The Interpretation and Estimation of Effective Sample Size

H. J. THIÉBAUX

Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 4H8, Canada

F. W. ZWIERS

Department of Mathematics, University of Saskatchewan, Saskatoon, Saskatchewan S7N 0W0, Canada

(Manuscript received 8 July 1983, in final form 7 March 1984)

ABSTRACT

Statistical and dynamical relationships between observed values of a geophysical system or model effectively reduce the number of independent data. This reduction is expressible in terms of the covariance structure of the process and, in some instances, it is reasonable to devise a measure of the "effective sample size" in terms of sample statistics. Here we discuss the concept of "effective sample size," and, having settled upon one of several possible definitions, examine various methods of estimating this quantity. It is found that "effective sample size" is quite difficult to estimate reliably. However, a procedure is described which we feel could be used successfully; it is noted that the concept could be extended to spatial arrays of data, in some circumstances.

1. Introduction

The concept of "effective number of degrees of freedom" or "effective sample size" (ESS) was discussed by Laurmann and Gates (1977). However, there, and in its use elsewhere, the quantity is not uniquely defined; this has lead to some confusion about its correct interpretation. We believe that a careful examination of this concept, of alternative formulations for its evaluation, and of some key statistical properties of these estimates may be helpful to researchers who wish to use the concept in the analysis of time series or spatially coherent processes.

In Section 2 we derive a measure of effective sample size, as it has been previously defined. In Section 3 we discuss the phraseology which has been used to describe the ESS and argue that the usual formulation for its evaluation is only one of a family of possible estimators. Section 4 describes three methods of estimating ESS; Section 5 compares these methods using Monte Carlo experiments; and Section 6 considers and rejects the use of ESS in testing for the mean of a time series; an alternative test is proposed.

2. Derivation of "effective sample size" for a time series

Non-negligible correlations between successive values of a stochastic process preclude independence of the components of a vector of any N sequential observations. This is a concomitant of predictability and its measure, either in terms of Leith's (1973) "time between effectively independent samples" or the reciprocal "effective number of independent samples" in a fixed time span; the latter provide useful means of quantifying the degree of time coherence in the process. The "time between effectively independent samples" or simply the "effective sample size" (ESS), may be defined by equating the ensemble mean square of a time-averaged mean, say $\sigma_{\bar{X}}^2$, to the standard formula for the variance of the mean of independent samples. The solution for the so-called number of independent samples is one measure of ESS. In this derivation the ensemble mean square is determined by the autocovariance function (acf), which thus gives explicit expression to time coherence of successive values, as shown in the following.

For N time-sequential observations, say $X_1, \ldots, X_N$, of a stationary stochastic process with variance σ^2, the ensemble mean square of the observed mean may be written in terms of the variance and autocorrelation function. Following Anderson (1971, p. 448),

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \langle[\bar{X}-\mu]^2\rangle = \langle[\sum_{i=1}^{N}(X_i-\mu)/N]^2\rangle \\ &= \sum_{i,j=1}^{N}\langle(X_i-\mu)(X_j-\mu)\rangle/N^2 \\ &= \sum_{i,j=1}^{N} C(i-j)/N^2 \\ &= \sum_{\tau=-(N-1)}^{+(N-1)}\{N-|\tau|\}C(\tau)/N^2 \\ &= \sigma^2\sum_{\tau=-(N-1)}^{+(N-1)}\left(1-\frac{|\tau|}{N}\right)\rho(\tau)/N,\end{aligned}$$

where $C(\tau)$ is the symmetric covariance between observations lagged by a time interval of length τ, and $\rho(\tau)$ is the corresponding lag-correlation function. If in fact $\bar{X}$ were the mean of N' *independent* sample values, this would be σ^2/N'. Hence by equating $\sigma^2(\bar{X})$ and σ^2/N' we obtain a measure of the effective number of independent samples:

$$N' = \sigma^2/\sigma_{\bar{X}}^2 = N\left[\sum_{\tau=-(N-1)}^{+(N-1)}\left(1 - \frac{|\tau|}{N}\right)\rho(\tau)\right]^{-1}. \quad (2.1)$$

3. Interpretation of "effective sample size" for a time series

Some expressions which have been used to describe the ESS (see, for example, Laurmann and Gates, 1977, and Livezey and Chen, 1983) are misleading, namely the phrases "equivalent number of degrees of freedom" and "number of effective degrees of freedom", which are sometimes associated with N'. Their applications to the construction of t statistics are also erroneous because, regardless of the extent of time coherence, a sample of N observations from *any* nonsingular Gaussian stochastic process has N degrees of freedom. A suitable linear transformation of the observation vector, for which the transformation is determined by the covariance structure of the process, has N independent components. The phrase "effective sample size" is less troublesome; it manages to convey the notion that the N pieces of information in the sample are "smeared" across the N observations by the time-durations of their influences.

The phraseology used in the literature may appear to imply that the ESS is uniquely defined. However, there are several equally valid functional definitions which we will examine here. Expression (2.1) above was derived by equating the variance of the time average of a sample of N observations from a correlated time series with the variance of a sample of N' observations from an uncorrelated series. However, there are other statistical characteristics of sampled sequences whose comparison could be used analogously to generate a measure of "effective sample size." For example, the expected number of runs above and below the mean, in a sample of length N from a correlated time series, could be equated with the number for a sample of length N' from an uncorrelated series. In the latter case, given that the sample contains N_1' observations above the mean, it can be shown that the expected number of runs R is

$$E(R) = 2N_1'(N' - N_1')/N' + 1$$

$$= 2N'p(1 - p) + 1,$$

where p is the proportion of observations above the mean (Mood and Graybill, 1963, p. 412). For large samples this can be simplified to

$$E(R) = N'/2 + 1,$$

since p converges to 0.5 with probability one, as N' goes to infinity. By equating this expression with the expected number of runs for a sample of length N from a correlated time series and solving for N', we get a different expression for ESS.

Since it is not possible to derive the expected number of runs for samples from correlated time series analytically, even for the simple case of a Markov process, we illustrate the non-equivalence of the two derivations of ESS with the results of a Monte Carlo experiment. Ensembles of 5000 samples of length 90 and 180 were created from a simulated Markov process, with a lag-one correlation coefficient of 0.75. These samples were used to estimate the expected number of runs above and below the mean. Table 1 summarizes the ESS estimates derived by equating the expected number of runs obtained for the Markov process to the expected number of runs for an uncorrelated sequence of length N' and compares them with those derived by equating variances. The descrepancies between the values in the last two columns make it clear that the two techniques for evaluating ESS yield substantially different results. Consequently, equating these characteristics of sample statistics for samples from correlated time series with those for samples from uncorrelated series does not yield a unique measure. Needless to say, there are other characteristics which could reasonably be compared to generate measures of "effective sample size." However, in the remainder of this paper we consider only the quantity defined by (2.1), arrived at by equating variances, and we reserve the term "effective sample size" and its abbreviation ESS to denote it.

