This is a preprint of an article published by Taylor & Francis in Communications in Statistics - Simulation and Computation,
available online: http://www.tandfonline.com/10.1080/03610918.2014.889161.

Estimation of the Predictive Ability of Ecological Models

KOHJI YAMAMURA

National Institute for Agro-Environmental Sciences, Tsukuba 305-8604, Japan

Received July 30, 2013; Accepted January 16, 2014
Address: National Institute for Agro-Environmental Sciences, 3-1-3 Kannondai, Tsukuba 305-8604, Japan
E-mail: yamamura@affrc.go.jp

The conventional criteria for predictive model selection do not indicate the absolute goodness of models. For example, the quantity of Akaike Information Criterion (AIC) has meanings only when we compare AIC of different models for a given amount of data. Thus, the existing criteria do not tell us whether the quantity and quality of data is satisfactory, and hence we cannot judge whether we should collect more data to further improve the model or not. To solve such a practical problem, we propose a criterion R_D that lies between 0 and 1. R_D is an asymptotic estimate of the proportion of improvement in the predictive ability under a given error structure, where the predictive ability is defined by the expected logarithmic probability by which the next data set (2nd data set) occurs under a model constructed from the current data set (1st data set). That is, the predictive ability is defined by the expected logarithmic probability of the 2nd data set evaluated at the model constructed from the 1st data set. Appropriate choice of error structures is important in the calculation of R_D. We illustrate examples of calculations of R_D by using a small data set about the moth abundance.

Keywords error structure; fixed dispersion parameter; generalized linear model; GLMM; model selection; predictive ability;

1. Introduction

A wide range of statistical models are used to predict ecological variables such as animal abundance (e.g., Yamamura et al., 2006; Yamamura et al., 2008). Simultaneously, statistical models are used to understand the principal mechanism that determines the ecological variables. We can use various criteria, such as AIC, AICc, BIC, and C_p, to select the best model for prediction among several candidate models (Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Konishi and Kitagawa, 2008). AIC (Akaike Information Criterion) was proposed to measure the closeness between the true probability distribution and the model's predicted probability distribution by measuring the Kullback-Leibler divergence that derives from Shannon's lemma about information (Kullback and Leibler, 1951; Akaike, 1973; Miyagawa, 1979). AICc was proposed by Sugiura (1978) as an unbiased version of AIC for special cases, i.e., fixed-effect models having a single normal error. On the other hand, BIC (Bayesian Information Criterion) adopts a model that maximizes the average likelihood for the current data under the assumption that the parameters are fluctuating by following prior distributions (Schwarz, 1978).

The conventional criteria of model-selection do not indicate the absolute goodness of models; they only indicate the relative goodness of models among candidate models for a given amount of data. For example, the quantity of AIC has no meaning by itself; it has meanings only when

we compare AIC of different models for a given amount of data. Anderson (2007, p.32) described the current situation of model-selection as follows: "In model selection, we are really asking which is the best model *for a given sample size*." Thus, the existing criteria do not tell us whether the quantity and quality of data is satisfactory, and hence we cannot judge whether we should collect more data to further improve the model or not.

In this paper, we propose a criterion R_D that enables the evaluation of the absolute goodness of models in their predictive ability. R_D is an estimate of the proportion of improvement in the predictive ability of models under a given error structure. A simple R function is provided in electronic appendices so that we can calculate R_D for generalized linear models.

Table 1
Number of captured male adults of the Oriental leafworm moth, *Spodoptera litura*

Trap number	Month			
	5	6	7	8
1	8	16	55	341
2	16	48	112	874

2. Problems in the Conventional Criteria

Table 1 indicates a portion of data on the number of male adults of the Oriental leafworm moth, *Spodoptera litura* (Fabricius) (Lepidoptera: Noctuidae), captured by pheromone traps (Wakamura et al., 1992; Yamamura, 2009). We use this small data set for the purpose of illustration so that we can easily perform the calculation. *S. litura* is one of the major pest insects of soybean and many other crops in the western part of Japan. These leafworms start their reproduction from small populations each year. Then, they quickly grow in number from May to September by means of rapid reproduction over several generations. Simultaneously, they disperse over a wider area of Japan. The seasonal change in their dispersal ability has been well studied in mark-recapture experiments (Yamamura, 2002).

Two factors are included in the data of Table 1: trap (2 levels) and month (4 levels). We treat the factor of trap as a nominal variable while treating the factor of month as a continuous variable. We first transform the number of moths by the logarithmic transformation to enhance the additive property and homoscedasticity. The logarithmic transformation for a variable y is generally given by

$$\log_e(y + w/2), \tag{1}$$

where w is the width of discreteness of y (Yamamura, 1999); we have $w = 0$ if y is a continuous variable. For the count data of the number of moths in Table 1, we have $w = 1$. We will discuss the meanings of the transformation in detail in a later section. We use the following model:

$$\log_e(x_{ij} + 0.5) = a_0 + a_i + b_0 M_j + b_i M_j + e_{ij}, \quad e_{ij} \sim N(0, \phi), \tag{2}$$

where x_{ij} is the number of moths captured in the ith trap in the jth month; M_j is the month (M_j = 5, 6, 7, and 8); a_0 is the intercept; a_i is the effect of the ith trap (i = 1, 2); b_0 is the coefficient for month; b_i is the coefficient for the interaction between trap and month; and e_{ij} is the random variable following a normal distribution having a variance ϕ. We define the hierarchical family as the series of models in which the interaction terms are included in the model only if all of the corresponding lower-order terms are included in the model. We should use only the models that belong to the hierarchical family because the interaction terms are usually defined as the components that are not explained by the corresponding lower-order terms. Hence, we compare the five models of a hierarchical family having the following sets of parameters: (a_0, a_i), (a_0, b_0), (a_0, a_i, b_0), (a_0, a_i, b_0, b_i), and the null model (a_0). We refer to these models as Models A, B, C, D, and E, respectively. In order to simplify the analysis for the purpose of illustration, we do

not consider quadratic terms of the effect of month, although higher order terms are also important in these data (Yamamura, 1993).

Table 2
Calculation of R_D for the moth data listed in Table 1

Factors	df	SS	F	P	$\hat{\phi}$	k	R_D	AICc	BIC
Model A: (a_0, a_i)									
Trap	1	1.44	0.53	0.496	2.73	2	0.061	40.43	34.66
Error	6	16.36							
Model B: (a_0, b_0)									
Month	1	15.68	44.50	5.5×10^{-4}	0.35	2	0.846	24.06	18.30
Error	6	2.11							
Model C: (a_0, a_i, b_0)									
Trap	1	1.44	10.57	0.023					
Month	1	15.68	115.45	1.2×10^{-4}	0.14	3	0.907	24.31	11.29
Error	5	0.68							
Model D: (a_0, a_i, b_0, b_i)									
Trap	1	1.44	8.52	0.043					
Month	1	15.68	93.08	6.5×10^{-4}	0.17	4	0.889	42.91	13.31
Trap×Month	1	0.01	0.03	0.868					
Error	4	0.67							
Model E: (a_0)									
Error	7	17.80			2.54	1	0	35.50	33.26

Five models were fitted to the logarithmic number of captured moth: $\log_e(x + 0.5)$. The results of the Type I ANOVA are also shown. df is the degree of freedom. SS is the sum of squares for the corresponding factors and errors. k is the number of fixed-effect parameters.

The quantity of AICc became smallest (24.06) in Model B, which contains only the influence of month (b_0) (Table 2). In the conventional procedure of model-selection by AICc, we usually adopt the model that indicates the smallest AICc while discarding all other models. Thus the adopted model is

$$\log_e(x_{ij} + 0.5) = -4.018 + 1.252M_j + e_{ij}, \quad e_{ij} \sim N(0, 0.352). \tag{3}$$

The quantity of BIC became smallest (11.29) in Model C, which contains the main effects of trap (a_i) and month (b_0). If we use a zero-sum constraint so that the intercept is not influenced by the addition of factors, the adopted model is

$$\log_e(x_{ij} + 0.5) = -4.018 - 0.424d_i + 1.252M_j + e_{ij}, \quad e_{ij} \sim N(0, 0.136), \tag{4}$$

where d_i is a dummy variable: $d_1 = 1$ and $d_2 = -1$.

