A guide to Bayesian model checking for ecologists

Paul B. Conn[1,5], Mevin B. Hooten[2,3,4], Devin S. Johnson[1], and Peter L. Boveng[1]

[1]*National Marine Mammal Laboratory, NOAA, National Marine Fisheries Service, Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, WA 98115 USA*

[2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523 USA*

[3]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523 USA*

[4]*Department of Statistics, Colorado State University, Fort Collins, CO 80523 USA*

Abstract. Checking that models adequately represent data an essential component of applied statistical inference. Ecologists increasingly use hierarchical, Bayesian statistical models in their research. The appeal of this modeling paradigm is undeniable, as researchers can build and fit models that embody complex ecological processes while simultaneously controlling for potential biases arising from sampling artifacts. However, ecologists tend to be less focused on checking model assumptions and assessing potential lack-of-fit when applying Bayesian methods than when they apply frequentist methods such as maximum likelihood. There are also multiple ways of assessing goodness-of-fit for Bayesian models, each of which has strengths and weaknesses. For instance, in ecological applications, the "Bayesian p-value" is probably the most widely used approach for assessing lack of fit. Such p-values are relatively easy to compute, but they are well known

[5]Email: paul.conn@noaa.gov

to be conservative, producing p-values biased towards 0.5. Alternatively, lesser known approaches to model checking, such as prior predictive checks, probability integral transforms, and pivot discrepancy measures may produce more accurate characterizations of goodness-of-fit but are not as well known to ecologists. In addition, a suite of visual and targeted diagnostics can be used to examine violations of different model assumptions and lack-of-fit at different levels of the modeling hierarchy, and to check for residual temporal or spatial autocorrelation. In this review, we synthesize existing literature in order to guide ecologists to the many available options for Bayesian model checking. We illustrate methods and procedures with several ecological case studies, including i) explaining variation in spatio-temporal counts of bearded seals in the eastern Bering Sea, (ii) modeling the distribution of a herbaceous plant in in the Ozark Highlands of Missouri (USA), and (iii) using resource selection functions to model habitat preferences of XXX. We argue that model checking is an essential component of scientific discovery and learning that should accompany Bayesian analyses whenever they are used to analyze ecological datasets.

Bayesian p-value, Bayesian qq-plot, count data, goodness-of-fit diagnostic check, hierarchical model, model checking, occupancy, resource selection, pivot discrepancy, predictive distribution, probability interval transform, resource selection, spatio-temporal model

Introduction

Ecologists increasingly use Bayesian methods to analyze complex hierarchical models for natural systems (Hobbs and Hooten 2015). Adoption of a Bayesian perspective requires that one specify prior distributions for model parameters, a process some have criticized for

introducing unneeded subjectivity into the scientific process (Lele and Dennis 2009). However, there are clear advantages of adopting a Bayesian mode of inference. For instance, one can entertain models that were previously intractable using common modes of frequentist statistical inference (e.g., maximum likelihood). Ecologists are using Bayesian modes of inference to fit richer classes of models to their datasets, allowing them to model features such as temporal or spatial autocorrelation, individual level random effects, hidden states, and to separate the effects of process and measurement error (Link et al. 2002, Clark and Bjørnstad 2004, Cressie et al. 2009). Applying Bayesian calculus also results in posterior probability distributions for parameters of interest; used together with posterior model probabilities, these can provide the basis for mathematically coherent decision and risk analyses (Link and Barker 2006, Berger 2013).

Ultimately, the reliability of inferences from a fitted model (Bayesian or otherwise) are dependent on how well the model approximates reality. There are multiple ways of assessing a model's performance in representing the system being studied. A first step is often to examine diagnostics that compare observed data to model output to pinpoint if and where any systematic differences occur. This process, which we term *model checking*, is an integral part of statistical inference, as it helps diagnose assumption violations and illuminate places where a model might be amended to more faithfully represent gathered data. Following this step, one might proceed to compare the performance of alternative models embodying different hypotheses using any number of model comparison or out-of-sample predictive performance metrics (see Hooten and Hobbs 2015, for a review) to gauge the support for alternative hypotheses or optimize predictive ability (Fig. 1). Note that scientific inference can still proceed if models do not fit the data well, but conclusions need to be tempered; one approach in such situations is to estimate a variance inflation

factor to adjust precision levels downward (e.g. Cox and Snell 1989, McCullagh and Nelder 1989).

Non-Bayesian statistical software often include a suite of goodness-of-fit diagnostics that allow practitioners to assess how well different models fit their data. For instance, when fitting generalized linear (McCullagh and Nelder 1989) or additive (Wood 2006) models in the R programming environment (R Core Team 2013), one can easily access diagnostics such as quantile-quantile, residual, and leverage plots. These diagnostics allow one to assess the reasonability of the assumed probability model, to examine whether there is evidence of heteroskedasticity, and to pinpoint outliers. Likewise, in capture-recapture analysis, there are established procedures for assessing overall fit as well as departures from specific model assumptions which are codified in user-friendly software such as U-CARE (Choquet et al. 2009). Results of such goodness-of-fit tests are routinely reported when publishing analyses in the ecological literature.

The implicit requirement that one conduct model checking exercises is not often adhered to when reporting results of Bayesian analyses in the ecological literature. For instance, a search of recent volumes of Ecology indicated that only 25% of articles employing Bayesian analysis on real datasets reported any model checking or goodness-of-fit testing (Fig. 2). We can think of several reasons why this might be the case. First, it likely has to do with momentum; the lack of precedent in ecological literature may lead some authors looking for templates on how to publish Bayesian analyses to conclude that model checking is unnecessary. Second, when researchers seek to publish new statistical methods, applications may be presented more as proof-of-concept exhibits than as definitive analyses that can stand up to scrutiny on their own. In such studies (and textbooks; see e.g., Royle and Dorazio 2008), topics like goodness-of-fit and model checking

are often reserved for future research, presumably in journals with less impact . We (the authors) are certainly culpable of presenting our research in this fashion. Third, all of the articles we examined did a commendable job in reporting convergence diagnostics to support their contention that Markov chains from MCMC runs had reached their stationary distribution. Perhaps there is a mistaken belief among authors and reviewers that convergence to a stationary distribution, combined with a lack of prior sensitivity, implies that a model fits the data? Finally, it may just be that those publishing Bayesian analyses in ecological literature ". . . like artists, have the bad habit of falling in love with their models" (to borrow a quote attributed to G.E.P. Box and referenced by Link and Barker (2010) with regard to model checking). We are certainly guilty of this fault as well; indeed this monograph can be viewed as a partial atonement for unrequited love.