4. Estimating the "effective sample size"

In practice the parameters σ^2 and $C(\tau)$, and therefore $\rho(\tau)$ as well, are generally not known. Various approaches to the estimation of these parameters and hence of N' are available. However, as we demonstrate in the following, none of them is entirely satisfactory.

The most straightforward approach is to estimate $C(\tau)$ and $\sigma^2 = C(0)$ *directly* from the data. Since the ensemble mean of the stochastic process will be generally unknown, the sample mean $\bar{X}$ is substituted for μ in the computational formulas for covariance estimation, as follows:

TABLE 1. Point and interval estimates of $E(R)$ based on 5000 samples of length N from a Markov process with lag-one correlation coefficient 0.75. The "runs comparison" values of ESS appear in column 4 and the "variance comparison" values, analytically derived with Expression (2.1), are given in column 5.

N	$E(R)$	95% C.I.	$N'(R)$	$N'(V)$
90	22.46	(22.32, 22.6)	43	13
180	42.95	(42.76, 43.14)	84	26

$$\hat{C}(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} (X_i - \bar{X})(X_{i+\tau} - \bar{X}),$$

and for the sample variance $S^2 = \hat{C}(0)$. However, these estimates of lag-covariance $C(\tau)$ and of variance σ^2 are not unbiased. Following Anderson (1971, 438–464), the statistical ensemble average of the lag-τ covariance estimate is

$$\langle \hat{C}(\tau) \rangle = C(\tau) - \frac{1}{N}\left\{C(0) + 2\sum_{r=1}^{\tau}\left[1 - \frac{r\tau}{N(N-\tau)}\right]C(r) + 2\sum_{r=\tau+1}^{N-\tau-1}\left[1 - \frac{r\tau}{N(N-\tau)} - \frac{r-\tau}{N-\tau}\right]C(r) + 2\sum_{r=N-\tau}^{N-1}\frac{(N-r)\tau}{N(N-\tau)}C(\tau)\right\}$$

for τ in the interval $(1, N-\tau-1)$, with a comparable expression for $\tau > (N-1)/2$. If it were possible to increase the length of the observed sequence indefinitely, the ensemble means of the sample covariances would converge to the corresponding values of the autocovariance function, (acf) provided the series $\sum_{-\infty}^{\infty} C(\tau)$ converges. Thus $\hat{C}(\tau)$ is asymptotically an unbiased estimate of $C(\tau)$ for all time lags. However, for an observation period which is likely to be significantly *non*-infinite with respect to the time coherence of the observed process, it is not realistic to assume that the bias factor is near zero. Accordingly, it appears that $\sigma_{\bar{X}}^2$ and σ^2, and hence N', cannot be reliably estimated by direct use of sample covariance and variance values.

In addition to the problem with bias, (2.1) requires estimates of $\rho(\tau)$ at large lags when the acf is effectively zero. Hence the denominator of (2.1) needlessly includes many estimates of zero. Therefore, in addition to being biased, the resulting estimates of the ESS also have large variances. A simple solution to this problem may be truncation of the summation in the denominator of (2.1). Both the untruncated and truncated approaches are presented in the Monte Carlo experiments described in the following section: The estimator of the ESS which uses the sample autocovariance values directly to estimate the autocovariance function of the process is denoted by DIRECT. The estimator which employs the truncated sample autocovariance estimator is denoted by DIRECT2. As will be discussed below, estimation of the ESS is essentially equivalent to estimation of the spectral density function at the origin. The approach denoted by DIRECT2 employs a crude "weighted covariance" estimator of the spectral density function. A more refined weighted covariance estimator of the spectral density function is employed in another estimator of the ESS, namely "BART," which also is discussed below.

Another approach to reducing the variance of ESS estimates is through the use of a time series model. For example, an autoregressive moving average (ARMA) model such as Box and Jenkins (1976) describe, may be fitted to the sequence of N observations and the autocovariance function of the fitted model used to estimate N'. However, once again, the estimates of the coefficients of the ARMA model are only asymptotically unbiased. For *finite* observation sequences, the estimated autocovariance function will be biased and, again, N' will not be estimated accurately. However, the autocovariance function of the fitted model has the property that it converges to zero as τ becomes large so that the large lag terms in (2.1) are effectively eliminated. Therefore this technique may be expected to produce estimates of the ESS which have lower variance.

The estimator utilizing the autocovariance function of autoregressive moving-average models is denoted by ARMA. In the Monte Carlo experiments, autoregressive models were fitted to sample sequences using the method of least squares, and Akaike's information criterion (AIC) (Akaike, 1973, 1976) was used to evaluate the order of the process. The AIC criterion is described by

$$\text{AIC} = N \ln \hat{\sigma}_\epsilon^2 + 2p,$$

where N is the length of the sample sequence, p is the order of the autoregressive model which is fitted, and $\hat{\sigma}_\epsilon^2$ is the estimated variance of the random uncorrelated innovations of the model. Essentially, the fitted model minimizes the sum of squared errors, with the exception that a "penalty" is imposed for "extravagant use of parameters."

As a second approach, we consider estimating EES from the *power spectrum* of the observed sequence. To see how this may be achieved we note that for large sample sizes we have the following approximation for the variance of the sample mean:

$$\sigma_{\bar{X}}^2 \approx 2\pi f_{XX}(0)/N,$$

where $f_{XX}(\lambda)$ is the spectral density function of the observed time series. Thus, if we equate the right-hand-side with the expression for the variance of N' independent observations and solve for N', we obtain

$$N' \approx \frac{N\sigma^2}{2\pi f_{XX}(0)},$$

and the problem of estimating N' reduces to estimating the spectrum of the process at the origin. There will be biases in this estimation, as well, and a trade-off must be made between the inherent variability of the spectral estimator and its bias (See Jones, 1975, or Koopmans, 1974, for discussions of the consequences of using smoothed periodogram- or similar spectral-estimators). When the true spectrum has a relative maximum at the origin, such as that of a red noise process, we should expect the estimator of $f_{XX}(0)$ to

be *negatively* biased and hence estimates of N' to be *positively* biased. On the other hand, when the true spectrum has a trough at the origin, such as those of many second-order processes which are important for explaining spatial relationships in the atmosphere (Thiébaux, 1976, 1981), we should expect the estimator of $f_{XX}(0)$ to be *positively* biased and hence the resulting estimates of N' to be *negatively* biased. Thus, it is clear that the problem of estimating EES is embedded in the problem of estimating the variance of the time-averaged mean, as considered by Jones (1975). In this connection Jones discusses various methods of fitting time series models and estimating the spectrum at the origin. Although any method which yields acceptable estimates of the variance of the time-averaged mean might also be expected to yield acceptable estimates of EES, this bridge must be approached with caution: Amounts of bias which are relatively inconsequential when making estimates of the variance, result in considerable bias in estimates of EES because they are proportional to the inverses of the variance estimates.

