Model selection using information criteria, but is the 'best' model any good?

Article *in* Journal of Applied Ecology · November 2017

DOI: 10.1111/1365-2664.13060

CITATIONS
31

READS
455

4 authors, including:

Ralph Mac Nally
University of Canberra
314 PUBLICATIONS **11,232** CITATIONS

SEE PROFILE

Richard P Duncan
University of Canberra
200 PUBLICATIONS **10,672** CITATIONS

SEE PROFILE

Jian D L Yen
University of Melbourne
36 PUBLICATIONS **449** CITATIONS

SEE PROFILE

Received: 19 September 2017 | Accepted: 21 November 2017

DOI: 10.1111/1365-2664.13060

COMMENTARY

Journal of Applied Ecology | BRITISH ECOLOGICAL SOCIETY

Model selection using information criteria, but is the "best" model any good?

Ralph Mac Nally[1,2] | Richard P. Duncan[1] | James R. Thomson[3] | Jian D. L. Yen[4]

[1]Institute for Applied Ecology, The University of Canberra, Bruce, ACT, Australia

[2]Department of Ecology, Environment and Evolution, La Trobe University, Bundoora, Australia

[3]Department of Environment, Land, Water and Planning, Arthur Rylah Institute for Environmental Research, Melbourne, Vic., Australia

[4]School of Biosciences, The University of Melbourne, Parkville, Vic., Australia

Correspondence
Ralph Mac Nally
Email: ralph.macnally@gmail.com

Handling Editor: Akira Mori

Abstract

1. Information criteria (ICs) are used widely for data summary and model building in ecology, especially in applied ecology and wildlife management. Although ICs are useful for distinguishing among rival candidate models, ICs do not necessarily indicate whether the "best" model (or a model-averaged version) is a good representation of the data or whether the model has useful "explanatory" or "predictive" ability.
2. As editors and reviewers, we have seen many submissions that did not evaluate whether the nominal "best" model(s) found using IC is a useful model in the above sense.
3. We scrutinized six leading ecological journals for papers that used IC to compare models. More than half of papers using IC for model comparison did not evaluate the adequacy of the best model(s) in either "explaining" or "predicting" the data.
4. *Synthesis and applications*. Authors need to evaluate the adequacy of the model identified as the "best" model by using information criteria methods to provide convincing evidence to readers and users that inferences from the best models are useful and reliable.

KEYWORDS

cross-validation, ecological models, external validation, goodness-of-fit, internal validation, model adequacy, model averaging, model selection

1 | THE USE OF INFORMATION CRITERIA FOR ECOLOGICAL INFERENCE

Many ecologists have become wary of relying solely on *p*-values arising from null-hypothesis tests to make inferences from statistical models, a sentiment reflected in recent statements by the American Statistical Association and some journal editors (Baker, 2016; Krausman, 2017). In response, a now widely adopted approach in applied ecology and wildlife management is to use information criteria (IC; e.g. AIC [and the oft-used small-sample-size corrected AIC_c], BIC, DIC, WAIC) (Burnham & Anderson, 2004; Vehtari, Gelman, & Gabry, 2016; Watanabe, 2013) to estimate the relative predictive capacity of two or more competing models. Information criteria have been used extensively to rank competing models, to select the "best-performing" model(s) from the set of competing alternatives, and to make inferences on variable importance using the nominal best-performing model(s) (Symonds & Moussalli, 2011). Discussions surrounding the appropriate use of *p*-values and IC emphasized the need to consider an array of statistical characteristics when evaluating the performance of a model (Baker, 2016; Stephens, Buskirk, Hayward, & Martinez Del Rio, 2005; Zuur & Ieno, 2016).

2 | THE "BEST" MODEL (OF THOSE CONSIDERED) NEEDS TO BE ASSESSED

Despite clear guidelines on the appropriate use of IC by their proponents (e.g. Burnham & Anderson, 2004), as journal editors and reviewers we increasingly see submissions in which authors do not present an absolute measure of the goodness-of-fit of a model selected by IC.

