File copy

THE INSTITUTE OF STATISTICS

UNIVERSITY OF NORTH CAROLINA SYSTEM

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Time Series Realizations Obtained According to An Experimental Design

by

A. R. Gallant, Thomas M. Gerig, and J. W. Evans

Institute of Statistics
Mimeograph Series No. 913
Raleigh - March 1974

TIME SERIES REALIZATIONS OBTAINED ACCORDING TO AN EXPERIMENTAL DESIGN

A. R. Gallant, Thomas M. Gerig, and J. W. Evans*

*Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27607. This work was supported in part by the United States Department of Agriculture Research Agreement No. 12-18-04-8-1491-X.

ABSTRACT

The paper presents a data analysis methodology consisting of a synthesis of experimental design methods and spectral methods of time series analysis which is appropriate in exploratory situations where the recording process generates a long sequence of correlated observations.

1. INTRODUCTION

The paper considers the study of experimental material which exhibits two characteristics. The first, replication subject to a particular configuration of factors thought to affect the experimental material is possible. The second, the process of recording the phenomena under study generates a long sequence of correlated observations. Situations where the suggested methodology is appropriate are likely to be exploratory studies where knowledge of the experimental material has not advanced to the point where a few statistics can be computed from these correlated observations which are sufficient to represent the phenomena under study. One example of the type of situation we have in mind would be the study of noise in an electronic circuit where selected components of the circuit are to be varied. A second example would be an exploratory study of the effect of various processing plant configurations on an index of product quality spanning an appreciable length of time. On the other hand, the methods presented would not apply to the study of most economic time series due to the impossibility of replication.

The paper is arranged as follows. In Section 2 we suggest that a spectral or frequency domain approach is appropriate to the study of data of the type described in the previous paragraph. We also point out the statistical properties of the Schuster periodogram which dictate its use in this context. In Section 3 we define sequences indexed by frequency which correspond to analysis of variance statistics such as treatment means, F-statistics, and estimators of variance components which are appropriate to the experimental design chosen. In Sections 4 and 5 we discuss how these sequences of analysis of variance statistics may be used to study the experimental material in much the same fashion as they are used in univariate analysis of variance. Section 6 contains

comments on methods of performing the computations. Section 7 presents an example of the application of the proposed methodology in an aerial crop identification study.

2. SPECTRAL METHODS AND THE SCHUSTER PERIODOGRAM

The observed portion $\{y(t)\}_{t=0}^{n-1}$ of a time series $\{Y(t)\}_{t=-\infty}^{\infty}$ is an n dimensional random vector. However, standard multivariate statistical methods are seldom employed in time series analysis because n is usually so large as to make the computations infeasible and reasonable first and second moment assumptions allow the use of computationally feasible alternative methods.

The first moment of the time series is

$$\mu(t) = \mathcal{E}(Y(t)) \ .$$

Quite frequently, an examination of the data or a priori considerations suggest a regular and periodic trend in the data of the form

$$\mu(t + A) = \mu(t) \ .$$

A natural representation [2, p. 92] of such a function is a low order Fourier series expansion

$$\mu(t) = \mu + \Sigma_{i=1}^{p} \alpha_i \cos(\omega_i^* t + \beta_i) \ ,$$

where μ, α_i, β_i, and ω_i^* are the unknown parameters. This model is nonlinear in the parameters ω_i^*.

This paper considers the case when a Fourier series representation of the trend $\mu(t)$ is appropriate. We suggest the following approach in the case when an additional non-periodic component $\tau(\underset{\sim}{x}_t, \underset{\sim}{\theta})$ depending on inputs $\underset{\sim}{x}_t$ appears to be present in the data. Estimate $\underset{\sim}{\theta}$ by least squares obtaining an estimate $\hat{\underset{\sim}{\theta}}$ for each "cell" in the experimental design. Use standard multivariate experimental design methods to interpret the information contained in the multivariate random

variable $\hat{\underset{\sim}{\theta}}$. Use the methods suggested in this paper to extract the information remaining in the residuals

$$e(t) = y(t) - \tau(\underset{\sim}{x}_t, \hat{\underset{\sim}{\theta}}) .$$

The bulk of the useful information is contained in the second moments

$$\mathcal{E}(Y(t + h) - \mu(t + h))(Y(t) - \mu(t))$$

in most time series analysis situations. A reasonable simplifying assumption [5, p. 3 ff.] is that these covariances depend only on the gap h and not on the position t in time. Such a series is called weakly covariance stationary with autocovariance function defined by

$$\gamma(h) = \mathcal{E}(Y(t + h) - \mu(t + h))(Y(t) + \mu(t)) .$$

We shall further assume that these covariances decline to zero as $|h|$ tends to infinity sufficiently fast to cause the series $\Sigma_{h=-\infty}^{\infty} |\gamma(h)|$ to be finite. This assumption insures the existence and continuity of the Fourier series

$$f(\omega) = (2\pi)^{-1} \Sigma_{h=-\infty}^{\infty} \gamma(h) e^{-i\omega h}, \quad -\pi \leq \omega \leq \pi$$

which is called the spectral density of the time series $\{Y(t)\}_{t=-\infty}^{\infty}$. Interest is often focused on the spectral density $f(\omega)$ $(-\pi \leq \omega \leq \pi)$ rather than the autocovariance function $\gamma(h)$ $(h = 0, \pm 1, \ldots)$ in applications because it is generally easier to interpret estimators of the spectrum than estimators of the autocovariances; moreover, the spectrum has a direct physical meaning in some applications [5, p. 7]. As we shall see later, ease of interpretation becomes particularly relevant when there is a possibility of a periodic trend $\mu(t)$ in the data.

Historically, the Schuster periodogram

$$P(\omega_s) = (\tfrac{2}{n}) \left| \Sigma_{t=0}^{n-1} y_t e^{-i\omega_s t} \right|^2 ,$$

where

$$\omega_s = (2\pi/n) \cdot s ,$$

$$s = 0, 1, \ldots, m = \begin{cases} \frac{n}{2}, & n \text{ even} \\ \frac{n-1}{2}, & n \text{ odd} \end{cases},$$

was first considered in connection with the analysis of time series suspected of having a deterministic component of the form

$$\mu(t) = \mu + \Sigma_{i=1}^{p} \alpha_i \cos(\omega_i^* t + \beta_i)$$

and a second moment function $\gamma(h) = \sigma^2, h = 0$ and $= 0$ for $h \neq 0$ [4, Chapt. 6]. However, under the second moment assumptions stated previously and assuming that each $\omega_i^* = \omega_s$ for some s, one can show that

$$\mathcal{E}(P(\omega_s)) = \begin{cases} (4\pi)g(0) + 2n\mu^2 & \omega_s = 0 \\ (4\pi)g(\omega_i^*) + \frac{n}{2}\alpha_i^2 & \omega_s = \omega_i^* \\ (4\pi)g(\omega_s) & \omega_s \neq 0, \omega_i^* \end{cases},$$