We used four estimators of the power spectrum in the Monte Carlo experiments described below. The estimators we refer to as SPEC1 and SPEC5 are based on the Daniel, or smoothed periodogram estimator of the spectrum. In this case $f_{XX}(0)$ is estimated simply as

$$f_{XX}^{(N)}(0) = \frac{1}{m}\sum_{j=1}^{m} I_{XX}^{(N)}\left(\frac{2\pi j}{N}\right),$$

where $I_{XX}^{(N)}(\lambda)$ is the periodogram, namely, the squared modulus of the finite Fourier transform of the observations, and m is the number of ordinates of the periodogram which are averaged together. We used $m = 1$ for SPEC1 and $m = 5$ for SPEC5. The former will have low bias and the latter low variance. At the expense of a small increase in variance, the bias of the Daniel estimator can be reduced by employing a data window or "taper." Estimator SPECT5 uses the Daniel estimator with 20% of the observations tapered with a cosine taper. This is to say that the X_t observations are transformed as

$$Y_t = b(t/N)X_t \quad \text{for} \quad t = 1, \ldots, N,$$

where

$$b(\nu) = \begin{cases} \frac{1}{2}[1 - \cos(\pi\nu/a)] & \text{for } 0 \leq \nu \leq a \\ 1 & \text{for } a \leq \nu \leq 1 - a \\ \frac{1}{2}[1 - \cos\{\pi(1-\nu)/a\}] & \text{for } 1 - a \leq \nu \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The method designated by "BART" uses a weighted covariance estimator with a Bartlett lag-window, for which the estimator of $f_{XX}(0)$ is given by

$$f_{XX}^{(N)}(0) = \frac{1}{2\pi}\sum_{\tau=-(M-1)}^{(M-1)} (1 - |\tau|/M)C(\tau).$$

Here M determines the "bandwidth" or resolution of the estimator. Although a number of different lag-windows might have been used, all those with essentially the same bandwidth would give essentially the same estimates of $f_{XX}(0)$ (see Jenkins and Watts, 1968). The bandwidth for the BART estimator is

$$\beta = 3\pi/M.$$

In our experiments we choose M to be the square root of sequence length N, thereby insuring that the bandwidth of the estimator, and therefore its bias, decrease with increasing N. Since the bandwidth of weighted covariance estimators is generally relatively wide compared to that of Daniel estimators, there is only a slight reduction in bias when a taper is used (see Koopmans, 1974). Consequently we did not employ a taper with this estimator.

5. Monte Carlo experiments

Monte Carlo experiments were carried out with simulated first- and second-order autoregressive processes for which the corresponding spectral density functions have a relative maximum and a relative minimum at the origin, respectively.

Four first-order autoregressions

$$X_t = \rho X_{t-1} + \epsilon_t \tag{5.1}$$

were examined with $\rho = 0.3, 0.45, 0.6$ and 0.75. These models have corresponding "times between effectively independent samples" of about 1.86, 2.64, 4 and 7 time units, respectively (by rigorous computation from model properties) and are useful for describing a broad range of meteorological processes with fairly short memories. In addition, the second-order autoregression

$$X_t = 1.2X_{t-1} - 0.6X_{t-2} + \epsilon_t, \tag{5.2}$$

for which the "time between effectively independent samples" is 1.76 time units, is examined. In both (5.1) and (5.2) the ϵ_t represent sequences of uncorrelated random variables. The spectral density functions for two of these models are shown in Fig. 1.

For each model, experiments were conducted by generating 1000 sequences of length N for various values of N. For each sequence generated, ESS was estimated by the three approaches discussed above; Tables 2–6 describe the outcomes in terms of parameters of the resulting Monte Carlo distributions. In each table the true "effective sample sizes," analytically derived for the corresponding models and lengths of sequences, are given in the column headed "ESS," and estimated ESS values are compared for sequence lengths N of 30, 60, 90, 120 and 240. Table 2 presents mode values for the 1000 sequences generated, Table 3 the medians, Table 4 the means of the distributions, Table 5 standard deviation values, and Table 6 rms errors.

In the foregoing experiments the sample autocovariance function was truncated at lag-10 for the pur-

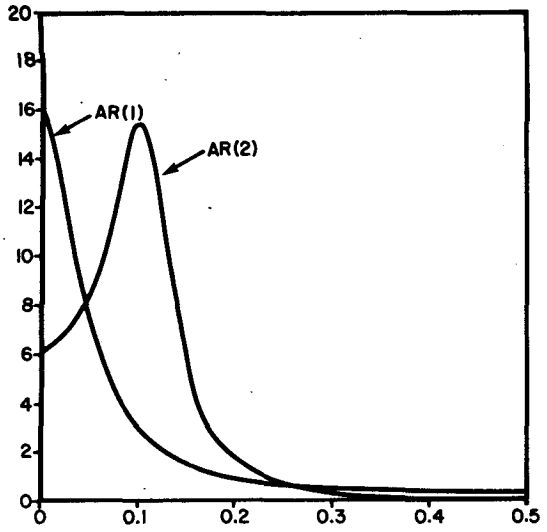

FIG. 1. Power spectra of the AR(1) process, $X_t = 0.75X_{t-1} + \epsilon_t$; and the AR(2) process, $X_t = 1.2X_{t-1} - 0.6X_{t-1} + \epsilon_t$, with $\sigma_\epsilon^2 = 1.0$. Note the relative maximum at the origin of the spectral density function for the AR(1) process and the relative minimum at the origin for the AR(2) process.

poses of estimator DIRECT2. This was felt to be adequate because, in the case of the autoregressive-one models, at most 5.6% of the contribution of the denominator of (2.1) is made by terms at lags greater than 10. The program UNIMAR from the TIMSAC-78 package (Akaike *et al.*, 1979) was used to find the best-fitting model satisfying the AIC criterion. For sequences generated for both first- and second-order processes, we allowed the UNIMAR program to select amongst models of orders up to five.

