Ecological Modelling, 44 (1989) 299–306
Elsevier Science Publishers B.V., Amsterdam – Printed in The Netherlands

MEAN SQUARED ERROR OF PREDICTION AS A CRITERION FOR EVALUATING AND COMPARING SYSTEM MODELS

D. WALLACH and B. GOFFINET

Laboratoire de Biométrie, Institut National de la Recherche Agronomique (INRA), BP 27, 31326 Castanet Tolosan Cedex (France)

(Accepted 7 June 1988)

ABSTRACT

Wallach, D. and Goffinet, B., 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol. Modelling*, 44: 299–306.

Model evaluation is an essential aspect of the process of development of system models. When the main purpose of the model is prediction, a reasonable criterion of model quality is the mean squared error of prediction. This criterion is defined here, and it is shown how it can be estimated from available data in a number of situations, including the situation where the parameters of the model are adjusted to the data. An example of the use of this criterion for choosing between alternative models is presented.

INTRODUCTION

Systems modelling of ecological or agronomic systems has become a widespread and valuable tool. The evaluation of model quality is obviously an essential aspect of the modelling activity, since it is necessary in order to know how much confidence one can have in the model results, and also in order to choose between alternative models. However, if one considers the modelling literature, one finds few indications as to how to evaluate models.

The quantitative approaches to model evaluation that have been proposed involve statistical tests of the model in question. The most commonly used test is of the hypothesis that the regression line of observed versus predicted values passes through the origin and has a slope of unity (Dent and Blackie, 1979; Carter, 1986). Other tests have been proposed for particular situations. Feldman et al. (1984) propose a test of the hypothesis that the model is unbiased, applicable to the case where the individual measurements are not independent. For example, they might represent measurements of insect populations, with sequantial measurements in a given field being related.

Reynolds and Deaton (1982) propose a series of tests for stochastic models. The hypothesis in this case is that the distribution of the model predictions is the same as the distribution of true values, for each set of values of the explanatory variables used in the model. Finally, another test described in Dent and Blackie (1979) tests the hypothesis that the overall (i.e. averaged overall values of the explanatory variables) distribution of results of the model is the same as the overall distribution of true values.

In the present paper we consider the particular, though quite common, case in which the goal of the model is to evaluate some particular output of the system (e.g. wheat yield, milk yield over a lactation period, etc.). The tests referred to above are not really satisfactory here, because they are not directly related to predictive accuracy. It seems more reasonable then, when prediction is the goal, to attempt to quantify directly the predictive accuracy of the model, and to use this as the criterion of model quality.

The measure of predictive accuracy that is commonly used in the statistical literature is the mean squared error of prediction. The purpose of the present paper is to introduce this notion to the modelling literature, and to show in particular how it can be used to compare between alternative models. In Section 1 the criterion is defined, and the calculation of the mean squared error of prediction in various situations is discussed. In Section 2 the use of this criterion in choosing between models is discussed and illustrated by an example. Conclusions are presented in Section 3.

1. MEAN SQUARED ERROR OF PREDICTION

The mean squared error of prediction of a model, denoted MSEP($\hat{p}$) is defined as follows:

$$\text{MSEP}(\hat{p}) = \mathscr{E}\left[\left(y - f(X, \hat{p})^2 \mid \hat{p}\right)\right] \tag{1}$$

where $\mathscr{E}$ indicates an expectation (over the population of interest), y is the quantity to be predicted, for example wheat yield, milk production, etc., and $f(X, \hat{p})$ is the prediction given by the model f. The model involves certain explicative variables, noted X (for example daily meteorological variables, soil characteristics, etc.), and various parameters, noted $\hat{p}$. (A hat on a quantity indicates an estimate; the parameters therefore are noted $\hat{p}$ since they are normally estimated quantities rather than known constants.) As the name implies, then, MSEP($\hat{p}$) is simply the average squared difference between the quantity of interest, and the model prediction of that quantity. It depends of course on the particular values of the parameters which are used in the model, as the notation indicates. To emphasize that the expectation is not over possible values of $\hat{p}$, the usual statistical notation $\mid \hat{p}$ is used, which is to be read 'for fixed $\hat{p}$'.

The value of MSEP$(\hat{p})$ is, by construction, a measure of the predictive accuracy of the model $f(X, \hat{p})$. Since this value refers to the entire population of interest, it cannot in general be measured directly. The problem then is to obtain as estimate, to be noted $\widehat{\text{MSEP}}(\hat{p})$ (using the hat notation), based on the available observations of y and the associated X. The observed data, or test sample, will be noted S, and the number of observations N. Thus $S = (y_i, X_i)$, $i = 1, N$. Different estimators of MSEP$(\hat{p})$ will be appropriate for different situations. Various common situations which arise in practice are discussed below. (More details can be found in Wallach and Goffinet, 1987.)