Each constructed model involves a certain amount of uncertainty. To ameliorate the uncertainty of a single model, an increasing number of researchers are adopting a model-averaging approach in which the AICc (or BIC) of several models are utilized as a weight in creating an averaged model (e.g., Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Wheeler and Bailer, 2009). Several recent versions of statistical software are supporting certain types of model-averaging procedures (SAS Institute Inc., 2010a, b). In the model-averaging procedure in JMP software, for example, all non-null models including non-hierarchical family are averaged. Then, if we use AICc-weight given by exp(−0.5×AICc), we obtain the following averaged model for the moth data.

$$\log_e(x_{ij} + 0.5) = -4.016 - 0.138d_i + 1.252M_j - 0.019d_iM_j. \tag{5}$$

The quantities of AICc (and BIC) are not influenced by the re-parameterization of the model. On the other hand, the averaged model changes depending on the re-parameterization in several cases. This indicates the logical inconsistency of the model-averaging approach. For example, let us express Eq. 2 by a "centered form":

$$\log_e(x_{ij}+0.5) = a_0 + a_i' + b_0 M_j + b_i (M_j - 6.5) + e_{ij}, \quad e_{ij} \sim N(0,\phi), \tag{6}$$

where $a_i' = a_i + 6.5 b_i$. In this parameterization, the averaged model weighted by AICc-weight is

$$\log_e(x_{ij}+0.5) = -4.016 - 0.198 d_i + 1.252 M_j - 0.000 d_i M_j. \tag{7}$$

Thus, the averaged model constructed from Eq. 6 is different from that constructed from Eq. 2, although Eqs. 2 and 6 are completely the same model. A model-averaging approach is also applicable for BIC. The weight created by BIC has a clearer meaning than AICc-weight; it corresponds to the posterior probability of the model, because BIC indicates the average posterior probability that is based on the assumption that all parameters are fluctuating by following prior distributions.

We cannot judge whether the current best model such as Eq. 3 is sufficiently good in their predictive ability or not. We can construct a better model if we collect a larger amount of data, such as by using a larger size of traps. Similarly, we may be able to construct a better model if we provide other explanatory variables, such as the temperature and wind speed. However, we cannot judge whether we should increase the quantity and quality of data to further improve the model or not. To enable the evaluation of the absolute goodness of models in their predictive ability, we should first discuss the definition of the predictive ability and the true model. The definitions will be logical rather than mathematical.

3. Derivation of the Estimate of Predictive Ability

3.1. Definition of Predictive Ability

We can define the predictive ability by using various arbitrary scoring rules (e.g., Gneiting and Raftery, 2007). However, we think that the predictive ability should be most directly defined by the probability that the predicted event actually occurs in the future. Thus, the predictive ability should be based on a probability such as "the probability that the next data set (2nd data set) occurs under the predictive model constructed from the current data set (1st data set)." We should evaluate the total probability that will result after repeating many independent predictions. Hence, we should use the geometric mean of probability instead of the arithmetic mean of probability. Furthermore, we should use the logarithm of the geometric mean instead of the geometric mean itself, because a multiplicative process becomes an additive process by a logarithmic transformation. Let us imagine the simplest case where we have only two kinds of prediction: good predictions having a high probability of hits and bad predictions having a low probability of hits. Let p be the proportion of good predictions. Let q_1 and q_2 be the probability of hits in a good prediction and a bad prediction, respectively. Then, the logarithm of the geometric mean of probability of hits is given by $p\log_e(q_1) + (1-p)\log_e(q_2)$. It is expressed by $\log_e(q_2) + p(\log_e(q_1) - \log_e(q_2))$. The quantity changes linearly with increasing the p. If we calculate the logarithm of the geometric mean of probability of hits after a long sequence of predictions, therefore, the quantity increases linearly with increasing the proportion of good predictions. Hence, we can evaluate the goodness of prediction more appropriately by using the logarithmic form. The logarithm of the geometric mean of probability is identical to the expectation of logarithmic probability. Hence, we define the predictive ability by the expected logarithmic probability of the next data set (2nd data set) evaluated at the predictive model constructed from the current data set (1st data set) with given dispersion parameters. We will denote the predictive ability by $E_2(l_1)$ in later sections. A unit change in $E_2(l_1)$ can be interpreted as a unit change in the proportion of some kind of "good" prediction. We later define R_D as an estimate of the relative quantity of $E_2(l_1)$.

We should note that this definition of predictive ability, $E_2(l_1)$, has no relation with "information" such as Shannon information and Kullback-Leibler divergence, although the equations are almost identical. Historically, Shannon (1948) adopted the logarithmic form in

defining the amount of "information" by somewhat similar reasons about the additive property. However, the measure of information was constructed to quantify the rarity of a given sequence of events. Shannon (1948) said that the measure of information is "a measure of how much 'choice' is involved in the selection of the event or of how uncertain we are of the outcome." Thus, the meaning of the measure of information is essentially different from our measure of predictive ability, although the equations are almost identical. Shannon information is alternatively called "entropy", but Shannon information has again no direct relation with thermodynamic entropy, except that the equations are almost identical. Hence, the physical interpretation of entropy is not applicable to Shannon information. Tribus and McIrvine (1971, p180) described the history as follows. "In 1961, one of us (Tribus) asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' " "

3.2. Laplace Definitions of True Probability and True Model

We use the philosophical definition given by Laplace (1825): the true probability is defined as the components that we cannot predict by a model that includes all knowledge we can use. Laplace (1825) considered that the probability arises from our lack of knowledge. Laplace said that "We ought then to consider the present state of the universe as the effect of its previous state and as the cause of that which is to follow. An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up . . . For such an intelligence nothing would be uncertain, and the future, like the past, would be open to its eyes." Even if an event is deterministically generated by a specific explanatory factor, we must define the event as a true random component if we have not measured the specific explanatory factor due to our lack of knowledge. The probability of occurrence of events is an objective probability, not a subjective probability, because the form of the probability model is uniquely determined under a given set of knowledge. The true probability and the true model change with the progress of science. Laplace (1825) gave the example of Halley's Comet. The emergence of Halley's Comet was a probabilistic event in ancient times under the existing knowledge, but it later became a deterministic event due to the progress of our knowledge; the component of true probability became smaller and finally it nearly disappeared along with the progress of science. The stochastic model for Halley's Comet in ancient times and the deterministic model for Halley's Comet under the current knowledge are both the true models at the time they were constructed. Similarly, we are currently using probability forecasting for rainfall, but the probability component of weather forecasting may become much smaller in the future due to the progress of our technology. The true probability, as well as the true model, continuously changes along with the advances in technology. Thus, no absolute true model exists in nature; the true model is determined by the current knowledge based on the limitations of the current science. Laplace (1825) described that "probability is relative in part to this ignorance and in part to our knowledge."

In this Laplace definition of true probability, we must define the true probability as the variability that is not explained by the saturated model or the most complex model, where the saturated model is defined by the model that includes all parameters we can use. Simultaneously, we must define the true model as the saturated model or the most complex model. The term "model" in true model is used as the same manner as the saturated model, null model, and maximal model. It discusses only the separation between systematic components and random components. Hence, we must use other knowledge about the distribution of random components and the link function to perform appropriate inference. The random components are then estimated from the true probability that is given by the residuals of the true model. If we did not use appropriate knowledge about the distribution, the inference about the model will

be inappropriate. The appropriate choice of error structure is thus important in calculating the predictive ability. We will later discuss this issue in section 4.2.

3.3. Estimation of Predictive Ability

It is extremely important for us to understand that a true model does not provide the highest predictive ability when the parameters of the model are estimated from a limited amount of data. A simpler model, which is a false model because it ignores some factors in the true model, provides the highest predictive ability in many cases. We can estimate the predictive ability by using a simple form approximately as follows.

