

AD-A079 737

WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER

F/G 12/1

A STUDY OF REAL DATA.(U)

OCT 79 G CHEN, G E BOX

DAA629-75-C-0024

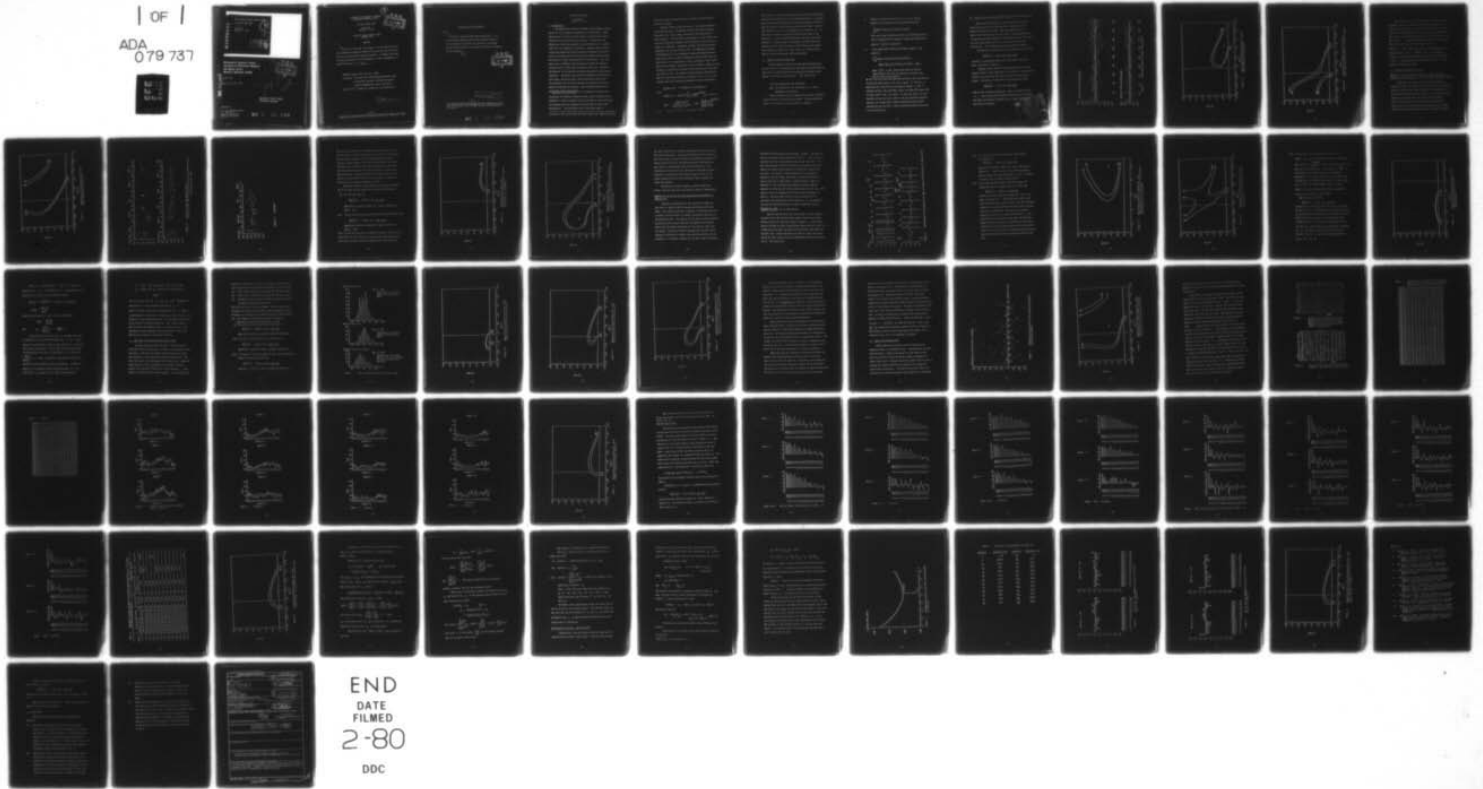
UNCLASSIFIED

MRC-TSR-2002

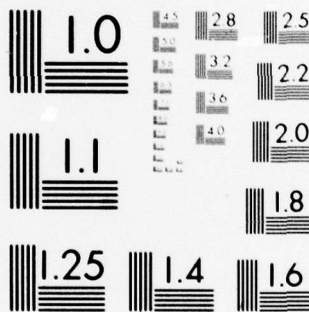
NL

| OF |

ADA
079 737



END
DATE
FILMED
2-80
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ADA 079737