If we accept the premise that Bayesian models in ecology should be routinely checked for compatibility with data, a logical next question is how best to conduct such checks. Unfortunately, there is no single best answer. Most texts in ecology (e.g., King et al. 2009, Link and Barker 2010, Kéry and Schaub 2012) focus on posterior predictive checks, as pioneered by Guttman (1967), Rubin (1981, 1984), and Gelman et al. (1996) (among others). These procedures are also the main focus of popular Bayesian analysis texts (e.g., Cressie and Wikle 2011, Gelman et al. 2014) and are based on the intuitive notion that data simulated from the posterior distribution should be similar to the data one is analyzing. However, "Bayesian p-values" generated from these tests tend to be conservative (biased towards 0.5) because the data are in effect used twice (once to fit the model and once to test the model; Bayarri and Berger 2000, Robins et al. 2000). By contrast, other approaches less familiar to ecologists (such as prior predictive checks, probability integral transforms, and pivot discrepancy measures) may produce more

accurate characterizations of goodness-of-fit but may require extra data for out-of-sample prediction or may be more difficult to implement.

In this monograph, we have collated relevant statistical literature with the goal of providing ecologists with a practical guide to Bayesian model checking. We start by defining a consistent notation that we use throughout the paper. Next, we work to compile a bestiary of Bayesian model checking procedures, providing positives and negatives associated with each approach. After describing several ways in which model checking results can sometimes be misleading (as with hierarchically centered models), we illustrate Bayesian model checking using three case studies. These include a species distribution model (SDM) developed from bearded seal counts (*Erignathus barbatus*) in the Chukchi Sea, an SDM developed from presence-absence data of a herbaceous plant (*Genus species*) in Missouri, and analysis of animal telemetry data. We conclude with several recommendations on how model checking results should be presented in the ecological literature.

Background and notation

Before describing specific model checking procedures, we first establish common notation. Bayesian inference seeks to describe the posterior distribution, $[\boldsymbol{\theta}|\mathbf{y}]$, of model parameters, $\boldsymbol{\theta}$, given data, $\mathbf{y}$. Here and throughout the paper, we use bold lowercase symbols to denote vectors. Matrices will be represented with bold, uppercase symbols, while roman (unbolded) characters will be used for scalars. The bracket notation $[\ldots]$ denotes a probability distribution or mass function, and a bracket with a vertical bar '|' denotes that it is a conditional probability distribution.

The posterior distribution is often written as

$$[\boldsymbol{\theta}|\mathbf{y}] \quad = \quad \frac{[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\mathbf{y}]}, \tag{1}$$

where $[\mathbf{y}|\boldsymbol{\theta}]$ is the assumed probability model for the data, given parameters (i.e., the likelihood), $[\boldsymbol{\theta}]$ denotes the joint prior distribution for parameters, and $[\mathbf{y}]$ is the marginal distribution of the data. In Bayesian computation, the denominator $[\mathbf{y}]$ is frequently ignored because it is a fixed constant that does not affect inference (although it is needed when computing Bayes factors for model comparison and averaging; Link and Barker 2006). The exact mechanics of Bayesian inference are well reviewed elsewhere (e.g., King et al. 2009, Link and Barker 2010, Hobbs and Hooten 2015), and we do not attempt to provide a detailed description here. For the remainder of this treatment, we assume that the reader has familiarity with the basics of Bayesian inference, including Markov chain Monte Carlo (MCMC) as a versatile tool for sampling from $[\boldsymbol{\theta}|\mathbf{y}]$.

In describing different model checking procedures, we will often need to reference data simulated under an assumed model. We use $\mathbf{y}_i^{rep}$ to denote a single, simulated dataset under the model that is being checked. In some situations, we may indicate that the dataset was simulated using a specific parameter vector, $\boldsymbol{\theta}_i$; in this case, denote the simulated dataset as $\mathbf{y}_i^{rep}|\boldsymbol{\theta}_i$. We use the notation $T(\mathbf{y}, \boldsymbol{\theta})$ to denote a discrepancy function that is dependent upon data and possibly the parameters $\boldsymbol{\theta}$. For instance, we might compare the the discrepancy $T(\mathbf{y}, \boldsymbol{\theta})$ calculated with observed data to a distribution obtained by applying $T(\mathbf{y}^{rep}, \boldsymbol{\theta})$ to multiple replicated data sets. Examples of candidate discrepancy functions are provided in Table 2.

Model checking procedures

Posterior predictive checks

Posterior predictive checks are the dominant form of Bayesian model checking advanced in statistical texts read by ecologists (e.g., King et al. 2009, Link and Barker 2010, Kéry and Schaub 2012, Gelman et al. 2014). Although sample size was small ($n = 25$), our survey of recent *Ecology* volumes indicated that posterior predictive checks are also the dominant form of Bayesian model checking being reported in ecological literature (if any checking is reported at all; Fig. 2). Posterior predictive checks are based on the commonsense notion that data simulated under a fitted model should be comparable to the real world data the model was fitted to. If observed data differ from simulated data in a systematic fashion (e.g, excess zeros, increased skew, lower kurtosis), it is good indication that model assumptions are not being met.

Posterior predictive checks can be used to look at differences between observed and simulated data graphically, or can be used to calculate "Bayesian p-values" (Alg. 1). Bayesian p-values necessarily involve application of a discrepancy function, $T(\mathbf{y}, \boldsymbol{\theta})$, for comparing observed and simulated data. There are several omnibus discrepancy measures that can be employed to examine overall lack-of-fit, and targeted discrepancy measures can be used to look for specific data features that systematically differ between simulated and observed data (Table 2).

Posterior predictive checks are straightforward to implement. Unfortunately, Bayesian p-values based on these checks tend to be conservative in the sense that the distribution of p-values calculated under a null model (i.e., when the data generating model and estimation model are the same) tends to be dome shaped (e.g., Fig. 3) instead of the

uniform distribution expected of frequentist p-values (Robins et al. 2000). This feature arises because data are used twice: once to approximate the posterior distribution and to simulate the reference distribution for the discrepancy measure, and a second time to calculate the tail probability (Bayarri and Berger 2000). As such, the power of posterior predictive Bayesian p-values to detect significant differences in the discrepancy measure is overstated. Evidently, the degree of conservatism can vary across data, models, and discrepancy functions, making it difficult to interpret or compare Bayesian p-values across models.

Another possible criticism of posterior predictive checks is that they rely solely on properties of simulated and observed data. Given that a lack of fit is observed, it may be difficult to diagnose where misspecification is occurring within the modeling hierarchy (e.g., poorly specified priors, errant mean structure, underdispersed error distribution, etc.). Further, a poorly specified mean structure may still result in reasonable fit of the model if the model is made sufficiently flexible through individual-level random effects (see *Avoiding potential traps with model checking*).

These cautions do not imply that posterior predictive checks are completely devoid of value. Indeed, given that tests are conservative, small (e.g., < 0.05) or very large (e.g., > 0.95) p-values are strongly suggestive of lack-of-fit. Further, graphical displays (see *Graphical techniques*) and targeted discrepancies (Table 2) may help pinpoint common assumption violations (e.g., lack of independence, zero inflation, overdispersion). However, it is less clear how to interpret p-values and discrepancies that are mildly indicative of lack-of-fit (e.g., a p-value of 0.15 or 0.2). Several researchers have developed approaches for calibrating Bayesian p-values so that they are asymptotically uniform (e.g., Dey et al. 1998, Bayarri and Berger 1999). However, these approaches can be computationally intensive

and/or difficult to implement (i.e., custom code is required for each unique application).