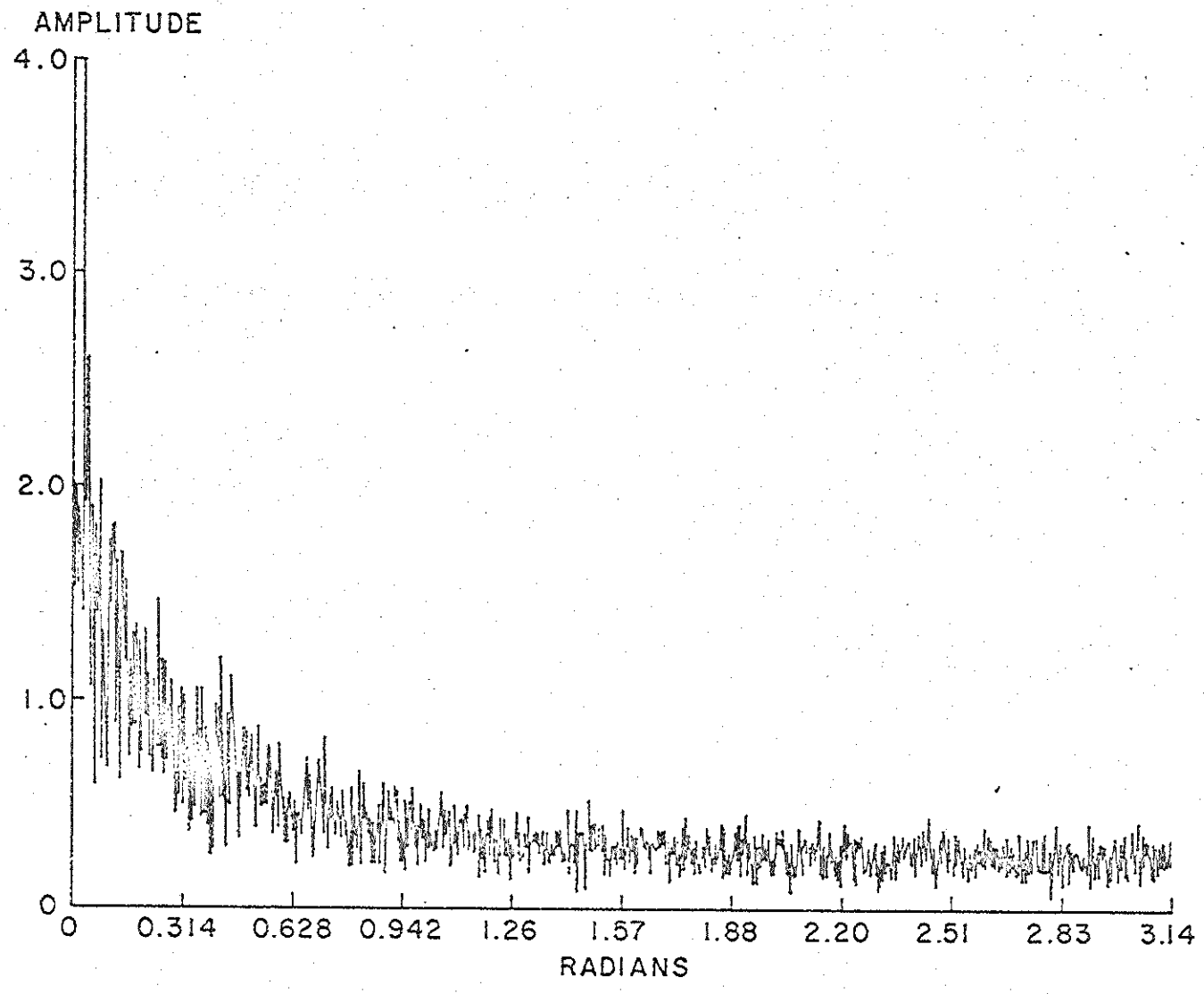
where

$$g(\omega) = (2\pi)^{-1} \Sigma_{h=-(n-1)}^{(n-1)} \frac{n-|h|}{n} \gamma(h) e^{-i\omega h}$$

and that the asymptotic distribution of $P(\omega_s)/(2\pi f(\omega_s))$ is a two degree-freedom chi-squared provided $\omega_s \neq 0, \omega_i^*$ $(i = 1, \ldots, p)$ [4, Chapt. 6]. Violation of the assumption that $\omega_i^* = \omega_s$ for some s is of negligible consequence in applications [2, p. 136-158, p. 387].

Except for sampling variation, one would expect a plot of the periodogram ordinates $P(\omega_s)$ against frequency ω_s to have a shape proportional to the spectral density $f(\omega)$ of the time series and to have spikes at those frequencies ω_i^* which are important in a low order Fourier series representation of the first moment of time series $\mu(t)$. See, for example, Figure A. Thus, the periodogram summarizes in its first moment the information contained in both the first and all second order moments of the time series. Moreover, for large time series lengths n (say $n > 512$), the sampling distribution of each periodogram ordinate is proportional to a two degree-freedom chi-squared. These are the properties -

A. TRANSFORMED PERIODOGRAM OF AN ORANGE ORCHARD SCAN



condensation of the first and second order moment information in the time series and known sampling distribution - which make the periodogram a useful representation of a time series in an exploratory study.

3. EXPERIMENTAL DESIGN CONSIDERATIONS

We have examined the considerations which indicate that the Schuster periodogram is a natural representation of time series data in an exploratory situation. However, the normal distribution is more appropriate in an analysis of variance context than a chi-squared with two degrees-freedom. For this reason, we propose to transform each periodogram ordinate by the function $\varphi_\alpha(z) = z^\alpha$, where $0 < \alpha < 1$. By means of a Taylor series expansion of $\varphi_\alpha(z)$, we argue that $\varphi_\alpha(P(\omega_s))$ will retain the desirable first moment properties of $P(\omega_s)$ and for $\alpha = 1/3$ one expects approximate normality to obtain [6, Section 12.7]. (In the experiment reported in Section 7 we found $\alpha = 1/4$ to be a more satisfactory choice.)

Subsequent to the performance of the experiment, one's data consists of a time series realization $\{y(t)\}_{t=0}^{n-1}$ for each "cell" of the experimental design. Each time series is then replaced by its transformed Schuster periodogram $\{\varphi_\alpha(P(\omega_s))\}_{s=0}^{m}$. The next step is to compute and retain the statistics which are appropriate for the design chosen for $s = 0, 1, \ldots, m$. The result of this process is one sequence of length m indexed by s corresponding to each statistic.

This process is easier to visualize from an example. Suppose that the indexing scheme for a univariate experiment carried out according to the experimental design chosen is

$x_{ijk\ell}$ $(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K;\ \ell = 1, \ldots, L)$.

The transformed periodograms of the time series realizations may be referenced

according to this same indexing scheme; viz.,

$\{\varphi_\alpha(P_{ijk\ell}(\omega_s))\}_{s=0}^{m}$ $(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K;\ \ell = 1, \ldots, L)$.

Now, fix the index s at s° and make the assignment

$$x_{ijk\ell} = \varphi_\alpha(P_{ijk\ell}(\omega_{s^\circ}))$$

and carry out the analysis of variance computations corresponding to the experimental design. Repeat this process letting s° successively assume the values $s^\circ = 0, 1, \ldots, m$ and retain: (i) estimates of treatment means which would be of interest in the corresponding univariate design, if any; (ii) F test statistics corresponding to hypotheses which would be of interest in the corresponding univariate design, if any; (iii) estimates of variance components which would be of interest in the corresponding univariate design, if any.

