd (University of Tasmania) for generously offering use of the coralline algae data.

Author Contributions

Conceptualization: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Data curation: Christopher E. Cornwall.

Formal analysis: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Investigation: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Methodology: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Software: Jiaxu Zeng, Peter W. Dillingham.

Visualization: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Writing – original draft: Jiaxu Zeng, David Fletcher, Peter W. Dillingham.

Writing – review & editing: Jiaxu Zeng, David Fletcher, Peter W. Dillingham, Christopher E. Cornwall.

References

1. Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. The American Statistician. 1990; 44(3):214–217. https://doi.org/10.1080/00031305.1990.10475722
2. Chatfield C. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A. 1995; 158(3):419–466. https://doi.org/10.2307/2983440
3. Burnham K, Anderson D. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer; 2002.
4. Lukacs PM, Burnham KP, Anderson DR. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics. 2010; 62(1):117. https://doi.org/10.1007/s10463-009-0234-4
5. Fletcher D. Model Averaging. Berlin: Springer Briefs in Statistics; 2018.
6. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics. 1997; 53(2):603–618. https://doi.org/10.2307/2533961
7. Fletcher D, Dillingham PW. Model-averaged confidence intervals for factorial experiments. Computational Statistics & Data Analysis. 2011; 55(11):3041–3048. https://doi.org/10.1016/j.csda.2011.05.014
8. Kabaila P, Welsh A, Abeysekera W. Model-averaged confidence intervals. Scandinavian Journal of Statistics. 2016; 43(1):35–48. https://doi.org/10.1111/sjos.12163
9. Hjort NL, Claeskens G. Frequentist model average estimators. Journal of the American Statistical Association. 2003; 98(464):879–899. https://doi.org/10.1198/016214503000000828
10. Fletcher D, Turek D. Model-averaged profile likelihood intervals. Journal of Agricultural, Biological, and Environmental Statistics. 2012; 17(1):38–51. https://doi.org/10.1007/s13253-011-0064-8
11. Turek D, Fletcher D. Model-averaged Wald confidence intervals. Computational Statistics & Data Analysis. 2012; 56(9):2809–2815. https://doi.org/10.1016/j.csda.2012.03.002
12. Kabaila P, Welsh A, Mainzer R. The performance of model averaged tail area confidence intervals. Communications in Statistics-Theory and Methods. 2017; 46(21):10718–10732. https://doi.org/10.1080/03610926.2016.1242741
13. Kabaila P. On the minimum coverage probability of model averaged tail area confidence intervals. Canadian Journal of Statistics. 2018; 46(2):279–297. https://doi.org/10.1002/cjs.11349
14. Stanley TR, Burnham KP. Information-theoretic model selection and model averaging for closed-population capture-recapture studies. Biometrical Journal. 1998; 40(4):475–494. https://doi.org/10.1002/(SICI)1521-4036(199808)40:4%3C475::AID-BIMJ475%3E3.0.CO;2-%23
15. Turkheimer FE, Hinz R, Cunningham VJ. On the undecidability among kinetic models: from model selection to model averaging. Journal of Cerebral Blood Flow & Metabolism. 2003; 23(4):490–498. https://doi.org/10.1097/01.WCB.0000050065.57184.BB
16. Poeter E, Anderson D. Multimodel ranking and inference in ground water modeling. Groundwater. 2005; 43(4):597–605. https://doi.org/10.1111/j.1745-6584.2005.0061.x
17. Namata H, Aerts M, Faes C, Teunis P. Model averaging in microbial risk assessment using fractional polynomials. Risk Analysis: An International Journal. 2008; 28(4):891–905. https://doi.org/10.1111/j.1539-6924.2008.01063.x

