Technical Report No. 286

STATISTICAL TECHNIQUES FOR VALIDATING COMPUTER SIMULATION MODELS

Mike Garratt

Natural Resource Ecology Laboratory
Colorado State University
Fort Collins, Colorado

GRASSLAND BIOME
U.S. International Biological Program

June 1975

TABLE OF CONTENTS

ABSTRACT

The purpose of this paper is to suggest a set of statistical tests useful in the validation of simulation models of real world systems. Multivariate techniques can be employed to test two hypotheses: agreement of model predictions and empirical observation and agreement of the system dynamics or "shape" of the predictive traces and data. The methods described employ multivariate analysis of variance (MANOVA), permutation tests, and nonparametric ranking tests. For illustrative purposes the validation tests were applied to the primary producer and decomposition sections of the ecosystem level model (ELM) of a grassland community being developed by the Grassland Biome, U.S. International Biological Program.

1.0 INTRODUCTION

The widespread use of computer simulation models to study real-world systems has necessitated a parallel birth of validation techniques. In practice, the physical sciences and engineering disciplines have rarely used the term validation in their literature. While the validation procedures proposed by these disciplines are often quite similar to those conceived in the economic, social, and biological sciences, biologists have brought the concepts and use of the term, validation, to the popular usage most often encountered today. Apparently, the physical sciences are less attracted to the concept of validation because these models are based in more well-defined laws than either the biological or social sciences.

The validation tests appropriate to a particular model are dependent upon the nature of the model and the goals and objectives of the modeling study, regardless of discipline. Models can be thought of as deterministic or stochastic, as well as mechanistic or empirical in nature. These two dimensions can be visualized as a pair of colinear axes on a plane as shown in Fig. 1. Empirical implies the model outputs are based on observed, experimental data. Mechanical models presuppose that natural processes can be mechanistically defined and are capable of complete explanation by the laws of chemistry and physics. In the second dimension we define a stochastic model as one involving variables that may take on any one of a specified set of values with a specified probability at each point in time. A deterministic model yields one and only one set of unique outputs for each set of fixed input values. Thus a model such as the one represented by point A in Fig. 1 would tend to produce output of a deterministic rather than stochastic nature and the functions generating the outputs are based more on empirical data,

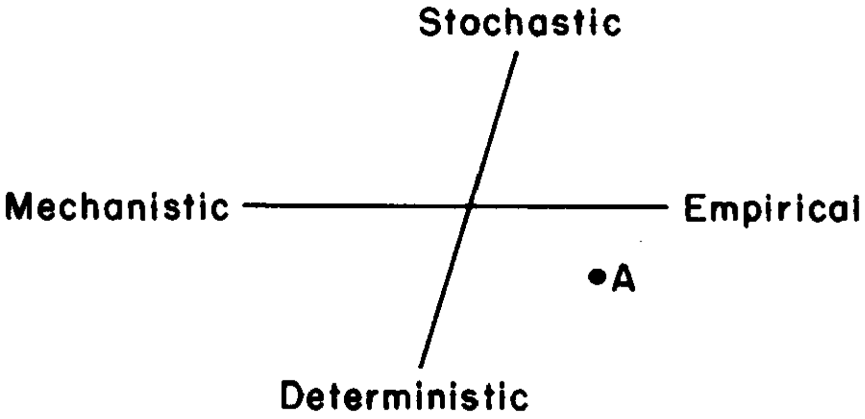

Fig. 1. Colinear modeling axes.

such as regression functions, rather than on functions based on known physical and chemical laws. Empirical models are more typical of the social, economic, and biological sciences while mechanistic models are more characteristic of the engineering and physical sciences.

Models can be developed for predictive or descriptive purposes. A predictive model must only produce accurate predictions of the output variables in the system. In this case the modeler is not concerned with exact replication in the model of the interactions between the variables in the system. The relationships employed in the model to generate the predicted output need not conform to the mechanisms in the real system which lead to the same outputs. In contrast, descriptive models must not only generate predictions in agreement with real system output, but the intermediate relationships employed in the model must also be realistic representations of the real processes which generate the observed outputs.

Validation of a predictive model would be concerned only with outputs while a descriptive model's validation would also be concerned with the accuracy with which the subsystems within the overall system were modeled. Validation of a predictive model would only attempt to determine that the model outputs and the observed outputs agree. On the other hand, validation of a descriptive model would require that the submodels of the subprocesses be validated prior to any validation of the overall model of the system. The validation of descriptive models is obviously a more involved undertaking than the validation of predictive models.

Most simulation models generate a functional trace of any of a number of variables exhibited over time as illustrated in Fig. 2. The observed data to be used in the validation process ordinarily consist of replicated

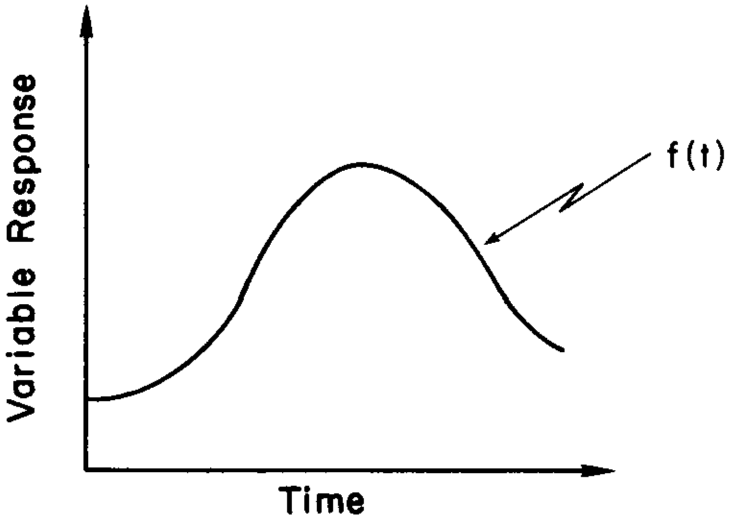

Fig. 2. Model generated trace.

observations at a number of points in time (see Fig. 3). Validation then becomes a process of checking the output of the model against empirical observations from field or laboratory measurements. Answering the question of how well the model output and the data agree suggests the need for generalized lack of fit tests. The aim of this paper is to suggest possible statistical solutions to the validation problem in both the univariate and multivariate situations.

Within this framework section 2.0 will outline the philosophy of the validation process and enumerate some of the validation techniques for simulation models as suggested in the literature. Section 3.0 will conceptualize the validation problem and present a series of possible solutions. Sections 4.0 and 5.0 will delve into the properties and behavior of these solutions using actual data sets and suggest conclusions and topics for further study.

2.0 LITERATURE REVIEW

There has been much effort devoted to developing the philosophy of modeling and the processes by which we attempt to verify or validate the truthfulness or correctness of computer simulation models through our knowledge and observation of the "real world." Many authors use the words validation and verification synonymously, yet some authors, notably Nolan (1972) and Wright (1971), make a distinction.

Verification concerns itself with the establishment of the correctness of a model. The verification process includes: (i) tests of the correctness of the computer coding used in the model, (ii) tests to determine the accuracy or correctness of the assumptions and hypotheses upon which the

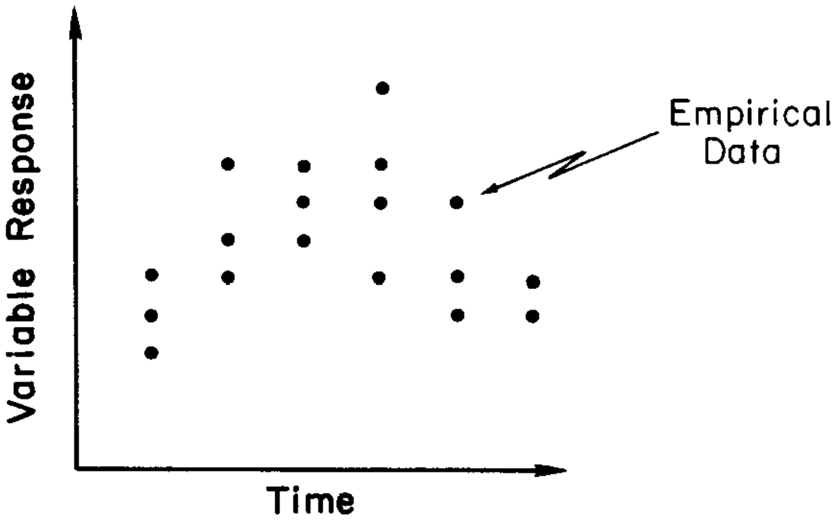

Fig. 3. Data employed in validation.

model is based, and (iii) tests of the agreement of observed outputs and model predictions when the model is run, using as inputs the data employed in the construction of the model. Thus the data used in the verification process is the same as that used in the development of the model. In a verified model the mechanisms and functions included in the model correctly simulate and realistically represent the assumptions upon which they are based. Tests (i) and (ii) can be used to verify purely mechanistic models designed without the aid of empirical data. A possible verification test of a purely empirical model would be to run the model using the observed empirical inputs employed in building the model and then determine that the model predictions generated agree with the observed outputs as in test (iii). If a model was both mechanistic and empirical in nature, all three tests could be applied.

On the other hand, validation is primarily concerned with determining the usefulness of a model as evidenced in the accuracy of its predictions. If a model of some system is an accurate representation of reality, then it should be able to predict future outputs under different sets of initial conditions. In determining the accuracy of model predictions, the best evidence or primary source of information available for validation is empirical data. This suggests that the usefulness of validation results is closely linked with the quality of the data employed in the validation process. Steinhorst (1973) suggests three minimal conditions that the data must meet.

(i) The validation data must not have been used in model development.

(ii) The data must be of sufficient precision to make the test meaningful.

(iii) The objectives of the modeling exercises must be kept in perspective.

Condition (i) implies that the data used in the validation process are independent of the data employed in the verification process. Condition (iii) implies that not all data are appropriate for use in validation. This occurs because data collection inherently requires some preconceptions of the phenomena being measured and these concepts may differ from those being modeled.

In elucidating this last point we note that since a data collection scheme presupposes some model of the system, from a pragmatic viewpoint, validation is tantamount to comparing the output of two models. In the case of data collection this implicit model is usually well accepted. At the same time other models of the same system may already exist. If these models have been validated with respect to empirical data, they can also be considered as another representation of reality, though of a less viable nature than empirical data. Therefore, a secondary source of information applicable to the validation process is the output of other validated models.

The fact that model output and data do indeed agree does not imply that the model is unequivocally correct as the data could be in error itself. This tenet can be supported by the fact that the experimental design used to collect the data assumes a model of the real world, and this model suffers from errors of omission and commission and the problems of interpretation common to the modeling realm. In statistical theory this error would be construed as a Type II error in hypothesis testing.

Various validation techniques for comparing model output and data have been suggested in the literature. These are applicable to both predictive or descriptive models and to models based on empirical or mechanistic functions. The critical dimension affecting the applicability of various techniques is that of the deterministic or stochastic nature of the output.

If the output of a deterministic model is to be validated against corresponding observed data, a number of techniques are available that are summarized in a paper by Wright (1972). These include point fit tests using regression analyses (Cohen and Cyert 1961), factor analysis (Cohen and Cyert 1961), and Theil's (1961) inequality coefficient and nonparametric distributional tests such as the chi-square goodness-of-fit test or the Kolmogorov-Smirnov test (Naylor and Finger 1967).

The point fit test suggested by Theil (1961) is a quantity designed to measure the agreement of prediction and observation called the inequality coefficient, U. If we let $P_1, \ldots, P_n$ be the n predictions and $A_1, \ldots, A_n$ be the actual observed outcomes, then

$$U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2} \Bigg/ \left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}P_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}A_i^2}\right],$$

where $0 \leq U \leq 1$. If $P_i = A_i$, for every $i = 1, \ldots, n$, we have complete agreement of prediction and observation and $U = 0$. Theil noted a rather serious drawback in the interpretability of this measure in that U is not invariant against location change. This indicates that the value of U calculated for P_i and A_i, $i = 1, \ldots, n$ would not be the same as that for the points P_i^* and A_i^* where

$$P_i^* = P_i + c$$

and

$$A_i^* = A_i + c \qquad \text{for } i = 1, \ldots, n$$

where c is a constant.

In a vein similar to Theil's inequality coefficient, Kapoor (1968) has suggested a similar measure of quality of a simulation model, the V inequality coefficient. Suppose we wish to validate a variable trace against a data record where the record consists of n_i replicates taken at each of $i = 1, \ldots, t$ times. Let a_i' be the maximum response and a_i'' be the minimum response of the n_i replicates at each of the t times. At each of the t times the model generates a prediction for the variable in question, P_i, $i = 1, \ldots, t$. For each of the $i = 1, \ldots, t$ times we define

$$e_i = \begin{cases} P_i - a_i' & \text{if } P_i > a_i' \\ 0 & \text{if } a_i'' \leq P_i \leq a_i' \\ a_i'' - P_i & \text{if } P_i < a_i'' . \end{cases}$$

Kapoor defines the V inequality coefficient as

$$V = \sqrt{\frac{1}{t} \sum_{i=1}^{t} e_i^2} \Bigg/ \sqrt{\frac{1}{t} \sum_{i=1}^{t} (a_i' - a_i'')^2} .$$

Dividing by the quantity in the denominator makes the measure scale invariant. One might note that V is always greater than zero and V is identically equal to zero if $a_i'' \leq P_i \leq a_i'$ for every $i = 1, \ldots, t$, that is, if the predictions at each of the t times falls between the maximum and minimum observed values for each time.

Cohen and Cyert (1961) suggest two other possible tests. In the first we assume that we observe n pairs of observations (P_i, A_i), $i = 1, \ldots, n$, where the P_i is the predicted value and the A_i is the observed value. We then estimate the parameters β_0 and β_1 in the equation $P = \beta_0 + \beta_1 A$ using conventional regression techniques and test the hypothesis

$$H_0: \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} .$$

If one does not reject the null hypothesis, then we may conclude that $A_i = P_i$, that is, the model predictions and data agree. The second method suggested is to compute the factor loadings for the variables' model outputs and the loadings for the observed data for the same variables. If the loading matrices for the model output and the data agree, this would be evidence of the viability of the simulation model.

