Biom. J. 26 (1984) 2, 173–184

University of Missouri – Columbia, U.S.A.

A Technique for Evaluating Forecasting Models

E. Xekalaki and S. K. Katti

Abstract

The paper presents a new methodology for evaluating the performance of a forecasting model. The evaluation-criterion utilizes a "credibility interval" centered at the model prediction. Given predicted and observed values, the length of the "credibility interval" is increased (or decreased) according as an observed value of the dependent random variable falls out of (or into) the interval. Based on that, various ways of assessing the rating of the model are discussed and illustrative examples are given.

Key words: Model evaluation, rating, credibility interval.

Paper prepared for the U.S. Department of Agriculture Project on "Test and Evaluation of Crop Yield Models; Development and Application of Methods" in partial fulfilment of Cooperative Research Agreement Number: 58-319T-0-0255X.

1. Introduction

The world is characterized by uncertainty about the future. As a result, the development of forecasting methods has become necessary and the interest of research workers in various fields, greatly stimulated by the work of Box and Jenkins (1976), has been oriented towards this direction. In parallel with the importance that forecasting models have in planning, the accuracy of forecasts has a vital part to play. The consequences of a poor forecast can be very costly in human, environmental and financial terms. Against this prospect, great concern has been shown in testing and evaluation of forecasting models and various techniques have been suggested. Among others, the papers of Stone (1974, 1978), Geisser (1975), Geisser and Eddy (1979), Snee (1977), Mitchell and Wilson (1979), Gass and Thompson (1980), Butler and Rothman (1980), Ramsey and Kmenta (1980) and Chow (1980) cover a substantial amount of work in this area. There is no doubt that much can be gained from the application of such techniques to practical problems in various fields.

The principal aim of this paper is of a dual nature. It focuses attention on the development of a model evaluation schema and of selection criteria to guide

one's choice among several alternative models with applications on crop yield forecasting models. The presumption is that the truth cannot be modeled accurately and that any model used is "wrong" in the sense that it is merely an approximation to the true state of nature. Hence the problem of evaluating a model is one of searching for evidence of and assessing the model's inadequacy. Consequently, the selection of the "best" model from a set of several alternative models will be done on the basis of "least inadequacy".

The next section describes a new methodology for assessing a model's performance based on the sequential construction of what we term "credibility intervals". In section 3 some scoring rules are suggested and a selection mechanism is proposed in section 4. The entire schema is then applied to simulated as well as real data (section 5). Finally a brief discussion follows in section 6.

2. Evaluation Methodology

In attempting to evaluate any given model the research worker is, in the present authors' opinion, faced with a fascinating statistical problem which by its nature does not admit a deterministic solution or an approach in the framework of classical statistical methods. The researcher by his experience with the experimental material and his knowledge of the subject must resort to his judgement in designing the evaluation schema.

Before we proceed with the description of our technique let us doscribe the sort of models to which it is intended to apply. These are among the models being considered by the United States Department of Agriculture for predicting crop yields.

Consider the model

$$Y_t = X_t\beta + e_t \tag{2.1}$$

where Y_t is an $l_t \times 1$ vector of yearly observations on the dependent random variable (crop yield), X_t is an $l_t \times m$ matrix of known coefficients, $l_t \geqq m$, $|X_t'X_t| \neq 0$ (trend and weather variables), β is an $m \times 1$ vector of regression coefficients and e_t is an $l_t \times 1$ vector of normal error random variables with mean $\mathrm{E}(e_t) = 0$ and dispersion matrix $V(e_t) = \sigma_t^2 I_t$. Here I_t is an $l_t \times l_t$ identity matrix.

From the model one predicts the yield of the $(t+1)$-th year to be $\hat{Y}_{t+1}^0 = X_{t+1}^0 \hat{\beta}_t$ where $\hat{\beta}_t$ is the least squares estimator of β at time $t+1$ given by $\hat{\beta}_t = (X_t'X_t)^{-1} X_t'Y_t$ and X_{t+1}^0 is a $(1 \times m)$ vector of values of the regressors for the $(t+1)$ th year. The variance of $\hat{Y}_{t+1}^0$ is then given by

$$\mathrm{V}(\hat{Y}_{t+1}^0) = \sigma_t^2 \{X_{t+1}^0 (X_t'X_t)^{-1} X_{t+1}^{0\prime} + 1\} \tag{2.2}$$

