Austral Ecology (2000) **25**, 327–331

Swapping space for time and unfair tests of ecological models

ANDREW J. TYRE,[1]* BRIGITTE TENHUMBERG,[1] MICHAEL A. MCCARTHY[2] AND HUGH P. POSSINGHAM[1]

[1]*Department of Applied and Molecular Ecology, University of Adelaide, Waite Campus, Private Bag 1, Glen Osmond, South Australia 5064, Australia (Email: dtyre@roseworthy.adelaide.edu.au) and* [2]*National Center for Ecological Analysis and Synthesis, UC Santa Barbara, Santa Barbara, California, USA*

Abstract Testing ecological models for management is an increasingly important part of the maturation of ecology as an applied science. Consequently, we need to work at applying fair tests of models with adequate data. We demonstrate that a recent test of a discrete time, stochastic model was biased towards falsifying the predictions. If the model was a perfect description of reality, the test falsified the predictions 84% of the time. We introduce an alternative testing procedure for stochastic models, and show that it falsifies the predictions only 5% of the time when the model is a perfect description of reality. The example is used as a point of departure to discuss some of the philosophical aspects of model testing.

Key words: adaptive management, power test, *Sorghum brachypodum*, standard deviate, stochastic population model.

INTRODUCTION

Ecological processes commonly operate on timescales of decades or longer. Consequently, testing models of these processes is problematic. When land management decisions need to be made immediately, then the issue of model testing becomes pressing. This situation calls for an active adaptive management (AAM) approach where the management process facilitates the test of the model (Parma *et al.* 1998).

Lonsdale *et al.* (1998) faced the problem of needing a management model quickly in their study of the effects of wet season burns on savannah vegetation, particularly the grass *Sorghum brachypodum*. They were interested in determining how best to use wet season burns to reduce the risk to infrastructure and revegetation sites from more destructive dry season fires, a clear management objective with a specific control option. Unfortunately, there was little information available on the ecological effects of wet season burns on *Sorghum*. An AAM strategy (Parma *et al.* 1998) recognises this problem and seeks to use management actions as experiments to increase ecological knowledge about the system being managed. Ecological models, experimentation and monitoring are crucial components of the AAM process.

Lonsdale *et al.* (1998) have all three of these components in their study and are to be congratulated on taking an AAM approach to their land management problem. A crucial part of the approach is testing the predictions of the model to gain confidence in the predictions and hence the management decision. While Lonsdale *et al.* (1998) did test the prediction of their model, we would like to comment on their testing procedure and suggest a better alternative. We find that their test of the model predictions is strongly biased. We suggest how future tests might be carried out and comment on the importance of model testing in the AAM framework.

*Corresponding author.
Accepted for publication October 1999.

METHOD AND RESULTS

The Watkinson model for *Sorghum* population dynamics

The model of annual plant population growth Lonsdale *et al.* (1998) used was developed by Watkinson *et al.* (1989) for another species of *Sorghum*:

$$N_{t+1} = \frac{\lambda N_t}{(1 + aN_t)^b + m\lambda N_t}, \lambda = s \bullet d \qquad (1)$$

where N_t is population density (units of m^{-2}) t years since a fire, a is the reciprocal of the density at which competition begins to take effect (0.0051 m^2), b describes the efficiency of resource uptake (0.73), m is the reciprocal of the asymptotic density following self thinning (0.0085 m^2), s is the per capita seed output at low densities (4–18 seeds) and d is the fraction of

individuals surviving density independent mortality from all sources (0.08–0.21). Both *s* and *d* have a range of values because they are assumed to fluctuate randomly from year to year. Lonsdale *et al.* (1998) estimated the range for *d* by changing the values until the average long-term density predicted by the model matched that in unburned sites. Lonsdale *et al.* (1998) did not specify the probability distributions used, so we have assumed they followed Watkinson *et al.* (1989) and used uniform or rectangular distributions between the limits described above. Neither paper described the correlation structure between the random variables either in time or between sites. Therefore, we assumed that there was no correlation in environmental variation of *s* and *d* in either space or time.

We implemented the model in Microsoft Excel (Version 7), using the macro language to generate replicate runs (spreadsheet available from the first author on request). We used the parameters given above to replicate the results of Lonsdale *et al.* (1998) (Fig. 2 in their paper) using 100 replicate runs starting from either the minimum (0.7 m^{-2}) or maximum (6.8 m^{-2}) post-burn densities reported (Fig. 1). The coincidence of our results with theirs satisfied us that the models were the same, although the 95% confidence limits on our average densities were much smaller because we used more than eight times as many runs to calculate them (100 *vs* 12).