Naylor and Finger (1967) suggest the following technique. A model through its predictions generates a trace, f(t), over time for some variate, say X, as in Fig. 4. It is possible to project this curve in such a manner that it generates a cumulative distribution function for the random variable X, say $F_X(x)$. Referring to Fig. 4, the probability, p, of observing a value of the random variable X less than x can be found as follows:

$$p = P[X \leq x] = \frac{\text{Area A}}{\int_0^t f(t)\, dt} . \quad \text{(shaded area in Fig. 4)}$$

Note that

$$P[X \leq 0] = 0$$

and

$$P[X \leq x_0] = 1 .$$

Therefore, if we find $P[X \leq x]$ for every x such that $0 \leq x \leq x_0$, the resulting function is the cumulative distribution function, $F_X(x)$, as displayed in Fig. 5 for the random variable X. Application of the Kolmogorov-Smirnov

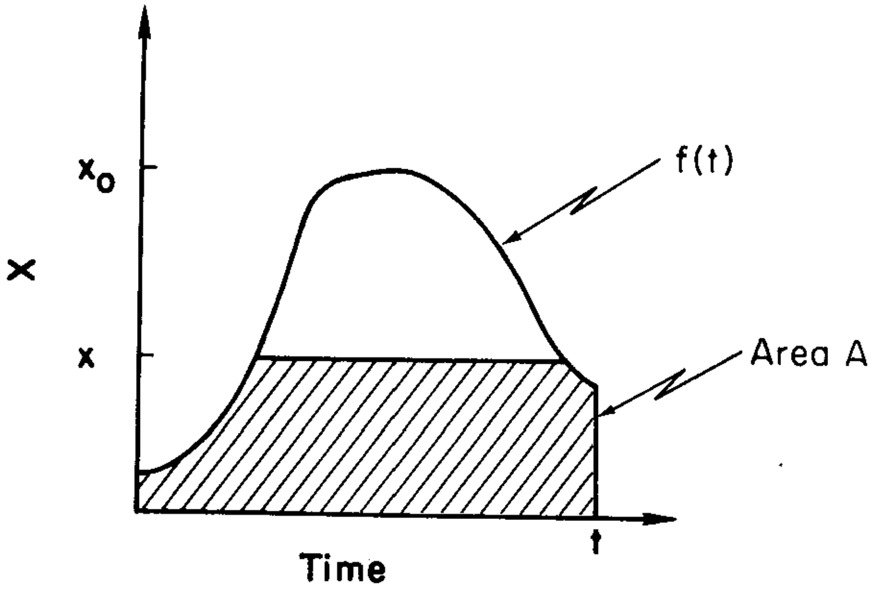

Fig. 4. Model output.

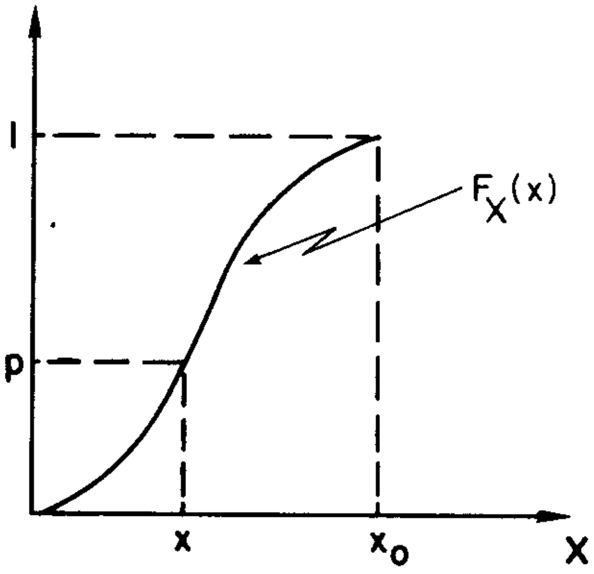

Fig. 5. Cumulative distribution function generated from model output in Fig. 4.

test would determine if the sample cumulative distribution function of the observed data is the same as the cumulative distribution function, $F_X(x)$, generated by the model output. Since the probability density function, $f_X(x)$, is directly related to the cumulative density function, $F_X(x)$, by

$$f_X(x) = \frac{d\ F_X(x)}{dx} \quad ,$$

we could employ a chi-square goodness-of-fit test to determine if the observed data came from the density, $f_X(x)$. Wright (1972) has noted that this method has a shortcoming in the fact that this approach destroys the time series nature of the data.

Another approach to the validation problem, applicable to models generating either stochastic or deterministic outputs employing time series analysis, has been suggested by Fishman and Kiviat (1967). If we observe the behavior of a process over an arbitrary but closed time interval, this constitutes a time series. If in addition the time series is covariance stationary, a method of comparing the time series generated by the model over some time interval with the time series associated with the observed data over the same time span is available through examining their respective spectral densities.

We define the random variable modeled, $X(t)$, to be time dependent, where

$$E(X(t)) = \mu_t$$

and

$$\text{Var}\ (X(t)) = \sigma^2 .$$

The covariance between the two random variables, $X(t)$ and $X(t+T)$, when $T \geq 0$ and $-\infty < t < \infty$, is defined as

$$\mathrm{Cov}(X(t),\ X(t+T)) = \sigma^2\ \rho(T)\ .$$

Note that $\rho(T)$ is a function of T alone and not t and T both. This implies that the covariance between X(t) and X(t+T) is not dependent on their relative location on the time scale but only on the distance between the two points. If this assumption is indeed true, then the time series associated with the random variable, X(t), is covariance stationary and $\rho(T)$ is called the autocorrelation function. The spectral density, $f(\lambda)$, is directly related to the autocorrelation function in the following manner,

$$\rho(T) = \int_0^\infty \cos \lambda r\ f(\lambda)d\lambda\ .$$

Within this structure it is possible to estimate the spectral density in an unbiased fashion if the times are equally spaced. In data collection the times are not always evenly spaced, but Kendall and Stuart (1966) point out that relatively small deviations create a relatively minor bias.

Comparison of the two spectral densities associated with the model output and the observed data can give some insight into the validity of the model. If prior knowledge indicates that the data and predictions do agree, then agreement of the estimated spectral densities lends further credence to model validity. On the other hand, similarity of the spectral densities does not necessarily imply that the predicted and observed time records will be in agreement. This last fact can detract from the overall usefulness of this technique.

When the empirical evidence is of such a nature that objective validation techniques are not applicable, one can resort to subjective validation procedures, such as graphical approaches. Naylor and Finger (1967) have suggested some criteria which the experimenter can apply in passing judgment on the validity of a model. The modeler can examine:

(1) the number of turning points,

(2) the timing of turning points,

(3) the direction of turning points,

(4) the amplitude of the fluctuations for corresponding time segments,

(5) the average amplitude over the whole series,

(6) the simultaneity of turning points for different variables,

(7) the average values of the variables, and

(8) the exact matching of values of variables.

Discrepancies in these measures between the model output and the data represent possible errors in the model.

3.0 STATISTICAL TECHNIQUES

3.1 Conceptualization of the Problem

A simulation model of an ongoing system generates a trace over time for each of the p variates in the model,

$$f_i(t); \; i = 1, \ldots, p$$

where the functions $f_i(\cdot)$ are usually continuous in the time domain. We desire to validate this generated function with respect to empirical evidence gathered on each of the p variates at each of t times in replicate form. We have the following observed data:

$$X_{ijk}; \quad \begin{array}{l} i = 1, \ldots, p \\ j = 1, \ldots, t \\ k = 1, \ldots, n_j \, . \end{array}$$

These data should be independent of the process of designing the model and also independent of any of the data employed in the verification process. Ideally we would like

$$X_{ijk} - f_i(j) = 0$$

for all $i = 1, \ldots, p$, $j = 1, \ldots, t$, and $k = 1, \ldots, n_j$, as this would imply that prediction and observation are identical. Obviously, this equality will rarely be true which implies the need for a statistical test to see if the difference is significant.

We next define

$$Y_{ijk} = X_{ijk} - f_i(j)$$

where Y_{ijk} is one of $k = 1, \ldots, n_j$ observations from the random variable Y_{ij}, $i = 1, \ldots, p$ and $j = 1, \ldots, t$. Let

$$\underline{Y}_j' = [Y_{1j}, Y_{2j}, \ldots, Y_{pj}];\ j = 1, \ldots, t$$

where $\underline{Y}_j$ has some unknown continuous p variate cumulative distribution function, $F_j(\underline{y})$. For statistical agreement between the data and model output we need to verify

$$H_0:\ F_1(\underline{y}) = F_2(\underline{y}) = \ldots = F_t(\underline{y}) = F_0(\underline{y}) \text{ for all } \underline{y}$$

where $F_0(\underline{y})$ is again some unknown continuous p variate cumulative distribution function with zero mean (more generally with zero location parameter). If we are interested in testing translation-type alternatives, we let

$$F_j(\underline{y}) = F_0(\underline{y} - \underline{\gamma}_j),\ \text{for } j = 1, \ldots, t\ .$$

Within this framework verifying the hypothesis given above is equivalent to testing the null hypothesis

$$H_0: \ \underline{\gamma}_j = \underline{0}, \text{ for } j = 1, \ldots, t \qquad (3.1.1)$$

versus the alternative hypothesis

H_1: at least one equality in H_0 is an inequality.

Often the model analyst is content to verify initially that the "shape" of the data and model output agree. This would occur if the modeler were primarily interested in the validation of model dynamics as contrasted to the exact fit of model output and data. In this case the fact that the absolute magnitude of the response is in error is not considered critical as this type of discrepancy can often be remedied simply by adding or subtracting the appropriate constant from the output for each variable. In this case the following hypothesis would be applicable:

$$H_0: \ \underline{\gamma}_1 = \underline{\gamma}_2 = \ldots = \underline{\gamma}_t \qquad (3.1.2)$$

versus the alternative

H_1: at least one equality in H_0 is an inequality.

In the following sections we present a number of different solutions to the problem of statistically testing hypotheses (3.1.1) and (3.1.2) under different sets of assumptions. There are two assumptions that are made throughout so they will be stated here. First implicit in the hypothesis

$$F_j(\underline{y}) = F_0(\underline{y} - \underline{\gamma}_j); \ j = 1, \ldots, t$$

is the fact that the scale parameters associated with the random variables, Y_{ij}, are the same, for $j = 1, \ldots, t$ within each $i = 1, \ldots, p$, that is, the scale parameter for Y_{ij} is, say σ_i^2, for $i = 1, \ldots, p$ and $j = 1, \ldots, t$. Secondly, we assume that the random variables are independent over time, that is,

$$\mathrm{Cov}\ (Y_{ij}, Y_{ij'}) = 0, \text{ for } j \neq j' = 1, \ldots, t \quad \text{and } i = 1, \ldots, p\ . \qquad (3.1.3)$$

This last assumption is not unreasonable if the t times are spaced sufficiently apart. The distance necessary to accomplish this depends, of course, on the nature of the variable and response. If the observations are equally spaced along the time dimension, Phillips (1971) has summarized a number of statistical techniques to check the validity of this assumption including the Anderson circular autocorrelation coefficient, the von Newmann ratio, "runs" tests, and a 2 × 2 contingency table independence test.

One should note that equation (3.1.3) does not make any assumptions about the relationship between the p responses at the j^{th} time. All of the techniques presented in this chapter incorporate the fact that the p responses may be correlated into their analyses. Therefore, the techniques are applicable regardless of the correlation structure of the p responses at each time.

3.2 MANOVA Methods

The first technique proposed employs a standard multivariate analysis of variance approach (MANOVA). Let

$$\underline{Y}_j' = [Y_{1j}, Y_{2j}, \ldots, Y_{pj}];\ j = 1, \ldots, t$$

be a p-variate response of some phenomena. $\underline{Y}_j$ is a p × 1 vector distributed as a multivariate normal with mean, $\underline{\gamma}_j$, and covariance matrix, $\underline{\Sigma}$. We have

$$\underline{Y}_j \sim MVN_p(\underline{y}_j : \underline{\gamma}_j, \underline{\Sigma})$$

where

$$\underline{\gamma}_j' = [\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{pj}]$$

and $\underline{\Sigma}$ is a p × p matrix with elements (σ_{mn}). Letting

$$\underline{Y}_i^{*'} = [Y_{i1}, Y_{i2}, \ldots, Y_{it}]; \; i = 1, \ldots, p \;\; ,$$

we define

$$\underline{Y}_i^{*} \sim MVN_t(\underline{y}_i^{*} : \underline{\gamma}_i, \sigma_i^2 \, \underline{I}_t)$$

where

$$\underline{\gamma}_i' = [\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{it}] \;\; ,$$

$\underline{I}_t$ is a t × t identity matrix, and $\sigma_i^2 = \sigma_{mn}$ with $m = n = i$.

If we replicate the observation of each random variable, Y_{ij}, n_j times, for $j = 1, \ldots, t$, then we can write

$$\underline{Y}_i' = [Y_{i11}, Y_{i12}, \ldots, Y_{i1n_1}, Y_{i21}, Y_{i22}, \ldots, Y_{i2n_2}, \ldots, Y_{it1}, Y_{it2}, \ldots, Y_{itn_t}], \text{ for } i = 1, \ldots, p \qquad (3.2.1)$$

and

$$\underline{Y}_i = \underline{X}\,\underline{\gamma}_i + \underline{e}_i \;\; = \begin{bmatrix} \underline{1}_{n_1} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{1}_{n_2} & \cdots & \underline{0} \\ \vdots & \vdots & & \vdots \\ \underline{0} & \underline{0} & \cdots & \underline{1}_{n_t} \end{bmatrix} \begin{bmatrix} \gamma_{i1} \\ \gamma_{i2} \\ \vdots \\ \gamma_{it} \end{bmatrix} + \underline{e}_i \, .$$

Letting $N = \sum_{j=1}^{t} n_j$, $\underline{X}$ is $N \times t$ and $\underline{1}_{n_j}$ is a $n_j \times 1$ vector of ones. Finally let

$$\underline{Y} = [\underline{Y}_1, \underline{Y}_2, \ldots, \underline{Y}_p]$$

where $\underline{Y}$ is $N \times p$.