We use the quantity of l_1 that is defined by the maximal logarithmic likelihood for the current data set (1st data set) with given dispersion parameters. Unbiased estimates of dispersion parameters are separately obtained from the true probability components that are the residuals from the true model which is given by the saturated model or the most complex model due to the Laplace definition. Let $l_{1\mathrm{AIC}}$ be the maximal logarithmic likelihood for the current data set (1st data set), in which the fixed-effect parameters and dispersion parameters are estimated simultaneously. It should be noted that we are using l_1 but not using $l_{1\mathrm{AIC}}$; this is an important difference. In the derivation of AIC or AICc, as we will see later, the quantity of $l_{1\mathrm{AIC}}$ is used instead of l_1. The comparison of $l_{1\mathrm{AIC}}$ does not indicate the comparison of probabilities although it indicates the comparison of likelihood. *We can compare the probabilities only if we use the same definition of probability for all models by fixing the dispersion parameters*.

We use the following well-known asymptotic result. We omit the proof which is somewhat complicated, but it is fully given in several studies (e.g., Takeuchi, 1976; Burnham and Anderson, 2002; Murata, 2005; Amari, 2007; Kitagawa, 2007; Konishi and Kitagawa, 2008). Let $E_2(l_1)$ be the expected logarithmic probability of the next data set (2nd data set) evaluated at the predictive model f constructed from the current data set (1st data set) with given dispersion parameters. $E_2(l_1)$ is the mathematical expression of predictive ability in the definition described above. Let $f(x|\boldsymbol{\theta})$ be the probability distribution of x for a model f with parameter vector $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ be the maximum likelihood estimates of $\boldsymbol{\theta}$ for an infinite amount of data under the model f. Let k be number of fixed-effect parameters in the model f. Then, $E_2(l_1)$ is asymptotically given by

$$E_2(l_1) = E_1(l_1) - \mathrm{tr}(\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)^{-1}), \tag{8}$$

where $E_1(l_1)$ is the expected logarithmic probability of the current data set (1st data set) evaluated at the predictive model constructed from the current data set (1st data set) with given dispersion parameters. $\mathbf{I}(\boldsymbol{\theta}_0)$ and $\mathbf{J}(\boldsymbol{\theta}_0)$ are the following $k \times k$ matrices $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ evaluated at the parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_0$:

$$\mathbf{I}(\boldsymbol{\theta}) = \int g(x) \frac{\partial \log_e f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log_e f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} dx, \tag{9}$$

$$\mathbf{J}(\boldsymbol{\theta}) = -\int g(x) \frac{\partial^2 \log_e f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} dx, \tag{10}$$

where $g(x)$ is the probability distribution of x for the true model. The difference between $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ is given by the following formula (Konishi and Kitagawa, 2008, p50):

$$\mathbf{I}(\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta}) = \int \frac{g(x)}{f(x|\boldsymbol{\theta})} \frac{\partial^2 f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} dx. \tag{11}$$

For the true model (i.e., the saturated model or the most complex model), we have $f(x|\boldsymbol{\theta}_0) = g(x)$. Hence, we have $\mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$ from Eq. 11. Then, we obtain $\mathrm{tr}(\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)^{-1}) = k$ in Eq. 8 under the regularity conditions, where k is the number of fixed-effect parameters in the model f. For the true model, therefore, we can estimate $E_2(l_1)$ by

$$\hat{E}_2(l_1) = l_1 - k. \tag{12}$$

The relation given by Eq. 12 is not applicable to false models, where the false models are constructed by dropping several parameters from the true model. Complicated calculations are required for exactly evaluating Eq. 8 in such cases (for example, Fujikoshi and Satoh, 1997). For simplicity, we use Eq. 12 for false models as well as the true model as a practical approximation in the next section. We will later discuss this issue by conducting simulation experiments in section 5.

3.4. Proportion of Improvement in the Predictive Ability

The "intelligence" of Laplace (1825) has the highest predictive ability because the "intelligence" completely knows the future data set in the definition of Laplace (1825). On the other hand, we have the lowest predictive ability when we have no explanatory variable. Hence, we consider the following measure, R_{pred}, that indicates the proportion of improvement in the predictive ability between these two extremes.

$$R_{\text{pred}} = \frac{E_2(l_1) - E_2(l_{1,\text{null}})}{E_2(l_{2,\max}) - E_2(l_{1,\text{null}})} = 1 - \frac{E_2(l_{2,\max}) - E_2(l_1)}{E_2(l_{2,\max}) - E_2(l_{1,\text{null}})}. \tag{13}$$

$E_2(l_1)$ is the expected logarithmic probability of the next data set (2nd data set) evaluated at the model constructed from the current data set (1st data set) with given dispersion parameters. $E_2(l_{1,\text{null}})$ is the quantity of $E_2(l_1)$ in which the null model is used for the predictive model. The null model is defined by the model that contains no explanatory variable; it contains only an intercept. In contrast, $E_2(l_{2,\max})$ is the expected logarithmic probability of the next data set (2nd data set) evaluated at the maximal model constructed from the next data set (2nd data set) with given dispersion parameters. The maximal model is defined by the model in which the number of fixed-effect parameters is the same as the number of observations. Thus, $E_2(l_{2,\max})$ is the *predictive ability of the best model constructed by the "intelligence" of Laplace (1825) under our error structure*. The quantity of Eq. 13 becomes 100% for the "intelligence" of Laplace (1825) while the quantity becomes 0% for a person having no explanatory variable.

If we are handling a single dependent variable, we estimate $E_2(l_1)$ and $E_2(l_{1,\text{null}})$ by $l_1 - k$ and $l_{1,\text{null}} - 1$, respectively, using Eq. 12; where $l_{1,\text{null}}$ is the observed logarithmic probability of the current data set (1st data set) evaluated at the null model constructed from the current data set (1st data set) with given dispersion parameters. It is important for us to notice that we have a logical equality, $E_2(l_{2,\max}) = E_1(l_{1,\max})$; where $E_1(l_{1,\max})$ is the expected logarithmic probability of the current data set (1st data set) evaluated at the maximal model constructed from the current data set (1st data set) with given dispersion parameters. Hence, we can estimate $E_2(l_{2,\max})$ simply by $l_{1,\max}$, that is, the observed logarithmic probability of the current data set (1st data set) evaluated at the maximal model constructed from the current data set (1st data set) with given dispersion parameters. Therefore, we obtain the estimate of R_{pred} by the following formula. We define this estimate as R_D.

$$R_D = 1 - \frac{l_{1,\max} - l_1 + k}{l_{1,\max} - l_{1,\text{null}} + 1}, \tag{14}$$

where dispersion parameters are estimated from the true probability that is given by the true model (i.e., the saturated model or the most complex model). If the number of dependent variables is θ, the denominator in Eq. 14 should be replaced by $l_{1,\max} - l_{1,\text{null}} + \theta$, because the number of intercepts is θ in such cases. As for a fixed-effect linear model having a single normal error or a generalized linear model having a single dispersion parameter (denoted by ϕ), we can calculate R_D by a simpler form by using the deviance where the dispersion parameter is estimated from the true probability.

$$R_D = 1 - \frac{D + 2k\hat{\phi}}{D_{\text{null}} + 2\hat{\phi}}, \tag{15}$$

where D is the deviance of the current data set, that is, $D = 2\hat{\phi}(l_{1,\max} - l_1)$. D_{null} is the quantity of D for the null model.

Electronic appendix A includes a simple R function for calculating R_D for generalized linear models by using Eq. 15. The function calculates R_D for all models that belong to the hierarchical family in which the interaction terms are included in the model only if all of the corresponding lower-order terms are included in the model. Then, it lists the models in descending order of R_D. The objects generated by the glm function in R are used in this function (R Development Core Team, 2011). Electronic appendix B includes a SAS macro for calculating R_D for generalized linear mixed models (GLMMs) by using Eq. 14. Proc GLIMMIX and Proc MIXED of SAS is used in this program (SAS Institute Inc., 2010b). Electronic appendix C describes a variant of R_D that is called RS_D. All electronic appendices are available from the following web site:
http://cse.naro.affrc.go.jp/yamamura/RD_criterion_program.html.

3.5. Relation with Conventional Criteria

The primary purpose of the calculation of R_D is to enable the evaluation of the quantity and quality of data. The purpose is slightly different from that of existing criteria which evaluate the relative goodness of models for a given amount of data. However, it will be meaningful for us to examine the relation between R_D and existing criteria. In several specific situations, the maximization of R_D coincides with the optimization of conventional criteria. In such cases, therefore, we can calculate the maximal quantity of R_D by using the optimal quantity of existing criteria.