MRC Technical Summary Report #2002

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

3

LEVEL

Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

DDC
RECEIVED
JAN 23 1980
E

[Handwritten signature]

October 1979

Received July 27, 1979

DDC FILE COPY

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

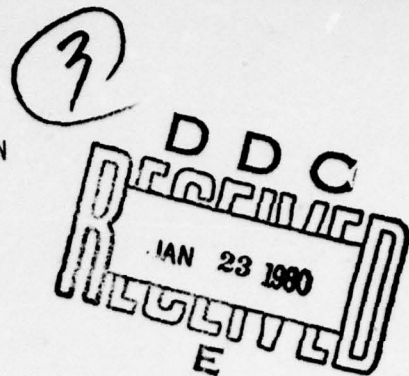
80 1 15 056

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

Technical Summary Report #2002
October 1979



ABSTRACT

→ Nine sets of real data are analyzed. The distributions within the contaminated exponential power family which best describe these data sets are obtained by maximum likelihood. It appears that heavy tailed distributions are often produced by secular inhomogeneity in mean and variance. → to page - B -

AMS(MOS) Subject Classification: 62G35

Key Words: real data sets, contaminated exponential power distribution, maximum likelihood estimation, secular inhomogeneity, serial correlation

Work Unit #4 - Probability, Statistics, and Combinatorics

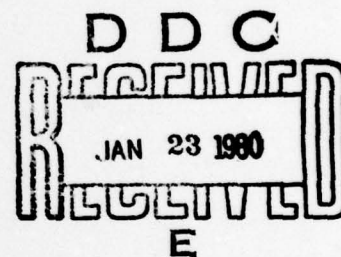
This document has been prepared for public release and under its distribution is unlimited.

-A-

SIGNIFICANCE AND EXPLANATION

cont.

A specific robust estimator would behave well if, in the real world, data occurred of the kind which favored it. In this paper more classical sets of data are considered and their characteristics are studied in relation to proposed robust estimators.



This document has been approved for public release and sale; its distribution is unlimited.

-B-

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.

80 1 15 056

A STUDY OF REAL DATA

Gina Chen
George E. P. Box

1. Introduction.

The appropriate procedure to adopt in data analysis depends heavily on how the data are supposed to be distributed. Exact distributions of data are rarely known since errors are usually contributed to by many different factors. The central limit theorem has often been appealed to by statisticians as a justification for the assumption of normality. However, more recently it has been argued that heavier-tailed distributions and distributions containing outliers might be more likely than the normal distribution to describe the behavior of real errors. New estimators have, therefore, been proposed which would be expected to behave well if these other assumptions were true. Studies of the behavior of the proposed estimators have usually been made by simulation. Typically, some feared distribution or distributions were chosen from which pseudo random samples have been drawn. The estimators have then been computed and various of their properties calculated and compared. Most frequently, efficiencies $(\frac{1/(n \text{ times Fisher information})}{\text{variance of the estimator}})$ (or asymptotic variances) were calculated for different estimators under various distribution assumptions, a robust estimator would be one having a high efficiency (or comparatively low asymptotic variance) over all the tested distributions. The difficulty here is that such study is

necessarily rather subjective since it depends on what distributions are included.

Stigler (1977) raised the question "Do robust estimators work with real data?" He applied different proposed estimators to several sets of real data, and measured the performance of each estimator. With the data he considered, the 10% trimmed mean was one of the best estimators and the ordinary sample mean came close as a competitor. Estimators selected from the best modern contenders did not come out very well. As commented by Stigler "the alternatives to an independent identically normally distributed sample that have been considered by modern workers are too restricted or too exaggerated to reflect accurately 'real' data." Later in this chapter, therefore, we shall examine a number of sets of data as a contribution towards finding out what data from the real world is like. As a means of approximately characterizing data sets and relating them to the results of TSR #1997, we shall consider again the contaminated exponential power distribution

$$P_c(y|\theta, \sigma, \beta, \alpha) = (1-\alpha)P(y|\theta, \sigma, \beta) + \alpha P(y|\theta, k\sigma, \beta)$$

with

$$P(y|\theta, \sigma, \beta) = w(\beta)\sigma^{-1} \exp\left[-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right] \quad -\infty < y < \infty$$

$$w(\beta) = \frac{\{\Gamma[\frac{3}{2}(1+\beta)]\}^{1/2}}{(1+\beta)\{\Gamma[\frac{1}{2}(1+\beta)]\}^{3/2}} \quad c(\beta) = \left\{\frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]}\right\}^{1/(1+\beta)}$$

and $\sigma > 0, \quad -\infty < \theta < \infty, \quad -1 < \beta \leq 1, \quad k > 1.$