As can be seen from Tables 2–5, the DIRECT estimates of ESS are extremely unreliable in the sense that they have large standard deviations for both first- and second-order generating processes and all sequence lengths, and their medians and means differ substantially from the true ESS values. Although the mode values are not generally greatly in error, comparisons of them with median and mean values show the distributions of DIRECT estimates to be strongly skewed to the right.

The DIRECT2 estimator represents a marked improvement over the DIRECT estimator. As can be seen from Tables 2–5, the strong bias and high variability of the DIRECT estimator are greatly reduced. This is probably best attributed to a reduction in the variance of the denominator of (2.1); a reduction in the frequency with which small estimates of the denominator are made will reduce the frequency with which extremely large ESS estimates are made and

TABLE 2. Mode of 1000 estimates of the "effective sample size" made with seven different estimators.

			Estimated ESS (mode)						
						Power spectrum			
Model	N	True ESS	DIRECT	DIRECT2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	40	24	13	10	16	16	19
	60	33	40	29	30	17	29	25	37
	90	49	—	55	52	25	—	37	47
	120	65	—	61	67	—	48	—	71
	240	130	201	142	133	—	113	98	—
AR(1) $\rho = 0.45$	30	12	10	13	—	6	15	14	15
	60	23	29	28	23	12	—	24	30
	90	35	102	27	—	—	—	31	40
	120	46	—	37	—	—	43	33	48
	240	91	142	90	90	56	91	—	97
AR(1) $\rho = 0.6$	30	8	7	10	8	5	13	13	12
	60	15	15	18	15	7	18	17	21
	90	23	26	21	23	13	23	24	30
	120	30	33	31	29	21	27	28	40
	240	60	50	62	57	29	52	46	68
AR(1) $\rho = 0.75$	30	5	8	5	7	3	12	12	11
	60	9	11	9	10	6	14	13	—
	90	13	14	16	12	6	17	15	—
	120	18	16	19	18	9	19	17	29
	240	35	29	—	38	—	30	35	44
AR(2)	30	16	14	14	19	8	11	11	13
	60	34	36	22	31	15	20	21	26
	90	51	—	36	47	38	37	41	43
	120	68	70	70	—	61	49	70	56
	240	136	—	156	149	—	101	184	—

TABLE 3. Median of 1000 estimates of the "effective sample size" made with seven different estimators.

			Estimated ESS (median)						
						Power spectrum			
Model	N	True ESS	DIRECT	DIRECT2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	63	36	15	22	19	19	22
	60	33	114.5	49	33	47	35	34	41
	90	49	166.5	66	51	64	50.5	50	59
	120	65	227	81.5	68	96	69	65	76
	240	130	>360	145	131	192	139	130	150
AR(1) $\rho = 0.45$	30	12	39	26	13	14	16	16	18
	60	23	80	37	26	32	26	26	31
	90	35	116.5	46	38	47	38	37	45
	120	46	164	58	50	65	48	47	56
	240	91	307.5	104.5	94	122	97	93	108
AR(1) $\rho = 0.60$	30	8	25	19	11	10	14	14	15
	60	15	55	26	18	21	20	19	24
	90	23	80	31	26	30	27	26	33
	120	30	102	39	33	40	33.5	32	42
	240	60	207	67.5	63	92.5	64.5	59	75
AR(1) $\rho = 0.75$	30	5	19	12	8	7	13	13	12
	60	9	33	14	12	11	15	15	18
	90	13	49	19	16	17	19	18	24
	120	18	62	23	20	22	22	22	29
	240	35	117	41	37	46	39	38	48
AR(2)	30	16	39	27	30	17	11	11	13
	60	34	102	43	38	41	23	24	28
	90	51	151	62	55	72	43	45	47
	120	68	210	79	73	94	63	65	64
	240	136	448	147	139	190	141	147	133

hence reduce the bias. Despite the improvement, the rms errors for this estimator are large compared to some of the others discussed below, except at the largest sample size considered. Unfortunately, there are theoretical limits to the improvement in quality of DIRECT2 estimates with increasing sample size. This is discussed further below when a comparison is made between estimates made by DIRECT2 and those made by the closely related estimator BART.

SPEC1 estimates behave similarly to DIRECT estimates. Specifically, their distributions are skewed to the right and have large variances. This behavior can be explained in terms of the method of estimation, in this case, since an unsmoothed ($m = 1$) periodogram estimate of the spectrum at the origin is used. The periodogram ordinate at frequency $2\pi/N$, which is that used to estimate $f_{XX}(0)$ when $m = 1$, is asymptotically proportional to a random variable with exponential distribution (see Brillinger, 1981). Correspondingly, estimates of $f_{XX}(0)$ near zero, and thus very large estimates of ESS, are relatively frequent.

SPEC5 estimates which are based on the smoothed periodogram with $m = 5$ are better behaved than the estimates described above. Their distributions are also somewhat skewed to the right, but not to the same extent, as can be seen by comparing results noted in Tables 3 and 4. From Table 5, we see that the variability of this estimator is much lower than that of DIRECT, DIRECT2 and SPEC1. Nonetheless, SPEC5 does not provide highly reliable estimates: In most cases the rms errors shown in Table 6, are more than 50% of true ESS values. [Note that in Tables 3 and 4 we see evidence of the effects of the biases in the estimates of $f_{XX}(0)$: In the AR(1) case we see that the bias is positive (and quite large, in relative terms) for small sequence lengths, while in the AR(2) case, the bias is negative for short sequences.]

The SPECT5 estimator which employs the smoothed periodogram estimator with $m = 5$ and a cosine taper is disappointing in that any reduction in bias in the estimates of $f_{XX}(0)$ seems to have been compensated by the corresponding increase in variance. As can be seen from Table 6, the rms errors of this estimator were about the same as those of SPEC5.

The BART estimator, which uses the weighted covariance estimator of $f_{XX}(0)$, has a mixed performance in the following sense. When the observations were generated from the AR(2) process, the estimator produced estimates of ESS with very low biases and low variances, *and* rms errors of estimation were relatively low (ranging from 20 to 29% of the true ESS). However, the estimator did not perform as well when observa-

TABLE 4. Mean of 1000 estimates of the "effective sample size" made with seven different estimators.