Authors advocating the use of IC methods make it crystal clear that multimodel comparisons based on IC have several components: identify plausible predictor variables, exclude implausible predictor combinations, calculate IC, evaluate model performance, conduct model averaging (if appropriate), and make inferences on the influence of predictor variables from the model. These authors emphasized that the evaluation of model performance should consider the absolute fit of the selected model, and that such an evaluation should occur prior to interpreting the results (and making inferences) of model comparisons using IC. For example, Symonds and Moussalli (2011, p. 19) wrote: "Because AIC is a relative measure of how good a model is among a candidate set of models given the data, it is particularly prone to poor choices of model formulation. You can have a set of essentially meaningless variables and yet the analysis will still produce a best model. It is therefore important to assess the goodness of fit (χ^2, R^2) of the model …." There almost always will be a best model from the set "offered" for evaluation, but the issue is whether than model is useful in an explanatory or predictive sense.

3 | THE BEST MODELS FREQUENTLY ARE NOT ASSESSED

To assess how well this advice has been followed, we examined all papers published in six ecological journals in the first part of 2017. The journals, volumes, issues and counts of papers were: (1) *Journal of Applied Ecology*, volume 54, issues 1–3 (108 papers); (2) *Diversity & Distributions*, volume 23, issues 1–6 (62 papers); (3) *Journal of Wildlife Management*, volume 81, issues 1–4 (102 papers); (4) *Ecological Applications*, volume 27, issues 1–5 (124 papers); (5) *Conservation Biology*, volume 31, issues 1–4 (63 "contributed papers"); and (6) *Biological Conservation*, volumes 205–209 (123 papers). We selected these journals because two of us are Associate Editors on the first two journals, and we are conscious that IC methods are recommended widely in wildlife management and conservation applications (Anderson, Burnham, & White, 1998).

From 582 papers, 119 papers employed IC in some form to assess model performance, with almost half (58) using the small-sample-size corrected version of AIC, AIC_c. The six journals differed substantially in the proportion of absolute-model-fit assessments, with *Conservation Biology* having the highest proportion (0.73) and *Journal of Wildlife Management* having the lowest proportion (0.28, Table 1). Many papers in *Conservation Biology* and *Ecological Applications* used bespoke quantitative and mechanistic models, which do not tend to use IC. Of the 119 papers using IC, only 55 (46%) evaluated the absolute goodness-of-fit of the "best-performing" model according to IC. Among those 55 papers, measures of explained variance ([undefined] R^2, conditional R^2, marginal pseudo R^2, Nagelkerke pseudo R^2, Nagakawa's and Schielzeth's R^2, Cohen's κ, Wald test, explained deviance) were used to assess absolute goodness-of-fit. Eight papers evaluated fit using area-under-curve (AUC) for logistic regression models, and eight used multi-fold, cross-validation. Five other papers did not do a formal evaluation of model fit but included the null (intercept only) model in the candidate set, which should be standard practice. One study found that the null model was the best-fitting model in two of four cases (Major, Buxton, Schacter, Conners, & Jones, 2017), suggesting little support for the ecological hypotheses of interest. Several papers examined the match between observed and fitted values for an averaged model (Edwards, Massam, Haugaasen, & Gilroy, 2017; Mitchell, Bakker, Vincent, & Davies, 2017), while another undertook cross-validation on a model-averaged result (Cavada et al., 2017). The other 64 papers did not evaluate model goodness-of-fit in an absolute sense. Inferences almost always were on the model or model-averaged coefficients (and standard errors), with criteria for importance of a predictor typically being whether the 95% confidence interval excluded 0. Overall, our review suggested that editorial and reviewer vetting is not always dealing with the need to assess goodness-of-fit of selected models.