If we wish to test hypothesis (3.1.1)

$$H_0: \ \underline{\gamma}_j = \underline{0}; \ j = 1, \ldots, t \ ,$$

then using multivariate least squares regression theory, we calculate the sum of squares under H_0

$$\underline{W} = \underline{Y}'(\underline{X} \ (\underline{X}'\underline{X})^{-1} \ \underline{X}') \ \underline{Y} \quad , \tag{3.2.2}$$

the error sums of squares

$$\underline{E} = \underline{Y}'(\underline{I} - \underline{X} \ (\underline{X}'\underline{X})^{-1} \ \underline{X}') \ \underline{Y} \tag{3.2.3}$$

and the total sums of squares

$$\underline{E} + \underline{W} = \underline{Y}'\underline{Y}$$

where $\underline{E}$ and $\underline{W}$ are $p \times p$ matrices.

An approximate F - statistic is given by Anderson (1958) as

$$F = \frac{1 - y}{y} \cdot \frac{h}{pt}$$

where

$$y = \left(\frac{|\underline{E}|}{|\underline{E} + \underline{W}|}\right)^{1/g},$$

$$g = \begin{cases} \left[\dfrac{(t-1)^2\, p^2 - 4}{(t-1)^2 + p^2 - 5}\right]^{\frac{1}{2}} & \text{if } (t-1)^2 + p^2 \neq 5 \\ 1 & \text{if } (t-1)^2 + p^2 = 5 \end{cases}$$

and

$$h = (N - (t-1) - \frac{p - t + 1}{2})\; g - \frac{pt}{2} + 1 \; .$$

Under H_0 the test statistic, F, is approximately distributed as an F random variable with pt and h degrees of freedom and the level of significance of the test is approximately

$$P \doteq \text{Prob}[F \geq F_{pt,h}] \; .$$

This is an exact test if p or t is 1 or 2.

If, on the other hand, we are interested in testing hypothesis (3.1.2), then the approach is altered slightly. We have

$$H_0\colon\; \underline{\gamma}_1 = \underline{\gamma}_2 = \ldots = \underline{\gamma}_t \; .$$

γ_{ij} can be thought of as the sum of two components, the first due solely to variate i, say μ_i, and the second part due to the time j and variable i, say α_{ij}. Therefore, hypothesis (3.1.2) is equivalent to

$$H_0\colon\; \mu_i + \alpha_{i1} = \mu_i + \alpha_{i2} = \ldots = \mu_i + \alpha_{it};$$
$$i = 1, \ldots, p \tag{3.2.4}$$

or

$$H_0: \ \alpha_{i1} = \alpha_{i2} = \ldots = \alpha_{it}; \ i = 1, \ldots, p \ .$$

This hypothesis is reparameterized to a full rank model as follows:

$$Y_{ijk} = \mu_i + \alpha_{ij} + e_{ijk}$$

$$= (\mu_i + \bar{\alpha}_{i.}) + (\alpha_{ij} - \bar{\alpha}_{i.}) + e_{ijk}$$

where

$$\bar{\alpha}_{i\circ} = \sum_{j=1}^{t} \frac{\alpha_{ij}}{t} \ .$$

Letting

$$\mu_i^* = \mu_i + \bar{\alpha}_{i.}$$

and

$$\alpha_{ij}^* = \alpha_{ij} - \bar{\alpha}_{i.} \ ,$$

we have

$$Y_{ijk} = \mu_i^* + \alpha_{ij}^* + e_{ijk} \ . \qquad (3.2.5)$$

Within this framework the hypothesis (3.2.4) is equivalent to

$$H_0: \ \alpha_{i1} - \bar{\alpha}_{i.} = \alpha_{i2} - \bar{\alpha}_{i.} = \ldots = \alpha_{it} - \bar{\alpha}_{i.}; \ i = 1, \ldots, p$$

or, using the notation introduced in equation (3.2.5),

$$H_0: \ \alpha_{i1}^* = \alpha_{i2}^* = \ldots = \alpha_{it}^* = 0; \ i = 1, \ldots, p$$

as $\alpha_{ij} - \bar{\alpha}_{i.} = 0$, for every $i = 1, \ldots, p$ and $j = 1, \ldots, t$ if H_0 is true.

If we define as before

$$\underline{Y}_i^{*\prime} = [Y_{i1}, Y_{i2}, \ldots, Y_{it}] ,$$

we have

$$\underline{Y}_i^* = \underline{X}^* \underline{\beta}_i^* + \underline{e}_i^*$$

$$= \begin{bmatrix} \underline{1}_{t-1} & \underline{I}_{t-1} \\ 1 & -\underline{1}'_{t-1} \end{bmatrix} \begin{bmatrix} \mu_i^* \\ \alpha_{i1}^* \\ \alpha_{i2}^* \\ \vdots \\ \alpha_{i,t-1}^* \end{bmatrix} + \underline{e}_i^* ,$$

as

$$\sum_{j=1}^{t-1} \alpha_{ij}^* = \alpha_{it}^* ,$$

or

$$\underline{Y}_i = \underline{X}\, \underline{\beta}_i^* + \underline{e}_i$$

$$\begin{bmatrix} \underline{1}_{n_1} & \underline{1}_{n_1} & \underline{0} & \cdots & \underline{0} \\ \underline{1}_{n_2} & \underline{0} & \underline{1}_{n_2} & \cdots & \underline{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \underline{1}_{n_{t-1}} & \underline{0} & \underline{0} & \cdots & \underline{1}_{n_{t-1}} \\ \underline{1}_{n_t} & -\underline{1}_{n_t} & -\underline{1}_{n_t} & \cdots & -\underline{1}_{n_t} \end{bmatrix} \begin{bmatrix} \mu_i^* \\ \alpha_{i1}^* \\ \alpha_{i2}^* \\ \vdots \\ \alpha_{i,t-1}^* \end{bmatrix} + \underline{e}_i$$

where $\underline{Y}_i$ is defined as in equation (3.2.1) for $i = 1, \ldots, p$. If we premultiply both sides by $(\underline{I}_N - \frac{1}{N} \underline{J}_N)$ where $\underline{J}_N$ is an $N \times N$ matrix of ones, then

$$(\underline{I}_N - \frac{1}{N}\underline{J}_N)\ \underline{Y}_i = (\underline{I}_N - \frac{1}{N}\underline{J}_N)\underline{X}\ \underline{\beta}_i^* + (\underline{I}_N - \frac{1}{N}\underline{J}_N)\ \underline{e}_i\ .$$

Let

$$\underline{A}_i = (\underline{I}_N - \frac{1}{N}\underline{J}_N)\ \underline{Y}_i^*$$

$$\underline{z}_i = (\underline{I}_N - \frac{1}{N}\underline{J}_N)\ \underline{e}_i^*$$

and

$$[\underline{0}\ \underline{B}] = (\underline{I}_N - \frac{1}{N}\underline{J}_N)\ \underline{X}$$

where $\underline{B}$ is N × t-1. Therefore,

$$\underline{A}_i = [\underline{0}\ \underline{B}]\begin{bmatrix}\mu^* \\ \underline{\alpha}_i^*\end{bmatrix} + \underline{z}_i$$

$$= \underline{B}\ \underline{\alpha}_i^* + \underline{z}_i$$

where

$$\underline{\alpha}^*{}_i' = [\alpha_{i1}^*, \alpha_{i2}^*, \ldots, \alpha_{i,t-1}^*]\ .$$

As before we let

$$\underline{A} = [\underline{A}_1, \underline{A}_2, \ldots, \underline{A}_p] \tag{3.2.6}$$

where $\underline{A}$ is N × p and employ multivariate least square regression theory to calculate the sum of squares under the null hypothesis

$$\underline{E} = \underline{A}'\ (\underline{B}\ (\underline{B}'\underline{B})^{-1}\ \underline{B}')\ \underline{A}\ \ , \tag{3.2.7}$$

the error sum of squares

$$\underline{W} = \underline{A}'\ (\underline{I}_N - \underline{B}\ (\underline{B}'\underline{B})^{-1}\ \underline{B}')\ \underline{A} \tag{3.2.8}$$

and the total sum of squares

$$\underline{E} + \underline{W} = \underline{A}'\underline{A}\ .$$

Again we can employ the approximate F statistic in Anderson (1958) in order to determine the significance under H_0 of the observed sums of squares. We calculate

$$F = \frac{1-y}{y} \cdot \frac{h}{p(t-1)}$$

where

$$y = \left[\frac{|\underline{E}|}{|\underline{E} + \underline{W}|} \right]^{1/g} ,$$

$$g = \begin{cases} \left[\frac{(t-1)^2 \; p^2 - 4}{(t-1)^2 + p^2 - 5} \right]^{\frac{1}{2}} & \text{if } (t-1)^2 + p^2 \neq 5 \\ 1 & \text{if } (t-1)^2 + p^2 = 5 \end{cases}$$

and

$$h = (N - (t-1) - \frac{p - t + 1}{2}) \; g - \frac{pt}{2} + 1 \; .$$

Under H_0 the test statistic, F, is approximately distributed as an F variate with p(t-1) and h degrees of freedom. The approximate level of significance, P, can be found by determining

$$\text{Prob } [F \geq F_{p(t-1),h}] \; .$$

Again this is an exact test if p = 1 or 2 or if t = 2 or 3.

If the model analyst is interested in validating one variate rather than p variates, the analysis reduces to a one-way analysis of variance. In testing hypothesis (3.1.1), we would have

$$H_0: \; \gamma_j = 0 \text{ for } j = 1, \ldots, t$$

versus

$$H_1: \; \gamma_j \neq 0 \text{ for at least one } j = 1, \ldots, t \; .$$

The observed data would be of the form

$$Y_{jk} = \gamma_j + e_{jk}; \; j = 1, \ldots, t$$
$$k = 1, \ldots, n_j$$

where

$$Y_{jk} \sim \text{Normal } (y_{jk}: \; \gamma_j, \sigma^2)$$

and

$$\text{Cov } (Y_{jk}, Y_{j'k'}) = 0 \text{ for } j \neq j' = 1, \ldots, t$$
$$\text{and } k \neq k' = 1, \ldots, n_j \; .$$

In this simplified situation the total sums of squares is

$$E + W = \sum_{j=1}^{t} \sum_{k=1}^{n_j} Y_{jk}^2$$

where E and W are now 1 × 1 matrices. The sums of squares under the null hypothesis is

$$W = \sum_{j=1}^{t} \left(\sum_{k=1}^{n_j} Y_{jk} \right)^2$$

and the error sums of squares is

$$E = \sum_{j=1}^{t} \left(\sum_{k=1}^{n_j} Y_{jk}^2 - \frac{\left(\sum_{k=1}^{n_j} Y_{jk} \right)^2}{n_j} \right) .$$

Under normal theory the quantity,

$$F = \frac{(N-t)W}{tE} ,$$

is distributed as a Snedecor's F random variable with t and N-t degrees of freedom. Therefore, the level of significance of the test of H_0 is

$$P = \text{Prob }[F \geq F_{t,N-t}] \; .$$

On the other hand, if we are interested in testing hypothesis (3.1.2), then we have

$$H_0\text{:} \quad \gamma_1 = \gamma_2 = \ldots = \gamma_t$$

or as defined above

$$H_0\text{:} \quad \mu + \alpha_1 = \mu + \alpha_2 = \ldots = \mu + \alpha_t$$

which is equivalent to testing

$$H_0\text{:} \quad \alpha_1 = \alpha_2 = \ldots = \alpha_t$$

versus

H_1: at least one equality in H_0 is an inequality.

In this case, the total sum of squares is

$$E + W = \sum_{j=1}^{t} \sum_{k=1}^{n_j} Y_{jk}^2 - \frac{\left(\sum_{j=1}^{t} \sum_{k=1}^{n_j} Y_{jk} \right)^2}{N} ,$$

the sums of squares under the null hypothesis is

$$W = \sum_{j=1}^{t} \frac{\left(\sum_{k=1}^{n_j} Y_{jk} \right)^2}{n_j} - \frac{\left(\sum_{j=1}^{t} \sum_{k=1}^{n_j} Y_{jk} \right)^2}{N} \qquad (3.2.9)$$

and the error sums of squares is

$$E = \sum_{j=1}^{t} \left(\sum_{k=1}^{n_j} Y_{jk}^2 - \frac{\left(\sum_{k=1}^{n_j} Y_{jk} \right)^2}{n_j} \right). \tag{3.2.10}$$

Again under normal theory the statistic,

$$F = \frac{(N-t)}{(t-1)} \frac{W}{E} \tag{3.2.11}$$

is distributed as a Snedecor's F random variable with t-1 and N-t degrees of freedom. The level of significance of this test is

$$P = \text{Prob}\left[F \geq F_{t-1,N-t}\right] .$$

3.3 Permutation Methods

3.3.1 *Permutation Test.* In spite of the fact that much evidence has been gathered to lend credence to the robustness of analysis of variance techniques under the normality assumptions, there are times when the assumptions are not admissible. To circumvent this problem, two areas have been developed, nonparametric tests and permutation tests. The latter will be the subject of this section.

Conceptually, permutation tests can be explained intuitively, for example, by the one-way analysis of variance. Say we have n_j replicate observations in each of t times or treatments,

$$Y_{jk} = \gamma_j + e_{jk}; \; j = 1, \ldots, t$$
$$k = 1, \ldots, n_j$$

and we wish to test

$$H_0: \ \gamma_1 = \gamma_2 = \dots = \gamma_t \qquad (3.3.1)$$

versus

H_1: at least one equality in H_0 is an inequality.

One can calculate the F statistic given in equation (3.2.11). Under normal theory this statistic has an F distribution, but without the normality assumptions the most we can say is that the null hypothesis tends to be false if the F statistic is "large" and true if the F statistic is "small."