For a fixed-effect model having a single normal error, the maximization of R_D is identical to the minimization of C_p of Mallows (1973; 1995), because Mallows' C_p is calculated by

$$C_p = \frac{D}{\hat{\phi}} + 2k - n\,, \tag{16}$$

where n is the number of observations. In this case, therefore, we can calculate the maximal quantity of R_D by using the minimal quantity of C_p.

Let $E_2(l_{1\text{AIC}})$ be the expected logarithmic likelihood of the next data set (2nd data set) evaluated at the predictive model f in which the fixed-effect parameters and dispersion parameters (including random-effect parameters and error variances) are estimated simultaneously from the current data set (1st data set). Let k_{AIC} be the number of parameters including fixed-effect parameters and dispersion parameters in the model f. AIC uses a similar asymptotic result as Eq. 12,

$$\hat{E}_2(l_{1\text{AIC}}) = l_{1\text{AIC}} - k_{\text{AIC}}\,, \tag{17}$$

assuming that Eq. 17 approximately applies to false models as well as the true model. We should again notice that the meaning of Eq. 17 is much different from that of Eq. 12; we should not confuse these two equations. AIC is then defined by

$$\text{AIC} = -2l_{1\text{AIC}} + 2k_{\text{AIC}}. \tag{18}$$

For a fixed-effect model having a Poisson error or a binomial error, the dispersion parameter is fixed at 1 beforehand, and hence we have $l_{1\text{AIC}} = l_1$ and $k_{\text{AIC}} = k$. For these cases, therefore, we can calculate the maximal quantity of R_D by using the minimal quantity of AIC.

For the maximum quasi-likelihood estimation, maximizing R_D is identical to minimizing QAIC of Burnham and Anderson (2002) except for the trivial difference between k and k_{AIC}. In this case, therefore, we can calculate the maximal quantity of R_D by using the minimal quantity of QAIC, where QAIC is defined by

$$\text{QAIC} = -2l_1 + 2k_{\text{AIC}}. \tag{19}$$

The quantity of l_1 in Eq. 19 is not the exact log-likelihood but the quasi-likelihood defined by McCullagh and Nelder (1989, p325), although we use the same notation for simplicity in this paper. The quasi-likelihood behaves like the log-likelihood in most cases, and hence the same

argument is applicable; we will illustrate it by using simulations in section 5.3. The dispersion parameter is estimated from the most complex model (Burnham and Anderson, 2002). Generalized Pearson chi-square statistics divided by the residual degrees of freedom are usually used for estimating the dispersion parameter (McCullagh and Nelder, 1989).

The coefficient of determination, R^2, is a classical index measuring the proportion of explanation of candidate models. R^2 indicates how well the model explains the current data set (1st data set) but it does not indicate how well the model predicts the next data set (2nd data set). R^2 is applicable only for normal fixed-effect models, and hence various extensions of R^2 have been proposed (e.g., Nagelkerke, 1991; Cox and Wermuth, 1992; Zheng, 2000; Xu, 2003; Liu et al., 2008; Orelien and Edwards, 2008; Recchia, 2010). Among these, the adjusted likelihood ratio index that was proposed by Ben-Akiva and Lerman (1985), which is also referred to as the adjusted McFadden's pseudo R^2, is somewhat similar to R_D. The adjusted McFadden's pseudo R^2 is defined by

$$R_{\mathrm{McF}} = 1 - \frac{l_{1\mathrm{AIC}} - k_{\mathrm{AIC}}}{l_{1\mathrm{AIC,null}}}, \tag{20}$$

where $l_{1\mathrm{AIC,null}}$ is $l_{1\mathrm{AIC}}$ evaluated at the null model (Long and Fre, 2000). R_{McF} becomes nearly equal to R_D if $l_{1,\max}$ is fixed at 0 and if the dispersion parameter is fixed at 1; for example, if we perform logistic regression for Bernoulli trials without considering overdispersion.

4. Example Calculation of R_D

4.1. Abundance of Moths

Table 2 includes the results of the Type I ANOVA for the data on the abundance of moths given in Table 1. The significance probabilities of the effect of trap were quite different between models. The effect of trap was significant at the 0.05 level in Models C and D (P = 0.02 and 0.04, respectively), but it was not significant in Model A (P = 0.50). The sum of squares (SS) of the effect of trap was the same for all models; it was 1.44 for Models A, C, and D. On the other hand, the estimate of the dispersion parameter in Model A was much larger than those for Models C and D: $\hat{\phi}$ = 2.73 for Model A, while $\hat{\phi}$ = 0.14 and 0.17 for Model C and D, respectively. This is because the estimate of the dispersion parameter in Model A was erroneously contaminated by the variability of the effect of month. Consequently, the quantity of F became much smaller in Model A. Hence, the result of the F-test in Model A is erroneous. This would have been the reason that Draper and Smith (1998) recommended the use of their "pure errors" in the F-test. The calculation of R_D is based on a procedure similar to that of Draper and Smith (1998); we always use the dispersion parameter that is calculated from the true probability that is given by the saturated model or the most complex model. In this case, we always use the dispersion parameter that was estimated from Model D: $\hat{\phi}$ = 0.17. In a fixed effect model having a single normal distribution of error, the deviance is identical to the residual sum of squares. Hence, the quantity of R_D in Model C, for example, is calculated by R_D = 1 − (0.68 + 2×3×0.17)/ (17.80+2×0.17) = 0.907 by using Eq. 15. The example R program for calculating R_D in Table 2 is given in electronic appendix A. When we use a *data frame* named `MothData` that contains columns of the dependent variable (`y`) and the explanatory variables (`trap` and `month`), for example, we can calculate R_D for the models that belong to the hierarchical family by using a program such as `RDcompare(log(y+0.5) ~ trap*month, data=MothData)`. The quantity of R_D was sufficiently high for Model C: there was a 90.7% improvement in the predictive ability. Hence, we can judge that the current quantity and quality of data was satisfactory for constructing a model for prediction.

4.2. Erroneous calculation of R_D

The quantity of R_D indicates the proportion of improvement in the predictive ability under the given error structure. Therefore, the appropriate choice of error structure is important in calculating the predictive ability. We previously analyzed the data on moth abundance. These data are so-called "count data." Therefore, several researchers may consider that these data

should be analyzed by using Poisson regression as described in the textbooks on generalized linear models. The results of calculation using Poisson errors are shown in Table 3. The quantity of R_D is about 0.98 for both Model C and Model D. This value is extremely large; the quantity of $1- R_D$ is only 2%. However, we should notice that the *P*-values are extremely small ($P < 2.2\times10^{-16}$) except for the interaction term in Model D. This constitutes a typical misuse of Poisson regression. This misuse belongs to a wider class of misuse that is called pseudoreplication; it was first pointed out by Hurlbert (1984, p205) for binomial distributions as well as for normal distributions.

The observed number of individuals will fluctuate following a Poisson distribution for a given expectation. However, the expectation itself will also usually fluctuate. The variance of the number of individuals (x) is generally given by

$$V(x) = V(\mu) + E(V_{\text{local}}(x)), \tag{21}$$

where E and V indicate the global expectation and global variance, while μ and $V_{\text{local}}(x)$ indicate the local expectation and local variance, respectively. We have the Poisson variability around the local expectation for count variables. Hence we can replace $V_{\text{local}}(x)$ by μ in Eq. 21 because the variance of a Poisson variable is equal to its mean. The calculation of R_D by using the Poisson regression will be appropriate only if the first term in the right-hand side of Eq. 21 is very small.

Table 3
Erroneous calculation of R_D for the moth data listed in Table 1

Factors	df	LR	*P*	*k*	R_D
Model A: (a_0, a_i)					
Trap	1	278.9	$<2.2\times10^{-16}$	2	0.108
Model B: (a_0, b_0)					
Month	1	2243.2	$<2.2\times10^{-16}$	2	0.872
Model C: (a_0, a_i, b_0)					
Trap	1	278.9	$<2.2\times10^{-16}$	3	0.979
Month	1	2243.2	$<2.2\times10^{-16}$		
Model D: (a_0, a_i, b_0, b_i)					
Trap	1	278.9	$<2.2\times10^{-16}$		
Month	1	2243.2	$<2.2\times10^{-16}$	4	0.979
Trap×Month	1	0.4	0.516		
Model E: (a_0)					
				1	0

Poisson errors were assumed for the number of captured moths. The results of the Type I analyses of deviance (ANODEV) are also shown. df is the degree of freedom. LR is the likelihood ratio chi-square. k is the number of fixed-effect parameters.