and is estimated by putting

$$S_t^2 = (Y_t - X_t\beta_t)' (Y_t - X_t\beta_t)/(l_t - m) \tag{2.3}$$

for σ_t^2. After the true crop yield Y_{t+1}^0 for the $(t+1)$-th year has been observed, the model to be used for predicting the crop yield of the $(t+2)$-th year becomes

$$Y_{t+1}=X_{t+1}\beta+e_{t+1}$$

where now the matrices X_{t+1} and Y_{t+1} are defined by

$$X_{t+1}=\begin{pmatrix} X_t \\ X_{t+1}^0 \end{pmatrix} \quad \text{and} \quad Y_{t+1}=\begin{pmatrix} Y_t \\ Y_{t+1}^0 \end{pmatrix}$$

with dimensions $(l_t+1)\times m$ and $(l_t+1)\times 1$ respectively.

Then, for the $(t+2)$th year the vector of regression coefficients β can be estimated by

$$\hat{\beta}_{t+1}=(X'_{t+1}X_{t+1})^{-1}X'_{t+1}Y_{t+1} \tag{2.4}$$

The evaluation schema that we are going to propose involves an n-stage technique that reflects the behavior of the model over the number n of years for which it was used. This technique consists of the following steps:

1. At time $t+1$, predict the crop yield $\hat{Y}_{t+1}^0$ for the $(t+1)$-th year using model (2.1).

2. Construct a "credibility interval" for Y_{t+1}^0, say C_{t+1}, thus

$$C_{t+1}=[\hat{Y}_{t+1}^0-k_tS_t,\ \hat{Y}_{t+1}^0+k_tS_t] \tag{2.5}$$

where S_t^2 is as given by (2.3) and k_t is a positive constant whose initial value is set by the experimenter.

3. Observe the true yield Y_{t+1}^0 for the $(t+1)$-th year.

4. Choose a scoring rule that assigns scores to the two complementary outcomes

$$\{Y_{t+1}^0\in C_{t+1}\} \quad \text{and} \quad \{Y_{t+1}^0\notin C_{t+1}\}.$$

5. Re-estimate the model's regression coefficients using (2.4) and predict the crop yield for the $(t+2)$-th year ($\hat{Y}_{t+2}^0$).

6. Construct the "credibility interval" of Y_{t+2}^0 as in step 2 with the constant k_{t+1} defined as

$$k_{t+1}=\begin{cases}(1-\alpha_{t+1})\,k_t & \text{if} \quad Y_{t+1}^0\in C_{t+1} \\ (1+\gamma_{t+1})\,k_t & \text{if} \quad Y_{t+1}^0\notin C_{t+1}\end{cases} \tag{2.6}$$

where α_t, γ_t are defined by the experimenter $(0<\alpha_t<1,\ 0<\gamma_t<1)$.

7. Repeat the process for as many times as the number of years the model was applied, say n. The average of the scores from step 4 over the n years is a possible choice of a final rating for the model reflecting its inadequacy.

To some extent, the technique just described bears an analogy to a two-stage method for the estimation of the mean of a distribution introduced by KATTI (1962) and extended to the multivariate case by WAIKAR and KATTI (1971).

The "credibility" p_{t+1} of the interval given by (2.5) is defined to be the probability with which the interval is expected to contain the actual crop yield, i.e.,

$$\begin{aligned} p_{t+1}&=P\,(Y_{t+1}^0\in C_{t+1}) \\ &=P\,(|\hat{Y}_{t+1}^0-Y_{t+1}^0|\leqq k_tS_t\mid k_t), \quad t=0,\,1,\,2,\,\ldots \end{aligned}$$

These probabilities can be evaluated using

$$p_{t+1} = -1 + 2T_{l_t - m}\left(\frac{k_t}{\sqrt{X^0_{t+1}(X_t'X_t)^{-1}\,X^{0'}_{t+1}+1}}\right)$$

where $T_\nu(\cdot)$ represents the cumulative distribution function of the t distribution with ν degrees of freedom.