If we assume Y_{jk} has some unknown distribution with location parameter, γ_j, and variance, σ^2, and

$$\text{Cov}\ (Y_{jk}, Y_{j'k'}) = 0 \text{ for } j \neq j' = 1, \dots, t$$
$$\text{and } k \neq k' = 1, \dots, n_j \ ,$$

then we can apply permutation theory results. If the null hypothesis is true, then all the γ_j are equal which implies

$$Y_{jk} = \gamma_1 + e_{jk};\ j = 1, \dots, t$$
$$k = 1, \dots, n_j \ .$$

In this case all the treatments or times would show equivalent yields so Y_{jk} could just as likely be an observation from treatment j as from some other treatment j' ≠ j = 1, ..., t. This implies that if the null hypothesis is true, we could just as readily randomly assign each of the N = $\sum_{j=1}^{t} n_j$ observations to any of the t treatments or times. We could calculate the F statistic in equation (3.2.11) for this random permutation of the observations, say F_1, and compare it to the original observed test statistic, say F_0.

If we repeated the permutation process, say M times, then we could compare each of these with the observed F-test statistic. If the number of F_i, $i = 1, \ldots, M$ less than F_0 is "large" this implies intuitively that F_0 was an extreme or unusual event, that is, not in accordance with the null hypothesis. On the other hand, if the number of F_i less than F_0 is "small," this lends credence to the truthfulness of the null hypothesis. This argument supports the tenet that the level of significance of the test of hypothesis (3.3.1) can be approximated by

$$P \doteq (\text{number of } F_i > F_0,\ i = 1, \ldots, M)/M \ .$$

One might note that the goodness of the approximation is directly related to the number of permutations selected. For example, if $M = 4$, the only possible levels of significance are 0.00, 0.25, 0.50, 0.75, and 1.00. The number of possible permutations is related to the sample size. For this test the number of permutations possible is N!, and the number of distinct values of the F statistic possible is

$$\binom{N!}{n_1!\; n_2!\; \ldots\; n_t!} = \frac{N!}{\prod_{j=1}^{t} n_j!} \ .$$

if the F statistic is calculated for all of the N! possible permutations of the observations or the $N!\Big/\prod_{j=1}^{t} n_j!$ combinations of the observations which lead to different distinct values of the F statistic, then the test described is exact. The level of significance is

$$P = (\text{number of } F_i > F_0,\ i = 1, \ldots, N!)/N!$$

if all possible permutations of the observations are found or

$$P = \left(\text{number of } F_i > F_0,\ i = 1, \ldots, N!\Big/\prod_{j=1}^{t} n_j!\right)\Bigg/\left(N!\Big/\prod_{j=1}^{t} n_j!\right)$$

if all possible combinations of the observations which give distinct values of the F statistic are enumerated. Often the number of permutations or combinations is prohibitively large, leading to the practice of choosing M of these at random for the approximation given in the preceding paragraph. It should be noted that nowhere in this discussion has an assumption been made about the distribution of the Y_{jk}.

We might note that the number of permutations, M, required to gain a suitably accurate estimate of P depends on the nature of the data involved. Consequently, from a computational viewpoint one might calculate

$$P_j = (\text{number of } F_i > F_0,\ i=1, \ldots, j)/j;\ j = 1, \ldots, M\ .$$

When the series (P_j) has stabilized sufficiently for the accuracy desired by the analyst, the process can be terminated. This technique is invaluable if the permutations are being generated by a computer as it necessitates only fixing an upper bound on the number of permutations. Often the number of permutations required to meet a ± 0.01 tolerance on the true level of significance is surprisingly small; for example, a value of $M \leq 50$ for permutation tests of MANOVA problems is often sufficient.

Generalizing to the multivariate problem, we are interested in applying a permutation test to the hypothesis (3.1.2)

$$H_0\colon\ \underline{\gamma}_1 = \underline{\gamma}_2 = \cdots = \underline{\gamma}_t$$

versus

$$H_1\colon\ \text{at least one equality in } H_0 \text{ is an inequality.}$$

We can proceed as in section 3.2 and calculate $\underline{E}$ and $\underline{W}$ by equations (3.2.7) and (3.2.8) for the observed data set. We can next calculate any one or all of the following multivariate criterion: the Willis-Bartlett likelihood ratio test statistic

$$\lambda_0 = \frac{|\underline{E}|}{|\underline{E} + \underline{W}|} ,$$

the Hotelling test statistic

$$T_0^2 = \text{trace } (\underline{W}\ \underline{E}^{-1})$$

or the Lawley-Pillai test statistic

$$T_0^* = \text{trace } [\underline{W}(\underline{E} + \underline{W})^{-1}] \ .$$

We next permute the N elements of each of the vectors, $\underline{Y}_i$, $i = 1, \ldots, p$, as defined in equation (3.2.1). This shuffling succeeds in assigning the N responses of each variate to the t times in a random fashion. This permutation process is repeated M times, where after each permutation we recalculate each of the multivariate test criterion, λ_i, T_i^2, and T_i^* for $i = 1, \ldots, M$.

Intuitively if the null hypothesis is true, we expect the likelihood ratio to be "large," the Hotelling statistic to be "small," and the Lawley-Pillai statistic to be "small." Therefore, the level of significance for the likelihood ratio criterion is approximately

$$P \doteq (\text{number of } \lambda_i < \lambda_0,\ i = 1, \ldots, M)/M$$

for the Hotelling criterion

$$P \doteq (\text{number of } T_i^2 > T_0^2,\ i = 1, \ldots, M)/M ,$$

and for the Lawley-Pillai criterion

$$P \doteq (\text{number of } T_i^* > T_0^*,\ i = 1, \ldots, M)/M \ .$$

In the univariate case, the procedure is identical, but the test criteria simplify in the following manner:

$$\lambda = \frac{E}{E + W} \ ,$$

$$T^2 = \frac{W}{E}$$

and

$$T^* = \frac{W}{E + W} = 1 - \lambda$$

where E and W are defined in equations (3.2.9) and (3.2.10).

Unfortunately, a permutation test for the hypothesis (3.1.1)

$$H_0\text{:}\ \ \underline{\gamma}_j = \underline{0} \text{ for } j = 1, \ldots, t$$

versus

$$H_1\text{:}\ \ \text{at least one equality in } H_0 \text{ is an inequality}$$

does not exist without making additional assumptions about the distribution of the Y_{jk}. This occurs because under the null hypothesis $\gamma_{ij} = 0$ or, in general, any hypothesis of the form $\gamma_{ij} = c_i$, where c_i is a constant, the permutation test is insensitive to the value of c_i. If one is willing to make the additional assumption that the distribution of Y_{jk} is symmetric, a technique does exist for testing this hypothesis. It is not presented here since, if one has sufficient prior knowledge to assume symmetry, the distribution of the means at the different times will be approximately normal for small sample sizes. In this case, the simpler MANOVA analyses would be quite adequate as the normality assumptions would not in all probability be grossly violated.

3.3.2 *Approximation to the Permutation Test.* An approximation to the permutation distribution test as described in section 3.3.1 has been suggested by Urquhart (1965). Using this approximation, the significance level of the test can be found without actually performing the permutations. Suppose we wish to test the hypothesis (3.1.2)

$$H_0: \underline{\gamma}_1 = \underline{\gamma}_2 = \dots = \underline{\gamma}_t$$

versus

H_1: at least one equality in H_0 is an inequality.

If we performed a permutation test of this hypothesis using the Lawley-Pillai test criterion

$$T_0^* = \text{trace}\ [\underline{W}(\underline{E} + \underline{W})^{-1}] \ ,$$

as described in section 3.3.1, we could calculate T_0^* for the observed data and then calculate T_i^* for each of the i = 1, ..., M permutations generated.

Since T_i^*/p is bounded between zero and one, we could think of the M values

$$\frac{T_i^*}{p} \ ; \ i = 1, \dots, M$$

as being observations from a Beta distribution with parameters α and β. We could then see at what percentile of the Beta distribution the observed T_0^*/p fell and thus determine the approximate level of significance of the test. If T_0^*/p is "large," then

$$P \doteq \text{Prob}\ [T_0^*/p > \text{Beta}(\alpha, \beta)]$$

would be "small" and we would reject H_0.

In essence Urquhart derives the first two moments of the permutation distribution, sets them equal to the first two moments of a Beta distributed random variable, and then solves these two equations, both of which are functions of α and β, simultaneously to derive estimates of α and β. The two parameters are in themselves sufficient to completely describe a Beta distribution.

To estimate these two parameters, we proceed to make the following calculations. We find

$$g = \sum_{i=1}^{N} \text{(square of the diagonal elements of } \underline{B}(\underline{B}'\underline{B})^{-1}\underline{B}'\text{)}$$

and

$$h = \sum_{i=1}^{N} \text{(square of the diagonal elements of } \underline{A}(\underline{A}'\underline{A})^{-1}\underline{A}'\text{)}$$

where $\underline{A}$ and $\underline{B}$ are defined as in equation (3.2.6). One might note that $\underline{J}_N\underline{A} = \underline{0}$ and $\underline{J}_N\underline{B} = \underline{0}$ where $\underline{A}$ is N × p and $\underline{B}$ is N × (t-1). Thus,

$$U = g - \frac{(t-1)(t+1)(N+1)}{N(N+1)} ,$$

$$V = h - \frac{p(p+2)(N-1)}{N(N+1)} ,$$

$$C_x = \frac{(N+1)(N)(N-1)}{(t-1)(N-t)(N-3)} \cdot U ,$$

$$C_y = \frac{(N+1)(N)(N-1)}{p(N-p-1)(N-3)} \cdot V ,$$

$$D = \frac{N-3}{2N(N-1)} C_x C_y ,$$

$$\delta = \frac{(Np-2)(N-1) - 2D(N-P-1)}{2(N-1)(N-p-1)(1+D)} ,$$

$$\alpha = \delta(t-1)$$

$$\text{and } \beta = \delta(N-t) .$$

Therefore, the significance level of H_0 is approximately

$$P \doteq \text{Prob}\ [T_0^*/p > \text{Beta}(\alpha,\beta)]\ .$$

Since there exists a one-to-one functional relationship between the Beta and F-distributions and the F-distribution is more widely tabulated than the Beta, one might wish to transform the Beta statistic to an F statistic. In this case,

$$P \doteq \text{Prob}\ [F > F_{m,n}]$$

where $m = 2\delta(t-1)$, $n = 2\delta(N-t)$, and $F = \dfrac{n\ T_0^*}{m(p-T_0^*)}$. In the univariate case the calculations are analogous to those given above except $\underline{A}$ is now a N × 1 matrix and

$$T_0^* = \frac{W}{E + W}$$

where E and W are calculated by the formulas given in equations (3.2.9) and (3.2.10).

In section 3.3.1 it was noted that a permutation test did not exist for testing the hypothesis (3.1.1) without making additional assumptions about the distribution of the Y_{ij}. Consequently, no approximation exists for calculating the significance of the null hypothesis under hypothesis (3.1.1) and the broad assumptions given in section 3.3.1.

This approximation appears to be quite invaluable when the sample size is large as the computing time necessary to generate the M permutations may become prohibitive. This technique again avoids making any distributional assumptions, and in fact the second permutation moment contains a term which expresses the variance from normality in $\underline{A}$ and $\underline{B}$.

3.4 Nonparametric Ranking Method

Another test criterion employing a nonparametric test based on ranks is also applicable to the validation problem. Let the cumulative distribution function associated with

$$\underline{Y}_j' = [Y_{1j}, Y_{2j}, \ldots, Y_{pj}];\ j = 1, \ldots, t$$

be denoted by $F_j(\underline{y})$. We wish to verify

$$H_0:\ F_1(\underline{y}) = F_2(\underline{y}) = \ldots = F_t(\underline{Y}) = F(\underline{y}) \text{ for all } \underline{y} \qquad (3.4.1)$$

where $F(\underline{y})$ is some unknown continuous p variate cumulative distribution function. We wish to test hypothesis (3.4.1) versus a location parameter type alternative, that is, versus

$$H_1:\ F_j(\underline{y}) = F(\underline{y} - \underline{\delta}_j) \text{ for } j = 1, \ldots, t \text{ and some } \underline{\delta}_j \neq \underline{0}$$

or equivalently test

$$H_0:\ \underline{\delta}_1 = \underline{\delta}_2 = \ldots = \underline{\delta}_t = \underline{0}$$

versus

H_1: at least one equality in H_0 is an inequality.

This hypothesis test is equivalent to the test given in hypothesis (3.1.2). One should note that only a translation type alternative is considered here and not a scale type alternative. This implies that

$$\mathrm{Var}(Y_{ij}) = \sigma_i^2 \text{ for } i = 1, \ldots, p \text{ and } j = 1, \ldots, t$$

or that the variances over time within each of the multivariate responses are equal as noted in section 3.1.