4.3. Multiplicative processes with Poisson errors

In performing an appropriate choice of error structure, we should consider the mechanism yielding the variability in the ecological variable. We previously calculated R_D after using a logarithmic transformation of the form $\log_e(x + 0.5)$. This procedure of calculation will be appropriate in analyzing the data on the population abundance that emerged from a multiplicative process in which each individual reproduces by its reproduction rate at each point of time. Let us assume that the abundance μ is determined by the instantaneous reproduction rate per capita at time t, denoted by r_t, by a form of $\exp(\int r_t dt)$. A multiplicative process becomes

an additive process in a logarithmic scale; the logarithm of population expectation, $\log_e(\mu)$, is determined by the sum of the instantaneous reproduction rate per capita by the form of $\int r_t dt$. Let us further consider that the instantaneous reproduction rate per capita (r_t) fluctuates by following a distribution with a fixed variance around a mean, while its mean differs depending on our experimental treatment or other conditions. Then, the distribution of $\log_e(\mu)$ after a fixed duration of reproduction follows a normal distribution having a fixed variance due to the central limit theorem, irrespective of the exact form of distribution of the fluctuation in the reproduction rate at each point of time, if the duration of reproduction is sufficiently long as compared with the self-correlated duration of reproduction rate. Therefore, the observed number of individuals follows a logarithmic Poisson generalized linear mixed model (logarithmic Poisson GLMM): the expectation of the number of individuals fluctuates by following a normal distribution with a fixed variance in a logarithmic scale, and the observed number of individuals fluctuates by following a Poisson distribution around the expectation. In this case, the global variance given by Eq. 21 is expressed as

$$V(x) = \psi^2[E(\mu)]^2 + E(\mu), \tag{22}$$

where ψ is the fixed CV of the lognormal distribution of μ.

4.4. Approximation to GLMM

Electronic appendix B provides a SAS macro to calculate R_D for the above type of GLMM. However, the calculation of R_D for GLMM generally requires a troublesome calculation including the numerical integration or the Laplace approximation. Hence, a practical approximation to GLMM will be also recommended. The first-term in the right-hand side of Eq. 22 (i.e., the variability of μ) becomes much larger than the second term (i.e., the variability around μ) if the global average of the population expectation $E(\mu)$ is sufficiently large. In this case, therefore, we can approximately use a linear model using the logarithmic transformation $\log_e(x + 0.5)$ as an approximation of the logarithmic Poisson GLMM, by ignoring the variability that corresponds to the second term in the right-hand side of Eq. 22. If the sampling effort is slightly different between experimental plots, we should use the number of individuals per sampling effort. Let s be the sampling effort in a plot and $s_{\min}$ be the minimal quantity. Then, we can use

$$\log_e\left(\frac{x}{s} + \frac{1}{2s_{\min}}\right). \tag{23}$$

The discrete width (w in Eq. 1) is set at the maximal value, $1/s_{\min}$, in this case. This approximation will be available only if the variability in the sampling effort is not so large.

5. Simulation experiments

We have uncertainty about whether R_D is a practical estimate of the proportion of improvement in the predictive ability, because we used Eq. 12 as approximations for false models as well as for true models in deriving R_D. The uncertainty will be larger for small data set due to the uncertainty in the asymptotic property and the uncertainty of the estimate of dispersion parameters. Therefore, we examined the behavior of maximal quantity of R_D under condition of a small amount of data by performing simulations for various combinations of design matrices, numbers of errors, and error distributions. In these simulations, we calculated $E_2(l_1)$ by the direct average of logarithmic probability density of 2nd data set evaluated at the model constructed from 1st data set by maximizing R_D. In calculating the proportion of improvement in the predictive ability, we directly calculated the quantity of Eq. 13. Then, we compared the average of maximal R_D with the quantity of Eq. 13 to judge whether the approximation by Eq. 14 is satisfactory or not. The simulations indicated that the maximal R_D is an appropriate estimate of the proportion of improvement in the predictive ability of the best model unless the sample size is too small. The simulations additionally indicated that the quantity of $E_2(l_1)$ of the best model constructed by the maximization of R_D is larger than that of the models constructed by other criteria, such as AIC, AICc, and BIC, when the maximization of R_D is different from

that of conventional criteria. This result implies the superiority of R_D over other criteria as a model-selection tool. However, the superiority as a model-selection tool seems to be trivial; the primary purpose of the calculation of R_D is to enable the evaluation of the quantity and quality of data, like the classical R^2 index, rather than the model-selection.

5.1. Fixed-effect models

We first used two-way design experiments defined by

$$y_{ijk} = c + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad e_{ijk} \sim N(0, \phi), \tag{24}$$

where y_{ijk} is the observation of the kth replication in the ith level of the first treatment in the jth level of the second treatment, c is the intercept, α_i is the effect of the ith level of the first treatment, β_j is the effect of the jth level of the second treatment, γ_{ij} is the interaction, and e_{ijk} is the error that independently follows a normal distribution of mean zero and variance ϕ. The model that contains all parameters, such as c, α_i, β_j, and γ_{ij}, is the saturated model in this case. It is a true model that includes all knowledge we can use. The true variance is estimated from the true probability that is given by the residuals of the true model. For simplicity, we used experiments with two levels: $i = 1, 2$ and $j = 1, 2$ in Eq. 24. We additionally compared $E_2(l_1)$ of the model constructed by maximizing R_D with that of the models constructed by AICc, AIC, and BIC. The modification, AICc, is defined by

$$\mathrm{AIC_c} = -2l_{1\mathrm{AIC}} + 2k_{\mathrm{AIC}} + \frac{2k_{\mathrm{AIC}}(k_{\mathrm{AIC}}+1)}{n - k_{\mathrm{AIC}} - 1}, \tag{25}$$

where n is the number of observations. For a fixed-effect model having a single normal error, AICc is an unbiased estimate of $-2E_2(l_{1\mathrm{AIC}})$ if f is the true model, that is, if $f(x|\mathbf{0_0}) = g(x)$ (see Sugiura, 1978). BIC is defined by

$$\mathrm{BIC} = -2l_{1\mathrm{AIC}} + k_{\mathrm{AIC}} \log_e(n). \tag{26}$$

We used a restriction on parameters to avoid the singularity: $\alpha_1 = 0$, $\alpha_2 = \alpha$, $\beta_1 = 0$, $\beta_2 = \beta$, $\gamma_{11} = \gamma_{22} = 0$, and $\gamma_{12} = \gamma_{21} = \gamma$. Then, we used two sets of parameters, $(c, \alpha, \beta, \gamma) = (1, 1, 1, 0.5)$ and $(1, 0.5, 0.5, 0.5)$. The number of replications for each combination of treatments was set at 3 and 12. The dispersion parameter was set at $\phi = 0.2$ and 1. Only the models that belong to the hierarchical family were compared: (c), (c, α), (c, β), (c, α, β), and $(c, \alpha, \beta, \gamma)$. We performed 1000 simulation runs for each set of parameters by using R. In each simulation run, we first generated new data set (1st data set) using a normal random number generator in which the true quantities of c, α, β, γ, and ϕ are used. Then, we obtained the unbiased estimate of dispersion parameter $\hat{\phi}$ from the true probability that is given by the saturated model that includes all parameters $(c, \alpha, \beta, \gamma)$. We next selected the set of fixed-effect parameters that yielded the largest R_D. The unbiased estimate of the dispersion parameter obtained from the saturated model was used throughout the procedure of model-selection. Simultaneously, we selected the set of parameters that yielded the smallest AICc, AIC, and BIC. In calculating AICc, AIC, and BIC, the dispersion parameter was estimated for each model by the maximum likelihood method. We next generated new data set (2nd data set) using a normal random number generator in which the true quantities of c, α, β, γ, and ϕ are used. Then, we calculated the logarithmic probability density of 2nd data set evaluated at the model constructed by each criterion from 1st data set. We performed 1000 simulation runs, and calculated the predictive ability, $E_2(l_1)$, by the direct average of logarithmic probability density of 2nd data set evaluated at the model constructed by each criterion from 1st data set. To calculate R_{pred} and RS_{pred}, the predictive ability of the null-model, $E_2(l_{1,\mathrm{null}})$, was calculated in a similar manner. The quantity of $E_2(l_{2,\max})$ and $E_2(l_{2,\mathrm{sat}})$ was also calculated in a similar manner except that $l_{2,\max}$ and $l_{2.\mathrm{sat}}$ were calculated by the logarithmic probability density of 2nd data set evaluated at the model constructed from 2nd data set under the dispersion parameters estimated from 1st data set.