Some practical suggestions may help to reduce the degree of conservatism of posterior predictive p-values. Lunn et al. (2013) suggest that the level of conservatism depends on the discrepancy function used; discrepancy functions that are solely a function of simulated and observed data (e.g., proportion of zeros, distribution of quantiles) may be less conservative than those that also depend on model parameters (e.g., summed Pearson residuals). Similarly, Marshall and Spiegelhalter (2003) suggest reducing the impact of the double use of data by iteratively resimulating random effects when generating posterior predictions for each data point (a procedure they term a "mixed predictive check"). For an example of this latter approach, see *Spatio-temporal bearded seal counts*.

Pivotal discrepancy measures

In addition to overstated power to detect model lack-of-fit, posterior predictive p-values are limited to examining systematic differences between observed data and data simulated under a hypothesized model. As such, there is little ability to examine lack-of-fit at higher levels of modeling hierarchy. One approach to conducting goodness-of-fit at multiple levels of the model is to calculate pivotal quantities (Johnson 2004, Yuan and Johnson 2012). Pivotal quantities are random variables that can be functions of data, parameters, or both, that have known probability distributions that are independent of parameters (see e.g., Casella and Berger 1990, section 9.2.2). For instance, consider a simple normal (Gaussian) model

$$y \sim \mathcal{N}(\mu, \sigma^2).$$

Recall from introductory statistics classes that $z = \frac{y-\mu}{\sigma}$ has a standard $f = \mathcal{N}(0,1)$ distribution; thus z is pivotal quantity in that it has a known distribution independent of μ or σ.

This suggests a potential strategy for assessing goodness-of-fit; for instance, in a Bayesian regression model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2), \tag{2}$$

where $\mathbf{X}$ represents a design matrix and $\boldsymbol{\beta}$ is a vector of regression coefficients, we might keep track of

$$z_{ij} = \frac{y_i - \mathbf{x}_i\boldsymbol{\beta}_j}{\sigma_j} \tag{3}$$

for each of $j \in 1, 2, \ldots, n$ samples from the posterior distribution (i.e., drawing each $[\boldsymbol{\beta}_j, \sigma_j] \sim [\boldsymbol{\theta}|\mathbf{y}]$). Systematic departures of z_{ij} from the theoretical $N(0,1)$ distribution can point to model misspecification. Note that although we have again focused on the data model in Eq. 2, this same approach could be used at higher levels of the modeling hierarchy.

In practice, there are several difficulties with using pivotal quantities as discrepancy measures in Bayesian model checking. First, the joint distribution of pivotal quantities calculated across $i \in 1, 2, \ldots, n$ samples from the posterior distribution are not independent because they depend on the same observed data, $\mathbf{y}$ (Johnson 2004). As with the Bayesian p-value calculated using a posterior predictive check, this latter problem can result in p-values that are conservative. Yuan and Johnson (2012) suggest comparing histograms of

a pivotal discrepancy function $T(\mathbf{y}, \boldsymbol{\theta}_i)$ to its theoretical distribution, f, to diagnose obvious examples of model misspecification. If an omnibus Bayesian p-value is desired, a test can be implemented by appealing to limiting distributions of order statistics (Johnson 2004), but these tests tend to have low power to detect lack of fit.

A second problem is that to apply these techniques, one must first define a pivotal quantity and ascertain its reference distribution. To assess normality is relatively straightforward using standardized residuals (e.g., Eq. 3), but pivotal quantities are not necessarily available for other distributions (e.g., Poisson). However, Yuan and Johnson (2012), building upon work of Johnson (2004) proposed an algorithm based on cumulative distribution functions (CDFs) that can apply to any distribution, and at any level of a hierarchical model (Alg. 4). For continuous distributions, this algorithm works by defining a quantity $w_{ij} = g(y_{ij}, \boldsymbol{\theta})$ (this can simply be $w_{ij} = y_{ij}$) with a known CDF, Θ. Then, according to the probability integral transformation, $\Theta(\mathbf{w})$ should be uniformly distributed if the the modeled distribution function is appropriate. Similarly, for discrete distributions, we can apply a randomization scheme (Smith 1985, Yuan and Johnson 2012) to generate what should be continuously distributed uniform variates. For example, when y_{ij} has integer valued support, we can define

$$w_{ij} \quad = \quad F(y_{ij} - 1|\boldsymbol{\theta}) + u_{ij} f(y_{ij}|\boldsymbol{\theta}),$$

where u_{ij} is a continuosly uniform random deviate on (0,1) and $F()$ and $f()$ are the cumulative mass and probability mass functions associated with $[\mathbf{y}|\boldsymbol{\theta}]$, respectively. In this case, w_{ij} will be uniformly and continuously distributed on (0,1) if the assumed distribution is reasonable; deviation from this distribution can point to model misspecification.

Note that while we have written Alg. 4 in terms of the data distribution $[\mathbf{y}|\boldsymbol{\theta}]$, the algorithm can be applied (without loss of generality) to any level of a hierarchical model. Further, Alg. 4 can be applied separately to different categories of mean response (e.g., low, medium, or high levels of predicted responses). These advantages are extremely appealing in that one can more thoroughly test distributional assumptions and look for places where lack-of-fit may be occurring, something that can be difficult to do with posterior predictive checks. We apply Alg. 4 to real data in *Examples* and provide R code for applying this approach to generic MCMC data in the R package `HierarchicalGOF` accompanying this paper (see *Software* for more information).

Prior predictive checks

Box (1980) argued that the hypothetico-deductive process of scientific learning can be embodied through successive rounds of model formulation and testing. According to his view, models are built to represent current theory and an investigator's knowledge of the system under study; data are then collected to evaluate how well the existing theory (i.e., model) matches up with reality. If necessary, the model under consideration can be amended, and the process repeats itself.

From a Bayesian standpoint, such successive rounds of *estimation* and *criticism* can be embodied through posterior inference and model checking, respectively (Box 1980). If one views a model, complete with all its set of assumptions and prior beliefs, as a working model of reality, then data simulated under a model should look similar to data gathered in the real world. This notion can be formalized through a prior predictive check, where

replicate data $\mathbf{y}^{rep}$ are simulated via

$$\begin{aligned} \boldsymbol{\theta}^{rep} &\sim [\boldsymbol{\theta}] \\ \mathbf{y}^{rep} &\sim [\mathbf{y}|\boldsymbol{\theta}^{rep}] \end{aligned} \tag{4}$$

and then compared to observed data $\mathbf{y}$ via a discrepancy function (Alg. 2).