Let $\underline{Y}_i$ be defined as in equation (3.2.1) and $\underline{Y}_i^{(R)} = \rho(\underline{Y}_i)$, for $i = 1, \ldots, p$. The rank function $\rho(\cdot)$ replaces the observed values, Y_{ijk}, with the observation's corresponding rank, that is, the smallest value is replaced by a one, the second smallest is replaced by a two, and so forth, until the largest value in the vector $\underline{Y}_i$ is replaced by an N. Let

$$\bar{y}_{ij}^{(R)} = \sum_{k=1}^{n_j} \frac{Y_{ijk}^{(R)}}{n_j} \text{ for } i = 1, \ldots, p \text{ and } j = 1, \ldots, t$$

and

$$\underline{\bar{Y}}_j^{(R)\prime} = [\bar{y}_{1j}^{(R)}, \bar{y}_{2j}^{(R)}, \ldots, \bar{y}_{pj}^{(R)}]; \; j = 1, \ldots, t .$$

Also

$$\bar{\bar{y}}^{(R)} = \sum_{j=1}^{t} \sum_{k=1}^{n_j} \frac{Y_{ijk}^{(R)}}{N} = \frac{1}{N} \sum_{i=1}^{N} i = \frac{1}{N}\left(\frac{N(N+1)}{2}\right) = \frac{N+1}{2} .$$

Note that $\bar{\bar{y}}^{(R)}$ is the same for $i = 1, \ldots, p$. Define

$$\underline{Y}^{(R)} = [\underline{Y}_1^{(R)}, \underline{Y}_2^{(R)}, \ldots, \underline{Y}_p^{(R)}]$$

and

$$\underline{V}^{(R)} = \frac{1}{N} \underline{Y}^{(R)\prime} [\underline{I}_N - \frac{1}{N} \underline{J}_N] \underline{Y}^{(R)} .$$

Lastly, we calculate the test statistic

$$L_p = \sum_{j=1}^{t} n_j (\underline{\bar{Y}}^{(R)} - \bar{\bar{y}}^{(R)} \underline{1}_p)' [\underline{V}^{(R)}]^{-1} (\underline{\bar{Y}}_j^{(R)} - \bar{\bar{y}}^{(R)} \underline{1}_p) .$$

Puri and Sen (1971) have shown that for n_j large L_p is asymptotically distributed as a chi-square random variable with p(t-1) degrees of freedom, that is,

$$L_p \sim \chi^2_{p(t-1)} \ .$$

The level of significance associated with this test of the hypothesis (3.4.1) is approximately equal to

$$P \doteq \text{Prob}\ [L_N \geq \chi^2_{p(t-1)}] \ .$$

When testing the univariate hypothesis

$$H_0\colon \ \delta_1 = \delta_2 = \ldots = \delta_t = 0$$

versus

$$H_1\colon \ \text{at least one equality in } H_0 \text{ is an inequality}$$

this test reduces to the well-known Kruskal-Wallis test, which is a nonparametric one-way analysis of variance using ranks. Since p = 1, we let $\underline{Y}_i = \underline{Y}$, where $\underline{Y}_i$ is defined as in equation (3.2.1). Similarly define

$$\underline{Y}^{(R)} = \rho(\underline{Y})$$

and

$$R_j = \sum_{k=1}^{n_j} Y^{(R)}_{jk}; \ j = 1, \ldots, t \ .$$

Therefore,

$$L_1 = \frac{12}{N(N+1)} \sum_{j=1}^{t} \frac{R_j^2}{n_j} - 3(N+1) \ .$$

For n_j large L_1 is asymptotically distributed as a χ^2 random variable with t - 1 degrees of freedom. The level of significance of the test in this case is

$$P \doteq P[L_1 \geq \chi^2_{(t-1)}] \ .$$

If the max $(n_j) \leq 5$, then tables of the exact distribution of L_1 exist (see Seigel 1956) and the exact level of significance of the test can be determined without resorting to the asymptotic result.

Since the cumulative distribution functions $F_j(\cdot)$ are continuous, this implies there should be no repeated observations in the data or ties in the ranks. This is not always the case as data measurements are only taken to a finite number of decimal points. When ties occur between two or more scores, usually each score is given the mean of the ranks for which it is tied. This practice in general causes the value of L_p to be underestimated, though not seriously if the number of ties is small. In the univariate case there exists a correction factor (C.F.) to compensate for this underestimation:

$$\text{C.F.} = \frac{\sum_{i=1}^{r} (t_i^3 - t_i)}{N^3 - N}$$

where t_i is the number of tied observations in the i^{th} group of tied scores and r is the number of tied groups of scores. Thus, the test statistic, L_1, adjusted for ties is

$$L_1 = \frac{L_1}{1 - \text{C.F.}} \ .$$

Often C.F. is near zero and negligible and hence can be ignored.

This test can be employed only to evaluate hypothesis (3.1.2). In order to create an intuitively sensible test statistic, we look at the sum of the squared deviations of the cell means, $\bar{y}_{ij}^{(R)}$, about the overall mean, $\bar{\bar{y}}^{(R)}$. This prohibits us from looking at hypothesis (3.1.1) unless we again make additional assumptions about the random variables, Y_{ij}, such as symmetry. If there is evidence to assume symmetry of the variables, for all practical purposes we will be near normality and shculd proceed under that assumption.

4.0 APPLICATION OF THE TECHNIQUES

4.1 Application to a Grassland Primary Producer Model

The statistical techniques discussed in the previous chapter were applied to two sections of an Ecosystem Level Model (ELM) of a grassland community in order to validate model output with respect to empirical observation. The section of ELM created by Sauer (1975) was designed to simulate biomass production by grassland primary producers. The primary producers were defined to be all photosynthetic plants which for simplicity of modeling were classified into five categories: warm season grasses, cool season grasses, forbs, shrubs, and cacti. The primary producer model was designed to generate deterministic outputs. The driving mechanisms of this descriptive model were based primarily on the literature, the expert opinion of those knowledgeable in the systems being modeled, and experimental evidence from process studies designed to provide specific production data, resulting in a more mechanistic than empirical model.

The output variables simulated by the primary producer model included various measures of biomass or organic material production. The model did not actually generate biomass predictions per se, but instead generated estimates of grams of carbon per square meter. The modelers opted for this carbon flow type model as the carbon content of the biomass of photosynthetic plants over time is approximately a constant. As determined by atomic weights, approximately 40% of the grams biomass (dry weight) is carbon. A conversion is readily facilitated by multiplying the model output by 2.5 in order to have predictions in the same units as the observed data, grams biomass (dry weight) per square meter.

Biomass data for validation were collected in the field by selecting randomly distributed square meter plots; clipping, uprooting, and gathering the organic materials; separating the harvest into preordained categories; and drying and weighing the results. Since this procedure leads to destructive sampling, new plots must be selected at each sampling date and the process repeated. The sampling scheme attempted to have the sampling dates be equally spaced on the time axis in order to get a representative view of the functions modeled, $f_i(t)$, $i = 1, \ldots, p$. Nevertheless, uncontrollable factors, such as unfavorable weather conditions, would occasionally interfere with data collection in the field, often making this goal unattainable.

Two variates were examined in the primary producer model validation study: aboveground biomass of warm season grasses and aboveground biomass of cool season grasses. Cool season grasses are those species which begin their photosynthetic activity in the "cool season" or early spring, while warm season grasses commence photosynthetic activity later in the spring. Each of these two categories encompasses a distinct collection or set of species of grasses. The model generated grams carbon per square meter for each of these two categories for each day in the years modeled. These values, corrected upward to represent biomass predictions, were compared to field data for these two categories by assigning the sampling results, species by species, to the two categories. The model predictions for that date could be compared with field observations for that date since the model was "run" with the actual weather record of that time period. At each of ten times approximately evenly distributed over the growing season approximately 12 plots were selected in replicate and clipped.

Letting variable 1 be warm season grasses biomass and variable 2 be cool season grasses biomass, we define

$$\underline{Y}_i' = [Y_{i1}', Y_{i2}', \ldots, Y_{i10}']; \ i = 1, 2$$

as in section 3.0. We then tested hypotheses (3.1.1) and (3.1.2) first for the bivariate case (p = 2) of warm season and cool season grasses together and then for the two univariate cases (p = 1) warm season and cool season grasses separately. Employing the tests discussed in section 3.0, we found ourselves rejecting all the null hypotheses. The tests and the levels of significance, P, for each are summarized in Table 1.

This result caused some concern which led to further analyses to determine if the tests were too restrictive and would always lead to rejection of the null hypotheses. Table 2 sheds light on this question. The means of the data and 95% confidence intervals on the means along with the corresponding predictions for each of the two variables at each of the ten times are given. Of the 20 predictions only one (or 5%) fell within the limits of the corresponding 95% confidence intervals. Further

Table 1. Results of validation tests on the grassland primary producer model.

Test	P
H_0: $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1,10}$; $\gamma_{21} = \gamma_{22} = \cdots = \gamma_{2,10} = 0$ (3.1.1)	
F-test	2.3×10^{-10}
H_0: $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1,10}$; $\gamma_{21} = \gamma_{22} = \cdots = \gamma_{2,10}$ (3.1.2)	
F-test	6.5×10^{-8}
Permutation test	0.00
Urguhart approximate to permutation test	5.3×10^{-8}
Nonparametric test	1.3×10^{-7}
H_0: $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1,10} = 0$ (3.1.1)	
F-test	1.0×10^{-21}
H_0: $\gamma_{11} = \gamma_{12} = \cdots = \gamma_{1,10}$ (3.1.2)	
F-test	1.0×10^{-3}
Permutation test	0.00
Urquhart approximate to permutation test	9.1×10^{-4}
Nonparametric test	8.5×10^{-5}
H_0: $\gamma_{21} = \gamma_{22} = \cdots = \gamma_{2,10} = 0$ (3.1.1)	
F-test	3.6×10^{-41}
H_0: $\gamma_{21} = \gamma_{22} = \cdots = \gamma_{2,10}$ (3.1.2)	
F-test	1.9×10^{-6}
Permutation test	0.00
Urquhart approximate to permutation test	1.6×10^{-6}
Nonparametric test	3.9×10^{-5}

Table 2. Means, predictions, and confidence intervals for the grassland primary producer model.

Time	$\overline{X}$	Prediction	95% confidence interval on μ
		Warm Season Grasses	
1	41.779	46.682	(33.267, 50.292)
2	43.321	63.409	(28.525, 58.117)
3	60.896	86.319	(36.085, 85.707)
4	57.350	96.567	(41.348, 73.352)
5	67.371	95.640	(53.414, 81.328)
6	71.533	93.055	(53.259, 89.807)
7	52.262	90.365	(42.620, 61.905)
8	54.112	86.767	(47.691, 60.534)
9	52.571	85.168	(38.932, 66.209)
10	45.787	96.149	(31.028, 60.547)
		Cool Season Grasses	
1	5.242	17.551	(2.459, 8.025)
2	12.796	25.196	(6.427, 19.165)
3	7.708	31.750	(4.391, 11.027)
4	17.323	32.214	(6.966, 27.680)
5	13.721	31.259	(7.177, 20.265)
6	4.779	30.602	(7.568, 6.990)
7	8.787	29.910	(5.275, 12.300)
8	7.708	28.812	(1.803, 13.614)
9	6.167	28.287	(0.669, 11.665)
10	6.629	25.969	(2.133, 11.125)

inspection revealed that the model overpredicted the amount of biomass at each of the 10 times for both variates. Also the model tended to overpredict more severely in the latter part of the growing season, that is, in time classes 7 through 10. Since the model overpredicted in a pattern that was not parallel to the observed means, the statistical tests rejected both hypotheses (3.1.1) and (3.1.2).

In order to determine within what range the predictions must lie in order to not reject H_0 in hypotheses (3.1.1) and (3.1.2), sets of predictions were "created." For each variate at each time the mean and standard error of the replicated observations were calculated,

$$\bar{y}_{ij} = \sum_{k=1}^{n_j} Y_{ijk/n_j}$$

and

$$\hat{\sigma}_{ij} = \left[\sum_{k=1}^{n_j} (Y_{ijk} \quad \bar{y}_{ij})^2 / n_j (n_j \quad 1) \right]^{\frac{1}{2}}$$

for i = 1, 2 and j = 1, 2, ..., 10 .

Seventeen predictions, P_{ijq}, were formed as follows for each time and variate:

$$\text{Prediction} = P_{ijq} = \bar{y}_{ij} \pm 0.5(q\ \hat{\sigma}_{ij})$$

where $q = 0, 1, 2, 3, 4, 5, 6, 7, 8$

for i = 1, 2 and j = 1, 2, ..., 10 .

For each q the F statistic for hypothesis (3.1.1) was calculated, and the F statistic, the Urquhart approximation to the permutation test, and the nonparametric ranks statistic were calculated for hypothesis (3.1.2).

The calculation of the four test statistics was performed for each of the nine sets of predictions for each of 2 years of data, 1970 and 1971. For each year the calculations were made for the bivariate case, warm season and cool season grasses biomass together, and also for each variate separately in the two univariate cases.

Since each of the statistics was calculated for each data set and each of the 17 sets of predictions, we could view these 17 values as points on a curve. We have four curves; one for each test statistic. These are plotted in Fig. 6 to 11 for each of two bivariate examples and the four univariate examples. On the ordinate we have the value of P, the level of significance of the test, and on the abscissa the predicted values as measured in number of standard errors from the mean.