It is interesting to remark here that due to the availability of fairly long time series, the behavior of S_t^2 mimics sufficiently closely that of σ_t^2. Hence assuming σ_t^2 known is purely a matter of detail. In such a case p_{t+1} can be evaluated using the standard normal tables. Alternatively, we may observe that $|\hat{Y}^0_{t+1} - Y^0_{t+1}|$ will have a folded normal distribution with mean $\mu_f = S_t\sqrt{\frac{2}{\pi}(1+X^0_{t+1}(X_t'X_t)^{-1}\,X^{0'}_{t+1})}$ and variance $\sigma_f^2 = \left(1-\frac{2}{\pi}\right)(1+X^0_{t+1}(X_t'X_t)^{-1}X^{0'}_{t+1})\,S_t^2$. This distribution was defined by LEONE et al. (1961) who provided tables for its cumulative distribution function, say $N_f(\cdot)$ for various values of μ_f/σ_f. Then, the credibilities can be evaluated by

$$p_{t+1} = N_f\left(\frac{k_t}{\sqrt{\left(1-\frac{2}{\pi}\right)(X^0_{t+1}(X_t'X_t)^{-1}\,X^{0'}_{t+1}+1)}}\right)$$

using their tables for $\mu_f/\sigma_f = \sqrt{\frac{2}{\pi-2}} = 1.3236$.

The effect of the experimenter's personal judgement becomes evident in steps 2 and 6. The length of the credibility interval for a given year is increased (or decreased) by an amount determined by the experimenter according as the observed crop yield of the previous year falls outside (or within) the corresponding credibility interval.

Starting with predicting the yield for year 1 ($t=0$) and repeating the process n times we end up with a constant $k_n = k_0\varphi(n, w_n)$ where

$$\varphi(n, w_n) = \prod_{i=0}^{w_n}(1-\alpha_i)\prod_{j=0}^{n-w_n}(1+\gamma_j) \tag{2.7}$$

which is to be used in an application of the technique for a further year ($(n+1)$-th year) if required. Here w_n represents the total number of times the observed crop yield falls within the credibility interval during the n years. By the nature of the problem it is obvious that the choice of α_i, γ_j should ensure that

$$\lim_{n\to+\infty}\lim_{w_n\to n}\varphi(n, w_n) = 0 \tag{2.8}$$

and

$$\lim_{n\to+\infty}\lim_{w_n\to 0}\varphi(n, w_n) = +\infty$$

We propose the following choice:

$$\alpha_i=\frac{1}{i+1},\quad i=1, 2, \ldots, w_n;\quad \gamma_j=\frac{1}{j+1},\quad j=1, 2, \ldots, n-w_n;\quad \alpha_0=\gamma_0=0.$$

Then $\varphi(n, w_n)=\frac{1}{2}\frac{n-w_n+2}{w_n+1}$ which obviously satifies (2.8). Moreover, $\lim_{n\to+\infty}$ $\lim_{w_n\to n/2}\varphi(n, w_n)=\frac{1}{2}$ which in turn implies a 50 % decrease in the initial value k_0 if in 50 out of 100 times the credibility region contains the actual crop yield. $\left(\text{In general } \lim_{n\to+\infty}\lim_{w_n\to np}\varphi(n, w_n)=\frac{1-p}{2p},\ 0<p\leqq 1.\right)$

3. Scoring Rules

To assess the inadequacy of the model in question we need to define a scoring rule. According to what we have proposed in step 4, the rule must be such that, for each year in the study, a score be assigned to the corresponding performance of the model. The final rating will be represented by the average score. In the sequel, some scoring rules are suggested.

The simplest possible scoring rule would amount to assigning a score r_{t+1} where

$$(3.1)\qquad r_{t+1}=\begin{cases}1 & \text{if} \quad Y^0_{t+1}\in C_{t+1}\\ 0 & \text{if} \quad Y^0_{t+1}\notin C_{t+1}\end{cases}$$

at each point t in time. Then, the final rating R of the model would be the average score obtained by the model over the n years, i.e., $R=\sum_i r_i/n$. Here, obviously, R represents the proportion of times the observed value fell within the credibility interval. Clearly, an average score very close to 0 will imply a highly inadequate model.

It is worth noting here that the rule in (3.1) does not depend on the length of the credibility interval. So at any time t, a model with a narrow interval will get the same score as another with a very wide credibility interval. To allow for a higher score to the model with the narrower credibility interval one may consider the rule

$$(3.2).\qquad r_{t+1}=\begin{cases}\dfrac{1}{S_t k_t} & \text{if} \quad Y^0_{t+1}\in C_{t+1}\\ 0 & \text{if} \quad Y^0_{t+1}\notin C_{t+1}\end{cases}$$

The higher the average score (final rating) is the lower the model inadequacy.