4 | SOME WAYS FORWARD FOR BEST-MODEL ASSESSMENT

We have touched on some of the approaches for assessing goodness-of-fit (e.g. various R^2 analogues). The most basic requirement should be to assess whether a model is adequately specified. Misspecified models can be identified with plots of residuals against fitted values, χ^2 tests or posterior predictive checks, which could be provided as supplementary material. There are appropriate residuals for all response distributions (McCullagh & Nelder, 1989), from which

TABLE 1 Breakdown of numbers of papers using IC without model-fit evaluation (IC−), papers using IC with some form of model-fit assessment (IC+), totals using IC (IC-all), the proportion of IC+/IC-all papers (i.e. papers that undertook absolute assessments of some kind), and the number of papers screened

Journal	IC−	IC+	IC-all	Proportion IC+	No. papers screened
Diversity and Distributions	10	5	15	0.33	102
Journal of Applied Ecology	10	6	16	0.37	108
Journal of Wildlife Management	13	5	18	0.28	62
Ecological Applications	11	19	30	0.63	124
Conservation Biology	3	8	11	0.73	63
Biological Conservation	17	12	29	0.41	123
Grand Total	64	55	119	0.46	582

R^2-analogue measures of variance or deviance explained can be calculated (Nakagawa & Schielzeth, 2013). Conventional (OLS) R^2-values have several interpretations: (1) explained variation, (2) improvement in model fit compared with the null model, or (3) the square of correlation between observed and predicted values (Long, 1997). Although pseudo-R^2 measures do not necessarily measure the percentage of variance explained (Hoetker, 2007), they mirror at least one of those interpretations and provide an absolute measure of goodness-of-fit. Where R^2 measures are not calculable, the inclusion of a null model, which should always be done, in an IC-based comparison allows for an assessment of model performance relative to a fixed baseline, which is similar to many pseudo-R^2 measures (Hoetker, 2007).

All goodness-of-fit statistics can be coupled with internal (or cross) validation to assess model predictive capacity (Roberts et al., 2017). Although cross-validation based on log-likelihoods is asymptotically equivalent to several IC (Watanabe, 2010, 2013), cross-validation can be based on R^2 measures, which provides an absolute measure of the predictive capacity of a model. Internal validation is relatively straightforward to implement with statistical computing environments such as R (R Core Team, 2016), and some fitting packages provide internal validation options (Yen, Thomson, Paganin, Keith, & Mac Nally, 2015). Predictive or external validation is a particularly stringent way to assess the usefulness of fitted models (Steyerberg, 2008), whereby model predictions are confronted with newly collected or different (analogous data collected elsewhere or by other workers) data. Models that pass this evaluation can be used with much greater surety than those passing the more basic, and generally more "optimistic," internal validation (Fleishman et al., 2018). The construction of an appropriate validation scheme often clarifies the main purpose of an analysis, which can help to identify an appropriate measure of goodness-of-fit.

5 | ASSESSMENT DEPENDS ON THE INTENDED USE OF MODEL

There is no universally agreed-upon measure for the goodness-of-fit of a model to data. The amount of variance (or deviance) that needs to be explained or predicted to make a model "good" is subjective, and depends on the ecological question and the intended use of the model. In the early stages of a research programme, modelling may be more diffuse than in more mature programmes and primarily seek to stimulate hypotheses or to narrow down sets of potential predictor variables. In mature programmes, modelling may be intended to generate specific predictions to increase ecological understanding or to inform management decisions. Notwithstanding these elements to the assessment of model fit, the use of IC demands model evaluation that extends beyond relative assessments of performance among candidate models to consider absolute measures of goodness-of-fit. Such measures allow editors, reviewers and ultimately readers to make their own judgements of: (1) whether the nominal best model is a meaningful fit to build data; and (2) model inferences, especially on the relative importance of predictor variables.