Unlike posterior predictive checks, p-values from prior predictive checks are uniformly distributed under the null model and have properly stated frequentist properties. The main problem with this approach is that the models being considered need to have considerable historical investment and proper prior distributions informed by expert opinion or data from previous studies. In many cases where Bayesian inference is employed, this is simply not the case. Still, this approach may be useful for hierarchical models that serve as an embodiment of current theory about a study system (e.g., population or ecosystem dynamics models). Alternatively, a subset of data (test data) can be withheld when fitting a model, and the posterior distribution $[\boldsymbol{\theta}|\mathbf{y}]$ can be substituted for $[\boldsymbol{\theta}]$ in Eq. 4. If used in this manner, prior predictive checks can be viewed as a form of cross validation, a subject we shall examine next.

Cross validation tests

Cross validation consists of leaving out one or more data points, rerunning analysis, and seeing how model predictions match up with actual observations. It is most often used to examine the relative predictive performance of different models (i.e., for model selection; see e.g. Arlot and Celisse 2010). However, it also possible to use cross validation techniques to examine model fit and diagnose outlier behavior. The major advantage of conducting

tests in this fashion is that there is no duplicate use of data (as with posterior predictive tests or pivotal discrepancy tests). The major disadvantage is that it can be computationally challenging for complicated hierarchical models.

One approach to checking models using cross validation in the cross-validated probability integral transform (PIT) test, which has long been exploited to examine the adequacy of probabilistic forecasts (e.g., Dawid 1984, Früiiwirth-Schnatter 1996, Gneiting et al. 2007, Czado et al. 2009). These tests work by simulating data at a set of times or locations, and computing the CDF of the predictions evaluated at the realized data (where realized data are not used to fit the model). This can be accomplished in a sequential fashion for time series data, or by withholding data (as with leave-one-out cross validation). In either case, divergence from a Uniform(0,1) distribution is indicative of a model deficiency. In particular, a U-shape suggests an underdispersed model, a dome shape suggests an overdispersed model, and skew (i.e., mean not centered at 0.5) suggests bias. Congdon (2014) suggests an algorithm for computing PIT diagnostic histograms for both continuous and discrete data in Bayesian applications (see Alg. **??**).

Cross-validation can also be useful for diagnosing outliers in spatial modeling applications. For instance, Stern and Cressie (2000) and Marshall and Spiegelhalter (2003) use Alg. **??** to identify regions that have inconsistent behavior relative to the model. Such outliers can either indicate that the model does not sufficiently explain variation in responses, that there are legitimate "hot spots" worthy of additional investigation Marshall and Spiegelhalter (2003), or both.

Residual tests

Lunn et al. (2013) suggest several informal tests based on distributions of Pearson and

deviance residuals. These tests are necessarily informal in Bayesian applications, as residuals all depend on $\boldsymbol{\theta}$ and are thus not truly independent as required in unbiased application of goodness-of-fit tests. Nevertheless, several rules of thumb can be used to screen residuals for obvious assumption violations. For example, standardized Pearson residuals for continuous data,

$$r_i = \frac{y_i - E(y_i|\boldsymbol{\theta})}{\sqrt{\text{Var}(y_i|\boldsymbol{\theta})}},$$

should generally take on values between -2.0 and 2.0. Values very far out of this range represent outliers. Similarly, for the Poisson and binomial distributions, an approximate rule of thumb is that the mean saturated deviance should approximately equal sample size for a well fitting model (Lunn et al. 2013).

For time series, spatial, and spatio-temporal models, failure to account for autocorrelation can result in bias and overstated precision (Lichstein et al. 2002). For this reason, it is important to look for evidence of residual spatio-temporal autocorrelation in analyses where data have a spatio-temporal index. There are a variety of metrics to quantify autocorrelation, depending upon the ecological question and types of data available (e.g. Perry et al. 2002). For Bayesian regression models, one versatile approach is to compute a posterior density associated with a statistic such as Moran's I (Moran 1950) or Getis-Ord G* (Getis and Ord 1992) on residuals. For example, calculating Moran's I for each posterior sample j relative to posterior residuals $\mathbf{Y} - \text{E}(\mathbf{Y}|\boldsymbol{\theta}_j)$, a histogram of I_j values can be constructed; substantial overlap with zero suggests little evidence of residual spatial autocorrelation. As calculation of Moran's I is dependent upon a a pre-specified distance weighting scheme, investigators might simulate a posterior sample of Moran's I at

several different choices of weights or neighborhoods to evaluate residual spatial autocorrelation at different scales.

Just build a bigger model! Tradeoffs between fit and prediction

cite Ver Hoef musings

Graphical techniques

Thought maybe I'd make this a separate section because these could potentially could be done with posterior or prior predictive checks, or even with PDMs. Some ideas: q-q plots, maps (for spatial data), residual and binned residual plots (Gelman et al. 2014).

Assessing path structure

Hierarchical statistical models can be represented using directed, acyclic graphs (DAGs). Such graphs represent the directed flow of probability through a model, with the assumption that connected nodes are stochastically dependent; unconnected nodes are assumed to be conditionally independent. In some models, the direction and connectivity among nodes is canonical and self-evident (for example, generalized regression models). However, in others (e.g., ecosystem models), there may be considerable uncertainty as to the appropriateness of a particular graph structure. In the latter case, a general assessment of the graph structure should also be a part of model checking exercises.

Shipley (2009) introduced a directional-separation test for assessing the path structure of DAGs. According to this test, each pair of nodes that are not directly connected constitute an independence claim (possibly conditional on the values of intervening nodes);

each such claim can be assessed via a statistical model. An overall p-value can then be calculated to assess the validity of the overall path structure.

too far afield? provide algorithm?

Avoiding potential traps with model checking

Mean structure vs. dispersion - not always obvious where misspecification occurs.

Hierarchical centering

Computing

Examples

Modeling the distribution of a herbaceous plant

Spatio-temporal bearded seal counts

Resource selection of XXX

Discussion

Focus on prior sensitivity, convergence diagnostics and sometimes model comparison (e.g. DIC or cross validation) - not as much focus on GOF.

GOF on most general model, then model selection/comparison/averaging (**?**).

Acknowledgments

Funding for Bering Sea aerial surveys was provided by the U.S. National Oceanic and

Atmospheric Administration and by the U.S. Bureau of Ocean Energy Management (Interagency Agreement M12PG00017). The views and conclusions in this article represent the views of the authors and the U.S. Geological Survey but do not necessarily represent findings or policy of the U.S. National Oceanic and Atmospheric Administration. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. Statistcs Surveys **4**:40–79.

Bayarri, M., and J. O. Berger. 2000. P values for composite null models. Journal of the American Statistical Association **95**:1127–1142.

Bayarri, M. J., and J. O. Berger, 1999. Quantifying surprise in the data and model verification. Pages 53–82 *in* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian Statistics 6. Oxford University Press, London.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics **162**:2025–2035.

Berger, J. O. 2013. Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Box, G. E. 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) pages 383–430.