Another possibly desirable feature of a scoring rule would be to give scores that depend on the distance between predicted and observed crop yield. One

such possibility might be the rule

$$r_{t+1}=\frac{|\hat{Y}^0_{t+1}-Y^0_{t+1}|}{S_t k_t} \tag{3.3}$$

which takes into account how close to (or how far from) the true yield the prediction is relative to the length of the corresponding credibility interval. Here, a low model inadequacy is reflected by an average score which is close to 0 ($R\cong 0$). The greater R is than 0 the more inefficient the model is.

Finally, we propose two further rules, namely

$$r_{t+1}=\begin{cases} -\ln p_{t+1} & \text{if} \quad Y^0_{t+1}\in C_{t+1} \\ 0 & \text{if} \quad Y^0_{t+1}\notin C_{t+1} \end{cases} \tag{3.4}$$

and

$$r_{t+1}=\begin{cases} \dfrac{p_{t+1}^{-1}}{\sqrt{p_{t+1}^{-2}+q_{t+1}^{-2}}} & \text{if} \quad Y^0_{t+1}\in C_{t+1} \\ 0 & \text{if} \quad Y^0_{t+1}\notin C_{t+1} \end{cases} \tag{3.5}$$

where $q_{t+1}=1-p_{t+1}$. In both cases the score assigned to the model performance at any time t is dependent upon the corresponding interval credibility. Again the higher the average score the lower the inadequacy of the model.

4. Model Selection

Selecting a model from a set of available alternative models will involve choosing a scoring rule and deciding on the basis of a final rating that reflects the least inadequacy. So, the problem of model selection reduces to that of deciding which rule to choose. The investigator has to make up his/her mind about the behavioral features of the model that, according to his/her qualitative "feel" for the situation will provide a "consensus" view of the inadequacy of the model. So choosing a scoring rule is purely a matter of personal judgement and depends on "intelligence" not available to any evaluation scheme.

The scoring rules suggested in section 3 are merely examples of assessing the merit of a model based on various features that may be to the researcher's interest. As it can be seen from the numerical results presented in the next section, what is considered to be the "best" model by one rule is not always the "best" by another rule.

Of course, the use of the scoring rules presupposes that the statistical behavior of the model in the future will be similar to its statistical behavior in the past. On this assumption, one can choose a model on the basis of the "best" final rating, i.e., one can choose the model with the highest or (lowest) final rating. In the case of rules (3.1), (3.2), (3.4) and (3.5), for instance, one can select the model i_0 ($i_0\in\{1, 2, \ldots, s\}$ for which

$$R^{(i_0)}=\max_{i\in\{1,2,\ldots,s\}}\sum_j r_j^{(i)}/n$$

(Here the super script i refers to the i-th model). Similarly in the case of scoring rule (3.3) we may select the i_0-th model with

$$R^{(i_0)} = \min_{i \in \{1,2,\ldots,s\}} \sum_j r_j^{(i)}/n$$

5. Some Applications

The methodology developed in sections 2, 3 and 4 has been tried out on some real data. Tables I and II present the results. In particular, these tables illustrate the performance of the model evaluation technique and of the model selection procedure on the basis of the forecasts of two different corn yield models and of the true yields reported by two crop reporting districts (*CRD s*) in the state of Indiana, USA for the years 1963–1980. The year by year rating of the model behavior is illustrated in terms of scoring rules (3.1) through (3.5). CL_t and CU_t denote the lower and upper end points of the credibility intervals respectively. The credibilities p_t of these intervals are given in the sixth column. These have been computed using the following approximation formula for the values of the standard normal cumulative distribution function due to HASTINGS (1965), p. 187)

$$p_{t+1} = 1 - \frac{1}{2}\left[1 + \sum_{i=1}^{6} a_i \frac{k_t^i}{\sqrt{2^i (X_{t+1}^0 (X_t' X_t)^{-1} X_{t+1}^{0'} + 1)^i}}\right]^{-16}$$

Table I

Evaluation results based on corn yield data as reported by CRD 20 in the State of Indiana for the years 1963–1980.