The six figures suggest certain generalizations. In testing hypothesis (3.1.1), here formulated as

$$H_0\colon\ \gamma_{ij} = 0;\ i = 1,\ 2 \text{ and } j = 1,\ 2,\ \ldots,\ 10$$

in the bivariate case and

$$H_0\colon\ \gamma_j = 0;\ j = 1,\ 2,\ \ldots,\ 10$$

in the univariate case versus

$$H_1\colon\ \text{at least one equality in } H_0 \text{ is an inequality}$$

and employing the F statistic described in section 3.2, the level of significance of the test was generally greater than 0.5 if the predictions were within one standard error of the mean. In testing hypothesis (3.1.2)

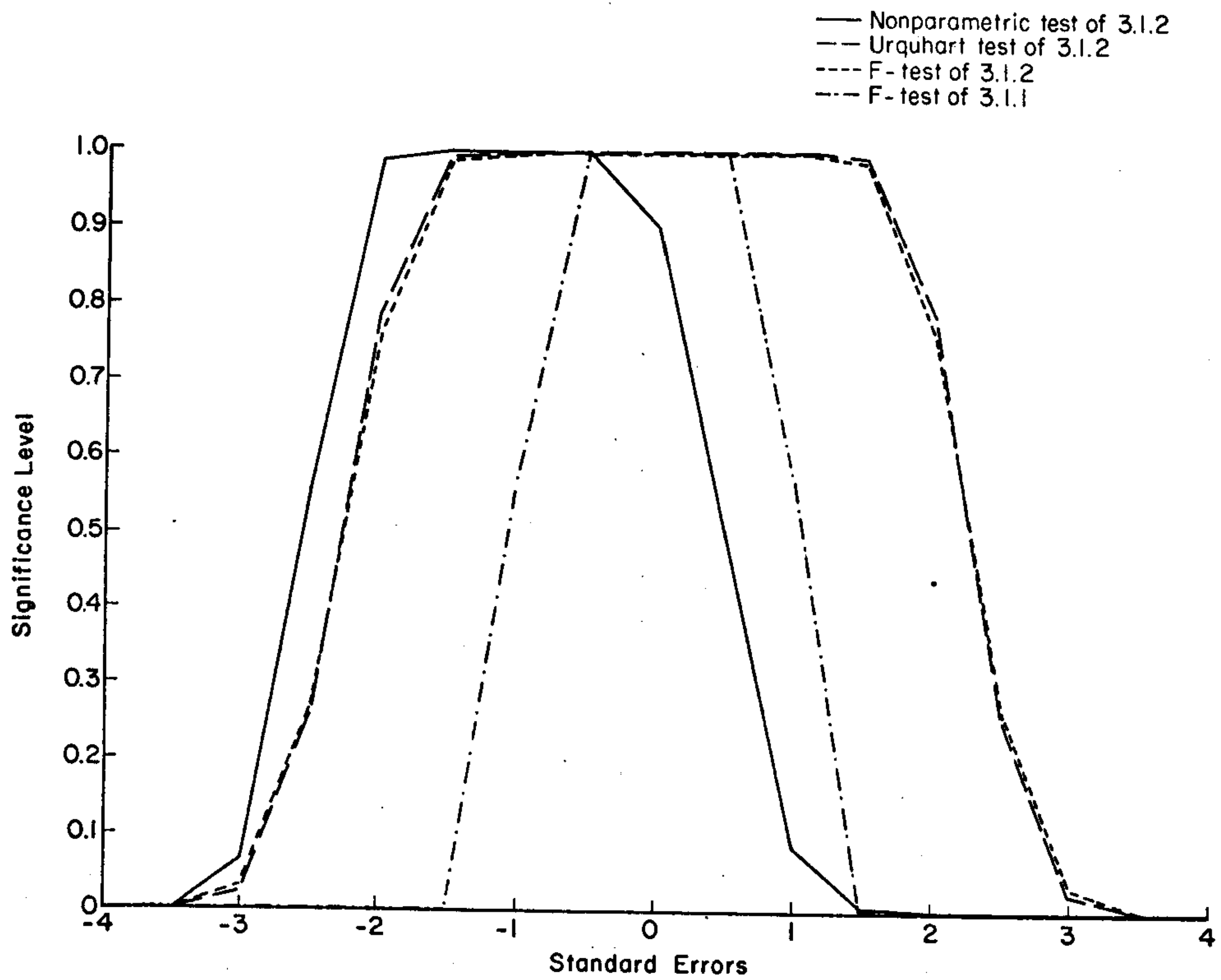

Fig. 6. Warm season and cool season grasses, 1970.

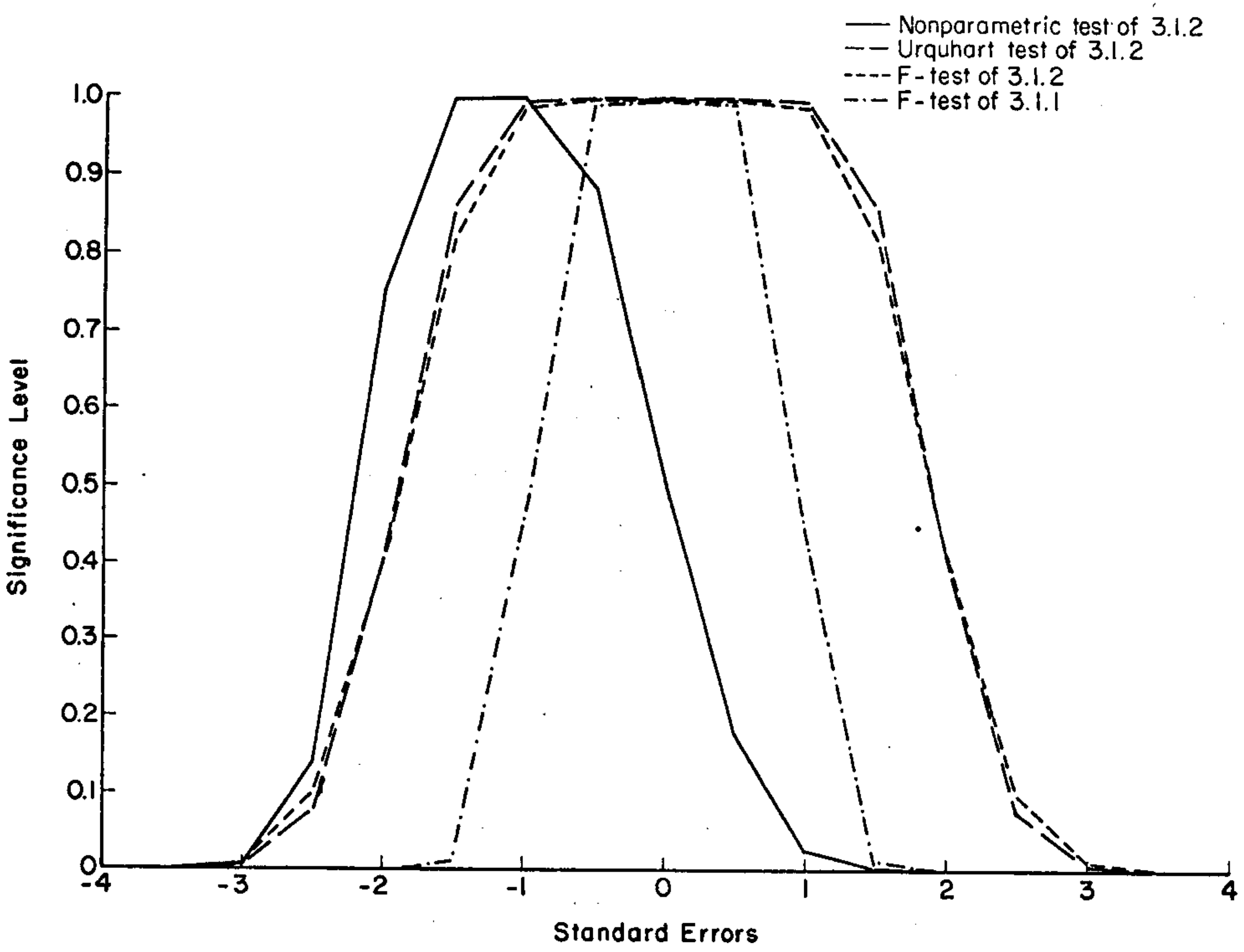

Fig. 7. Warm season grasses, 1970.

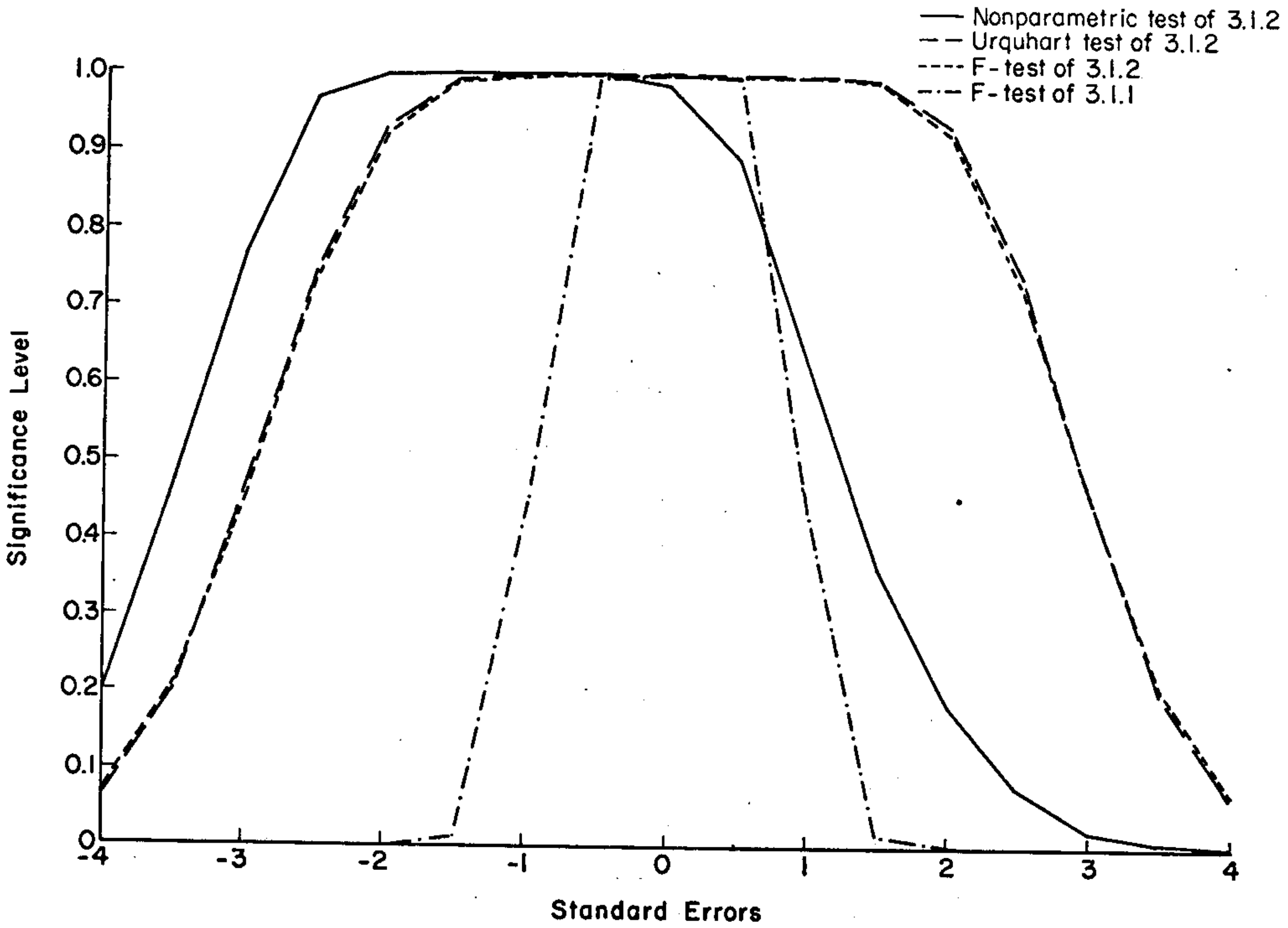

Fig. 8. Cool season grasses, 1970.

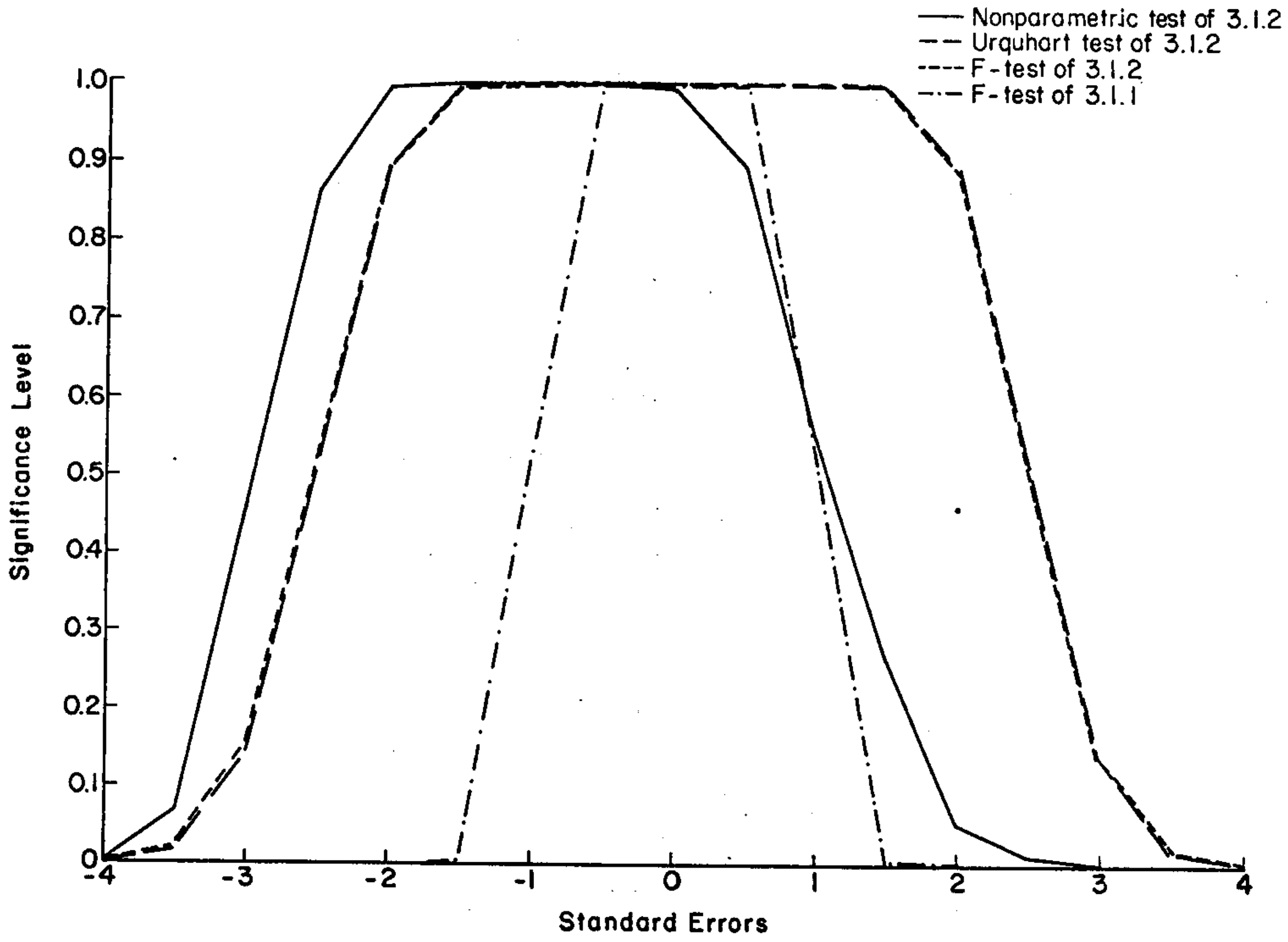

Fig. 9. Warm season and cool season grasses, 1971.