Lonsdale *et al.* (1998) concentrated on the average trajectory, calculated by averaging predicted density at each time over all replicate runs. Knowing the average density of *Sorghum* over 100 replicate simulations can be misleading. This is clear from 95% confidence intervals of the overall distribution of population sizes (i.e., the 2.5 percentile and 97.5 percentile) for the low density trajectory (Fig. 1). Despite the fact that the average density rises to 15 m^{-2} over the first 10 years post-burn, a population density of < 1 m^{-2} is not an unreasonable occurrence (i.e. it is inside the 95% confidence interval of the total distribution) starting from a population density of 0.7 m^{-2}. The oldest post-burn site in their sample (4 years) had a density this low and this was almost certainly part of the reason they falsified their prediction. This 4-year density is much lower than the average trajectory, but that does not mean that it is an impossible occurrence if the model is exactly true.

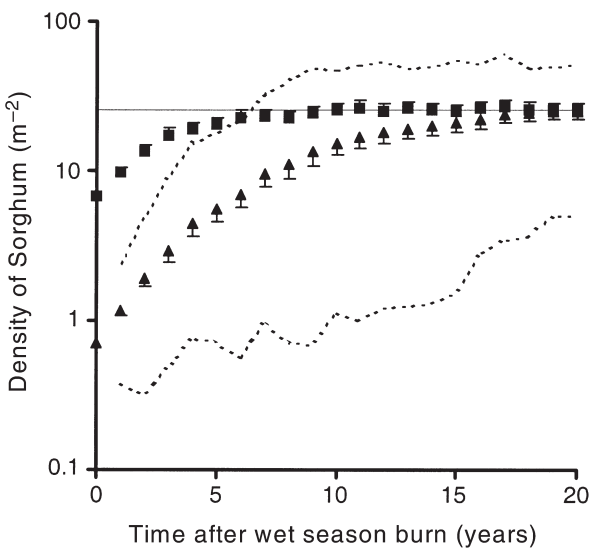

Fig. 1. Average return trajectories from the model starting from the highest initial density (■), and the lowest initial density (▲). Error bars are 95% confidence limits on the average density. The fine horizontal line is the estimated pre-burn stable density used to estimate the range for *d*. The dashed lines represent the 95% confidence interval for the entire distribution of the low initial density trajectory (triangles).

Model spurned by unfair test

Lonsdale *et al.* (1998) tested the prediction of the model that the average population density increases with time up to the long-term average density in unburnt patches. In other words, the (entirely reasonable) prediction is that there is a positive correlation between time post-burn and *Sorghum* population density. Their empirical data set consisted of five sites sampled at various post-burn times. They calculated the correlation coefficient between population density and time post-burn, and falsified the prediction because their empirical observations had a non-significant negative correlation coefficient. However, even if the model was a perfect description of reality, the correlation test they applied would almost always falsify the prediction. We showed this by using the model itself to generate sets of data, and then calculated the correlation test on this computer-generated data.

Using the model to generate data to test a statistic is a recommended standard technique for testing new statistics (Hilborn & Mangel 1997). Assuming that the model is a good representation of reality in order to generate test data for the statistic does not lead to concluding that the model is 'true'. The approach leads only to a test of whether or not a particular statistic is biased, given the model is a good representation.

Each run of the model generated a sequence of population densities starting from a specific initial density. We generated 1000 sets of five runs, one run for each study site in the study of Lonsdale *et al.* (1998) study. Each run within a set had its own sequence of random values for *r* and *d*, because it was assumed that there was zero correlation in environmental stochasticity between sites. The starting densities for the five runs in each set were fixed at 0.7, 2.2, 3.7, 5.2 and 6.8 m^{-2}. Only the lowest and highest densities were provided by Lonsdale *et al.* (1998) in their paper. We chose the three intermediate densities to be equally spaced between the

minimum and maximum. We sampled data generated by the model for three sites at 1 year post-burn, one at 2 years, and the last at 4 years post-burn. The actual post-burn sample times were 0.25, 1.25 and 3.25 years post-burn (Lonsdale *et al.* 1998), but the model can only generate predictions for integer times because it is a discrete time model. They did not describe the population densities immediately post-burn, so we randomly assigned starting densities to sampling times for each set. For each of these five points, we then calculated the Pearson correlation coefficient between density and the number of years post-burn. This gave us 1000 replicate tests of the model when the model predictions were perfect, because the data were generated directly by the model. The power of the test was determined by the number of times, as proportion of total, that a significant positive correlation was obtained. The prediction was rejected if the correlation was not significant or was significantly negative.