Model A

Index	year t	pred $\hat{Y}_t^0$	STD S_t	CL_t	CU_t	p_t	yield Y_t^0	score (3.1)	score (3.2)	score (3.3)	score (3.4)	score (3.5)	k_t
0	–	–	–	–	–	–	–	–	–	–	–	–	2.00
1	1963	48.2	2.125	43.950	52.450	.63	52.300	1.000	.235	.965	.457	.502	1,00
2	1964	56.6	2.065	54.535	58.665	.61	44.600	.000	.000	5.812	.000	.000	1.50
3	1965	42.6	2.152	39.372	45.828	.69	56.800	.000	.000	4.398	.000	.000	2.00
4	1966	49.3	2.361	44.578	54.022	.78	46.800	1.000	.212	.529	.244	.267	1.33
5	1967	42.2	2.309	39.121	45.279	.70	50.900	.000	.000	2.825	.000	.000	1.67
6	1968	50.5	2.379	46.534	54.466	.79	55.000	.000	.000	1.135	.000	.000	2.00
7	1969	60.9	2.369	56.162	65.638	.79	59.500	1.000	.211	.295	.235	.256	1.50
8	1971	56.3	2.318	52.823	59.777	.70	62.800	.000	.000	1.870	.000	.000	1.75
9	1972	60.6	2.313	56.552	64.648	.84	61.800	1.000	.247	.296	.174	.187	1.40
10	1973	58.9	2.269	55.724	62.076	.79	62.300	.000	.000	1.070	.000	.000	1.60
11	1974	59.0	2.255	55.392	62.608	.84	38.900	.000	.000	5.570	.000	.000	1.80
12	1975	64.1	3.314	58.135	70.065	.75	62.400	1.000	.168	.285	.293	.322	1.50
13	1976	57.8	3.254	52.919	62.681	.81	67.600	.000	.000	2.008	.000	.000	1.67
14	1977	63.9	3.377	58.272	69.528	.84	62.600	1.000	.178	.231	.172	.185	1.43
15	1978	65.4	3.321	60.655	70.145	.83	61.700	1.000	.211	.780	.190	.204	1.25
16	1979	67.3	3.297	63.179	71.421	.83	68.400	1.000	.243	.267	.186	.200	1.11
17	1980	63.7	3.247	60.093	67.307	.80	54.000	.000	.000	2.689	.000	.000	1.22
Average score								.471	.100	1.825	.115	.125	

Model B

Index	year t	pred $\hat{Y}_t^0$	STD S_t	CL_t	CU_t	p_t	yield Y_t^0	score (3.1)	score (3.2)	score (3.3)	score (3.4)	score (3.5)	k_t
0	–	–	–	–	–	–	–	–	–	–	–	–	2.00
1	1963	50.8	2.656	45.488	56.112	.95	52.300	1.000	.188	.282	.046	.047	1.00
2	1964	48.3	2.618	45.682	50.918	.81	44.600	.000	.000	1.413	.000	.000	1.50
3	1965	48.9	2.640	44.939	52.861	.90	56.800	.000	.000	1.995	.000	.000	2.00
4	1966	48.3	2.882	42.536	54.064	.96	46.800	1.000	.173	.260	.041.	.042	1.33
5	1967	51.4	2.843	47.609	55.191	.87	50.900	1.000	.264	.132	.143	.152	1.00
6	1968	53.4	2.798	50.602	56.198	.81	55.000	1.000	.357	.572	.214	.232	.80
7	1969	56.1	2.765	53.888	58.312	.76	59.500	.000	.000	1.537	.000	.000	1.00
8	1971	59.9	2.774	57.126	62.674	.82	62.800	.000	.000	1.045	.000	.000	1.20
9	1972	59.9	2.770	56.576	63.224	.86	61.800	1.000	.301	.572	.151	.161	1.00
10	1973	61.4	2.745	58.655	64.145	.81	62.300	1.000	.364	.328	.214	.232	.86
11	1974	57.5	2.709	55.178	59.822	.76	38.900	.000	.000	8.011	.000	.000	1.00
12	1975	60.5	3.698	56.802	64.198	.82	62.400	1.000	.270	.514	.203	.220	.87
13	1976	62.1	3.659	58.898	65.302	.79	67.600	.000	.000	1.718	.000	.000	1.00
14	1977	61.4	3.703	57.697	65.103	.81	62.600	1.000	.270	.324	.214	.232	.89
15	1978	63.3	3.659	60.047	66.553	.80	61.700	1.000	.307	.492	.229	.249	.80
16	1979	67.0	3.622	64.102	69.898	.77	68.400	1.000	.345	.483	.264	.289	.73
17	1980	66.2	3.584	63.593	68.807	.73	54.000	.000	.000	4.680	.000	.000	.82
Average score								.588	.167	1.433	.101	.109	

Table 11

Evaluation results based on corn yield data as reported by CRD 30 in the State of Indiana for the years 1963–1980.