Table 4 indicates that the averages of the maximal R_D were fairly close to the R_{pred} of the best model. However, the difference was larger than 0.1 in Models 5 and 7, which have the following disadvantageous conditions: the number of replications is small, and the dispersion

parameter ϕ is large. The calculated R_{pred} became negative in Model 7 due to the fluctuation in the estimate of the dispersion parameter. The quantity of $E_2(l_1)$ of the model constructed by the maximization of R_D (which in this case is identical to the predictive ability of the model constructed by C_p) was larger than that of the models constructed by AICc, AIC, and BIC. As for the comparison between AICc and AIC, the model constructed by AICc was superior to that constructed by AIC when the number of replications was small and the dispersion parameter ϕ was large (i.e., in Models 5 and 7), as Sugiura (1978) indicated for small samples.

5.2. Mixed models

We next examined the models that have more than one normal random component. Such models are referred to as mixed models if they contain at least one fixed-effect parameter. In the analysis of mixed models, the restricted maximum likelihood method (REML) is usually used as the default method in estimating variances in most of the statistical software, such as SAS (SAS Institute Inc., 2010b), JMP (SAS Institute Inc., 2010a), Stata (StataCorp, 2009), and SPSS (IBM Corp., 2011). We can select the parameters of the variance components by AIC or AICc. However, we cannot use AIC or AICc in selecting the fixed-effect parameters if we use REML, because only the likelihood concerning random-effect parameters is maximized in REML. We must use the maximum likelihood method if we want to select the fixed-effect parameters by AIC or AICc. Thus, there is an inconsistency; the variance structure is selected by REML but the quantity of variance is subsequently estimated by the maximum likelihood method instead of using REML (see, for example, Verbeke and Molenberghs, 1997). In contrast, in the procedure of model-selection using R_D, no such inconsistency exists. We first use REML or the moment method to obtain unbiased estimates of all variances from the true probability that is given by the true model (i.e., the saturated model or the most complex model). Then, we estimate the fixed-effect parameters for candidate models by using the maximum likelihood method by treating the variances estimated by REML (or the moment methods) as known nuisance parameters; the variances estimated from the true probability that is given by the true model are used throughout the process of model-selection.

We examined the behavior of R_D and RS_D in a split-plot design experiment that is one of the simplest cases of mixed models:

$$y_{ijk} = c + \alpha_i + e_{ik} + \beta_j + \gamma_{ij} + e_{ijk}\,, \ (i = 1,2; j = 1,2),$$
$$e_{ik} \sim N(0, \phi_1), \quad e_{ijk} \sim N(0, \phi_2)\,, \tag{27}$$

where y_{ijk} is the observation for the jth secondary factor (subplot factor) of the kth replicate of the ith primary factor (main plot factor), α_i is the effect of the ith primary factor, β_j is the effect of the jth secondary factor, e_{ik} is the error between replications that follows a normal distribution with mean zero and variance ϕ_1, and e_{ijk} is the error within replications (residual error) that follows a normal distribution with mean zero and variance ϕ_2. We again used a restriction on parameters to avoid the singularity: $\alpha_1 = 0$, $\alpha_2 = \alpha$, $\beta_1 = 0$, $\beta_2 = \beta$, $\gamma_{11} = \gamma_{22} = 0$, and $\gamma_{12} = \gamma_{21} = \gamma$. Then, we used two sets of parameters, $(c, \alpha, \beta, \gamma) = (1, 1, 1, 0.5)$, and $(1, 0.5, 0.5, 0.5)$. The number of replications for each combination of treatments was set at 3 and 12. The dispersion parameters were set at $\phi_1 = \phi_2 = 0.2$ and 1. Only the models that belong to the hierarchical family were compared: (c), (c, α), (c, β), (c, α, β), and $(c, \alpha, \beta, \gamma)$. We performed 1000 simulation runs for each set of parameters. We additionally compared $E_2(l_1)$ of the models constructed by the maximization of R_D with that of the models constructed by AICc, AIC, and BIC.

Table 5 indicates that the averages of the maximal R_D were fairly close to the R_{pred} of the best model. The difference between R_D and R_{pred} was larger than 0.1 in Models 11, 13, and 15, which have the following disadvantageous conditions: the number of replications is small while the dispersion parameters, ϕ_1 and ϕ_2, are large. The quantity of $E_2(l_1)$ of the model constructed by the maximization of R_D was larger than that of the models constructed by AICc, AIC, and BIC, except in the case of Model 15. As for the comparison between AIC and AICc, the model constructed by AICc was superior to that constructed by AIC only when the number of replications was small (i.e., in Models 9, 11, 13, and 15).

5.3. Quasi-likelihood models

We next examined the behavior of R_D in the inference using quasi-likelihood. We used an assumption that the variance is proportional to the mean; it corresponds to an overdispersed Poisson distribution having a constant dispersion parameter ϕ (McCullagh and Nelder, 1989). We additionally compared $E_2(l_1)$ of the model constructed by the maximization of R_D with that of the model constructed by QAICc defined by

$$\mathrm{QAIC_c} = -2l_1 + 2k_{\mathrm{AIC}} + \frac{2k_{\mathrm{AIC}}(k_{\mathrm{AIC}}+1)}{n-k_{\mathrm{AIC}}-1}, \tag{28}$$

where l_1 is the quasi-likelihood although we use the same notation for simplicity. We assumed that the observation fluctuates by following a Poisson distribution while the mean of the Poisson distribution independently fluctuates by following a gamma distribution. Then, the observation fluctuates by following a negative binomial distribution. We used a restriction on the parameters of the gamma distribution so that the dispersion parameter is kept constant; the scale parameter of the gamma distribution was kept constant while the shape parameter was changed to yield different means. This type of negative binomial model was called NB1 by Cameron and Trivedi (1998). Then, in selecting the best model, we can approximately use the quasi-likelihood of the overdispersed Poisson distribution in place of the logarithmic likelihood of the negative binomial distribution. We again used a two-way factorial design for the systematic components: i = 1,2 and j = 1,2 in Eq. 24. We used a restriction on parameters to avoid the singularity: $\alpha_1 = 0$, $\alpha_2 = \alpha$, $\beta_1 = 0$, $\beta_2 = \beta$, $\gamma_{11} = \gamma_{22} = 0$, and $\gamma_{12} = \gamma_{21} = \gamma$. The dispersion parameter should be larger than 1 for an overdispersed Poisson distribution, and hence we used ϕ = 1.2 and 2. We used two sets of fixed-effect parameters, (c, α, β, γ) = (8, 6, 4, 2) and (4, 3, 2, 1), which yield moderate amounts of R_D. The number of replications for each combination of treatments was set at 3 and 12. We again compared only the models that belong to the hierarchical family: (c), (c, α), (c, β), (c, α, β), and (c, α, β, γ). We performed 1000 simulation runs for each set of parameters.

We did not use the likelihood of actual distribution, a negative binomial distribution, in calculating R_D; we only used the knowledge that the variance is proportional to the mean. Nonetheless, the averages of the maximal R_D were fairly close to the R_{pred} of the best model unless the number of replications was too small (Table 6). The difference between R_D and R_{pred} was larger than 0.1 in Models 19, 21, and 23, which have the following disadvantageous conditions: the number of replications is small, and the dispersion parameter ϕ is large while the effects of factors are small. The quantity of $E_2(l_1)$ of the model constructed by the maximization of R_D (which in this case is identical to the predictive ability of the model constructed by QAIC) was larger than that of the model constructed by QAICc.