To be more specific, suppose that our example is a one way layout with the index i corresponding to treatments i = 1, ..., I and the remaining indices j, k, ℓ representing levels of subsampling within treatments. The model for the univariate design is

$$x_{ijk\ell} = \tau_i + \alpha_{ij} + \beta_{ijk} + \gamma_{ijk\ell},$$

where we assume $\mathcal{E}(x_{ijk\ell}) = \tau_i$ and that the errors are $NID(0,\sigma_\alpha^2)$, $NID(0,\sigma_\beta^2)$, and $NID(0,\sigma_\gamma^2)$, respectively. Natural statistics to retain would be treatment means $\hat{\tau}_i$, the F-statistic for the hypothesis H: $\tau_1 = \tau_2 = \ldots = \tau_I$ against A: $\tau_i \neq \tau_{i'}$ for some $i \neq i'$, and estimates of the variance components $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_\beta^2$, and $\hat{\sigma}_\gamma^2$. By means of the process of performing the computations to obtain these statistics for each s, we generate the corresponding sequences $\{\hat{\tau}_i(\omega_s)\}$, $\{F(\omega_s)\}$, $\{\hat{\sigma}_\alpha^2(\omega_s)\}$, $\{\hat{\sigma}_\beta^2(\omega_s)\}$, and $\{\hat{\sigma}_\gamma^2(\omega_s)\}$ where in each case the index s ranges from 0 to m.

In the remaining sections of the paper we will discuss how these sequences may be used to play the same roles in an exploratory analysis as do their univariate counterparts. These roles are estimation of treatment effects, detection of treatment differences, and identification of sources of variation.

Let us summarize the computational process before going further. The time series realizations $\{y(t)\}_{t=0}^{n-1}$ have been collected according to an experimental design. Each series is replaced by its Schuster periodogram $\{P(\omega_s)\}_{s=0}^{m}$ which is, in turn, transformed to obtain the sequences $\{\varphi_\alpha(P(\omega_s))\}_{s=0}^{m}$, one such sequence for each "cell" in the design. Then, depending on the experimental design chosen, these sequences are replaced by sequences corresponding to treatment means $\{\hat{\tau}_i(\omega_s)\}$, F-statistics $\{F(\omega_s)\}_{s=0}^{m}$, and variance component estimates $\{\hat{\sigma}_\alpha^2(\omega_s)\}_{s=0}^{m}$. On the basis of these sequences we propose to study the data.

4. USES OF TREATMENT MEANS AND F-STATISTICS

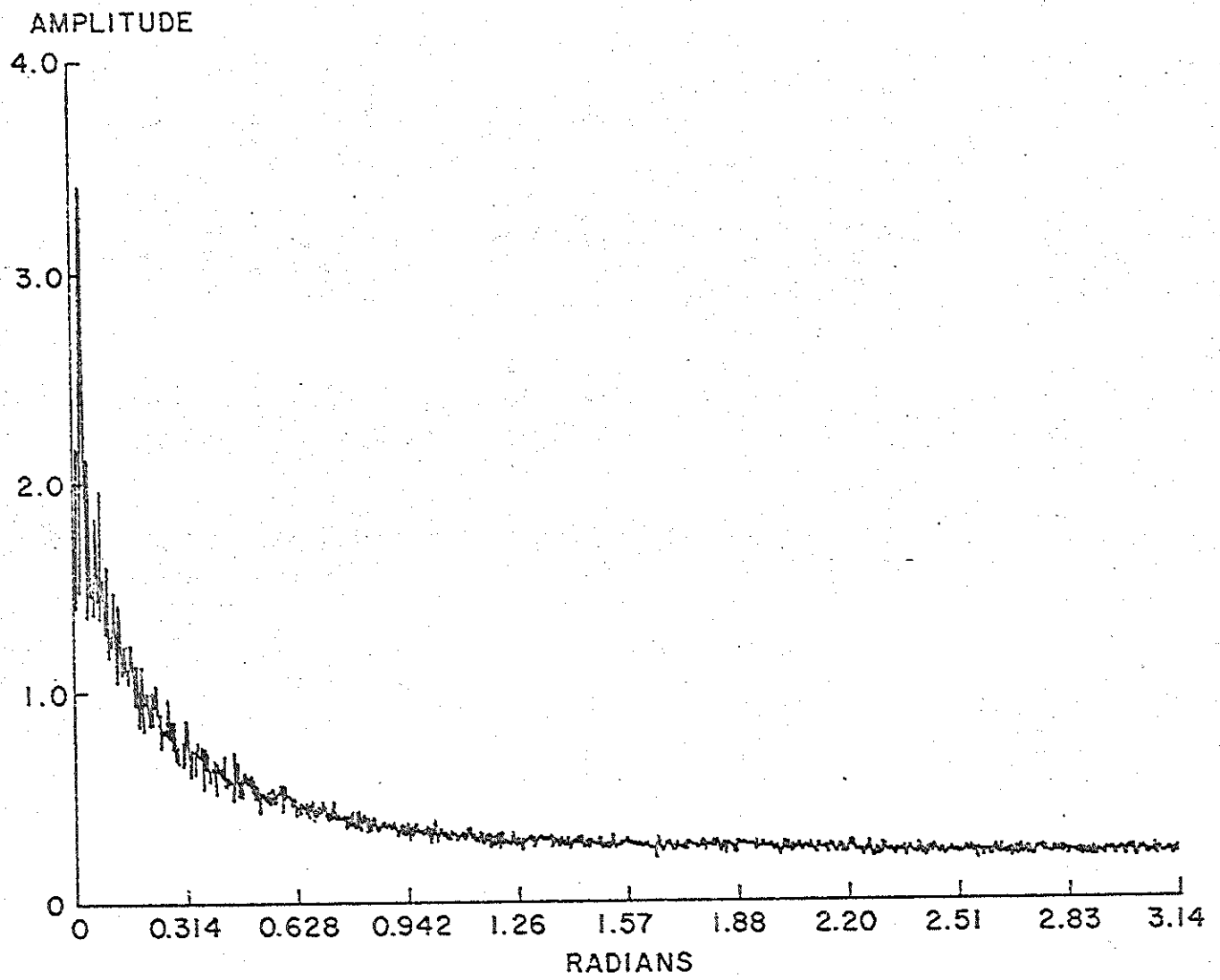
The statistics associated with the fixed effects in a univariate design are used to estimate treatment means and decide whether or not these means differ. Analogously, a set of sequences of treatment means $\{\hat{\tau}_i(\omega_s)\}$ $(i = 1, 2, \ldots, I)$ and the associated F-statistics corresponding to the null hypothesis of no treatment differences may be used to estimate the spectral density for treatment i, detect those frequencies ω^* which are important in a low order Fourier series representation of the first moment $\mu(t)$ of a realization subjected to treatment i, and decide whether or not the true means $\{\tau_i(\omega)\}$ differ on some interval $[\lambda_1, \lambda_2]$.