The correlation test rejected the prediction 84% of the time, and nearly one-third of all correlations were negative (Fig. 2). Nearly 9% of data sets had correlations more negative than that observed by Lonsdale *et al.* (1998) for their empirical data ($r = -0.58$). Only positive correlations above 0.88 were significantly greater than 0. More samples and/or greater spread of post-burn times would probably improve the power of the correlation. The primary reason why the test rejected the predictions of a good model was that there was inadequate power: it relied on getting a positive correlation between time post-burn and density, which paradoxically means rejecting a null hypothesis of no correlation. With only five data points, rejecting the null hypothesis of no correlation was unlikely.

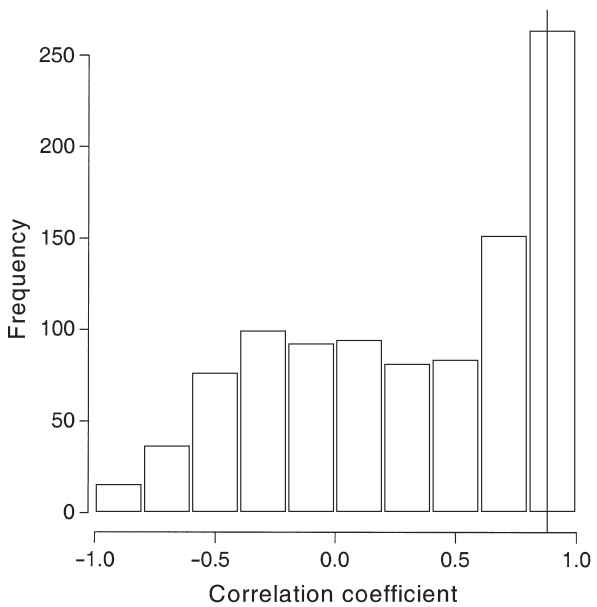

Fig. 2. Distribution of correlation coefficients for 1000 replicate sets of data generated by the model. Correlations to the right of the fine vertical line are significantly different from zero.

The secondary reason why this test failed is more intriguing: the 'space for time' swap. Placing sites burned at different times along a single time-axis did not reproduce the average trajectory because each site had been following an independent path and there was no temporal auto-correlation between the sites (as there would be for a single site followed through time). Lonsdale *et al.* did in fact suggest this as a reason for why the prediction failed, but considered it a less-likely explanation than changes in the parameter values of the model post-burn.

The standard deviate test: equity for models

The second problem with the approach used by Lonsdale *et al.* (1998) is that it assumes that the properties of the mean trajectory are what should be tested. A stochastic model predicts both a mean and a variance. Both properties are important and should be tested. Furthermore, the 'space for time swap' could be avoided by a procedure that compares each point with the distribution of population sizes that could be observed at that time, given a particular starting point. This can be done by standardizing the observed density by the predicted average and predicted standard deviation:

$$x_{transformed} = \frac{x_{observed} - \bar{x}_{predicted}}{s_{predicted}} \qquad (2)$$

where x (either observed, predicted or transformed) is population density and $s_{predicted}$ the standard deviation predicted by the model. These standard deviates will be normally distributed with a mean of 0 and a variance of 1, if the model is correct (Sokal & Rohlf 1981). Thus, both mean and variance predicted by the model can be tested. This approach has been suggested for testing forestry models (Reynolds *et al.* 1981) and used to test 'null models' in community ecology (Gilpin & Diamond 1982). To our knowledge it has not been widely employed in population ecology.

We generated predicted averages and standard deviations for 1–4 years post-burn from each of the five starting densities described above, with 200 replicates at each density. As with the correlation test, we then generated 1000 sets of 'real' data sampled at the same five points in time described by Lonsdale *et al.* (1998) and with sample times and starting densities randomly ordered. We used the predictions to transform each set of 'real' data (Eqn 2) and then tested the hypothesis that the average of the transformed values was zero and the variance was 1. We used a standard *t*-test for the hypothesis that the mean was zero:

$$t_s = \frac{x_{transformed} - \mu}{s_{\bar{x}}} \quad (3)$$

where t_s is the sample statistic, distributed as a t-variate with n-1 degrees of freedom ($n = 5$ here), s_x is the standard deviation of $x_{transformed}$, and μ is the theoretical expected value of zero (Sokal & Rohlf 1981). The hypothesis that the variance of the standard deviates was 1 can be tested with:

$$X^2 = (n-1)s^2/\sigma^2 \quad (4)$$

where the sample statistic X^2 is distributed as a χ^2 variate with $n-1$ degrees of freedom, s^2 is the sample variance of the transformed observations, and σ^2 is the theoretical expected value of 1 (Sokal & Rohlf 1981). If these were fair tests then the model would be rejected 5% of the time (i.e. have a Type I error rate of 5%) when the data have been generated by the model itself.