Model	N	True ESS	Estimated ESS (mean)						
						Power spectrum			
			DIRECT	DIRECT2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	107	88	30	99	21	22	24
	60	33	185	96	42	251	41	39	44
	90	49	277	112	54	279	58	57	63
	120	65	355	135	72	595	78	76	81
	240	130	698	176	136	983	161	149	155
AR(1) $\rho = 0.45$	30	12	76	60	20	71	18	18	20
	60	23	143	73	29	131	29	29	34
	90	35	204	76	41	358	43	41	47
	120	46	271	82	53	235	55	53	59
	240	91	512	118	98	1130	113	110	113
AR(1) $\rho = 0.6$	30	8	54	56	15	41	15	15	16
	60	15	100	47	21	154	22	21	25
	90	23	140	52	29	201	30	29	35
	120	30	182	52	35	140	37	36	44
	240	60	344	77	65	705	74	67	77
AR(1) $\rho = 0.75$	30	5	48	36	15	80	13	13	13
	60	9	73	24	14	80	16	16	19
	90	13	89	28	17	117	20	20	25
	120	18	114	28	21	88	25	25	30
	240	35	213	45	38	222	43	43	50
AR(2)	30	16	67	70	26	55	11	12	14
	60	34	150	92	43	201	27	27	30
	90	51	228	116	60	309	50	52	50
	120	68	314	137	78	382	73	75	68
	240	136	513	188	145	1196	162	171	138

tions were generated from the AR(1) processes. In the latter case its performance was roughly on a par with that of SPEC5 and SPECT5. Its performance was slightly better than the latter estimators when ρ was small and the sample length long. The relatively good performance of the BART estimator when observations were generated from the AR(2) process is, as can be seen from Fig. 1, partly attributable to the sharp peak of the AR(1) spectrum at the origin, which makes it difficult to estimate $f_{XX}(0)$ accurately. While the bandwidth of the weighted covariance estimator is sufficiently narrow to permit accurate estimation when $f_{XX}(\lambda)$ is varying slowly near the origin, it is apparent that the bandwidth is too wide to permit accurate estimation when $f_{XX}(\lambda)$ varies rapidly near the origin. Decreasing the bandwidth does not improve the situation due to the corresponding increase in variance.

As was indicated earlier, estimators DIRECT2 and BART are related in that both employ weighted covariance estimators of the spectral density function of the observed process. The bandwidth of the spectral estimator used in BART is

$$\beta = 3\pi N^{-1/2},$$

where N is the sample length; the corresponding parameter for DIRECT2 tends to

$$\beta = \pi/10$$

as N becomes large, and is somewhat larger than this for N small. For all sample lengths examined the bandwidth of the spectral estimator employed with DIRECT2 is less than that for BART. This accounts for the fact that the variance and consequently the rms error of the DIRECT2 estimates is much higher than that of the BART estimates. As the sample length increases beyond 240 the bandwidth of the spectral estimator used with BART will become smaller than that for DIRECT2 and hence the variance of the latter will be lower. However, the rms error of estimator DIRECT2 cannot be expected to become asymptotically less than that of BART. This is a consequence of the fact that the spectral estimator used with BART was chosen to be consistent. Hence BART is also consistent, meaning that its rms error is asymptotically zero. On the other hand, the spectral estimator used with DIRECT2 is asymptotically biased because its bandwidth does not go to zero and therefore the rms error of DIRECT2 also does not tend to zero with increasing sample length.

The ARMA estimator, which employs an autocovariance function derived from a fitted time series model, appears to be the best of those considered. The biases are about the same as those of the SPEC5,

TABLE 5. Standard deviation of 1000 estimates of the "effective sample size" made with seven different estimators.

			Estimated ESS (standard deviation)						
						Power spectrum			
Model	N	True ESS	DIRECT	DIRECT2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	125.9	270.3	61.2	453	7.5	9.3	8.5
	60	33	215.8	219.6	34.5	1247	21.9	20.4	15.2
	90	49	301.6	237.8	20.8	1367	32.0	27.8	21.2
	120	65	432.7	250.9	24.1	5233	40.7	41.2	25.2
	240	130	730.5	141.5	30.7	5057	91.5	77.4	41.3
AR(1) $\rho = 0.45$	30	12	99.9	151.0	24.8	481.5	4.9	5.7	7.1
	60	23	188.0	153.9	14.6	747.2	11.0	14.2	12.0
	90	35	235.2	171.5	17.8	2692	19.2	19.1	15.1
	120	46	294.9	129.8	18.5	1139	27.5	26.3	16.4
	240	91	560.8	65.6	21.8	7114	64.9	60.5	30.2
AR(1) $\rho = 0.60$	30	8	79.9	225.0	15.6	195.4	3.6	4.3	4.8
	60	15	130.7	129.5	11.3	2093	7.3	7.5	6.8
	90	23	170.7	127.8	13.0	1761	11.4	12.6	10.7
	120	30	210.1	63.0	12.7	640.9	15.2	16.3	12.1
	240	60	446.1	36.2	16.1	3187	39.6	33.8	19.4
AR(1) $\rho = 0.75$	30	5	82.3	120.1	63.6	1402	2.7	3.1	4.0
	60	9	117.5	43.9	9.3	1342	4.4	5.2	4.8
	90	13	121.0	57.2	7.8	882.2	6.5	7.5	6.3
	120	18	160.8	24.4	8.3	355.4	9.5	10.5	6.8
	240	35	290.3	16.8	9.7	867.0	19.7	20.0	11.6
AR(2)	30	16	103.8	206.9	24.9	162	1.1	2.5	2.5
	60	34	140.4	264.5	24.0	939	15.0	11.6	7.0
	90	51	224.2	259.9	26.2	1698	28.7	29.6	15.0
	120	68	323.6	283.9	28.7	1866	38.2	38.2	19.8
	240	136	692.5	201.2	37.5	1476	87.0	99.7	32.6

SPECT5 and BART methods, as can be seen by comparing the entries of Tables 3 and 4. Table 5 shows that the variances of the estimates remain fairly constant as a function of sequence length; a desirable behavior which can be attributed to the fact that a fitted time series model provides a consistent estimator of the true model used to generate the data. From this it must follow that the derived estimator of the autocovariance function provides a consistent estimator of the true autocovariance function. Table 6 shows that the rms errors for ARMA make it competitive with the other estimators for samples of length 90 or more, and that the rms errors decrease quickly with increasing sequence length regardless of the process generating the data.