5.4. Poisson or binomial models

Finally, we examined the behavior of R_D for Poisson errors or binomial errors without overdispersion. The model constructed by the maximization of R_D is identical to the model constructed by AIC in this case. We again used a two-way factorial design for the systematic components: i = 1,2 and j = 1,2 in Eq. 24. We used a restriction on parameters to avoid the singularity: $\alpha_1 = 0$, $\alpha_2 = \alpha$, $\beta_1 = 0$, $\beta_2 = \beta$, $\gamma_{11} = \gamma_{22} = 0$, and $\gamma_{12} = \gamma_{21} = \gamma$. We again compared only the models that belong to the hierarchical family: (c), (c, α), (c, β), (c, α, β), and (c, α, β, γ). For Poisson errors, we used two sets of fixed-effect parameters, (c, α, β, γ) = (8, 6, 4, 2) and (4, 3, 2, 1). The number of replications for each combination of treatments was set at 3 and 12. For binomial errors, we used two sets of fixed-effect parameters, (c, α, β, γ) = (0.2, 0.2, 0.2, 0.2) and (0.1, 0.1, 0.1, 0.1). The number of replications for each combination of treatments was set at 3 and 12. The number of cases (n) for each replication was fixed at 20. We performed 1000 simulation runs for each set of parameters.

For Poisson errors, the averages of the maximal R_D were close to the R_{pred} of the best model unless the sample size was too small (Table 7). The difference between R_D and R_{pred} in Table 7 is smaller than that of the corresponding overdispersed Poisson errors given in Table 6, probably because the dispersion parameter is known in Table 7. For binomial errors, the

averages of the maximal R_D were also close to the R_{pred} of the best model unless the sample size was too small (Table 7). The quantity of $E_2(l_1)$ of the model constructed by the maximization of R_D (which in this case is identical to the predictive ability of the model constructed by AIC) was equal to or larger than that of the models constructed by AICc and BIC in these simulations, although Anderson (2007) highly recommended the use of AICc.

6. Conclusions

We proposed a criterion R_D as an estimate of the proportion of improvement in the predictive ability of ecological models. R_D indicates the absolute goodness of models while other existing criteria, such as AIC, BIC, C_p, indicate the relative goodness among candidate models for a given amount of data. For the data on the abundance of moths, the quantity of maximal R_D was about 0.9, and hence the predictive ability of the model was judged to be sufficiently high (Table 2), although we need further discussion on the sufficient quantity of R_D. If the largest quantity of R_D is still small, such as 0.4, we should increase the quantity and quality of data by collecting more data and by preparing new explanatory variables.

We sometimes use models to understand the principal mechanism that yields the observation, rather than for purposes of prediction. The criterion R_D will also be useful for this purpose; R_D enables us to identify an appropriate model that summarizes the mechanism that yields the observation. For example, a researcher may consider that a model achieving an 80% improvement in predictive ability is an appropriate model for summarizing the mechanism yielding the observation. Such a researcher should adopt a model that has an R_D near 0.8, instead of adopting the model that has the largest R_D. In the example about the abundance of moths, the R_D of Model B (0.846) was closest to 0.8 (Table 2). In this case, therefore, such a researcher should consider that the number of moths is principally determined by a single factor, the month. However, the appropriate quantity of R_D for the summarization, as well as the appropriate quantity of R_D for the prediction, will change depending on the purpose of research. This should be discussed in future studies.

Acknowledgments

The author would like to thank the anonymous referee for the comments that greatly helped to improve the manuscript.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In: *Second International Symposium on Information Theory*, Petrov, B. N., Csáki, F. eds., Budapest, Hungary: Akadémiai Kiadó, pp. 267–281.

Amari, S. (2007). Akaike information criterion, AIC: concepts and new development, In: *Akaike Information Criterion AIC: Modeling, Prediction and Knowledge Discovery*, Murota, K., Tsuchiya, T. eds., Tokyo: Kyoritu, pp. 52–78 (in Japanese).

Anderson, D. R. (2007). *Model Based Inference in the Life Sciences: A Primer on Evidence*, New York: Springer.

Ben-Akiva, M. E., Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge: MIT Press.

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

Cameron, A. C., Trivedi, P. K. (1998). *Regression Analysis of Count Data*, Cambridge, UK: Cambridge University Press.

Claeskens, G., Hjort, N. L. (2008). *Model Selection and Model Averaging*, Cambridge, UK: Cambridge University Press.

Cox, D. R., Wermuth, N. (1992). A comment on the coefficient of determination for binary responses, *Am. Stat.*, 46:1–4.

Draper, N. R., Smith, H. (1998). *Applied Regression Analysis, 3rd edition*, New York: Wiley.

Fujikoshi, Y., Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression *Biometrika*, 84:707–716.

Gneiting, T., Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, 102:359–378.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments, *Ecol. Monogr.*, 54:187–211.

IBM Corp. (2011). *IBM SPSS Statistics 20 Command Syntax Reference*, Chicago: IBM Corp.

Kitagawa, G. (2007). Information criterion and statistical modeling, In: *Akaike Information Criterion AIC: Modeling, Prediction and Knowledge Discovery*, Murota, K., Tsuchiya, T. eds., Tokyo: Kyoritu, pp. 79–109 (in Japanese).

Konishi, S., Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, New York: Springer.
Kullback, S., Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, 22:79–86.
Laplace, P. S. (1825). *A Philosophical Essay on Probabilities (Translated from the Fifth French Edition of 1825 by Andrew I. Dale)*, New York: Springer.
Liu, H., Zheng, Y., Shen, J. (2008). Goodness-of-fit measures of R^2 for repeated measures mixed effect models, *J. Appl. Statist.*, 35:1081–1092.
Long, J. S., Fre, J. (2000). Scalar measures of fit for regression models, *Stata Technical Bulletin*, 56:34–40.
Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, 15:661–675.
Mallows, C. L. (1995). More comments on C_p, *Technometrics*, 37:362–372.
McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models, 2nd edition*, London: Chapman and Hall.
Miyagawa, H. (1979). *Information Theory*, Tokyo: Conona Publishing (in Japanese).
Murata, N. (2005). *Basics of Information Theory*, Tokyo: Saiensu-sha (in Japanese).
Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination, *Biometrika*, 78:691–692.
Orelien, J. G., Edwards, L. J. (2008). Fixed-effect variable selection in linear mixed models using R^2 statistics, *Comput. Stat. Data Anal.*, 52:1896-1907.
R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.
Recchia, A. (2010). R-squared measures for two-level hierarchical linear models using SAS, *Journal of Statistical Software*, 32:1–9.
SAS Institute Inc. (2010a). *JMP® 9 Modeling and Multivariate Methods*, Cary: SAS Institute Inc.
SAS Institute Inc. (2010b). *SAS/STAT® 9.22 User's Guide*, Cary, NC: SAS Institute Inc.
Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Stat.*, 6:461–464.
Shannon, C. (1948). A mathematical theory of communication, *Bell System Technical Journal*, 27:379–423 and 623–656.
StataCorp (2009). *Stata Longitudinal Data/Panel Data Reference Manual Release 11*, College Station: Stata Press.
Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Statist. - Theory and Methods*, A7:13–26.
Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting, *Suri-Kagaku (Mathematical Sciences)*, 153:12–18 (in Japanese).
Tribus, M., McIrvine, E. C. (1971). Energy and information, *Sci. Am.*, 225:179–188.
Verbeke, G., Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-oriented Approach*, New York: Springer.
Wakamura, S., Kozai, S., Kegasawa, K., Inoue, H. (1992). Population dynamics of adult *Spodoptera litura* (Fabricius) (Lepidoptera: Noctuidae): Estimation of male density by using release-recapture data, *Appl. Entomol. Zool.*, 27:1–8.
Wheeler, M. W., Bailer, A. J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation *Environ. Ecol. Statist.*, 16:37–51.
Xu, R. (2003). Measuring explained variation in linear mixed effects models, *Stat. Med.*, 22:3527–3541.
Yamamura, K. (1993). On the choice of multiple comparison procedures, *Shokubutsu Boeki (Plant Protection)*, 47:370–375 (in Japanese).
Yamamura, K. (1999). Transformation using (x + 0.5) to stabilize the variance of populations, *Res. Popul. Ecol.*, 41:229–234.
Yamamura, K. (2002). Dispersal distance of heterogeneous populations, *Popul. Ecol.*, 44:93–101.
Yamamura, K., Yokozawa, M., Nishimori, M., Ueda, Y., Yokosuka, T. (2006). How to analyze long-term insect population dynamics under climate change: 50-year data of three insect pests in paddy fields, *Popul. Ecol.*, 48:31–48.
Yamamura, K., Matsuda, H., Yokomizo, H., Kaji, K., Uno, H., Tamada, K., Kurumada, T., Saitoh, T., Hirakawa, H. (2008). Harvest-based Bayesian estimation of sika deer populations using state-space models, *Popul. Ecol.*, 50:131–144.
Yamamura, K. (2009). Generalized linear models and model selection, *Shokubutsu Boeki (Plant Protection)*, 63:324–329 (in Japanese).
Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data, *Stat. Med.*, 19:1265–1275.