18. Zwane E, Van der Heijden P. Capture-recapture studies with incomplete mixed categorical and continuous covariates. Journal of Data Science. 2008; 6:557–572.
19. Symonds MR, Moussalli A. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology. 2011; 65(1):13–21. https://doi.org/10.1007/s00265-010-1037-6
20. Piegorsch WW, An L, Wickens AA, Webster West R, Peña EA, Wu W. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. Environmetrics. 2013; 24 (3):143–157. https://doi.org/10.1002/env.2201 PMID: 24039461
21. Ritz C, Gerhard D, Hothorn LA. A unified framework for benchmark dose estimation applied to mixed models and model averaging. Statistics in Biopharmaceutical Research. 2013; 5(1):79–90. https://doi.org/10.1080/19466315.2012.757559
22. Piegorsch WW. Model uncertainty in environmental dose–response risk analysis. Statistics and Public Policy. 2014; 1(1):78–85. https://doi.org/10.1080/2330443X.2014.937021
23. Hall HI, Song R, Gerstle JE III, Lee LM. Assessing the completeness of reporting of human immunodeficiency virus diagnoses in 2002–2003: capture-recapture methods. American Journal of Epidemiology. 2006; 164(4):391–397. https://doi.org/10.1093/aje/kwj216 PMID: 16772373
24. Xu R, Mehrotra DV, Shaw PA. Incorporating baseline measurements into the analysis of crossover trials with time-to-event endpoints. Statistics in Medicine. 2018; 37(23):3280–3292. https://doi.org/10.1002/sim.7834 PMID: 29888552
25. Fletcher D, Webster R. Skewness-adjusted confidence intervals in stratified biological surveys. Journal of Agricultural, Biological, and Environmental Statistics. 1996; 1(1):120–130. https://doi.org/10.2307/1400564
26. Augustin N, Sauerbrei W, Schumacher M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. Statistical Modelling. 2005; 5(2):95–118. https://doi.org/10.1191/1471082X05st089oa
27. Hall P. The bootstrap and Edgeworth expansion. New York: Springer Series in Statistics; 1997.
28. Wang H, Zhou SZ. Interval estimation by frequentist model averaging. Communications in Statistics. 2013; 42(23):4342–4356. https://doi.org/10.1080/03610926.2011.647218
29. Efron B. Estimation and accuracy after model selection. Journal of the American Statistical Association. 2014; 109(507):991–1007. https://doi.org/10.1080/01621459.2013.823775 PMID: 25346558
30. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge University Press; 1997.
31. Davison AC, Hinkley DV, Young GA. Recent developments in bootstrap methodology. Statistical Science. 2003; 18(2):141–157. https://doi.org/10.1214/ss/1063994969
32. Caldeira K, Wickett ME. Anthropogenic carbon and ocean pH. Nature. 2003; 425(6956):365. https://doi.org/10.1038/425365a PMID: 14508477
33. Sabine CL, Feely RA, Gruber N, Key RM, Lee K, Bullister JL, et al. The oceanic sink for anthropogenic CO2. Science. 2004; 305(5682):367–371. https://doi.org/10.1126/science.1097403 PMID: 15256665
34. Kroeker KJ, Kordas RL, Crim R, Hendriks IE, Ramajo L, Singh GS, et al. Impacts of ocean acidification on marine organisms: quantifying sensitivities and interaction with warming. Global Change Biology. 2013; 19(6):1884–1896. https://doi.org/10.1111/gcb.12179 PMID: 23505245
35. Cornwall CE, Hepburn CD, McGraw CM, Currie KI, Pilditch CA, Hunter KA, et al. Diurnal fluctuations in seawater pH influence the response of a calcifying macroalga to ocean acidification. Proceedings of the Royal Society B: Biological Sciences. 2013; 280(1772):20132201. https://doi.org/10.1098/rspb.2013.2201 PMID: 24107535
36. Cornwall CE, Hepburn CD, Pilditch CA, Hurd CL. Concentration boundary layers around complex assemblages of macroalgae: Implications for the effects of ocean acidification on understory coralline algae. Limnology and Oceanography. 2013; 58(1):121–130. https://doi.org/10.4319/lo.2013.58.1.0121
37. Hurd CL, Cornwall CE, Currie K, Hepburn CD, McGraw CM, Hunter KA, et al. Metabolically induced pH fluctuations by some coastal calcifiers exceed projected 22nd century ocean acidification: a mechanism for differential susceptibility? Global Change Biology. 2011; 17(10):3254–3262. https://doi.org/10.1111/j.1365-2486.2011.02473.x
38. Cornwall CE, Boyd PW, McGraw CM, Hepburn CD, Pilditch CA, Morris JN, et al. Diffusion boundary layers ameliorate the negative effects of ocean acidification on the temperate coralline macroalga Arthrocardia corymbosa. PloS one. 2014; 9(5):e97235. https://doi.org/10.1371/journal.pone.0097235 PMID: 24824089
39. Crain CM, Kroeker K, Halpern BS. Interactive and cumulative effects of multiple human stressors in marine systems. Ecology Letters. 2008; 11(12):1304–1315. https://doi.org/10.1111/j.1461-0248.2008.01253.x PMID: 19046359

40. Boyd P, Dillingham P, McGraw C, Armstrong E, Cornwall C, Feng Yy, et al. Physiological responses of a Southern Ocean diatom to complex future ocean conditions. Nature Climate Change. 2016; 6(2):207. https://doi.org/10.1038/nclimate2811

41. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. 2018. https://www.R-project.org.

42. Martin MA, Roberts S. Bootstrap model averaging in time series studies of particulate matter air pollution and mortality. Journal of Exposure Science and Environmental Epidemiology. 2006; 16(3):242–250. https://doi.org/10.1038/sj.jea.7500454 PMID: 16106257

43. Holländer N, Augustin N, Sauerbrei W. Investigation on the improvement of prediction by bootstrap model averaging. Methods of Information in Medicine. 2006; 45(1):44–50. https://doi.org/10.1055/s-0038-1634035 PMID: 16482369

44. Buchholz A, Holländer N, Sauerbrei W. On properties of predictors derived with a two-step bootstrap model averaging approach—a simulation study in the linear regression model. Computational Statistics & Data Analysis. 2008; 52(5):2778–2793. https://doi.org/10.1016/j.csda.2007.10.007

45. Breiman L. Bagging predictors. Machine Learning. 1996; 24(2):123–140. https://doi.org/10.1023/A:1018054314350

46. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer Series in Statistics; 2009.

47. Kabaila P, Leeb H. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association. 2006; 101(474):619–629. https://doi.org/10.1198/016214505000001140