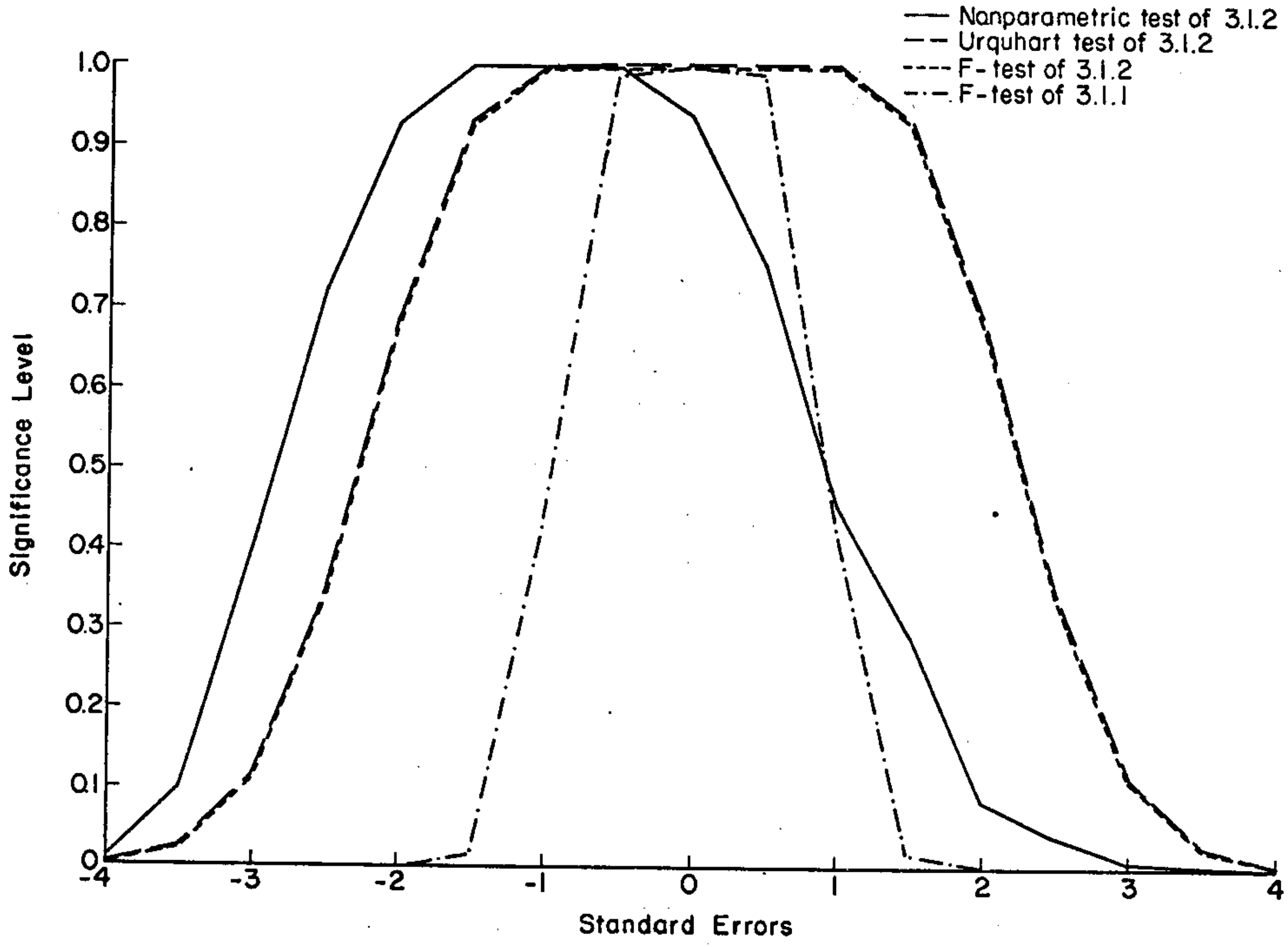

Fig. 10. Warm season grasses, 1971.

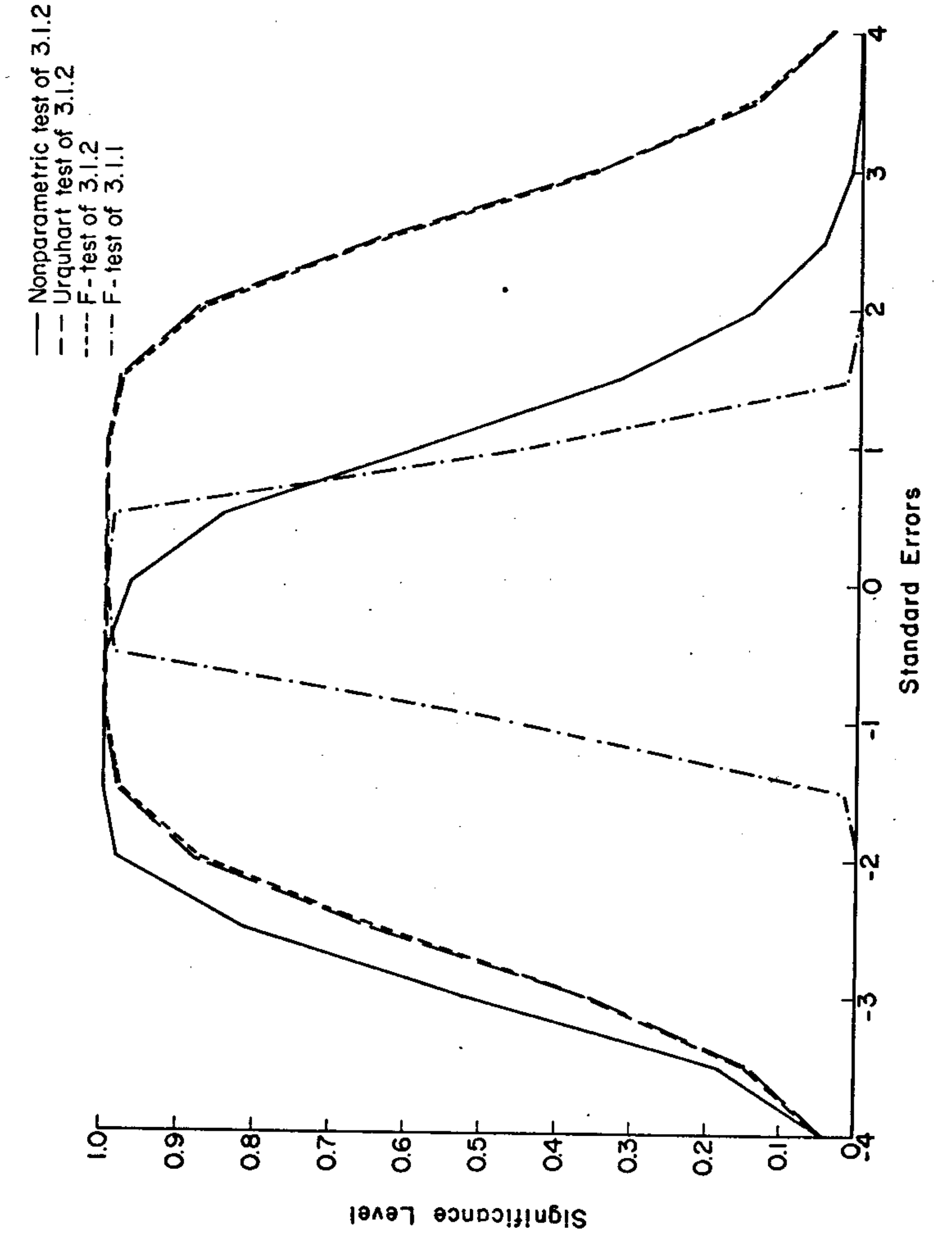

Fig. 11. Cool season grasses, 1971.

$$H_0\colon\ \gamma_{i1} = \gamma_{i2} = \cdots = \gamma_{i10};\ i = 1, 2$$

in the bivariate case and

$$H_0\colon\ \gamma_1 = \gamma_2 = \cdots = \gamma_{10}$$

in the univariate case versus

H_1: at least one equality in H_0 is an inequality,

the F test and the Urquhart approximation to the permutation test given in section 3.3 gave similar results. If the predictions were within two standard errors of the mean of the observed data at each time, the null hypothesis would not be rejected. In testing hypothesis (3.1.2) with the nonparametric ranking test described in section 3.4, the evidence indicated that the nonparametric test would accept the null hypothesis if the predictions were within two standard errors of the median of the observed replications.

Many biologists are often skeptical of the normality assumptions necessary for MANOVA analyses since biomass data are quite typically of a nonnormal nature. Since there are usually a few large observations, this causes the distributions of observed biomass to be skewed to the right. This is reflected in the fact that the median is often one standard error less than the mean of the observed data. Analysts are interested in what manner the normality assumptions affect the ability of the test to accept the null hypothesis. The close agreement of the curves in Fig. 6 through 11 for the F statistic and the Urquhart approximation to the permutation test, coupled with the fact that the Urquhart approximation does not contain any normality assumptions, lends credence to the tenet that MANOVA techniques are generally quite robust to the normality assumptions.

4.2 Application to a Grassland Decomposer Model

Another example of an application of these principles and the validation tests previously described was provided in the interest of validating the decomposer section of ELM. A description of this portion of ELM is included in an article by Hunt (1975). This descriptive model was designed to simulate the belowground decomposition of dead material in a grassland ecosystem. One set of model predictions was based on tracing the decomposition of 2 grams (dry weight) of bluestem hay buried on some preset initial date. The model outputs are the amount of material not yet decomposed as predicted on a day-to-day basis.

In order to create multiple observations for validation, the following experiment was run. On an initial date 15 litter bags, each containing 2 grams of dry bluestem hay, were buried. The container bags were designed in such a fashion that they would not interfere with the decomposition process. On five dates with approximately equal spacing throughout the "decomposition season" three of the bags were recovered and the nondecomposed material was collected and weighed. This procedure provided three replicate observations of the decomposition process at each of five different times.

Since predictions were available on a daily basis, it was possible to validate the model predictions for the 5 days on which the litter bags were recovered against the observed outputs of the litter bag experiment.

In Table 3 the means of the observed data on each of the five dates and the corresponding model predictions are presented. The difference between the means and the predictions is given and 95% confidence intervals on the observed means are provided. The results are presented graphically in Fig. 12. Note that the predicted values fell within the bounds of the 95% confidence intervals 40% of the time.

The validation tests described in section 3.0 were run on these data, the results being presented in Table 4. Hypotheses (3.1.1) and (3.1.2) were tested and the significance level, P, of the test calculated. Since the predictions were less than the observed means on all five dates, hypothesis (3.1.1) of exact agreement between the model output and observed data was easily rejected with $\alpha = 0.01$. A more interesting fact to the experimenter was the similarity of the shape of the lines generated by the observed means and the model predictions. The results presented by the test of hypothesis (3.1.2) of the similarity of shape of the two curves produced P values ranging from 0.081 to 0.131, making rejection of hypothesis (3.1.2) at the 0.1 level questionable. The experimenter noticed that the shapes of the two curves were almost identical for $t = 1, 2, 3$, but the decomposition rate for the observed means decreased more rapidly than the decomposition rate of the predictions for $t = 4, 5$. This occurrence was traceable to an overestimation of the decomposition rate of the "hard" or slowly decomposing materials in the system. The majority of the material decomposing at $t = 1, 2, 3$ was "soft" or easily decomposed matter while at $t = 4, 5$ the majority of the material was "hard." Because of this last result, the validation procedure provided a source of valuable information for the experimenter as far as understanding the model was concerned.

Table 3. Means, predictions, and confidence intervals for the grassland decomposer model.