ACKNOWLEDGEMENTS

The article was much improved by two rounds of comments from four reviewers (including G. Hayward, S. W. Buskirk and M. Gillingham) and the recommendations of the Associate Editor A. Mori, for which the authors are very grateful. R.M. and R.P.D. appreciate the appointments of Centenary Professorships from The University of Canberra, and J.D.L.Y. recognizes the award of a John McKenzie Fellowship from The University of Melbourne.

AUTHORS' CONTRIBUTIONS

R.M. initiated the Commentary, conducted all the scrutiny of the journals appearing in Table 1 and wrote the first version of the MS. All authors contributed to the writing of subsequent drafts and provided technical input into the content, especially J.D.L.Y. in the revision.

DATA ACCESSIBILITY

Data have not been archived because all data are present in the manuscript (see Table 1).

ORCID

Ralph Mac Nally http://orcid.org/0000-0002-4473-1636

REFERENCES

Anderson, D. R., Burnham, K. P., & White, G. C. (1998). Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, *25*, 263–282. https://doi.org/10.1080/02664769823250

Baker, M. (2016). Statisticians issue warning on P values: Statement aims to halt missteps in the quest for certainty. *Nature*, *531*, 151.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. https://doi.org/10.1177/0049124104268644

Cavada, N., Ciolli, M., Rocchini, D., Barelli, C., Marshall, A. R., & Rovero, F. (2017). Integrating field and satellite data for spatially explicit inference on the density of threatened arboreal primates. *Ecological Applications*, *27*, 235–243. https://doi.org/10.1002/eap.1438

Edwards, D. P., Massam, M. R., Haugaasen, T., & Gilroy, J. J. (2017). Tropical secondary forest regeneration conserves high levels of avian phylogenetic diversity. *Biological Conservation*, *209*, 432–439. https://doi.org/10.1016/j.biocon.2017.03.006

Fleishman, E., Yen, J. D. L., Thomson, J. R., Mac Nally, R., Dobkin, D. S., & Leu, M. (2018). Identifying spatially and temporally transferrable surrogate measures of species richness. *Ecological Indicators*, *84*, 470–478. https://doi.org/10.1016/j.ecolind.2017.09.020

Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, *28*, 331–343. https://doi.org/10.1002/(ISSN)1097-0266

Krausman, P. R. (2017). P-values and reality. *The Journal of Wildlife Management*, *81*, 562–563. https://doi.org/10.1002/jwmg.21253

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Major, H. L., Buxton, R. T., Schacter, C. R., Conners, M. G., & Jones, I. L. (2017). Habitat modification as a means of restoring crested auklet

colonies. *The Journal of Wildlife Management*, *81*, 112–121. https://doi.org/10.1002/jwmg.21175

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman and Hall. https://doi.org/10.1007/978-1-4899-3242-6

Mitchell, R. M., Bakker, J. D., Vincent, J. B., & Davies, G. M. (2017). Relative importance of abiotic, biotic, and disturbance drivers of plant community structure in the sagebrush steppe. *Ecological Applications*, *27*, 756–768. https://doi.org/10.1002/eap.1479

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, https://doi.org/10.1111/ecog.02881

Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Martinez Del Rio, C. (2005). Information theory and hypothesis testing: A call for pluralism. *Journal of Applied Ecology*, *42*, 4–12. https://doi.org/10.1111/j.1365-2664.2005.01002.x

Steyerberg, E. (2008). *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer Science & Business Media.

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. https://doi.org/10.1007/s00265-010-1037-6

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Yen, J. D. L., Thomson, J. R., Paganin, D. M., Keith, J., & Mac Nally, R. (2015). Function regression in ecology and evolution: FREE. *Methods in Ecology and Evolution*, *6*, 17–26. https://doi.org/10.1111/2041-210X.12290

Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, *7*, 636–645. https://doi.org/10.1111/2041-210X.12577

How to cite this article: Mac Nally R, Duncan RP, Thomson JR, Yen JDL. Model selection using information criteria, but is the "best" model any good? *J Appl Ecol*. 2017;00:1–4. https://doi.org/10.1111/1365-2664.13060