Table 4
Results of simulations for the two-way factorial design experiment having a single normal error

Model number	1	2	3	4	5	6	7	8
Fixed effect parameters (c, α, β, γ)	(1,1,1,0.5)	(1,1,1,0.5)	(1,0.5,0.5,0.5)	(1,0.5,0.5,0.5)	(1,1,1,0.5)	(1,1,1,0.5)	(1,0.5,0.5,0.5)	(1,0.5,0.5,0.5)
Dispersion parameter ϕ	0.2	0.2	0.2	0.2	1	1	1	1
Number of replications	3	12	3	12	3	12	3	12
Average of R_D	0.662	0.717	0.363	0.446	0.273	0.320	0.130	0.119
R_{pred}	0.657	0.718	0.316	0.449	0.164	0.309	−0.051	0.080
Average of RS_D	0.790	0.942	0.501	0.831	0.404	0.743	0.205	0.418
RS_{pred}	0.784	0.943	0.460	0.840	0.267	0.732	−0.103	0.326
Average of $\hat{\phi}$	0.200	0.200	0.200	0.200	1.001	1.001	1.001	1.001
Predictive ability, $E_2(l_1)$								
R_D	−11.606	−32.302	−11.981	−32.304	−21.499	−71.249	−21.115	−71.725
AICc	−14.853	−32.699	−14.993	−32.701	−23.352	−71.751	−21.810	−72.287
AIC	−14.381	−32.693	−14.585	−32.695	−23.915	−71.607	−23.373	−72.035
BIC	−14.454	−32.782	−14.720	−32.795	−23.931	−72.344	−23.209	−73.070

The average of maximal R_D is compared to R_{pred} of the best model. The model constructed by the maximization of R_D is identical to the model constructed by C_p in this case. The row of predictive ability shows the direct average of logarithmic probability density of 2nd data set evaluated at the model constructed by each criterion from 1st data set.

Table 5
Results of simulations for the split-plot design experiment having normal errors

Model number	9	10	11	12	13	14	15	16
Fixed effect parameters (c, α, β, γ)	(1,1,1,0.5)	(1,1,1,0.5)	(1,0.5,0.5,0.5)	(1,0.5,0.5,0.5)	(1,1,1,0.5)	(1,1,1,0.5)	(1,0.5,0.5,0.5)	(1,0.5,0.5,0.5)
Variance between replication ϕ_1	0.2	0.2	0.2	0.2	1	1	1	1
Variance within replication ϕ_2	0.2	0.2	0.2	0.2	1	1	1	1
Number of replications	3	12	3	12	3	12	3	12
Average of R_D	0.621	0.644	0.364	0.389	0.275	0.247	0.160	0.093
R_{pred}	0.560	0.637	0.232	0.382	0.066	0.213	−0.106	0.040
Average of RS_D	0.754	0.919	0.492	0.790	0.399	0.658	0.241	0.345
RS_{pred}	0.700	0.920	0.361	0.799	0.119	0.614	−0.231	0.193
Average of $\hat{\phi}_1$	0.215	0.207	0.215	0.200	1.074	1.033	1.074	1.033
Average of $\hat{\phi}_2$	0.189	0.198	0.189	0.200	0.944	0.988	0.944	0.988
Predictive ability, $E_2(l_1)$								
R_D	−17.967	−46.125	−18.513	−46.315	−27.773	−85.455	−27.380	−85.659
AICc	−21.635	−46.645	−19.656	−46.861	−28.139	−86.078	−26.823	−86.260
AIC	−22.394	−46.622	−22.853	−46.827	−31.767	−85.870	−31.107	−86.052
BIC	−22.462	−46.730	−22.885	−47.090	−31.446	−86.735	−30.630	−86.565

The average of maximal R_D is compared to R_{pred} of the best model. The row of predictive ability shows the direct average of logarithmic probability density of 2nd data set evaluated at the model constructed by each criterion from 1st data set.

Table 6

Results of simulations for the two-way factorial design experiment having an overdispersed Poisson error

Model number	17	18	19	20	21	22	23	24
Fixed effect parameters (c, α, β, γ)	(8,6,4,2)	(8,6,4,2)	(4,3,2,1)	(4,3,2,1)	(8,6,4,2)	(8,6,4,2)	(4,3,2,1)	(4,3,2,1)
Dispersion parameter ϕ	1.2	1.2	1.2	1.2	2	2	2	2
Number of replications	3	12	3	12	3	12	3	12
Average of R_D	0.384	0.440	0.237	0.261	0.276	0.311	0.178	0.181
R_{pred}	0.287	0.429	0.076	0.246	0.155	0.296	−0.020	0.141
Average of RS_D	0.534	0.831	0.363	0.689	0.411	0.734	0.281	0.570
RS_{pred}	0.434	0.823	0.137	0.668	0.260	0.712	−0.041	0.488
Average of $\hat{\phi}$	1.188	1.187	1.150	1.207	1.937	1.982	1.884	1.989
Predictive ability, $E_2(l_1)$								
R_D	−40.433	−121.467	−36.823	−104.614	−30.843	−83.612	−26.797	−73.461
QAICc	−41.480	−121.583	−37.225	−104.819	−31.557	−83.757	−26.872	−73.662

The average of maximal R_D is compared to R_{pred} of the best model. The model constructed by the maximization of R_D is identical to the model constructed by QAIC in this case. The row of predictive ability shows the direct average of quasilikelihood of 2nd data set evaluated at the model constructed by each criterion from 1st data set.

Table 7

Results of simulations for the two-way factorial design experiment having a Poisson error or a binomial error (n = 20) without overdispersion

Model number	25	26	27	28	29	30	31	32
Error distribution	Poisson	Poisson	Poisson	Poisson	binomial	binomial	binomial	binomial
Fixed effect parameters (c, α, β, γ)	(8,6,4,2)	(8,6,4,2)	(4,3,2,1)	(4,3,2,1)	(0.2,0.2,0.2,0.2)	(0.2,0.2,0.2,0.2)	(0.1,0.1,0.1,0.1)	(0.1,0.1,0.1,0.1)
Number of replications	3	12	3	12	3	12	3	12
Average of R_D	0.382	0.475	0.237	0.299	0.610	0.683	0.322	0.421
R_{pred}	0.340	0.479	0.122	0.289	0.610	0.689	0.257	0.428
Average of RS_D	0.544	0.852	0.370	0.729	0.750	0.935	0.470	0.825
RS_{pred}	0.496	0.851	0.210	0.706	0.760	0.935	0.398	0.832
Predictive ability, $E_2(l_1)$								
R_D	−35.058	−132.256	−30.787	−115.436	−27.938	−105.028	−27.005	−98.661
AICc	−35.884	−132.318	−31.203	−115.569	−28.559	−105.028	−28.246	−98.665
BIC	−35.275	−132.621	−30.982	−116.276	−27.984	−105.028	−27.267	−98.700

The average of maximal R_D is compared to R_{pred} of the best model. The model constructed by the maximization of R_D is identical to the model constructed by AIC in this case. The row of predictive ability shows the direct average of logarithmic probability density of 2nd data set evaluated at the model constructed by each criterion from 1st data set.