Model A

Index	year t	pred $\hat{Y}_t^0$	STD S_t	CL_t	CU_t	p_t	yield Y_t^0	score (3.1)	score (3.2)	score (3.3)	score (3.4)	score (3.5)	k_t
0	–	–	–	–	–	–	–	–	–	–	–	–	2.00
1	1963	42.8	2.863	37.074	48.526	.63	51.200	.000	.000	1.467	.000	.000	3.00
2	1964	54.8	2.794	46.418	63.182	.80	39.300	.000	.000	1.849	.000	.000	4.00
3	1965	32.8	2.912	21.151	44.449	.92	52.600	.000	.000	1.700	.000	.000	5.00
4	1966	51.0	3.288	34.562	67.438	.97	48.300	1.000	.061	.164	.028	.029	2.50
5	1967	44.7	3.208	36.679	52.721	.87	46.400	1.000	.125	.212	.140	.149	1.67
6	1968	46.9	3.132	41.680	52.120	.77	54.200	.000	.000	1.399	.000	.000	2.00
7	1969	54.4	3.138	48.124	60.676	.78	52.800	1.000	.159	.255	.249	.273	1.50
8	1971	57.7	3.069	53.097	62.303	.68	55.100	1.000	.217	.565	.386	.426	1.20
9	1972	57.1	3.006	53.493	60.707	.75	59.300	1.000	.277	.610	.282	.310	1.00
10	1973	59.9	2.954	56.946	62.854	.73	58.100	1.000	.338	.609	.318	.351	.86
11	1974	49.9	2.903	47.412	52.388	.68	35.300	.000	.000	5.868	.000	.000	1.00
12	1975	50.7	3.242	47.458	53.942	.68	52.400	1.000	.308	.524	.386	.426	.87
13	1976	58.1	3.185	55.313	60.887	.68	62.800	.000	.000	1.686	.000	.000	1.00
14	1977	60.7	3.162	57.538	63.862	.73	64.900	.000	.000	1.328	.000	.000	1.12
15	1978	60.3	3.145	56.761	63.839	.78	60.900	1.000	.283	.170	.243	.266	1.00
16	1979	66.4	3.094	63.306	69.494	.77	67.500	1.000	.323	.356	.256	.280	.90
17	1980	65.1	3.047	62.358	67.842	.76	65.100	1.000	.365	.000	.279	.307	.82
Average score								.588	.145	1.104	.151	.166	

Model B

Index	year t	pred $\hat{Y}_t^0$	STD S_t	CL_t	CU_t	p_t	yield Y_t^0	score (3.1)	score (3.2)	score (3.3)	score (3.4)	score (3.5)	k_t
0	–	–	–	–	–	–	–	–	–	–	–	–	2.00
1	1963	48.6	3.242	42.117	55.083	.95	51.200	1.000	.154	.401	.053	.055	1.00
2	1964	45.5	3.207	42.293	48.707	.80	39.300	.000	.000	1.933	.000	.000	1.50
3	1965	45.4	3.298	40.453	50.347	.90	52.600	.000	.000	1.455	.000	.000	2.00
4	1966	48.6	3.435	41.730	55.470	.96	48.300	1.000	.146	.044	.038	.039	1.33
5	1967	48.0	3.380	43.494	52.506	.88	46.400	1.000	.222	.355	.133	.140	1.00
6	1968	48.0	3.335	44.665	51.335	.81	54.200	.000	.000	1.859	.000	.000	1.25
7	1969	50.8	3.422	46.523	55.077	.87	52.800	1.000	.234	.468	.140	.149	1.00
8	1971	54.4	3.385	51.015	57.785	.82	55.100	1.000	.295	.207	.203	.220	.83
9	1972	55.6	3.338	52.818	58.382	.77	59.300	.000	.000	1.330	.000	.000	1.00
10	1973	57.3	3.339	53.961	60.639	.82	58.100	1.000	.300	.240	.204	.221	.86
11	1974	51.1	3.295	48.275	53.925	.77	35.300	.000	.000	5.594	.000	.000	1.00
12	1975	48.8	3.924	44.876	52.724	.81	52.400	1.000	.255	.917	.206	.223	.87
13	1976	60.8	3.908	57.381	64.219	.79	62.800	1.000	.292	.585	.240	.262	.78
14	1977	58.7	3.870	55.690	61.710	.77	64.900	.000	.000	2.060	.000	.000	.89
15	1978	59.1	3.924	55.612	62.588	.80	60.900	1.000	.287	.516	.228	.248	.80
16	1979	62.8	3.886	59.691	65.909	.77	67.500	.000	.000	1.512	.000	.000	.90
17	1980	64.1	3.899	60.590	67.610	.79	65.100	1.000	.285	.285	.237	.259	.82
Average score								.588	.145	1.162	.099	.107	