A sequence of treatment means $\{\hat{\tau}_i(\omega_s)\}$ is, of course, the average of the transformed periodograms $\{\varphi_\alpha(P(\omega_s))\}$ of all realizations which have been subjected to the same treatment. For the example of the previous section, we have

$$\hat{\tau}_i(\omega_s) = (JKL)^{-1} \Sigma_j \Sigma_k \Sigma_\ell \varphi_\alpha(P_{ijk\ell}(\omega_s)) .$$

The effect of this averaging is to reduce variance and gain a better estimator of the true mean $\tau_i(\omega_s)$. A visual impression of this variance reduction can be obtained by comparing Figures A and B which are plots of a single transformed

B. AVERAGE OF TRANSFORMED PERIODOGRAMS FOR ORANGES.



periodogram and the average of twelve transformed periodograms for the same treatment. The variance reduction takes place in the vertical direction in these plots and its visual effect is to "smooth" them. This type of estimator is similar to Bartlett's estimator of the spectrum [4, p. 228 ff.]. The essential difference between the estimator $\{\hat{\tau}_i(\omega_s)\}$ and Bartlett's estimator is that, here, the average is taken over transformed periodograms which are independently and identically distributed due to the design structure rather than being taken over periodograms formed from contiguous segments of the same realization which are, of course, correlated. Thus, one achieves more variance reduction than is the case with Bartlett's estimator due to the absence of correlation. In the study reported in Section 7, we found that the variance reduction due to averaging over the transformed periodograms within a treatment achieved sufficient "smoothing" for our purposes. One may achieve additional variance reduction, if desired, by moving average smoothing as explained in [4, Chapter 6].

Plots of the pairs $(\omega_s, \hat{\tau}(\omega_s))$ are interpreted in the same manner as plots of estimated spectra since φ_α is a strictly increasing function. Jenkins and Watts [5] discuss the interpretation of estimated spectra in their book, especially in Chapter 7. We give an example in Section 7 of this paper.

As one compares the plotted estimates $(\omega_s, \hat{\tau}_i(\omega_s))$ for the various treatments ($i = 1, 2, \ldots, I$) one may notice differences in shape and height in some interval $[\lambda_1, \lambda_2]$ of $[0, \pi]$. A natural question is whether or not the differences in this interval represent actual differences or sampling variation. The sequence of F-statistics $\{F(\omega_s)\}$ each with numerator degrees freedom v_1 and denominator degrees freedom v_2 may be used to answer this question.

If $\omega_s \neq \omega_{s'}$, then, asymptotically, $\varphi_\alpha(P(\omega_s))$ is independent of $\varphi_\alpha(P(\omega_{s'}))$ under reasonable assumptions [4, p. 211]. The experience of practitioners indicates that periodogram ordinates do, in fact, have an appearance of

independence in applications. This "approximate" or asymptotic independence will, of course, carry over to the sequence of F-statistics $\{F(\omega_s)\}$. Let s_1 and s_2 be such that ω_s for $s = s_1, s_1 + 1, \ldots, s_2$ are those frequencies satisfying $\omega_s \in [\lambda_1, \lambda_2]$ from the set of frequencies $\omega_s = (2\pi/m)s$, where $s = 0, 1, \ldots, m$. Suppose the hypothesis H: $\tau_i(\omega_s) = \tau_{i'}(\omega_s)$ holds for all treatment pairs $i, i' = 1, 2, \ldots, I$ and all $\omega_s \in [\lambda_1, \lambda_2]$. Then $\{F(\omega_s)\}_{s=s_1}^{s_2}$ is a sequence of independent and identically distributed central F-statistics each with v_1 numerator and v_2 denominator degrees freedom. On the other hand, if the alternative A: $\tau_i(\omega_s) \neq \tau_{i'}(\omega_s)$ holds for some $i \neq i'$ and some $\omega_s \in [\lambda_1, \lambda_2]$ then the sequence $\{F(\omega_s)\}_{s=s_1}^{s_2}$ is a mixture of central and non-central F-statistics. Any goodness of fit test which is sensitive to this departure from an assumed random sample from the central F-distribution is an appropriate test of H against A. The chi-squared goodness of fit test [10, p. 126] is a candidate. Tables of the F-distribution such as in [10, p. 529] which give a full range of lower and upper significance points may be used to set up the cell boundaries for the test.

As pointed out in Section 2, if ω_i^* is important in a low order Fourier series representation of $\mu(t)$ one expects to see a "spike" in a plot of the pairs $(\omega_s, P(\omega_s))$. This will carry over to plots of the pairs $(\omega_s, \hat{\tau}(\omega_s))$. See Figures A and B for an illustration. Thus, inspection of the plots yields a visual means of estimating the nonlinear parameters ω_i^*. Methods of confirming these visual impressions by tests of hypotheses and methods of estimating the remaining parameters of $\mu(t)$ are given in [2, Chapt. 4].

5. USES OF VARIANCE COMPONENT ESTIMATES

The statistics associated with the random effects in a univariate design are used to isolate sources of variation and estimate variances of linear combinations of the observations. Analogously, a set of sequences of variance

component estimators such as $\{\hat{\sigma}^2_\alpha(\omega_s)\}$, $\{\hat{\sigma}^2_\beta(\omega_s)\}$, and $\{\hat{\sigma}^2_\gamma(\omega_s)\}$ for the example of Section 3 may be used to isolate sources of sampling variation and estimate variances of linear combinations of transformed periodograms by frequency.

In the previous section, we mentioned that variance reduction in sequences of treatment means in addition to that achieved by the design computations could be obtained by moving average smoothing. This same consideration applies to sequences of variance estimators as well. Our experience indicates that some smoothing of variance component estimates is desirable. This is accomplished as follows. Let $\{\hat{\sigma}^2(\omega_s)\}_{s=0}^{m}$ be a sequence of variance components to be smoothed where negative values of $\hat{\sigma}(\omega_s)$ have been replaced by zeroes. The smoothed sequence $\{\tilde{\sigma}^2(\omega_s)\}$ is defined by

$$\tilde{\sigma}^2(\omega_s) = \frac{1}{2K+1} \Sigma_{j=-K}^{K} \hat{\sigma}^2(\omega_{s-j}) ,$$

where $\omega_{s-j} = \omega_{s+j}$ if $s - j < 0$ and $\omega_{s-j} = \frac{\pi}{2} + \frac{2\pi}{m} j$ if $s - j > m$. The number of points $2K + 1$ to be included in the moving average is found by choosing the smallest K such that a plot of the pairs $(\omega_s, \tilde{\sigma}^2(\omega_s))$ has a visual appearance of stability.

The smoothed estimates may be used to estimate variances of linear combinations of the transformed periodograms $\{\varphi_\alpha(P(\omega_s))\}$ such as treatment means $\{\hat{\tau}_i(\omega_s)\}$ at each frequency ω_s. An example is afforded by Figure C which is a plot of the pairs $(\omega_s, \tilde{\sigma}^2(\omega_s))$, where $\tilde{\sigma}^2(\omega_s)$ is an estimate of the variance of a single transformed periodogram $\varphi_\alpha(P_{ijk\ell}(\omega_s))$ for the design we have used for illustration, namely