For this model x was significantly different from zero in 59 out of 1000 tests, not significantly different from the expected 5% (χ^2 test, $p = 0.23$). The variance of the transformed observations was significantly different from one in 66 out of 1000 tests, which was significantly different from the expected 5% (χ^2 test, $p = 0.04$). The distribution of the variances did not quite match the theoretical expectation of the χ^2 with four degrees of freedom (Fig. 3), suggesting that our assumption that the $x_{transformed}$ are normally distributed was not quite correct with this model. Although this is not perfect, the standard deviate test still performed much better than the correlation test used by Lonsdale *et al.* (1998). It also avoided the 'space for time' swap problem because it treated each observation independently, obviating any need to assume that variation in space reflects variation in time.

DISCUSSION

While it is very important to test ecological models before using them for management, it is equally important to use tests that are not biased in one direction or another. An unfair test gives models (and modellers) a bad name. In this case, the model prediction made by Lonsdale *et al.* (1998) may have been unfairly rejected. The model may still make poor predictions, but the test Lonsdale *et al.* (1998) used is heavily biased towards rejection. By contrast, the standard deviate test not only is much fairer, but could also help to identify the situations under which the model is most incorrect. If the model makes poor predictions, Lonsdale *et al.* 's (1998) discussion about the reasons why is still perfectly valid.

One disadvantage of the standard deviate test is that it requires some assumptions about the starting densities of the populations in order to make reasonable predictions. Lonsdale *et al.* (1998) did not have (or did not report) this information. Therefore, we assumed evenly spaced densities over a reasonable range. Alternatively, we could assume that all sites started at the average density in the three sites sampled at 1 year post-burn. Even better, if those five sites were to be sampled again in 1999, the comparison of the observed densities with the model using the standard deviate test would be quite powerful because accurate estimates of the starting densities were obtained in 1994.

Returning to the primary reason why Lonsdale *et al.* (1998) unfairly rejected the model predictions, there