6. A test for the mean of a time series

One situation in which the ESS has been employed or suggested for use is in connection with the evaluation of climate change experiments with general circulation models. (Again, see Laurmann and Gates, 1977, and Livezey and Chen, 1983.) Clearly, in order to evaluate the outcomes of such experiments it is necessary to have a statistical procedure for testing hypotheses concerning means of sequences of time-correlated observations. In particular, Laurmann and Gates proposed the statistic

$$Y = \frac{\hat{N}'^{1/2}(\bar{x} - \mu_0)}{S} \tag{6.1}$$

to test the hypothesis

$$H_0{:}\mu = \mu_0.$$

In (6.1), $\hat{N}'$ is an estimate of the ESS, $\bar{X}$ is the sample mean, and S^2 is the sample variance computed as

$$S^2 = \sum_{j=1}^{N} (X_j - \bar{X})/N.$$

As discussed in Section 3, the terminology which has been used to describe ESS suggests that a sample of length N contains just N' independent pieces of information and, therefore, that Y should be distributed as a Student's t variate with $N' - 1$ degrees of freedom when $\mu = \mu_0$. However, the distribution of the statistic given by (6.1) can be markedly different from the Student's t distribution. To illustrate this we conducted a Monte Carlo experiment by generating samples of 1000 Y values using the SPEC5 estimate of ESS. Values of the test statistic were computed from samples of various lengths taken from both the AR(1) and AR(2)

TABLE 6. Root-mean-square error of 1000 estimates of the "effective sample size" computed as a proportion of the true ESS.

			Estimated ESS (rms error)						
						Power spectrum			
Model	N	True ESS	DIRECT	DIRECT2	ARMA	SPEC1	SPEC5	SPECT5	BART
AR(1) $\rho = 0.30$	30	17	9.09	16.4	3.68	27.1	0.51	0.62	0.65
	60	33	8.00	6.92	1.08	38.4	0.70	0.64	0.58
	90	49	7.72	5.02	0.43	28.3	0.68	0.59	0.52
	120	65	8.01	4.01	0.39	80.9	0.66	0.66	0.46
	240	130	7.12	1.14	0.24	39.5	0.74	0.61	0.37
AR(1) $\rho = 0.45$	30	12	9.90	13.2	2.18	40.4	0.62	0.68	0.88
	60	23	9.70	7.04	0.69	32.8	0.55	0.67	0.70
	90	35	8.28	5.04	0.54	77.5	0.59	0.58	0.56
	120	46	8.07	2.93	0.43	25.1	0.63	0.59	0.45
	240	91	7.71	0.78	0.25	79.0	0.75	0.70	0.41
AR(1) $\rho = 0.60$	30	8	11.54	28.76	2.11	24.8	0.99	1.06	1.14
	60	15	10.40	8.89	0.84	140	0.66	0.65	0.82
	90	23	9.00	5.70	0.62	77.0	0.58	0.61	0.70
	120	30	8.65	2.22	0.45	21.7	0.56	0.58	0.63
	240	60	8.81	0.67	0.28	54.2	0.70	0.58	0.43
AR(1) $\rho = 0.75$	30	5	18.6	24.8	12.88	281	1.69	1.92	1.79
	60	9	14.9	5.09	1.17	149	0.92	0.97	1.23
	90	13	11.0	4.55	0.67	68.3	0.73	0.79	1.04
	120	18	10.4	1.47	0.49	20.1	0.66	0.70	0.77
	240	35	9.73	0.56	0.29	25.1	0.61	0.62	0.54
AR(2)	30	16	7.23	13.36	1.68	10.4	0.32	0.29	0.20
	60	34	5.34	7.96	0.75	28.1	0.49	0.40	0.24
	90	51	5.60	5.26	0.54	33.7	0.56	0.58	0.29
	120	68	5.98	4.30	0.45	27.8	0.57	0.57	0.29
	240	136	5.90	1.53	0.28	13.4	0.67	0.78	0.24

models described previously. To check the t-distribution conjecture we estimated the probability of rejecting null hypothesis H_0, when true, with a test conducted at the nominal 5% significance level. That is, we estimated the true significance level of the test which is given by

$$\alpha' = P[|Y| > t_{(\hat{N}'-1),0.975}],$$

where $t_{(\hat{N}'-1),0.975}$ is the point on the Student's t distribution with $\hat{N}' - 1$ degrees of freedom, which is exceeded with probability 0.025. That is,

$$P[|t| > t_{(\hat{N}'-1),0.975}] = 0.05$$

when t truly does have Student's t distribution with $\hat{N}' - 1$ degrees of freedom. The results of the experiment are summarized in Table 7. Similar results will be obtained if other estimators of ESS are employed. In general we see that the distribution of statistic (6.1) is different from that conjectured, and that its distribution will depend upon how ESS is estimated.

Regardless of the method of estimating ESS it would be preferable to compare the value of the test statistics Y against a critical value chosen from its true asymptotic distribution, where the latter may be determined from the N' estimation equation. In the case of the SPEC5 estimator this is

$$N' = \frac{N\sigma^2}{2\pi f_{XX}^{(m)}(0)} \quad \text{with} \quad m = 5,$$

so that

$$Y = N^{1/2}(\bar{X} - \mu_0)[2\pi f_{XX}^{(m)}(0)]^{-1/2},$$

which is asymptotically distributed as Student's t with $2m = 10$ degrees of freedom (Brillinger, 1981). With this approach we may be certain of having the test's limiting characteristics correct (at least).

The reasons for the "sample size adjustment" not having the desired effect with respect to the distribution of the test statistic, are elaborated in the Appendix.

TABLE 7. Proportion of 1000 samples for which $|t| > t_{\hat{N}'-1,0.975}$, where $\hat{N}'$ is estimated using the SPEC5 estimator of the ESS.

	Data source	
Sample length	AR(1) ($\rho = 0.75$)	AR(2)
30	0.288	0.042
60	0.175	0.040
90	0.140	0.070
120	0.108	0.063
240	0.104	0.061

7. Summary and discussion of a spatial analog

We have shown that the ESS should be defined and interpreted carefully. It is difficult to estimate this quantity reliably, in that procedures which one would intuitively follow lead to estimates of ESS which do not rapidly converge to true "effective sample size." We have attempted to estimate ESS by a variety of techniques, with mixed success. The precisions of the estimates which are obtainable with these techniques have been seen to depend on the stochastic structure of the observed process, but only insofar as the shape of the spectral density function of the process is concerned, and in some cases on the length of sample available. As can be seen from Table 6, the rms error of estimate is not very sensitive to the value of the parameter of the AR(1) generating process. Similar sensitivity experiments were conducted with an AR(2) generating process with similar results. It appears that the precision obtainable primarily depends upon whether the power spectrum of the generating process has a ridge or a trough at the origin. We feel that we should only make the following rather limited recommendations: If it is known that the power spectrum of the observed process varies rather slowly near the origin then an estimator similar to BART appears to be a good choice. However, if this is not the case, and if a long sample is not available, it appears that the ESS cannot be estimated reliably by any of the techniques described above.