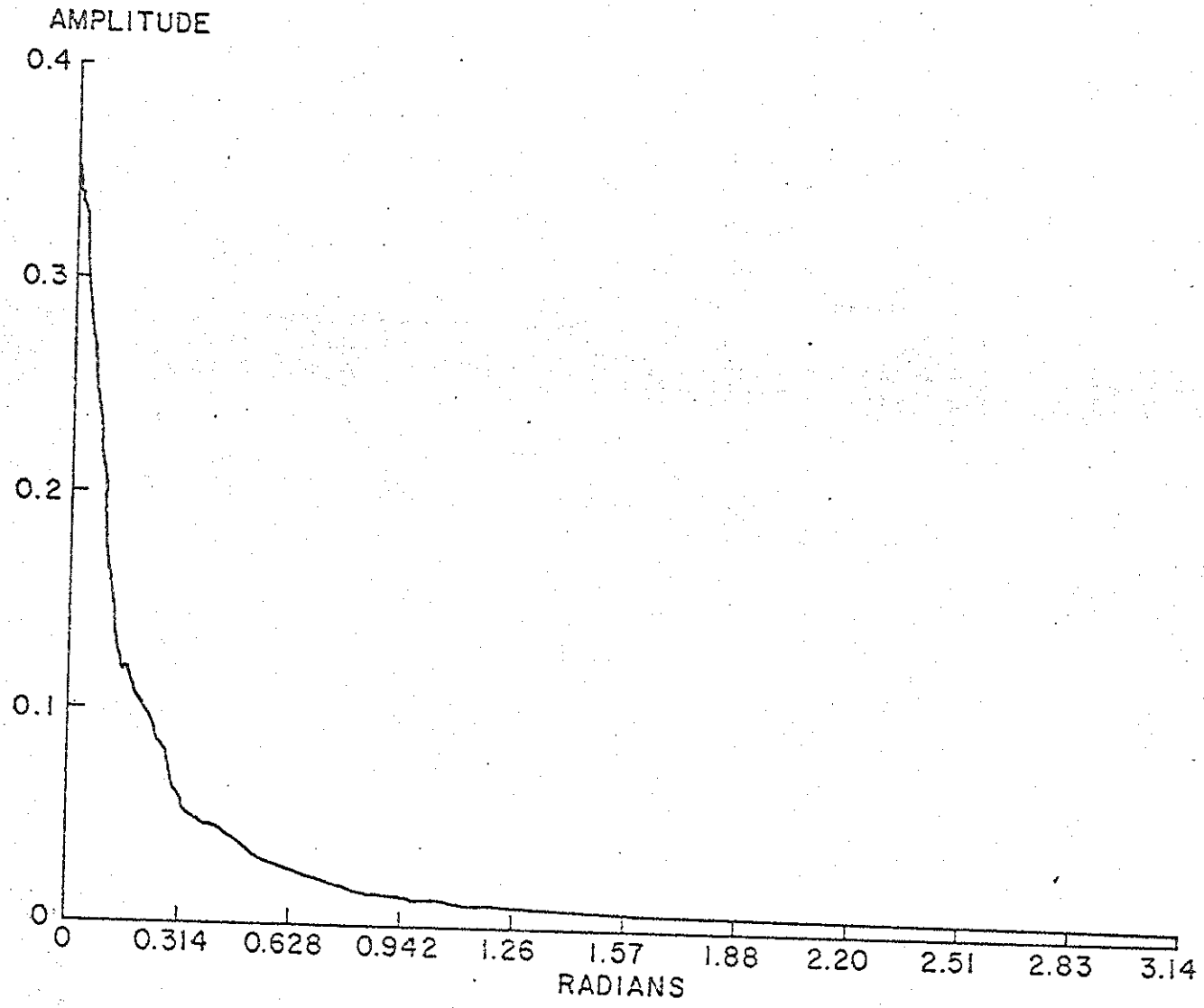
$$\varphi_\alpha(P_{ijk\ell}(\omega_s)) = \tau_i(\omega_s) + \alpha_{ij}(\omega_s) + \beta_{ijk}(\omega_s) + \gamma_{ijk\ell}(\omega_s) .$$

The estimator of this variance is

$$\tilde{\sigma}^2(\omega_s) = \tilde{\sigma}^2_\alpha(\omega_s) + \tilde{\sigma}^2_\beta(\omega_s) + \tilde{\sigma}^2_\gamma(\omega_s),$$

where the tildas over the estimators on the right hand side indicate the smoothing described in the previous paragraph ($K = 41$). As one would expect from the

C. ESTIMATED VARIANCE OF THE TRANSFORMED PERIODOGRAM OF A SCAN.



discussions in Section 2, this variance varies with frequency.

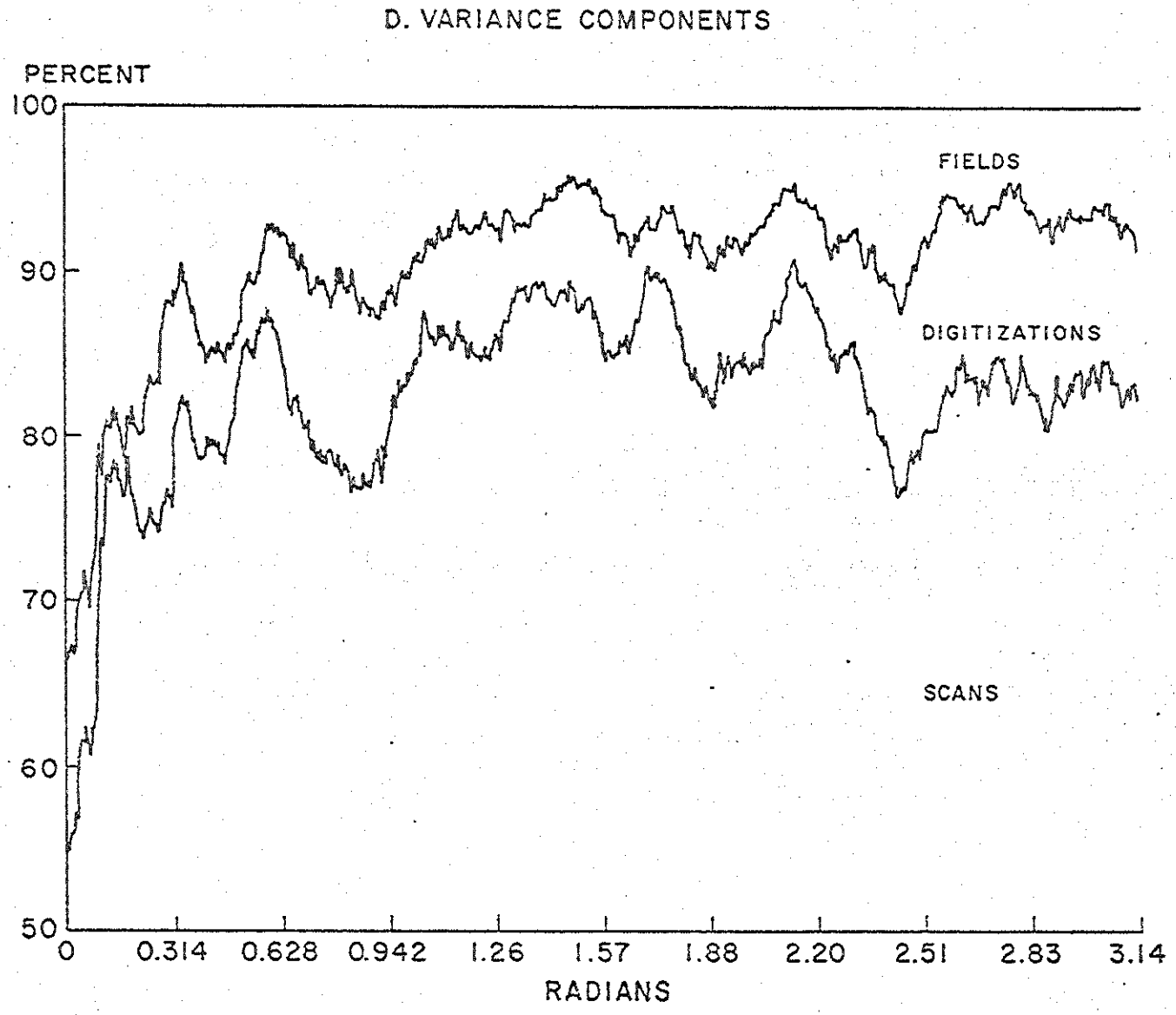
Sources of variation at various frequencies can be isolated by building up estimators of the variance of a linear combination of interest and plotting the additive contribution of each source of variation on the same grid. Figure D is an example of this sort of plot. Of interest was the relative contribution of the components $\sigma_\alpha^2(\omega_s)$, $\sigma_\beta^2(\omega_s)$, and $\sigma_\gamma^2(\omega_s)$ to their sum $\sigma^2(\omega_s)$ which is the variance of a single transformed periodogram. Plots of $(\omega_s, \tilde{\sigma}_\alpha^2(\omega_s)/\tilde{\sigma}^2(\omega_s))$, $(\omega_s, (\tilde{\sigma}_\alpha^2(\omega_s) + \tilde{\sigma}_\beta^2(\omega_s))/\tilde{\sigma}^2(\omega_s))$, and $(\omega_s, (\tilde{\sigma}_\alpha^2(\omega_s) + \tilde{\sigma}_\beta^2(\omega_s) + \tilde{\sigma}_\gamma^2(\omega_s))/\tilde{\sigma}^2(\omega_s) \equiv 1)$ on the same grid help identify the proportion of the total variance contributed by each source of variation in the experimental procedure.