Casella, G., and R. L. Berger. 1990. Statistical Inference. Duxbury Press, Belmont, CA.

Choquet, R., J.-D. Lebreton, O. Gimenez, A.-M. Reboulet, and R. Pradel. 2009. U-CARE: Utilities for performing goodness of fit tests and manipulating CApture–REcapture data. Ecography **32**:1071–1074.

Clark, J. S., and O. N. Bjørnstad. 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology **85**:3140–3150.

Congdon, P. 2014. Applied Bayesian modelling. John Wiley & Sons.

Cox, D. R., and E. J. Snell. 1989. Analysis of binary data. CRC Press.

Cressie, N., C. Calder, J. Clark, J. Ver Hoef, and C. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications **19**:553–570.

Cressie, N., and C. K. Wikle. 2011. Statistics for spatio-temporal data. Wiley, Hoboken, New Jersey.

Czado, C., T. Gneiting, and L. Held. 2009. Predictive model assessment for count data. Biometrics **65**:1254–1261.

Dawid, A. P. 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society. Series A (General) pages 278–292.

Dey, D. K., A. E. Gelfand, T. B. Swartz, and P. K. Vlachos. 1998. A simulation-intensive approach for checking hierarchical models. Test **7**:325–346.

414 Früiiwirth-Schnatter, S. 1996. Recursive residuals and model diagnostics for normal and
415 non-normal state space models. Environmental and Ecological Statistics **3**:291–309.

416 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. Bayesian data analysis,
417 Third edition. Taylor & Francis.

418 Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model
419 fitness via realized discrepancies. Statistica Sinica **6**:733–760.

420 Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance
421 statistics. Geographical analysis **24**:189–206.

422 Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. Probabilistic forecasts, calibration
423 and sharpness. Journal of the Royal Statistical Society: Series B (Statistical
424 Methodology) **69**:243–268.

425 Guttman, I. 1967. The use of the concept of a future observation in goodness-of-fit
426 problems. Journal of the Royal Statistical Society. Series B (Methodological) pages
427 83–100.

428 Hobbs, N. T., and M. B. Hooten. 2015. Bayesian Models: A Statistical Primer for
429 Ecologists. Princeton University Press.

430 Hooten, M., and N. Hobbs. 2015. A guide to Bayesian model selection for ecologists.
431 Ecological Monographs **85**:3–28.

432 Johnson, V. E. 2004. A Bayesian χ^2 test for goodness-of-fit. Annals of Statistics pages
433 2361–2384.

Kéry, M., and J. A. Royle. 2016. Applied Hierarchical Modeling in Ecology. Elsevier, London.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press.

King, R., B. Morgan, O. Gimenez, and S. Brooks. 2009. Bayesian analysis for population ecology. CRC Press, Boca Raton, Florida.

Lele, S. R., and B. Dennis. 2009. Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain? Ecology **19**:581–584.

Lichstein, J., T. Simons, S. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs **72**:445–463.

Link, W., and R. Barker. 2010. Bayesian Inference with Ecological Applications. Academic Press, London, U.K.

Link, W., E. Cam, J. Nichols, and E. Cooch. 2002. Of BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. Journal of Wildlife Management **66**:277–291.

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

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2013. The BUGS book: A practical introduction to Bayesian analysis. Chapman & Hall/CRC, Boca Raton, Florida.

454 Marshall, E. C., and D. J. Spiegelhalter. 2003. Approximate cross-validatory predictive
455 checks in disease mapping models. Statistics in Medicine **22**:1649–1660.

456 McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Chapman and Hall,
457 New York.

458 Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. Biometrika .

459 Perry, J., A. Liebhold, M. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and
460 S. Citron-Pousty. 2002. Illustrations and guidelines for selecting statistical methods for
461 quantifying spatial pattern in ecological data. Ecography **25**:578–600.

462 R Core Team, 2013. R: A Language and Environment for Statistical Computing. R
463 Foundation for Statistical Computing, Vienna, Austria. URL
464 `http://www.R-project.org`.

465 Robins, J. M., A. van der Vaart, and V. Ventura. 2000. Asymptotic distribution of P values
466 in composite null models. Journal of the American Statistical Association **95**:1143–1156.

467 Royle, J., and R. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Academic
468 Press, London, U.K.

469 Rubin, D. B. 1981. Estimation in parallel randomized experiments. Journal of Educational
470 and Behavioral Statistics **6**:377–401.

471 Rubin, D. B., et al. 1984. Bayesianly justifiable and relevant frequency calculations for the
472 applied statistician. The Annals of Statistics **12**:1151–1172.

473 Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. Ecology
474 **90**:363–368.

475 Smith, J. Q. 1985. Diagnostic checks of non-standard time series models. Journal of
476 Forecasting **4**:283–291.

477 Stern, H. S., and N. Cressie. 2000. Posterior predictive model checks for disease mapping
478 models. Statistics in Medicine **19**:2377–2397.

479 Wood, S. N. 2006. Generalized additive models. Chapman & Hall/CRC, Boca Raton,
480 Florida.

481 Yuan, Y., and V. E. Johnson. 2012. Goodness-of-fit diagnostics for Bayesian hierarchical
482 models. Biometrics **68**:156–164.

483 ALGORITHMS

Algorithm 1 Posterior predictive check algorithm for computing a Bayesian p-value, P using m samples from the posterior distribution. A selection of discrepancy measures $T(\mathbf{y}, \boldsymbol{\theta})$ are provided in Table 2.

$P \leftarrow 0$
for $i \in 1 : m$ **do**
  Draw $\boldsymbol{\theta}_i \sim [\boldsymbol{\theta}|\mathbf{y}]$
  Draw $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}_i]$
  Calculate $T_i^{rep} = T(\mathbf{y}_i^{rep}, \boldsymbol{\theta}_i)$
  Calculate $T_i^{obs} = T(\mathbf{y}_i, \boldsymbol{\theta}_i)$
  if $T_i^{obs} < T_i^{rep}$ **then**
    $P \leftarrow P + 1$
  end if
end for
$P \leftarrow P/m$

Algorithm 2 Prior predictive check algorithm for computing a Bayesian p-value, P using m samples from the posterior distribution. A selection of discrepancy measures $T(\mathbf{y}, \boldsymbol{\theta})$ are provided in Table 2.