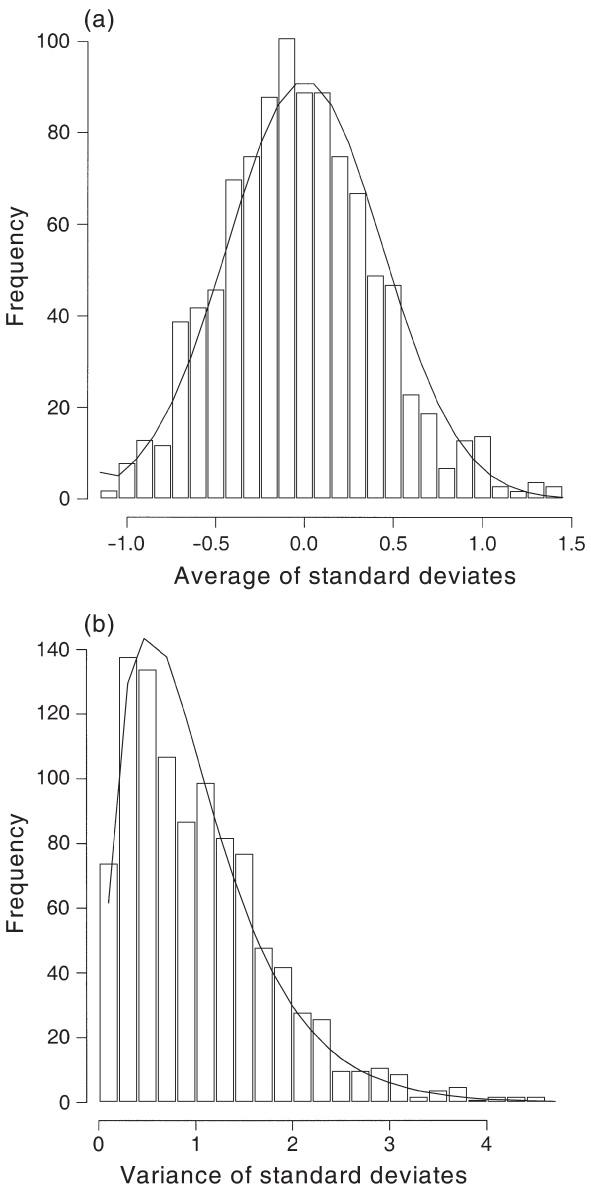

Fig. 3. Distribution of (a) $\bar{x}_{transformed}$ and (b) the variance of $\bar{x}_{transformed}$ from comparing 1000 'real' data sets with the predictions from the model. The normal distribution with $\mu = 0$, ▲ = 0.43 is marked on (a). The theoretical standard deviation of $\overline{SD}$ is 1/sqrt(5) = 0.45. The χ^2 distribution with 4 degrees of freedom is marked on (b).

is a clear philosophical difference between the two methods of testing models explored in this comment. The correlation test used by Lonsdale *et al.* (1998) associates model correctness with a significant result of a statistical test. Therefore, the power of the test ($1-p$(Type II error)) dictates fairness. The standard deviate test associates model correctness with a non-significant result of the statistical test. Therefore, it is the Type I error rate that makes the test fair or unfair. For further information, Rykiel (1996) provides a good entry point into the more philosophical background of model testing and validation.

A final philosophical issue is the approach of testing just one model. For applied science (rather than pure) there is no use subjecting just one management model to testing: it is much better to have alternatives (Hilborn & Mangel 1997). In the example explored here, if we had thrown out the model, then there is no model to manage the system with: what actions do management take then? The alternative is to propose two or more models, even if one of them is unrealistic, such as a model that proposes that the abundance of *Sorghum* is constant or increases linearly. Once the alternatives are proposed we can use a likelihood approach to assign degrees of belief in the alternatives. We can either proceed with the most likely model, or make decisions weighted by the likelihood of alternatives. Either way the manager has something to act on while we gather more information (Possingham 1998).

There is much more that could be done using the general approach outlined here. For example, it may be possible to detect departures from model assumptions by plotting standard deviates as a function of observed density or time post-burn. More importantly, it is possible to determine the ability of the standard deviate test to detect any particular departure from the model assumptions, in other words, to calculate its power. This is beyond the scope of our intent here, but it is both feasible and necessary if firm conclusions about the quality of a model are to be drawn.

We would like to emphasize that the approach of Lonsdale *et al.* (1998), when coupled with good tests of model predictions, is exactly the way we feel Australian applied ecology should be moving.

ACKNOWLEDGEMENTS

AJT is supported by a Strategic Partnerships in Research and Training (SPIRT) grant to HPP. BT is supported by a Large Australian Research Council grant to Michael Keller and HPP. MAM is supported by a SPIRT grant to HPP and David B. Lindenmayer.

REFERENCES

Gilpin M. E. & Diamond J. M. (1982) Factors contributing to nonrandomness in species co-occurrences on islands. *Oecologia* **52,** 75–84.

Hilborn R. & Mangel M. (1997) *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, NJ.

Lonsdale W. M., Braithwaite R. W., Lane A. M. & Farmer J. (1998) Modelling the recovery of an annual savanna grass following a fire-induced crash. *Austr. Ecol.* **23,** 509–13.

Parma A., Amarasekare P., Mangel M., Moore J., Murdoch W. W., Noonburg E., Pascual M. A., Possingham H. P., Shea K., Wilcox C. & Yu D. (1998) What can adaptive management do for our fish, forests, food, and biodiversity? *Integrative Biol.* **1,** 16–26.

Possingham H. P. (1998) Active Adaptive Management: a solution to the management/research and experiment/model nexus. Current Issues in Australian Limnology: controlling and managing algal blooms, river regulation and rehabilitation, and quantitative ecology, pp. 61–5. Australian Society for Limnology, Abbotsford, Victoria.

Reynolds M. R., Burkhart H. E. & Daniels R. F. (1981) Procedures for statistical validation of stochastic simulation models. *For. Sci.* **27,** 349–64.

Rykiel E. J. (1996) Testing ecological models: the meaning of validation. *Ecol. Modelling* **90,** 229–44.

Sokal R. R. & Rohlf F. J. (1981) *Biometry*. W. H. Freeman, New York.

Watkinson A. R., Lonsdale W. M. & Andrew M. H. (1989) Modelling the population dynamics of an annual plant: *Sorghum intrans* in the wet-dry tropics. *J. Ecol.* **77,** 162–81.