Time	$\overline{X}$	Prediction	$\overline{X}$-Prediction	95% Confidence Interval on μ
1	1.557	1.440	.117	(1.505, 1.608)
2	1.327	1.240	.087	(1.275, 1.378)
3	1.253	1.140	.113	(1.126, 1.381)
4	1.273	1.080	.193	(0.979, 1.568)
5	1.247	1.040	.207	(1.167, 1.327)

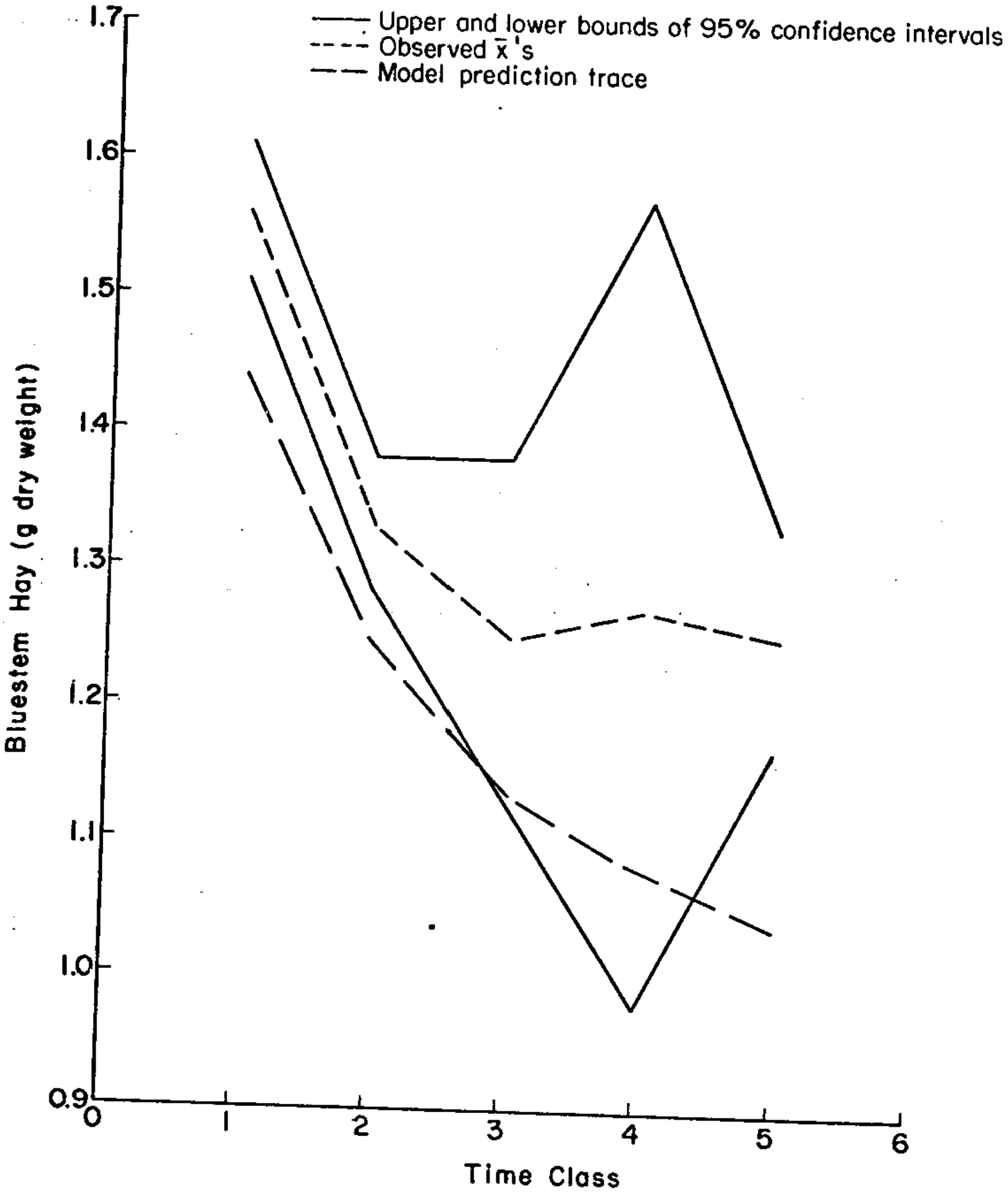

Fig. 12. Validation of decomposition model.

Table 4. Results of validation tests on the grassland decomposer model.

Test	P
$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_5 = 0$	(3.1.1)
F-test	0.00009
$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_5$	(3.1.2)
F-test	0.131
Permutation test	0.090
Urquhart approximation to permutation test	0.113
Nonparametric test	0.081

5.0 DISCUSSION

The validation process has suffered from an overdependence on subjective procedures and a lack of available applicable objective procedures. Hopefully, the statistical tests presented in section 3.0 will fill part of this need for ecosystem simulation models though much more work in this area is still needed to bridge the gap.

To date, ecosystem modeling has concentrated on producing deterministic rather than stochastic models. This philosophy has been fostered by the belief of biological modelers that, if all the driving variables could be defined and all the complex interactions inherent in biological systems explained, a set of inputs would uniquely determine the correct outputs. At the same time the high noise levels of certain biological phenomena as reflected in the large variability of their data imply that the real world systems modeled are too complex to be treated in a purely deterministic fashion. The innate variability of some biological data can be decreased by placing more controls over the experiment designed to provide the desired information, as in the case of process studies. Yet, as we proceed to place more controls on the experiment, we also lose more of the "reality" that we wish to model. The result is a trade off between variability and the maintenance of "real world" conditions in the experiment. A possible method of confronting this problem may be to employ stochastic models in biological simulations. Since a controlled amount of variability can be built into a stochastic model, this approach might provide a more appropriate and realistic representation of the system modeled.

The problems of validation are complicated by the fact that every model is unique in conception and application, resulting in validation problems unique to each model. Many of the validation techniques defined in the literature were designed to solve the problem of validating one particular type of model. The problems encountered in validating economic models of a firm or econometric models are often quite different from those problems encountered in validating biological ecosystem models because of the nature of the subject matter under study.

The statistical techniques discussed in this paper have been directed toward the problem of validating systems level models of biological communities. Since these models generate predictive traces over time, the data gathered for validation must also be collected over time. In the theory we assume that the data collected at each time is independent of those observations taken at other times. This assumption may be unnecessarily restrictive. Rather than assuming a covariance structure for each of the i = 1, ..., p variates of the form

$$\mathrm{Cov}(Y_{ij}, Y_{ij'}) = \begin{cases} \sigma_i^2 \text{ if } j = j' = 1, \ldots, t \\ 0 \text{ otherwise,} \end{cases}$$

we might more realistically assume

$$\mathrm{Cov}(Y_{ij}, Y_{ij'}) = \begin{cases} \sigma_i^2 \text{ if } j = j' = 1, \ldots, t \\ \sigma_{ijj'}^2 \text{ if } j \neq j' = 1, \ldots, t \end{cases}$$

where $\sigma_{ijj'}^2$, could be some function of the length of the interval between j and j'. Any test with this type of covariance structure probably would ultimately require estimation of the covariances, $\sigma_{ijj'}^2$, which could become quite difficult as the sampling dates are often not equally spaced on the time axis.

One might well inquire what effect, if any, assuming the covariances are zero has when in reality they are not. If the covariances are not zero and, at the beginning of the time record, the predictions and data agree, then we would be more prone to accept the null hypothesis than we should. On the other hand, if at the beginning of the time record the predictions and data are not compatible, then the tendency would be to reject the null hypothesis more often than we should. If the time record is of sufficient length, these discrepancies should average out giving an unbiased estimate of the level of significance.

Being able to calculate one level of significance in a simultaneous test of several output variables modeled satisfies the need for an objective, quantitative validation test. Often quantitative tests based on subjective judgment are employed. One such criterion suggested by ecosystem modelers has been to accept the model as valid if for each variable the predictions are within the bounds of 95% confidence intervals on the means. There is an interesting similarity between this criterion and one of the results from section 4.1 which showed that for the MANOVA test or the Urquhart approximation to the permutation test of the hypothesis (3.1.2) we would not reject the null hypothesis if the predictions were within two standard errors of the means of the data. A 95% confidence interval would tend to extend approximately 2.5 standard errors on each side of the mean.

Since biological modelers are primarily interested in trying to match data means, it is intuitively viable to use confidence intervals to create a set of "gates" or "hurdles" through which one hopes the model-generated trace will pass. Since there are no data available other than that gathered at the t sampling dates, one can actually only validate the model at those t points. This allows the experimenter to devise various subjective validation tests by meeting a criterion of having the predictions fall within X% confidence intervals Y% of the time. This appears to lend support to the use of significance tests, possibly in conjunction with subjectively based quantitative tests, to validate simulation models.

The fact that the data medians tend to be one standard error less than the means could present a question about the approach of biological modelers. Most modelers generally concentrate on trying to predict or model the mean of the data. Since the median is less than the mean, this implies they are not modeling the most common event. This suggests a question of whether it is more viable scientifically to model the most common event, the median, or to model the average but not most often observed event, the mean.

6.0 ACKNOWLEDGMENTS

The author wishes to thank Dr. R. Kirk Steinhorst for sharing his comprehensive statistical knowledge and Drs. Melvin Dyer, William Hunt, and George Innis for much useful insight into the details of validating biological models.

7.0 LITERATURE CITED

Anderson, T. W. 1958. An introduction to multivariate statistical analysis. John Wiley & Sons, Inc., New York. 374 p.

Cohen, K. J., and R. M. Cyert. 1961. Computer models in dynamic economics. Quart. J. Econ. 75(1):112-127.

Fishman, G. S., and P. J. Kiviat. 1967. The analysis of simulation generalist time series. Manage. Sci. 13(7):525-557.

Hunt, H. W. 1975. A simulation model for decomposition in grasslands. US/IBP Grassland Biome Preprint No. 155. Colorado State Univ., Fort Collins. 62 p. (Submitted to Ecol. Monogr.).

Kapoor, U. K. 1968. On the validation of simulation experiments: A review of existing techniques and a proposed technique for the Wisconsin River water quality simulation model. M.S. Thesis. Univ. Wisconsin, Madison. 28 p.

Kendall, M. G., and A. Stuart. 1966. The advanced theory of statistics. Vol. 3. Charles Griffin & Co., London. 552 p.

Naylor, T. H., and J. M. Finger. 1967. Verification of computer simulation models. Manage. Sci. 14(1):92-101.

Nolan, R. L. 1972. Verification/validation of computer simulation models, p. 1254-1265. *In* Proceedings of the 1972 Summer Computer Simulation Conference. Simulation Councils, Inc., La Jolla, California.

Phillips, J. B. 1971. Satistical methods in systems analysis, p. 34-52. *In* J. B. Dent and J. R. Anderson [ed.] Systems analysis in agricultural management. John Wiley & Sons, Inc., New York.

Puri, M. L., and P. K. Sen. 1971. Nonparametric methods in multivariate analysis. John Wiley & Sons, Inc., New York. 440 p.

Sauer, R. H. 1975. A simulation model for grassland primary producer phenology and biomass dynamics. US/IBP Grassland Biome Preprint No. 152. Colorado State Univ., Fort Collins. 97 p. (Submitted to Ecol. Monogr.).

Seigel, S. 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill Book Co., Inc., New York. 312 p.

Steinhorst, R. K. 1973. Validation test for ecosystem simulation models. US/IBP Grassland Biome Preprint No. 94. Colorado State Univ., Fort Collins. 15 p. (Submitted to Biometrics).

Theil, H. 1961. Economic forecasts and policy. 2nd ed. North-Holland Publishing Co., Amsterdam. 567 p.

Urquhart, S. N. 1965. On the permutation distribution of a multivariate test statistic. Ph.D. Diss. Colorado State Univ., Fort Collins. 106 p.

Wright, A. 1971. Farm systems, models and simulation, p. 17-33. *In* J. B. Dent and J. R. Anderson [ed.] Systems analysis in agricultural management. John Wiley & Sons, Inc., New York.

Wright, R. D. 1972. Validating dynamic models--an evaluation of tests of predictive power, p. 1286-1294. *In* Proceedings of the 1972 Summer Computer Simulation Conference. Simulation Councils, Inc., La Jolla, California.

7.1 Additional References

Amstutz, A. E. 1967. Computer simulation of competitive market response. MIT Press, Cambridge, Mass. 457 p.

Bainstow, J. N. 1970. A review of systems evaluation packages. Computer Decisions 2(6):16-19.

Bayes, A. J. 1970. Statistical techniques for simulation models. Aust. Computer J. 2(4):180-184.

Bonini, C. P. 1963. Simulation of information and decision systems in the firm. Prentice-Hall, Inc., Englewood Cliffs, N. J. 160 p.

Burdick, D. S., and T. H. Naylor. 1966. Design of computer simulation experiments for industrial systems. Commun. of the ACM 9(6):329-339.

Churchman, C. W. 1965. Reliability of models in the social sciences, p. 23-28. *In* P. Langhoff [ed.] Models, measurement and marketing. Prentice-Hall, Inc., Englewood Cliffs, N. J.

Conway, R. W. 1963. Some tactical problems in digital simulation. Manage. Sci. 10(1):58-60.

Conway, R. W., B. M. Johnson, and W. L. Maxwell. 1959. Some problems of digital systems simulation. Commun. of the ACM 6(1):92-110.

Drymes, P. J., E. P. Howrey, S. H. Hyman, J. Kmenta, E. E. Leamer, R. E. Quandt, J. B. Ramsey, N. T. Shapiro, and V. Zarnowitz. 1972. Criteria for evaluation of ecometric models. Ann. Econ. Soc. Measurement 1(3):291-324.

Emshoff, J. R., and R. L. Sisson. 1970. Design and use of computer simulation models. Macmillan and Co., New York. 302 p.

Fishman, G. S. 1967. Problems in the statistical analysis of simulation experiments: The comparison of means and the length of sample record. Commun. of the ACM 10(2):94-99.

Forrester, J. W. 1961. Industrial dynamics. MIT Press, Cambridge, Mass. 464 p.

Gordon, G. 1969. System simulation. Prentice-Hall, Inc., Englewood Cliffs, N. J. 303 p.

Graybill, F. A. 1961. An introduction to linear statistical models. Vol. 1. Mcgraw-Hill Book Co., Inc., New York. 463 p.

Hermann, C. 1967. Validation problems in games and simulation. Behav. Sci. 12(3):216-230.

Huff, D. 1954. How to lie with statistics. W. W. Norton & Co., Inc., New York. 142 p.

Kempthorne, O. 1952. The design and analysis of experiments. John Wiley & Sons, Inc., New York. 631 p.

Kleignen, J., T. H. Naylor, and T. Seaks. 1972. The use of multiple ranking procedures to analyze simulations of management systems: A tutorial. Manage. Sci. 8:245-257.

McKenney, J. L. 1962. Critique of verification of computer simulation models. Manage. Sci. 14(2):102-103.

McMillan, C., and R. F. Gonzalez. 1965. System analysis. R. D. Irwin, Inc., Homewood, Ill. 336 p.

Mood, A. M., F. A. Graybill, and D. C. Boes. 1973. An introduction to the theory of statistics. McGraw Hill, Inc., New York. 564 p.

Morrison, D. F. 1967. Multivariate statistical methods. McGraw Hill Book Co., Inc., New York. 338 p.

Naylor, T. H., J. L. Balintfy, D. S. Burdick, and K. Chu. 1966. Computer simulation techniques. John Wiley & Sons, Inc., New York. 352 p.

Naylor, T. H., K. Wertz, and T. Wonnacott. 1967. Some methods for analyzing data generalist by computer simulation experiments. Commun. of the ACM 10(11):703-710.

Schrank, W., and C. Holt. 1967. Critique of verification of computer simulation models. Manage. Sci. 14(2):104.

Van Horn, R. L. 1969. Validation, p. 232-251. *In* T. H. Naylor [ed.] The design of computer simulation experiments. Duke Univ. Press, Durham, N. C.