$P \leftarrow 0$
for $i \in 1 : m$ **do**
  Draw $\boldsymbol{\theta}_i \sim [\boldsymbol{\theta}]$
  Draw $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}_i]$
  Calculate $T_i^{rep} = T(\mathbf{y}_i^{rep}, \boldsymbol{\theta}_i)$
  Calculate $T_i^{obs} = T(\mathbf{y}_i, \boldsymbol{\theta}_i)$
  if $T_i^{obs} < T_i^{rep}$ **then**
    $P \leftarrow P + 1$
  end if
end for
$P \leftarrow P/m$

Algorithm 3 Algorithm for conducting χ^2 discrepancy check to assess the distribution of modeled quantities. If distributional assumptions are reasonable, the cumulative distribution function associated with modeled quantities should be uniformly distributed ((Johnson 2004, Yuan and Johnson 2012)). Note that n denotes sample size and m denotes the number of posterior samples utilized. This method relies on binning the pivotal quantity ($w_{ij} = g(y_{ij}, \boldsymbol{\theta}_i)$ into $K \times L$ bins, where K and L are fixed by the investigator (bins should be chosen to achieve reasonable sample size in each of the KL bin combinations). We use Θ to denote the cumulative distribution function for the distribution of the pivotal quantity. Specific examples of $g()$ and Θ are provided in the text. As written, this algorithm assesses the fit of the data distribution $[\mathbf{y}|\boldsymbol{\theta}]$; however, note that it can be applied to other levels of a hierarchical model.

Set $b_l \leftarrow l/L$ for $l = 0, 1, \ldots, L$
Set $O_{ikl} \leftarrow 0 \ \forall \ i \in 1:m, k \in 1:K, l \in 1:L$
Set $n_{ik} \leftarrow 0 \ \forall \ i \in 1:m, k \in 1:K$
for $i \in 1:m$ **do**
  Draw $\boldsymbol{\theta}_i \sim [\boldsymbol{\theta}|\mathbf{y}]$
  for $j \in 1:n$ **do**
    $\mu_{ij} \leftarrow \mathrm{E}(y_j|\boldsymbol{\theta}_i)$
    $w_{ij} \leftarrow g(y_{ij}, \boldsymbol{\theta}_i)$
  end for
  Set $q_0 \leftarrow -\infty, q_K \leftarrow \infty$, and $q_h \leftarrow \text{quantile}_{h/K*100\%}(\mu_{ij})$ for $h \in 1:(K-1)$ and the quantile is taken over $j \in 1:n$)
  for $k \in 1:K$ **do**
    for $j \in 1:n$ **do**
      if $q_{k-1} \leq \mu_{ij} < q_k$ **then**
        $r_{ij} \leftarrow k$
        $n_{ik} \leftarrow n_{ik} + 1$
      end if
    end for
    for $l \in 1:L$ **do**
      if $\Theta(w_{ij}) \in (b_{l-1}, b_l]$ & $r_{ij} = k$ **then**
        $O_{ikl} \leftarrow O_{ikl} + 1$
      end if
    end for
    Set $T_{ik}(\mathbf{y}, \boldsymbol{\theta}_i) \leftarrow \sum_{l=1}^{L} \frac{(O_{ikl} - n_{ik}L^{-1})^2}{n_{ik}L^{-1}}$
  end for
  Set $T_i(\mathbf{y}, \boldsymbol{\theta}_i) \leftarrow \sum_{k=1}^{K} T_{ik}(\mathbf{y}, \boldsymbol{\theta}_i)$
end for
Test $T_{ik}(\mathbf{y}, \boldsymbol{\theta}_i) \sim \chi^2_{L-1}$ for targeted lack-of-fit
Test $T_i(\mathbf{y}, \boldsymbol{\theta}_i) \sim \chi^2_{K(L-1)}$ for omnibus lack-of-fit

Algorithm 4 Algorithm for conducting predictive probability integral transform (PIT) checks, as described by e.g. Früiiwirth-Schnatter (1996). This approach requires having "test" data; here we assume that a "leave-one-out" procedure is used, although other approaches are certainly possible (and may be preferable, especially when sample sizes are large). To this end, we define $\mathbf{y}_{-i}$ as the set of data for which the ith observation is missing, m to be the total number of observations, and n to be the number of posterior samples that are analyzed for each data set. The indicator function $I(A)$ takes on the value 1.0 if the statement A is true, and is 0.0 otherwise.

Set $u_j = 0 \forall j \in 1, 2, \ldots, n$
for $i \in 1 : m$ **do**
  for $j \in 1 : n$ **do**
    Simulate a draw $\boldsymbol{\theta}_{ij}$ from the posterior distribution $[\theta|\mathbf{y}_{-i}] \propto [\mathbf{y}_{-i}|\theta][\theta]$
    Simulate a posterior prediction $\tilde{y}_{ij}$ from the predictive density (or mass function), $[y_i|\boldsymbol{\theta}_{ij}]$
  end for
  if y_i has continuous support **then**
    Set $u_i = \sum_j I(\tilde{y}_{ij} \leq y_i)$
  end if
  if y_i has nonnegative integer support (i.e. for count data) **then**
    Set $u_i = \sum_j I(\tilde{y}_{ij} < y_i) + 0.5I(\tilde{y}_{ij} = y_i j)$
  end if
  if y_i has binary support **then**
    Set $u_i = \sum_j I(\tilde{y}_{ij} = y_i j)$
  end if
end for
Plot a histogram of u_i values. Divergence from a Uniform(0,1) distribution is indicative of lack of fit. Very high or very low values may indicate outliers.

484 TABLES

TABLE 1. Discrepancy functions and pivotal quantities useful for hierarchical model checking.

Name	Definition	Comments
A. Omnibus discrepancy functions		
χ^2	$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i \frac{(y_i - E(y_i\|\boldsymbol{\theta}))^2}{\text{var}(y_i\|\boldsymbol{\theta})}$	Often used for count data; suggested by Gelman et al. (2014) (among others).
Deviance (D)	$T(\mathbf{y}, \boldsymbol{\theta}) = -2\log[\mathbf{y}\|\boldsymbol{\theta}]$	used by King et al. (2009)
Likelihood ratio statistic	$T(\mathbf{y}, \boldsymbol{\theta}) = 2\sum_i y_i \log(\frac{y_i}{E(y_i\|\boldsymbol{\theta})})$	used by Lunn et al. (2013)
Freeman-Tukey Statistic	$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i (\sqrt{y_i} - \sqrt{\text{E}(y_i)})^2$	Less sensitive to small expected values than χ^2; suggested by Kéry and Royle (2016).
B. Targeted discrepancy functions		
Proportion of zeros	$T(\mathbf{y}) = \sum_i I(y_i = 0)$	Zero inflation check for count data
Skewness checks	$T(\mathbf{y}) = y_{p\%}$	Using the $p\%$ quantile can be useful for checking for over- or underdispersion.
C. Pivotal quantities		
$Y \sim \text{Exponential}(\lambda)$	$\lambda\bar{Y} \sim \text{Gamma}(n, n)$	Note n is sample size
$Y \sim \mathcal{N}(\mu, \sigma^2)$ (Gaussian)	$\frac{Y-\mu}{\sigma} \sim \mathcal{N}(0, 1)$	For mean μ and standard deviation σ
$Y \sim \text{Weibull}(\alpha, \beta)$	$\beta Y^\alpha \sim \text{Exponential}(1)$	
Y from *any* distribution	$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$	For large sample size (n), Z converges in distribution to a standard normal (Slutsky's theorem) and Z is termed an "asymptotically pivotal quantity."