6. COMPUTATIONS

The computations described in Sections 2 and 3 require an efficient means of computing the Schuster periodogram and the statistics associated with the experimental design used to obtain the time series realizations. The Fast Fourier Transform algorithm is the most efficient means to obtain the periodogram. The FORTRAN implementation of the algorithm by Singleton [11] may be used for a reasonably dense set of realization lengths n. Implementations of the Fast Fourier Transform algorithm for which n is restricted to some power of two are available at most computing centers.

The computation of the design statistics must be approached on an ad hoc basis according to the design chosen. The easiest approach is to write a subroutine in a scientific language such as FORTRAN or PL/1 which inputs data for a corresponding univariate design and outputs the statistics of interest. For our example,

$$x_{ijk\ell} = \tau_i + \alpha_{ij} + \beta_{ijk} + \gamma_{ijk\ell} ,$$


D. VARIANCE COMPONENTS
PERCENT
100
90
80
70
60
50
FIELDS
DIGITIZATIONS
SCANS
0
0.314
0.628
0.942
1.26
1.57
1.88
2.20
2.51
2.83
3.14
RADIANS

the inputs are the $x_{ijk\ell}$ ($i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$; $\ell = 1, 2, \ldots, L$) and the outputs are $\hat{\tau}_i$ ($i = 1, 2, \ldots, I$), the F-statistic for the null hypothesis of no treatment differences, $\hat{\sigma}^2_\alpha$, $\hat{\sigma}^2_\beta$, and $\hat{\sigma}^2_\gamma$. This subroutine is then placed in a loop whose index s runs from 0 to m with the assignments $x_{ijk\ell} = \varphi_\alpha(P_{ijk\ell}(\omega_s))$ made on input and the assignments $\hat{\tau}_i(\omega_s) = \hat{\tau}_i$, $F(\omega_s) =$ the F-statistic, $\hat{\sigma}^2_\alpha(\omega_s) = \hat{\sigma}^2_\alpha$, $\hat{\sigma}^2_\beta(\omega_s) = \hat{\sigma}^2_\beta$, and $\hat{\sigma}^2_\gamma(\omega_s) = \hat{\sigma}^2_\gamma$ made on output. Clearly, a simple balanced experimental design will lead to easier programming and reduced costs at this stage of the computations.

7. APPLICATION TO AERIAL CROP IDENTIFICATION

The ideas presented in the previous sections are illustrated in this section using an application to an aerial crop identification study as an example. The design according to which the realizations were collected is the same one-way layout with two levels of subsampling which we have used as an example throughout our discussion. The methods set forth in the previous sections were used to detect treatment differences for purposes of identifying statistics which can be used in discriminant analysis. A secondary objective, was to isolate the contribution to the total variance of each step of our data collection technique.

The feasibility of crop identification by sensing ground reflectance from an aerial platform is presently a subject of active investigation, notably at the Laboratory for Agricultural Remote Sensing, Purdue University, and the Forestry Remote Sensing Laboratory, University of California, Berkeley. We call the interested reader's attention to the bibliography prepared by NASA [9], the annual reports of the Purdue Laboratory [7,8], and the recent survey report by Colwell [3]. Methods of crop discrimination ranging from classification by trained inspectors using photographs taken from aircraft and spacecraft platforms to computer classification of digitally recorded optical scans are reported in

this literature. The latter approach is relevant to the present study and is perhaps best illustrated by the experiment reported in [7, p. 12]. Here, ground reflectance of selected fields was sensed over twelve bands (400-440 mμ, 440-460 mμ, 460,480 mμ, 480-500 mμ, 500-520 mμ, 520-550 mμ, 550-580 mμ, 580-620 mμ, 620-660 mμ, 660-720 mμ, 720-800 mμ, 800-1000 mμ) of the electromagnetic spectrum using a multispectral scanner mounted in an airplane. A grid of contiguous segments was sensed such that each represents approximately fifteen feet by fifteen feet of ground area. Each segment is then classified into crop catagories by means of standard multivariate discriminant methods [1., p. 126 ff.].

A considerable amount of ground detail is lost when the resolution of the sensing procedure is as coarse as the fifteen foot grid used above. The objective of this exploratory study is to determine the degree to which this lost information is relevant to crop discrimination. Specifically, this experiment was designed to determine the presence or absence of useful information other than mean field reflectance in a digital record of a high resolution optical scan.

The data were collected as follows. A designated target was photographed at an altitude of 2750 feet on Eastman 2443 color infrared film using an 8 1/4 inch focal length Ziess camera with a nine inch by nine inch format. The targets were located on the resulting transparencies and converted to a digital record by means of a microdensitometer. Each microdensitometer scan consisted of 2048 adjacent readings along a linear path at 45° to the crop rows if a row crop (or 45° to a field boundary if not). Each reading represents a ground area of 1.9 inches by 1.9 inches. The scan was repeated using red, green, and blue filters in order to be able to estimate the mean field reflectance over three bands (500-600 mμ, 600-700 mμ, 700-900 mμ) of the electromagnetic spectrum by averaging the filtered observations. The targets were selected according to the one way design shown in Table 1. Within each field there were two levels of subsampling; two

TABLE 1. Experimental Design

i	Crops	Number of Fields
1	Cotton	$j = 1, \ldots, J_1 = 2$
2	Grapes	$j = 1, \ldots, J_2 = 2$
3	Oranges	$j = 1, \ldots, J_3 = 3$
4	Almonds	$j = 1, \ldots, J_4 = 3$
5	Alfalfa	$j = 1, \ldots, J_5 = 2$
6	Corn	$j = 1, \ldots, J_6 = 2$
7	Walnuts	$j = 1, \ldots, J_7 = 3$

Subsampling within Fields

Digitizations: $k = 1, \ldots, K = 2$

Scans: $\ell = 1, \ldots, L = 2$

digitizations within each field and two scans within each digitizations. The subsampling was undertaken to isolate sources of error in the data collection technique to enable variance reduction in subsequent experiments.