where

$$a_1 = .0705230784, \quad a_2 = .0422820123, \quad a_3 = .0092705272$$

$$a_4 = .0001520143, \quad a_5 = .0002765672, \quad a_6 = .0000430638$$

The last column gives the values of k_t. Note that the initial value of k_t at $t=0$ is 2 for both models.

Looking closely at the final scores of the two models one can compare the performance of the two models. Consider for example Table I. If one were to choose a model to predict the yield for 1981 on the basis of scoring rule (3.1), one would choose model B. For scoring rules (3.2), (3.3), (3.4) and (3.5) the choice would be B, B, A and A respectively. Similarly, for Table II the final scores call for the selection of model B (or A), B (or A), A, A, and A. From a further inspection of Table II one may observe that the average scores assigned to the two models are quite close. This indicates that the statistical variability is small and some inference, with proper assessment of the reliability of the inference, is feasible. This problem will be the subject of future study. Moreover, it is worth noting that all the five scoring rules yielded similar ratings of the models. A look at the correlations among the five rules presented in Table III can verify this. It appears from the high correlations that the five scoring rules considered are consistent means of evaluation of these models. Table III indicates that the relationship between scores (3.1) (3.2) (3.4) and (3.5) is stronger than the relationship between any of these four rules and rule (3.3). This is probably due to the non-zero score that rule (3.3) assigns to the event $\{Y_{t+1}^0 \notin C_{t+1}\}$. Also, the negative signs that appear merely reflect the fact that by rule (3.3) a high score indicates a low performance.

Table III

Correlations among the scores of the performance of models A and B on the corn yield data of the State c Indiana based on the scoring rules (3.1) to (3.5).

	Model A					Model B				
Score	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)
Score					CRD 20					
(3.1)	1.00000	.98580	−.63134	.89675	.89259	1.00000	.95022	−.55368	.84595	.8398
(3.2)	.98580	1.00000	−.61791	.88828	.88373	.95022	1.00000	−.51671	.94885	.9454
(3.3)	−.63134	−.61791	1.00000	−.53812	−.53515	−.55368	−.51671	1.00000	−.45286	−.4492
(3.4)	.89675	.88828	−.53812	1.00000	.99994	.84595	.94885	−.45286	1.00000	.9999
(3.5)	.89259	.88373	−.53515	.99994	1.00000	.83989	.94540	−.44922	.99991	1.0000
					CRD 30					
(3.1)	1.00000	.86272	−.58783	.85888	.85476	1.00000	.94895	−.60951	.84703	.8407
(3.2)	.86272	1.00000	−.49276	.92593	.92484	.94895	1.00000	−.56790	.96574	.9625
(3.3)	−.58783	−.49276	1.00000	−.46915	−.46611	−.60951	−.56790	1.00000	−.48750	−.4833
(3.4)	.85888	.92593	−.46915	1.00000	.99995	.84703	.96574	−.48750	1.00000	.9999
(3.5)	.85476	.92484	−.46611	.99995	1.00000	.84074	.96254	−.48330	.99991	1.0000

It should perhaps be noted here that the average scores used as final ratings are meaningful only as a means of studying the relative performance of several models. Standardizing them (say, in the range (0,1)) may be desirable for the purpose of obtaining a more meaningful final rating for the performance of a single model.