The ESS should only be used as a "diagnostic" quantity which measures the degree and gives an indication of the effect of the stochastic structure of the observed process. We have pointed out that ESS is not a measure of degrees of freedom in the sense that it has been used for identification of the distribution of a "t ratio," and should not be so used. In this connection, the question of testing the mean of a time series remains an open question in the theory of statistical inference. Thiébaux and Zwiers (1981) and Zwiers (1981) have examined some procedures for testing the mean of a time series, and Zwiers (1983) has explored implications of their use.

Finally, we would like to note that the concept of "effective sample size" may, in some instances, be extended to spatial arrays of observations. The extension makes sense when it can be reasonably assumed that the observed process is homogeneous in space. When this assumption can be made, as it can for winter 500 mb heights over North America (after adjusting for spatial inhomogeneity of variance), the ESS can give a useful indication of the scale of the spatial correlation structure of the observed process. One subject area in which the scale correlation structure has important ramifications is in the interpretation of statistical tests comparing two mean fields using techniques such as those employed by Chervin and Schneider (1976). The related problem of ascribing global levels of significance to such arrays of statistical tests has been discussed recently by Hasselmann (1979), Storch (1982), Livezey and Chen (1983) and Zwiers (1983), among others.

When the concept of "equivalent sample size" can be applied spatially, its interpretation and estimation is analogous to that of time-sequential outcomes. However, since the usual methods of estimating the spatial correlation structure of an observed field involve removal of the mean field, the bias problems discussed above will not be present. The results of the experiments described in Section 5 suggest that the best estimates of ESS for spatial arrays could be made by modeling the spatial correlation structure of the observed field as has been done by Thiébaux (1976).

Acknowledgments. Research by H. J. Thiébaux was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada Grant A9182. Research by F. W. Zwiers was supported by NSERC Grant A5448 and by the Atmospheric Environment Service of Canada.

We very much appreciated the useful comments and suggestions made by the reviewers of an earlier version of this paper. Also, we appreciate the able programming assistance of Alan Kelm.

APPENDIX

An Examination of the Sample Size Adjustment

The reasons that the sample size adjustment does not have the desired result can be appreciated if the defining property of the Student's t statistic is considered. It is the ratio of two independent random variables. The numerator of the ratio is a standardized normal variable, i.e., the difference between a normal variable and its mean, divided by the square root of its variance, namely:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}},$$

and the denominator of the t ratio is the square root of a statistically independent chi-square variable χ^2 divided by its degrees of freedom df. Hence

$$t = Z(\chi^2/df)^{-1/2}.$$

The tabulated probabilities for each member of this family of distributions, depending (only) on the degrees of freedom parameter, have been derived from these three assumptions, namely:

1) the standard normal distribution of the numerator variable,

2) the chi-square distribution of the denominator variable,

3) the statistical independence of numerator and denominator.

If one or more of these are inaccurate in practice, then the tabulated probabilities for a Student's distribution

are an inaccurate characterization of the distribution of whatever ratio has been formed. Of particular relevance here is Scheffé's (1959) demonstration of the effect of serial correlation on inferences about means, and his conclusion that this can be serious.

The classical use of a Student's distribution is for a test of the mean of a sample of statistically-independent normally distributed random variables, $X_1, \ldots, X_N$, all with the same but unknown variance σ^2. *In that case* we have the following:

1) The variance of the sample mean is $\sigma_{\bar{X}}^2 = \sigma^2/N$, so that

$$Z = \frac{\bar{X} - \mu}{(\sigma^2/N)^{1/2}};$$

2) $\chi^2 = (N-1)S^2/\sigma^2$ has a chi-square distribution with $df = N - 1$, where

$$S^2 = \frac{1}{N-1}\sum_j (X_j - \bar{X})^2;$$

3) Z and χ^2, and hence Z and $(\chi^2/df)^{1/2}$, are statistically independent.

Consequently, σ^2 cancels in the t-ratio and we obtain

$$t = \frac{(\bar{X} - \mu)/(\sigma^2 N^{-1})}{\{[(N-1)S^2/\sigma^2](N-1)^{-1}\}^{1/2}} = \frac{N^{1/2}(\bar{X} - \mu)}{S}.$$

The situation we wish to consider here differs from the foregoing because of sequential correlation among the $X_1, \ldots, X_N$. This leads to violations of the assumptions underlying the Student's distribution, and thus invalidates comparison of the ratio

$$\frac{N^{1/2}(\bar{X} - \mu)}{S}$$

with that distribution. Specifically, it can be shown that:

1) The variance of the sample mean is

$$\sigma_{\bar{X}}^2 = \sigma^2 \sum_{i,j=1}^{N} \rho(i-j)/N^2,$$

rather than σ^2/N, where $\rho(i-j)$ is the lag $(i-j)$ correlation, so that the standard normal variable is

$$Z = (\bar{X} - \mu)[\sigma^2 \sum_{i,j} \rho(i-j)/N^2]^{-1/2}.$$

2) The relevant quadratic form with a chi-square distribution, which is statistically independent of $\bar{X}$, is

$$\chi^2 = (\mathbf{X} - \bar{X}\mathbf{e})' \, \mathbb{\Sigma}^{-1}(\mathbf{X} - \bar{X}\mathbf{e}),$$

rather than $\sum_j (X_j - \bar{X})^2/\sigma^2$, where $\mathbf{e}' = (1 \cdots 1)$, $\mathbf{X}' = (X_1 \cdots X_n)$ and $\mathbb{\Sigma}$ is the variance/covariance matrix

$$(\sigma_{ij}) = [\sigma^2 \rho(i-j)].$$

3) $\bar{X}$ and S are not statistically independent.

Note that it is only when $\mathbb{\Sigma}$ is proportional to the identity matrix that

$$\sigma_{\bar{X}}^2 = \sigma^2/N \quad \text{and} \quad \chi^2 = \sum (X_j - \bar{X})^2/\sigma^2.$$

Adjusting the value of N in the ratio $(\bar{X} - \mu) \times (SN^{-1/2})^{-1}$ could compensate for (1), but not for (2) and (3). The latter is perhaps easiest to see with regard to (2), in consequence of which the replacement of χ^2 by

$$\sum_j (X_j - \bar{X})^2/\sigma^2$$

cannot be accommodated by adjustment of a "degrees of freedom" constant. A valid t-test for the mean must utilize the actual variance/covariance structure for $\mathbf{X}$, namely $\mathbb{\Sigma}$, or an independent estimate of this lag-covariance array.