Most crops are arranged in rows of equal spacing so we would expect that the expectation $\mu(t)$ of a realization would have either a sinusoidal or square wave pattern. Either of these models can be adequately approximated by the low order Fourier series expansion

$$\mu(t) = \mu + \Sigma_{i=1}^{p} \alpha_i \cos(\omega_i^* t + \beta_i) .$$

Moreover, either a square wave or sinusoidal wave of the type we envisage would satisfy $\omega_i^* = \omega^* \cdot i$ and $\alpha_1 = \max\{\alpha_i\}$. An orchard with twenty foot spacing would yield $\omega^* = (2\pi/240)(1.9/\sqrt{2}) = .0352$ radians and corn planted in thirty inch rows would yield $\omega^* = (2\pi/30)(1.9/\sqrt{2}) = .281$ radians using the formulae given by Jenkins and Watts [5, p. 52] and Anderson [2, p. 387].

Following the methodology presented in the previous sections, the transformed periodograms $\{\varphi_{\frac{1}{4}}(P_{ijk\ell}(\omega_s))\}_{s=0}^{1024}$ were obtained from the realizations $\{y_{ijk\ell}(t)\}_{t=0}^{2047}$ and the sequences $\{\hat{\tau}_i(\omega_s)\}$, $\{F(\omega_s)\}$, $\{\hat{\sigma}_\alpha^2(\omega_s)\}$, $\{\hat{\sigma}_\beta^2(\omega_s)\}$, and $\{\hat{\sigma}_\gamma^2(\omega_s)\}$ appropriate to the model

$$\varphi_{\frac{1}{4}}(P_{ijk\ell}(\omega_s)) = \tau_i(\omega_s) + \alpha_{ij}(\omega_s) + \beta_{ijk}(\omega_s) + \gamma_{ijk\ell}(\omega_s)$$

were computed. The analysis of variance table for this model is given in Table 2. An example of a transformed periodogram for a single scan over an orange grove is afforded by Figure A. The average of the transformed periodograms over all orange groves, twelve in number, is plotted in Figure B. Notice the variance reduction in the vertical direction achieved by the design computation and the peak corresponding to 250 inch row spacing.

The smoothed estimate of the variance corresponding to the transformed periodogram of a single realization $\sigma^2(\omega)$ is plotted against frequency in Figure

TABLE 2. ANOVA

Source	d.f.		Expected Mean Square
Mean	1	= 1	
Crops	I-1	= 6	$\sigma_\gamma^2(\omega) + 2\sigma_\beta^2(\omega) = 4\sigma_\alpha^2(\omega) + 4/6\ \Sigma_{i=1}^{I} J_i[\tau_i(\omega)-\tau(\omega)]^2$
Fields	$\Sigma_{i=1}^{I}(J_i-1)$	= 10	$\sigma_\gamma^2(\omega) + 2\sigma_\beta^2(\omega) + 4\sigma_\alpha^2(\omega)$
Digitizations	$\Sigma_{i=1}^{I}J_i(K-1)$	= 17	$\sigma_\gamma^2(\omega) + 2\sigma_\beta^2(\omega)$
Scans	$\Sigma_{i=1}^{I}J_i K(L-1)$	= 34	$\sigma_\gamma^2(\omega)$
TOTAL	$\Sigma_{i=1}^{I}J_i KL$	= 68	

C and the relative contribution of the variance components $\sigma^2_\alpha(\omega)$, $\sigma^2_\beta(\omega)$, and $\sigma^2_\gamma(\omega)$ is shown in Figure D. (See Section 5 for a more complete description of these plots.) As seen from Figure D, in all but the lowest frequencies, about seventy-five percent of the total variance is due to variations in realizations at the lowest level of the design. Even at the lowest frequencies $\sigma^2_\alpha(\omega)$ accounts for at least fifty percent of the total variance. For our purposes this is a fortunate discovery as replication at the lowest level of the design is the most economical. It is only necessary to take multiple scans within a given target and average the resultant transformed periodograms to achieve substantial variance reduction.

A visual inspection of overlaid plots of the pairs $(\omega_s, \hat{\tau}_i(\omega_s))$ for $i = 1, 2, \ldots, 7$ indicated substantial differences in the location and height of maxima and more subtle differences in shape of the curves in the intervals $[0,.25]$ and $[.25,1.0]$. Also, the curves appear to essentially reach an asymptote at 1.0 and the asymptote appears to differ for a few curves. These visual impressions are confirmed by the chi-squared goodness of fit test described in Section 4.

As stated earlier, our objective is to discover additional information in a digitized optical scan which may be used to improve classification error rates. Based on this study, the following statistics appear relevant. For a given realization, denote the largest periodogram ordinate by $P(\omega_L)$ and the corresponding abscissa by ω_L. The information as to the presence or absence of a deterministic component (row crop) in a realization is contained in the statistic

$$T = P(\omega_L)/(\Sigma_{s=0}^{n/2} P(\omega_s) - P(\omega_L) - P(\omega_0))$$

[2, p. 102 ff.]. The estimated row spacing in inches of a row crop is given by the statistic

$$R = (2\pi/\omega_L)(1.9/\sqrt{2}) .$$

We propose to detect shape differences by fitting the segmented linear model

$$\Psi(\omega) = \begin{cases} a + S_1\omega & 0 \le \omega \le .25 \\ a + S_1(.25) + S_2(\omega - .25) & .25 \le \omega \le 1.00 \\ a + S_1(.25) + S_2(.75) + S_3(\omega - 1.00) & 1.00 \le \omega \le 3.14 \end{cases}$$

to $\varphi(P(\omega_s))$ by a robust method and retain the statistics a, S_1, S_2, and S_3. Our first choice of methods for fitting the model was to minimize the sum of absolute deviations. Because of its relatively high cost per fit and its anticipated application to large data sets, we rejected the method in favor of a more economical one. This consisted of fitting the model using least squares and setting observations with excessively large residuals equal to their predicted values and refitting to this adjusted data. We feel that this procedure substantially reduced the effect of extreme observations of the sort seen in Figure A.

One of the goals of this investigation was to demonstrate that use of the information available in high resolution scans can greatly decrease the rate of misclassification in crop discrimination. To effect the comparison, we extracted mean reflectance for each scan by averaging 2048 readings representing 3.61 square inches of ground area to obtain the equivalent of 51.3 square feet of ground area. This procedure was applied to scans obtained using a red, a green, and a blue filter, and, using no filter. Taken together, these four measurements represent the low resolution information. High resolution information is contained in T, R, a, S_1, S_2, and S_3 defined above.

Linear discriminant function methods were used to compare classification performance. Upper bounds for P_j = the probability of misclassifying an observation from crop j were estimated and are reported in Table 3. The estimated bounds were derived using the Bonferroni inequality and the methods described in Anderson [1, p. 126 ff.]. The bounds are given for low resolution information alone, high resolution information alone, and for the combined information.

Table 3 gives strong evidence that the combined data from high and low resolution scans substantially reduces the probabilities of misclassification as compared to that of the low resolution scans. Also of interest is the fact that the high resolution variables seem to outperform the low.