The simplest case occurs when the observations in S are mutually independent, and when the model is independent of the test sample (that is, the parameters of the model are derived from independent experiments, and are not adjusted to data of the test sample). In this case the natural estimate of MSEP$(\hat{p})$ is simply the average squared error for the observed sample, namely:

$$\widehat{\text{MSEP}}_1(\hat{p}) = \frac{1}{N} \sum_{i=1}^{N} \text{ERR2}_i \tag{2}$$

where

$$\text{ERR2}_i = (y_i - f(X_i, \hat{p}))^2$$

is the squared difference between the ith observed value of y and the corresponding model prediction. Equation (2) can be applied regardless of the size N of the test sample. However, clearly the reliability of the estimate of MSEP$(\hat{p})$ will be poor if N is small. It is important then to consider, along with the estimated value, the variance of that estimate, which measures its reliability. Since $\widehat{\text{MSEP}}_1(\hat{p})$ is simply an average, there is the usual simple expression for the estimated variance, namely:

$$\widehat{\text{VAR}}[\widehat{\text{MSEP}}_1(\hat{p})] = \frac{1}{N-1} \sum_{i=1}^{N} (\text{ERR2}_i - \widehat{\text{MSEP}}_1)^2 \tag{3}$$

A different situation arises if the model parameters are adjusted to the data in the test sample. It is generally recognized that in this case equation (2) gives an underestimate of MSEP$(\hat{p})$. This is intuitively clear. The model has now been specifically adjusted to the test data, and therefore reproduces those data more closely than it predicts on the average for the entire population of interest.

One possible approach in this situation is cross-validation, in which the test data are divided into two groups, one group being used for model adjustment and the second for model evaluation (Picard and Cook, 1984;

Snee, 1977). If this procedure is followed, then equation (2) can be used, based just on the observations in the second group. However, while this technique may be appropriate for fairly large sample sizes, it may not be reasonable for small N, for splitting the sample may then result in unacceptably poor estimates of the parameter values and of MSEP($\hat{p}$). (These will be poor in the sense of having large variances). An alternative approach is to use a resampling method in which the same data are used for adjustment and evaluation, but which takes this into account, at least approximately, in estimating MSEP($\hat{p}$). Efron (1983) used simulated data to compare various resampling techniques for the estimation of MSEP($\hat{p}$) in a particular situation, and found that the bootstrap method gave the best results, so that is the method described below.

The notion that equation (2) underestimates MSEP($\hat{p}$) if the parameters have been fitted to the test sample can be formalized by writing:

$$\mathrm{MSEP}(\hat{p}) = \widehat{\mathrm{MSEP}}_1(\hat{p}) + \mathrm{OP} \tag{4}$$

The correction term that should be added to $\widehat{\mathrm{MSEP}}_1(\hat{p})$ is denoted OP for optimistic, because it measures to what extent one is being overly optimistic concerning the predictive ability of the model when one estimates MSEP($\hat{p}$) by $\widehat{\mathrm{MSEP}}_1(\hat{p})$. The problem now has become one of estimating OP, and it is then an estimate of this quantity that the bootstrap technique provides.

The bootstrap calculation proceeds as follows:

(1) Draw a random sample, with replacement, of N data points from S. The bth such sample will be noted S^b. S^b will normally contain some of the observations two, three, etc. times, while others will not appear at all.

(2) Calculate the model parameters using the data values in S^b. These parameter values will be noted $\hat{p}^b$. These parameters are to be calculated in the same way as the real parameters of interest, $\hat{p}$, were calculated based on the test sample S. It should be noted that this step implies that one can explicitly describe how the parameters were calculated. The bootstrap technique cannot be used, for example, if the parameters have been adjusted subjectively to the data.

(3) Evaluate equation (2) with the sum over the data values in S^b, and with the parameters $\hat{p}^b$ in the model. The result will be noted $\widehat{\mathrm{MSEP}}{}_1^b(\hat{p}^b)$.

(4) Evaluate equation (2) with the sum over the data values in S and with the parameters $\hat{p}^b$ in the model. The result will be noted $\widehat{\mathrm{MSEP}}{}_1^*(\hat{p}^b)$.

(5) Calculate the bth estimate of OP as $\widehat{\mathrm{OP}}^b = \widehat{\mathrm{MSEP}}{}_1^*(\hat{p}^b) - \widehat{\mathrm{MSEP}}{}_1^b(\hat{p}^b)$.

(6) Repeat steps (1)–(5) B times, and than calculate:

$$\widehat{\mathrm{OP}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathrm{OP}}^b \tag{5}$$

B should be chosen sufficiently large that further increases in B have little effect on $\widehat{\text{OP}}$. Efron (1983) suggests that values of B in the range 25–200 should be sufficient.