TABLE 2. A summary of Bayesian model checking approaches. For each method, we describe whether each method allows for (1) computation of an overall p-value ("p-value?"), (2) whether the method tends to be conservative (i.e., has overstated power to detect goodness-of-fit; "conservative?"), (3) whether all levels of the modeling hierarchy can be evaluated ("all levels?"), and (4) whether out-of-sample data are needed to assess lack-of-fit ("out-of-sample?").

Method	p-value?	conservative?	all levels?	out-of-sample?
Pivotal discrepancy	Yes	Yes	Yes	No
Posterior predictive check	Yes	Yes	No	No
Prior predictive check	Yes	No	No	Yes
Predictive PIT tests	No	No	No	Yes
Graphical	No	Maybe	Yes	No

485 FIGURE CAPTIONS

486 FIGURE 1. A decision diagram describing the steps we suggest ecologists adopt when
487 reporting the results of Bayesian analyses in the literature, particularly when results will be
488 used for conservation and management or to inform ecological theory. The first step is to
489 formulate reasonable ecological models, ensuring that the model(s) and associated software
490 is free of errors and that convergence to the posterior distribution can be achieved (using
491 Markov chain Monte Carlo, for instance). Following this step, models should be checked
492 against observed data to diagnose possible model misspecification (the subject of this
493 article). Assuming no obvious inadequacies, various model comparison or averaging
494 techniques can be used to compare the predictive performance of alternative models that
495 embody different ecological hypotheses. Finally, we suggest conducting robustness analyses
496 (prior sensitivity analyses, simulation analyses where model assumptions are violated) to
497 gauge the importance of implicit parametric assumptions on ecological inference.

498 FIGURE 2. Type of model checking procedures used in $n = 31$ articles published in the
499 journal Ecology during 2014 and 2015. Articles were found via a Web of Science for articles
500 including the topic "Bayesian" (search conducted 10/1/2015). Six articles were determined
501 to be non-applicable (N/A) because they either (1) were simulation studies, or (2) used

approximate Bayesian computation, which is conceptually different than traditional Bayesian inference (see e.g. Beaumont et al. 2002). Of the remaining 25, 20 did not report any model checking procedures. Five articles reported specific model checking procedures, which included a combination of Bayesian cross validation (*Cross.val*,), frequentist software (*Non-Bayes*), posterior predictive p-values (*Pp.pval*), and posterior predictive graphical checks (*Pp.gc*). Some articles also investigated prior sensitivity which can be regarded as a form of model checking, but we do not report prior sensitivity checks here.

FIGURE 3. Summary of 10,000 Bayesian posterior predictive p-values generated using Algorithm 1 under a χ^2 discrepancy function. In each case, $n = 10$ counts were generated from a Poisson(λ) distribution where $\lambda \sim \text{Uniform}(1, 10)$ and Markov chain Monte Carlo was used to fit the correct (Poisson) model to the data. For an unbiased test, the observed histogram would be approximately uniform. In this case, the dome shape indicates the test tends to be too conservative (i.e., the probability of making a type I error is smaller than stated and the power of the test is overstated).

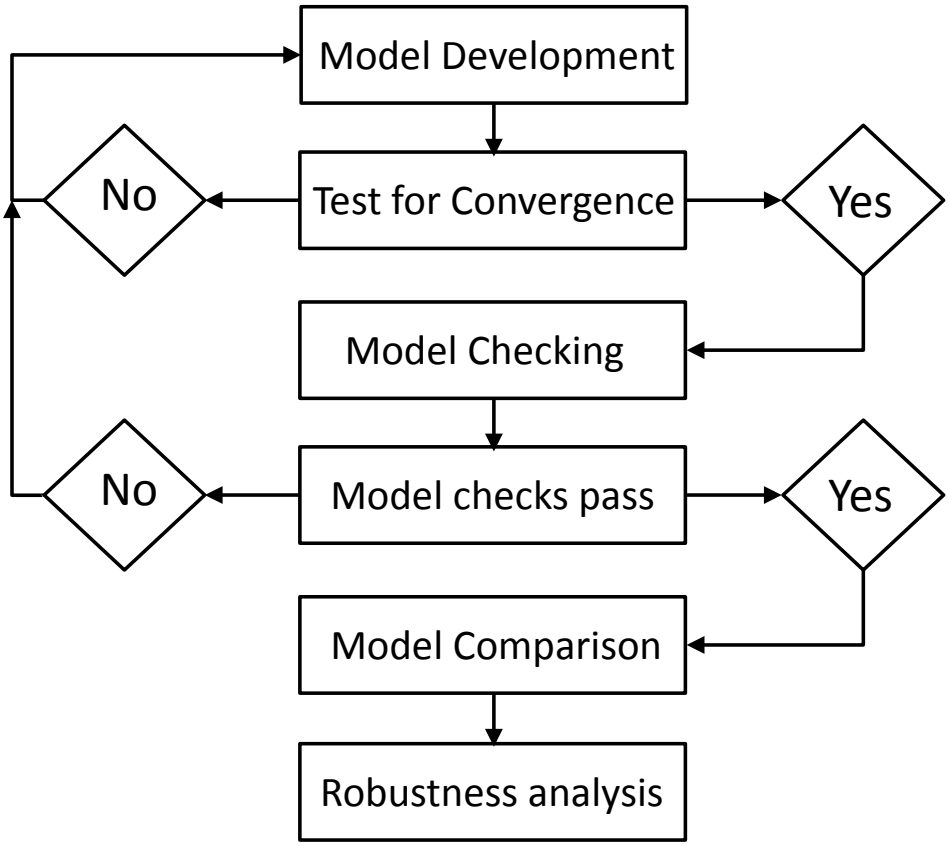
Model Development
Test for Convergence
No
Yes
Model Checking
Model checks pass
No
Yes
Model Comparison
Robustness analysis