TABLE 3 Classification Comparisons

Crop	Resolution		
	Low	High	Combined
Cotton	.573	.171	.042
Grapes	.528	.194	.059
Oranges	.580	.143	.058
Almonds	.572	.239	.077
Alfalfa	.474	.095	.015
Corn	.355	.085	.025
Walnuts	.419	.173	.053

ACKNOWLEDGEMENT

The authors wish to express their appreciation to Messrs. Harold Huddleston, William Wigton, and Donald Von Stein, Statistical Crop Reporting Service, United States Department of Agriculture, for their assistance and support of the study reported in Section 7.

REFERENCES

[1] Anderson, T.W. An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, Inc., 1958.

[2] Anderson, T.W. The Statistical Analysis of Time Series. New York: John Wiley and Sons, Inc., 1971.

[3] Colwell, R.N. Monitoring Earth Resources from Aircraft and Spacecraft. Washington, D.C.: U.S. Government Printing Office, 1971.

[4] Fuller, W.A. Introduction to Statistical Time Series. Ames, Iowa: University Bookstore, Iowa State University, 1971.

[5] Jenkins, G.M. and Watts, D.G. Spectral Analysis and Its Applications. San Francisco: Holden-Day, 1968.

[6] Kendall, M.G. The Advanced Theory of Statistics. London: Charles Griffin and Company Limited, 1943.

[7] Laboratory for Agricultural Remote Sensing. Remote Multispectral Sensing in Agriculture, Vol. 3. Lafayette, Indiana: Purdue University Agricultural Experiment Station, Research Bulletin No. 844, 1968.

[8] Laboratory for Agricultural Remote Sensing. Remote Multispectral Sensing in Agriculture. Lafayette, Indiana: Purdue University Agricultural Experiment Station, Research Bulletin No. 873, 1970.

[9] N.A.S.A. Remote Sensing of Earth Resources, a Literary Survey With Indexes. Springfield, Virginia: National Technical Information Service, 1970.

[10] Ostle, B. Statistics in Research, Second Edition. Ames, Iowa: Iowa State University Press, 1963.

[11] Singleton, R.C. An Algorithm for Computing the Mixed Radix Fast Fourier Transform. I.E.E.E. Transactions on Audio and Electroacoustics, AU-17, (June), 1969.

INSTITUTE OF STATISTICS

MIMEOGRAPH SERIES

(These are available at 1¢ per page)

846. SEN, P. K. Weak convergence of multidimensional empirical processes for stationary phi-mixing processes. 17 pp.
847. SMITH, WALTER J. and SUJIT BASU. General moment functions and density version of the central limit theorem. 31 pp.
848. CHAND, NANAK. Sequential tests of composite hypotheses. Thesis. 129 pp.
849. SUCHINDRAN, C. M. Estimators of parameters in biological models in human fertility. 96 pp.
850. GOODNIGHT, JAMES HOWARD. Quadratic unbiased estimation of variance components in linear models with an emphasis on the one-way classification. 54 pp.
851. CHATTERJEE, S. K. Rank procedures for some two-populations multivariate extended classification problems. 20 pp.
852. RAGHAVARAO, D. Boolian sums of sets of certain designs. 9 pp.
853. NIEDERREITER, H. and WALTER PHILIPP. Berry-Esseen bounds and a theorem of Erdos and Turan on uniform distribution Mod. 1. 26 pp.
854. EDELMAN, ANTHONY. Estimation of the mean and variance components in some random effects models with composite samples. Thesis 81 pp.
855. RAGHAVARAO, D. and RAJINDER SINGH. Applications of PBIB designs in cluster sampling. 10 pp.
856. BROOK, RICHARD J. On the use of a minimax regret function to set significance points in prior tests of estimation. Thesis 82 pp.
857. WEGMAN, EDWARD J. Computer graphics in undergraduate statistics. 18 pp.
858. BARIZI. An assessment and some applications of the empirical Bayes approach to random regression models. Thesis 1973 85 pp.
859. SHACHTMAN, RICHARD AND CURTIS MCLOUGHLIN. Mathematical programs for a mathematical model of the family planning process. 15 pp.
860. SHACHTMAN, RICHARD. An optimization scheme for variable capacity networks. 24 pp.
861. HUNTER, JEFFREY J. Renewal theory in two dimensions: basic results. 20 pp.
862. SEN, P. K. and MALAY GHOSH. A Chernoff-Savage representation of rank order statistics for stationary ϕ-mixing processes. 29 pp.
863. RAGHAVARAO, D. and K. J. . SMITH. Partial ridge regression. 8 pp.
864. KOCH, GARY G. and H. DENNIS TOLLEY. A generalized modified X^2 analysis of categorical data from a complex dilution experiment. 13 pp.
865. RAJPUT, BALRAM. Equivalent Gaussian measures whose R-N derivative is the exponential of a diagonal form. 26 pp.
866. JOHNSON, N .L. Robustness of certain tests of censoring of extreme sample values. 12 pp.
867. SEN, P. K. On unbiased estimation for randomized response model. 21 pp.
868. HUNTER, JEFFREY J. Renewal theory in two dimensions: asymptotic results. 29 pp.
869. CHAKRAVARTI, I. M. On random search using binary systems derived from finite projective planes. 10 pp.
870. CHAKRAVORTI, S. R. On some tests of growth curves models under Behren-Fisher's situation. 27 pp.
871. JOHNSON, N. L. Return to repetitions. 11 pp.
872. CASTILLO-MORALES, ALBERTO. Drawing an optimal tree from a distance matrix. Thesis.
873. JOHNSON, N. L. and S. KOTZ. A vector valued multivariate hazard rate — I.
874. GUALTIEROTTI, A. and S. CAMBANIS. Some remarks on the equivalence of Gaussian processes. 11 pp.
875. GALLANT, A. R. Inference for non-linear models.
876. JEFFCOAT, COLIN E. Some related queuing models with dependent service and inter arrival time.
877. FELDMAN, JOSEPH GERALD. Evaluating the effect of group therapy in the prognosis of coronary patients. Thesis 94 pp.
878. MUSTAFA, AHMED F.M. Fecundability differentials among acceptors and nonacceptors of family planning: A simulation experiment. Ph.D. Thesis. 90 pp.
879. NOUR, L. SAID. A stochastic model for the study of human fertility.
880. SYMONS, M. Bayes' modification of some clustering criteria.
881. MAHMOUD, MAHMOUD RIAD. Sequential decision procedures for testing hypotheses concerning general estimable parameters. Thesis
882. KOONG, LING JUNG and H. L. LUCAS.. A mathematical model for the joint metabolism of nitrogen and energy. Thesis
883. JOHNSON, N. L. and S. KOTZ. A vector-valued multivariate hazard rate: II.
884. DAVIES, H. I. and EDWARD J. WEGMAN. Sequential nonparametric density estimation.
885. O'FALLON, JUDITH. Discriminate analysis under truncation. (Ph.D. thesis)
886. LEADBETTER, M. R. On extreme values in stationary sequences.
887. LEADBETTER, H. R. Locating the maximum of a stationary normal process.
888. SIMONS, GORDON D. Generalized cumulative distribution functions II: The σ-lower finite case.
889. GALLANT, A. R. The power of the likelihood ratio test of location in nonlinear regression models.