(7) The bootstrap estimator of the mean squared error of prediction is finally:

$$\widehat{\text{MSEP}}_2(\hat{p}) = \widehat{\text{MSEP}}_1(\hat{p}) + \widehat{\text{OP}} \tag{6}$$

There is no obvious estimator of the variance of $\widehat{\text{MSEP}}_2(\hat{p})$. However, if $\widehat{\text{OP}}$ is small compared to $\widehat{\text{MSEP}}_1(\hat{p})$, then a rough estimator of the variance is given by equation (3).

A further type of situation arises when the data in S are not independent. For example, a common sampling scheme involves first choosing contexts at random (the context might be the year, or the region), and then several individuals (which might be individual plots, animals, plants, etc.) from each context. In this case the observations from the same context are not independent, and this should be taken into account in estimating $\text{MSEP}(\hat{p})$. This situation is treated in Wallach and Goffinet (1987).

2. COMPARING MODELS – AN EXAMPLE

As already noted, one of the fundamental uses of a criterion of model quality is as a basis for choosing between alternative models. For example suppose that one wishes to choose between two different models, $f(X, \hat{p})$ and $g(X', \hat{p}')$. (As indicated, the two models may involve different sets of explicative variables, and different parameters.) The simplest approach is to estimate the mean squared error of prediction for each model, and prefer the model with the smaller error.

A slightly different and more informative approach is to treat the problem of model comparison as one of estimating the difference in MSEP values between the two models. This difference is:

$$\Delta\text{MSEP}(\hat{p}, \hat{p}') = \mathscr{E}\left[(y - f(X, \hat{p}))^2 - (y - g(X', \hat{p}'))^2 \mid \hat{p}, \hat{p}'\right] \tag{7}$$

which reduces to

$$\begin{aligned}\Delta\text{MSEP}(\hat{p}, \hat{p}') = \mathscr{E}\big[&2y(g(X', \hat{p}') - f(X, \hat{p})) \\ &+ \left(f(X, \hat{p})^2 - g(X', \hat{p}')^2\right) \mid \hat{p}, \hat{p}'\big]\end{aligned} \tag{8}$$

If the estimated difference is positive, model g is preferred, while if it is negative model f has the smaller estimated prediction error. Once again, the variance of the estimated value is of importance. If the square root of the variance is much smaller than the estimated value of $\Delta\text{MSEP}(\hat{p}, \hat{p}')$ then one will have confidence that the estimated difference represents a real difference in predictive ability. On the other hand, if the square root of the variance and the estimated ΔMSEP are comparable, one cannot reliably decide which model is a better predictor, and one might then decide to rely on other criteria for choosing between the models (such as simplicity, or biological realism, etc.).

The various situations discussed above, which lead to different estimators of the mean squared error of prediction, are of course relevant here as well. The simplest situation occurs when the individual observations can be treated as independent of one another, and when both models are independent of the test data. In this case the estimator analagous to that of equation (2) is:

$$\hat{\Delta}\mathrm{MSEP}_1(\hat{p},\ \hat{p}') = \frac{1}{N}\sum_{i=1}^{N} V_i \tag{9}$$

where

$$V_i = 2y_i\left[g(X_i',\ p') - f(X_i,\ \hat{p})\right] + \left[f(X_i,\ \hat{p})^2 - g(X_i',\ \hat{p}')^2\right]$$

An estimate of its variance is:

$$\hat{\mathrm{V}}\mathrm{AR}\left[\hat{\Delta}\mathrm{MSEP}_1(\hat{p},\ \hat{p}')\right] = \frac{1}{N-1}\sum_{i=1}^{N}\left(V_i - \hat{\Delta}\mathrm{MSEP}_1(\hat{p},\ \hat{p}')\right)^2 \tag{10}$$

If either of the models to be compared is adjusted to the test data, then this must be taken into account in estimating ΔMSEP($\hat{p}$, $\hat{p}'$). If the observations are again mutually independent, the analogue of equation (6) is:

$$\hat{\Delta}\mathrm{MSEP}_2(\hat{p},\ \hat{p}') = \hat{\Delta}\mathrm{MSEP}_1(\hat{p},\ \hat{p}') + \hat{\Delta}\mathrm{OP} \tag{11}$$

The bootstrap calculation of $\hat{\Delta}$OP is the same as that described for $\hat{\mathrm{O}}$P in the preceding section, except with $\hat{\Delta}\mathrm{MSEP}_1$ in place of $\hat{\mathrm{M}}\mathrm{SEP}_1$.

As an example of the use of MSEP for comparing models, we consider the problem of predicting corn yield at the INRA research center. The observed yields are shown in Table 1. The various measurements in the same year represent plots with different preceding crops and different amounts of irrigation. When the within and between year variances of these data are calculated using analysis of variance estimators (see for example Searle, 1971, p. 474), the former is much larger than the latter. Therefore the year effect is ignored, and the data are treated as independent random observations from the population of interest.