Fig 1

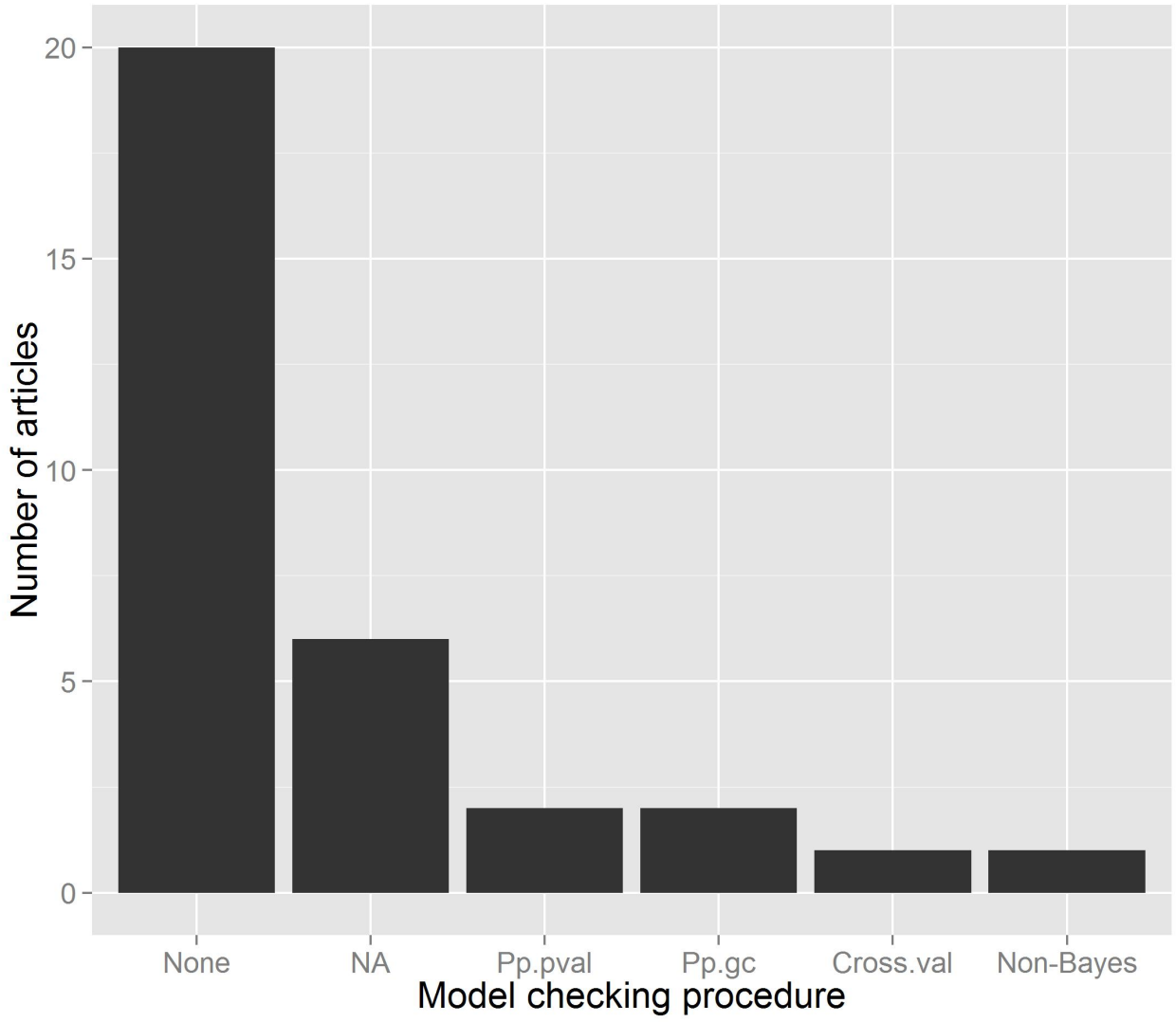
20
15
10
5
0
Number of articles
None
NA
Pp.pval
Pp.gc
Cross.val
Non-Bayes
Model checking procedure

FIG 2

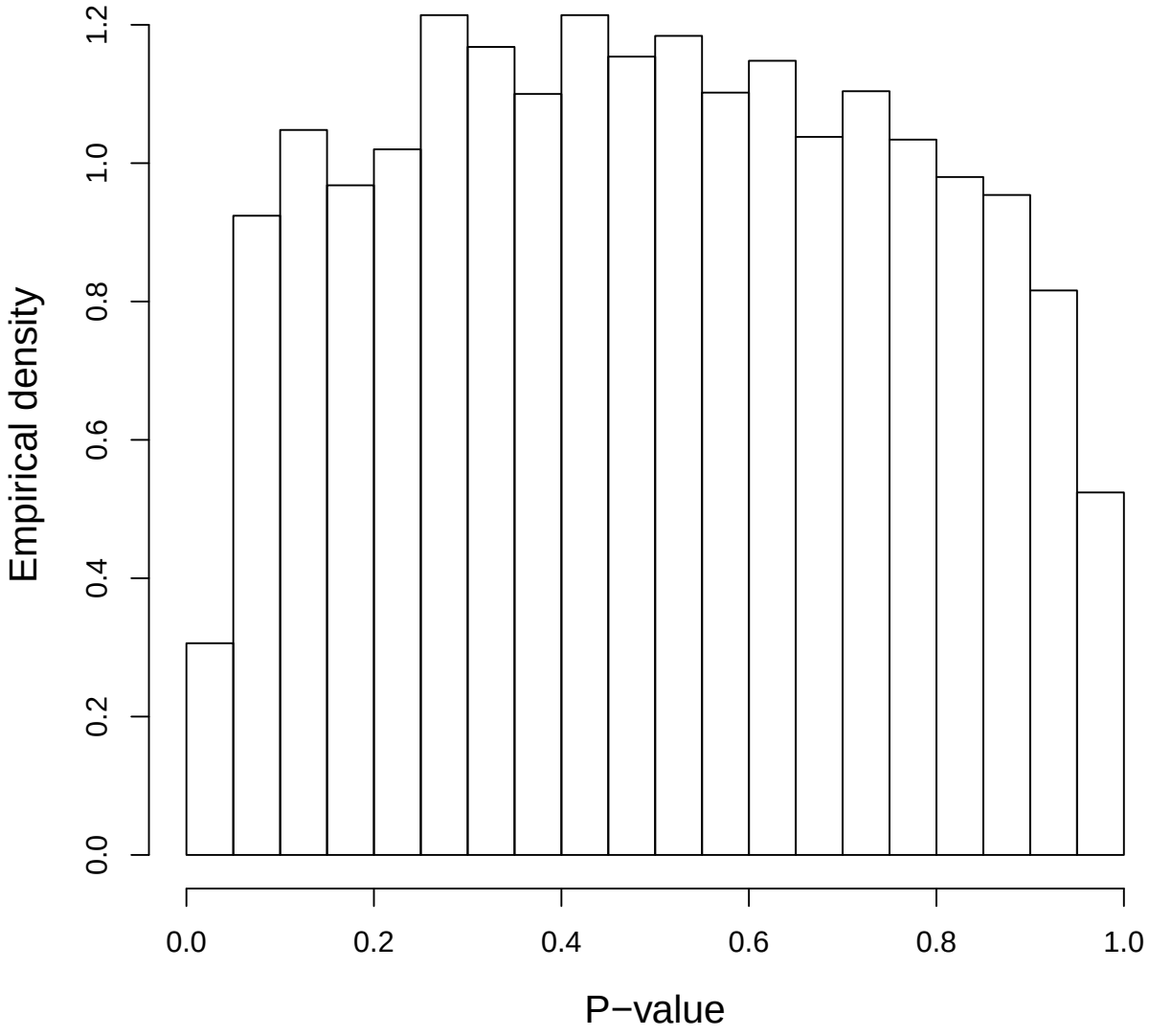

FIG 3