In order to construct a valid t statistic for a serially correlated sample, we note that since $\mathbb{\Sigma}$ is symmetric it may be written as a product

$$\mathbb{\Sigma} = \mathbf{M}\mathbf{M}'$$

whose factors define a useful transformation matrix, namely $\mathbf{M}^{-1}$. If we now consider the transformed sample vector, namely

$$\mathbf{Y} = \mathbf{M}^{-1}\mathbf{X},$$

this will be seen to have ensemble mean $\mu_Y = \mathbf{M}^{-1}\mathbf{e}\mu$ and variance/covariance matrix

$$\mathbb{\Sigma}_Y = \mathbf{M}^{-1}\mathbb{\Sigma}\mathbf{M}'^{-1} = I.$$

Thus

$$\chi^2 = (\mathbf{Y} - \bar{Y}\mathbf{e})'\mathbb{\Sigma}_Y^{-1}(\mathbf{Y} - \bar{Y}\mathbf{e})$$
$$= \sum (Y_j - \bar{Y})^2 = (N-1)S_Y^2.$$

Furthermore, this is independent of $\bar{Y}$. Hence

$$t = \frac{N^{1/2}(\bar{Y} - \mu_Y)}{S_Y},$$

where $\bar{\mu}_Y = \mathbf{e}'\mathbf{M}^{-1}\mathbf{e}\mu/N$ satisfies the assumptions required for a valid t-test for the mean.

The foregoing is a special case of a more general result for "F-ratios" of variance estimates (since a t statistic is the square root of an F), namely that

$$\frac{(S_1^2/\sigma_1^2)}{(S_2^2/\sigma_2^2)}$$

only has the distribution known as Fisher's F distribution if it is the ratio of *independent* chi-square variables, each divided by its degrees of freedom. Usually this ratio is formed by two independent estimates of the same σ^2, so that

$$F = \frac{[\Sigma(X_{1j} - \bar{X}_1)^2/\sigma^2](N_1 - 1)^{-1}}{[\Sigma(X_{2k} - \bar{X}_2)^2/\sigma^2](N_2 - 1)^{-1}} = \frac{S_1^2}{S_2^2}.$$

Following Anderson (1958), the only way to guarantee the independence of numerator and denominator is through the (joint) covariance structure of the sample values from which the variance estimates are computed. This is to say that if

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

does not have independent components, then again it is necessary to transform to a vector which does, and formulate the test statistic in terms of the new variable. Modification of the degrees of freedom constant(s) used to select the F-distribution for comparison, cannot bring a ratio of variance estimates into conformance with an F-distribution, if the numerator and denominator are not independent quadratic forms; that is, if the ratio cannot be written as

$$\frac{(\mathbf{X}_1 - \mu_1)'\mathbf{\Sigma}_1^{-1}(\mathbf{X}_1 - \mu_1)}{(\mathbf{X}_2 - \mu_2)'\mathbf{\Sigma}_2^{-1}(\mathbf{X}_2 - \mu_2)},$$

where the variance/covariance matrix of

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{pmatrix}.$$

In general, the (off-diagonal) covariance submatrices $\mathbf{\Sigma}_{12} = \langle \mathbf{X}_1 \mathbf{X}_2' \rangle$ and Σ_{12}' in

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \Sigma_{12} \\ \Sigma_{12}' & \mathbf{\Sigma}_2 \end{pmatrix}$$

are not $\mathbf{0}$; and, again, the symmetric factorization $\mathbf{\Sigma} = \mathbf{M}\mathbf{M}'$ must be invoked to transform

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

to a vector with uncorrelated components,

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},$$

prior to constructing a valid F-ratio.

REFERENCES

Akaike, H., 1973: Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika,* **60,** 255–265.

——, 1976: Canonical correlation analysis of time series and the use of an information criterion. *System Identification: Advances and Case Studies,* R. K. Mehrs and O. G. Lainiotis, Eds., Academic Press, 27–96.

——, G. Kitasawa, E. Arahata and F. Tada, 1979: TIMSAC-78. *Computer Sci. Monogr.,* No. 11, Inst. Stat. Math, Tokyo.

Anderson, T. W., 1958: *An Introduction to Multivariate Statistical analysis.* Wiley, 374 pp.

——, 1971: The sample mean, covariances and spectral density. *The Statistical Analysis of Time Series,* Wiley, 438–500.

Box, G. E. P., and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control.* Holden-Day, 553 pp.

Brillinger, D. R., 1981: *Time series: Data Analysis and Theory.* Holden-Day, 540 pp.

Chervin, R. M. and S. H. Schneider, 1976: On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.,* **33,** 405–412.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology over the Tropical Oceans,* B. D. Shaw, Ed., Roy. Meteor. Soc., 251–259.

Jenkins, G. M., and D. G. Watts, 1968: *Spectral Analysis and its Applications.* Holden-Day, 525 pp.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.,* **14,** 159–163.

——, 1976: On estimating the variance of time averages. *J. Appl. Meteor.,* **15,** 514–515.

Koopmans, L. H., 1974: *The Spectral Analysis of Time Series.* Academic Press, 366 pp.

Laurmann, J. A., and W. L. Gates, 1977: Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *J. Atmos. Sci.,* **34,** 1187–1199.

Leith, C. E., 1973: The standard error of time-average estimates of climate means. *J. Appl. Meteor.,* **12,** 1066–1069.

Livesey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Mood, A. M., and F. A. Graybill, 1963: *Introduction to the Theory of Statistics.* McGraw-Hill, 443 pp.

Scheffé, H., 1959: *The Analysis of Variance.* Wiley, 477 pp.

Storch, H. V., 1982: A remark on Chervin–Schneider's algorithm to test significance of climate experiments with GCM's. *J. Atmos. Sci.,* **39,** 187–189.

Thiébaux, H. J., 1976: Anisotropic correlation functions for objective analysis. *Mon. Wea. Rev.,* **104,** 994–1002.

——, 1981: The kinetic energy spectrum vis a vis a statistical model for geopotential. *Tellus,* **33,** 417–427.

——, and F. W. Zwiers, 1981: Some alternative approaches to testing for the significance of differences in climate model outcomes. *Preprints Seventh Conf. on Probability and Statistics in Atmospheric Science,* Monterey, Amer. Meteor. Soc., 34–39.

Zwiers, F. W., 1981: Simulation experiments with a test for the mean of a time series for use in climate change experiments. *Preprints Seventh Conf. on Probability and Statistics in Atmospheric Science,* Monterey, Amer. Meteor. Soc., 28–33.

——, 1983: Evaluating climate change experiments. *Preprints Second Int. Meeting on Statistical Climatology,* Lisbon, World Meteor. Org., 52–57.