The first model considered, noted f, is the EPIC model (Williams et al., 1984) as modified in France (Cabelguenne et al., 1986). It is assumed that the parameters in this model have not been adjusted to the test data. As Table 1 shows, the values predicted by this model, compared with the

TABLE 1

Observed and predicted yields

Plot	Year	Irrigation (mm)	Observed yield (g/m^2)	EPIC yield (g/m^2)	Model g yield (g/m^2)
B04	1984	155	981	881	1006
B05	1984	155	863	797	889
C02	1984	0	568	730	797
G06	1985	70	904	649	685
H05	1985	0	558	544	540
B04	1986	110	829	690	742
C03	1986	110	753	632	661
E02	1986	0	262	441	397
Average			715	670	715

observed yields, are on the average too low. One might imagine then a second model, denoted g, which has the form:

$$g(X', \hat{p}') = \hat{p}'_1 + \hat{p}'_2 f(X, \hat{p}) \tag{12}$$

This adjusted version of EPIC is just a simple linear model in the parameters $\hat{p}'_1$ and $\hat{p}'_2$. The usual least squares calculation gives $\hat{p}'_1 = -21.202$ and $\hat{p}'_2 = 1.382$. The yields predicted by the model g are shown in the last column of Table 1. The model g of course fits the data better than does model f. It has, after all, been specifically constructed to do so. However, it is not necessarily a better predictor. This is what we wish to test using the machinery described above.

We begin by calculating the first term in equation (11). The result is $\hat{\Delta}\mathrm{MSEP}_1(\hat{p}, \hat{p}') = 44.3$. The fact that this result is positive simply indicates that model g gives the better fit to the data, which we already knew would be the case. Next, the correction term $\hat{\Delta}\mathrm{OP}$ must be estimated. In the present case this is just equal to $-\hat{\mathrm{O}}\mathrm{P}$ for the model g, since the model f is assumed to be independent of the test data. The bootstrap calculation, with $B = 1000$, gives $\Delta\mathrm{OP} = -111.2$. The final result is then:

$$\hat{\Delta}\mathrm{MSEP}_2(\hat{p}, \hat{p}') = 44.3 - 111.2 = -66.9 \tag{13}$$

Since this value is negative the conclusion is that the model f, that is the unadjusted version of EPIC, is a better predictor of future corn yields for this environment than is the adjusted version.

3. CONCLUSIONS

The mean squared error of prediction criterion provides a logical framework for evaluating and comparing models. It should particularly be noted that this criterion takes into account, in a logical way, the extent to which a model has been adjusted to the test data. A tempting way of improving a model for a particular environment is to adjust it, by standard statistical techniques, to data from that environment. The technique here allows one to compare the adjusted and original model, as illustrated in the preceding section.

It should also be noted that the estimation of MSEP depends on very few assumptions. The major assumption concerns the representativity of the test data. The results presented here require that the test data represent a random sample from the population of interest.

REFERENCES

Cabelguenne, M., Charpenteau, J.L., Jones, C.A., Marty, J.R. and Rellier, J.P., 1986. Conduite des systèmes de grandes cultures et prévision des rendements: Tentatives de modélisation. II. Etalonage du modèle: resultats et perspectives. C. R. Acad. Agric. Fr., 72: 125–132.

Carter, N., 1986. Simulation modelling. In: G.D. McLean, R.G. Garret and W.G. Ruesink (Editors), Plant Virus Epidemics. Academic Press, Sydney, N.S.W., pp. 193–215.

Dent, J.B. and Blackie, M.J., 1979. Systems Simulation in Agriculture. Applied Science, London, 180 pp.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc., 78: 316–331.

Feldman, R.M., Curry, G.L. and Wehrly, T.E., 1984. Statistical procedure for validating a simple population model. Environ. Entomol., 13: 1446–1451.

Picard, R.R. and Cook, R.D., 1984. Cross validation of regression models. J. Am. Stat. Soc., 79: 575–583.

Reynolds, M.R., Jr. and Deaton, M.L., 1982. Comparison of some tests for validation of stochastic simulation models. Comm. Stat. Simul. Comput., 11: 769–799.

Snee, R.D., 1977. Validation of regression models: methods and examples. Technometrics, 19: 415–428.

Wallach, D. and Goffinet, B., 1987. Mean squared error of prediction in models for studying ecological and agronomic systems. Biometrics, 43: 561–573.

Williams, J.R., Jones, C.A. and Dyke, P.T., 1984. A modelling approach to determining the relationship between erosion and soil productivity. Trans. ASAE, 27: 129–144.