Some other forms for α_t and γ_t have also been considered giving similar results which for the sake of brevity are not included here.

6. Discussion

An approach has been suggested for the evaluation of the performance of a forecasting model before the predictions obtained are fed to the decision making system.

In designing the evaluation schema an attempt was made to move away from the classical statistical methodology. The innovation brought is the introduction of the credibility interval whose length is changed depending on the agreement between the actually observed and the predicted yield. So, instead of first fixing the probability with which one wishes the actual value to fall within an interval and then constructing this interval, we follow the opposite approach. We first define the interval which, according to our judgement of the particular situation, will provide a range of values of the actual yield that can be thought of as reflecting a "reasonable" model performance. Then, we evaluate how credible this range of values is. Applying the technique sequentially for a number of years for which

data are available and scoring the model behavior for every individual year we come up with a final assessment (rating) of the model's inadequacy in representing the truth. The evaluation methodology and the model selection procedure has been illustrated on some real crop yield data.

Of course this is only a first study of a new technique which is worth further exploration. It has been made with the hope of providing a means of monitoring and appraising the performance of a model so that its forecasts can be utilized in optimizing actions under an uncertain future. Though the application concerned crop yield models only, the technique described here can be, perhaps with some modifications, of equal value to users of linear models in other fields. So, it would be interesting to examine the application of this technique for evaluating the forecasting potential of, among others, economic models estimating national products, private investment, public expenditure etc. The models of GÓMEZ and TINTNER (1980), for instance, may provide an interesting case for such an investigation.

References

BOX, G. E. P. and JENKINS, G. M., 1976: Time series analysis. Forecasting and control. Holden-Day, San Francisco.

BUTLER, R. and ROTHMAN, E. D., 1980: Predictive intervals based on reuse of the sample. J. Amer. Statist. Assoc. **75**, 881–889.

CHOW, G. C., 1980: Evaluation of econometric models by decomposition and aggregation. In: Methodology of macroeconomic models, J. KMENTA and J. B. RAMSEY (Eds.). North-Holland Publ. 1980, in press.

GASS, S. I. and THOMPSON, B. W., 1980: Guidelines for model evaluation: An abridged version of the U.S. general accounting office exposure draft. Operations Research **28**, 431–439.

GEISSER, S., 1975: The predictive sample reuse method with applications. J. Amer. Statist. Assoc. **70**, 320–328.

GEISSER, S. and EDDY, W. F., 1979: A predictive approach to model selection. J. Amer. Statist. Assoc. **74**, 153–160.

GÓMEZ, G. L. and TINTNER, G., 1980: The application of diffusion processes in problems of developmental economic planning: A case study (Colombia). In: Studies in Economic Theory and Practice. JERZY LOS (ed.). North Holland Publ., pp. 177–194.

HASTINGS, C., Jr., 1955: Approximations for digital computers. Princeton University Press.

KATTI, S. K., 1962: Use of some a priori knowledge in the estimation of means from double samples. Biometrics **18**, 139–147.

LEONE, F. C., NELSON, L. S. and NOTTINGHAM, R. B., 1961: The folded normal distribution. Technometrics **3**, 543–550.

MITCHELL, T. J. and WILSON, D. G., 1979: Energy model validation: Initial perceptions of the process. Technical Report, Union Carbide Corporation, Oak Ridge.

RAMSEY, J. B. and KMENTA, J., 1980: Problems and issues in evaluating econometric models. In: Evaluation of Ecomometric Models, J. Kmenta and J. B. Ramsey (Eds). Academic Press, New York, pp. 1–11.

SNEE, R. D., 1977: Validation of regression models: Methods and examples. Technometrics **19**, 415–428.

STONE, M., 1974: Cross-validatory choice and assessment of statistical predictions J. Roy. Statist. Soc. B, **36**, 111–147.

STONE, M., 1978: Cross-validation: A review. Math. Operationsforsch. Statist., Ser. Statistics. **9**, 127–139.

WAIKAR, V. B. and KATTI, S. K., 1971: On a two-stage estimate of the mean. J. Amer. Statist. Assoc. **66**, 75–81.

Manuscript received: 25. 2. 1982

Author's address:

Dr. EVDOKIA XEKALAKI
Dept. of Mathematics
University of Crete
Heraklio, Crete
Greece