This provides a reasonably broad family of symmetric distributions containing both light-tailed and heavy-tailed members. To characterize the parent distribution for each set of data we shall fit this distribution to the data by maximum likelihood. There are four parameters $\theta, \sigma, \alpha, \beta$ altogether (k is fixed at 3). In this chapter, however, we are particularly interested in two parameters, α and β , which characterize the shape of the distribution. For each data set, therefore, maximum likelihood estimates for all four parameters will be obtained and approximate confidence regions for (β, α) with θ, σ fixed at their maximum likelihood estimates will be calculated.

2. Analyzing the Real Data Sets

In this study of searching for the parent distributions of real data sets, one must keep in mind that we are assuming the data set is a random sample from some particular parent distribution in the family $P_c(y|\theta, \sigma, \beta, \alpha)$. This implies that

- (i) The observations are independent.
- (ii) The observations are stationary, i.e., have a fixed mean and variance.

These assumptions must be checked, because their violation could invalidate the estimation procedure. If the assumptions appear to be satisfied, one can proceed as follows:

1. Maximum likelihood estimates for $\theta, \sigma, \beta, \alpha$ may be obtained by maximizing the log likelihood function,

$$\sum_{i=1}^n \log[(1-\alpha)P(y_i|\theta, \sigma, \beta) + \alpha P(y_i|\theta, 3\sigma, \beta)].$$

This may be done by evaluating this log likelihood function over a suitable grid of parameter values and obtaining the maximum $(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha})$ numerically.

2. A very approximate $100(1-\epsilon)\%$ confidence region is now given by

$$-2 \sum_{i=1}^n \{\log[(1-\alpha)P(y_i|\hat{\theta}, \hat{\sigma}, \beta) + \alpha P(y_i|\hat{\theta}, 3\hat{\sigma}, \beta)] - \log[(1-\hat{\alpha})P(y_i|\hat{\theta}, \hat{\sigma}, \hat{\beta}) + \hat{\alpha} P(y_i|\hat{\theta}, 3\hat{\sigma}, \hat{\beta})]\} < \chi_{\epsilon}^2(2)$$

where $\chi_{\epsilon}^2(2)$ is the 100ϵ percent point of the chi-square distribution with two degrees of freedom since we have fixed two parameters.

We will thus have some idea about what positions the distributions of real data might take in the (β, α) space. In TSR#1997, we mentioned that trade offs can be expected between α and β yielding oblique ridgy confidence regions oriented from upper left to lower right in the β, α plane. It turns out that it is also possible to have two local maxima in the log likelihood function surface, one representing a highly contaminated short-tailed distribution and the other a heavy-tailed distribution with small or no contamination.

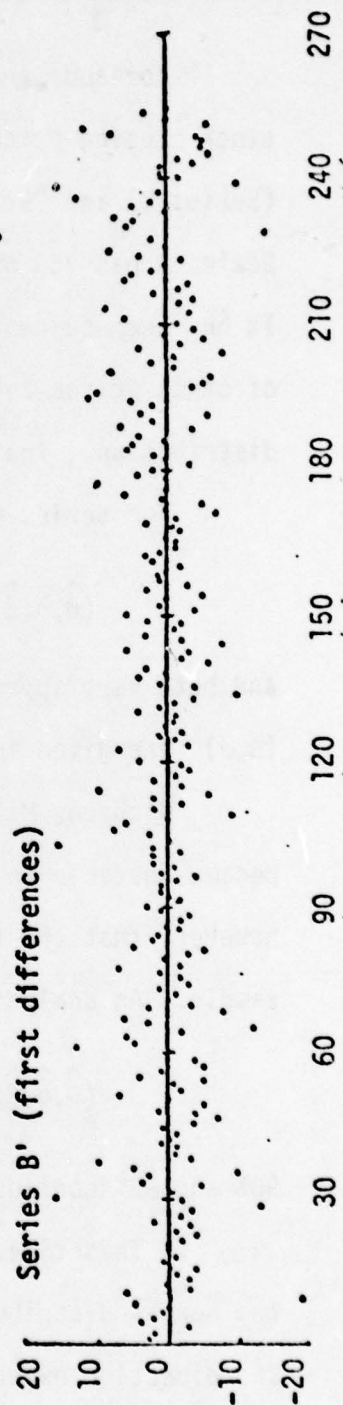
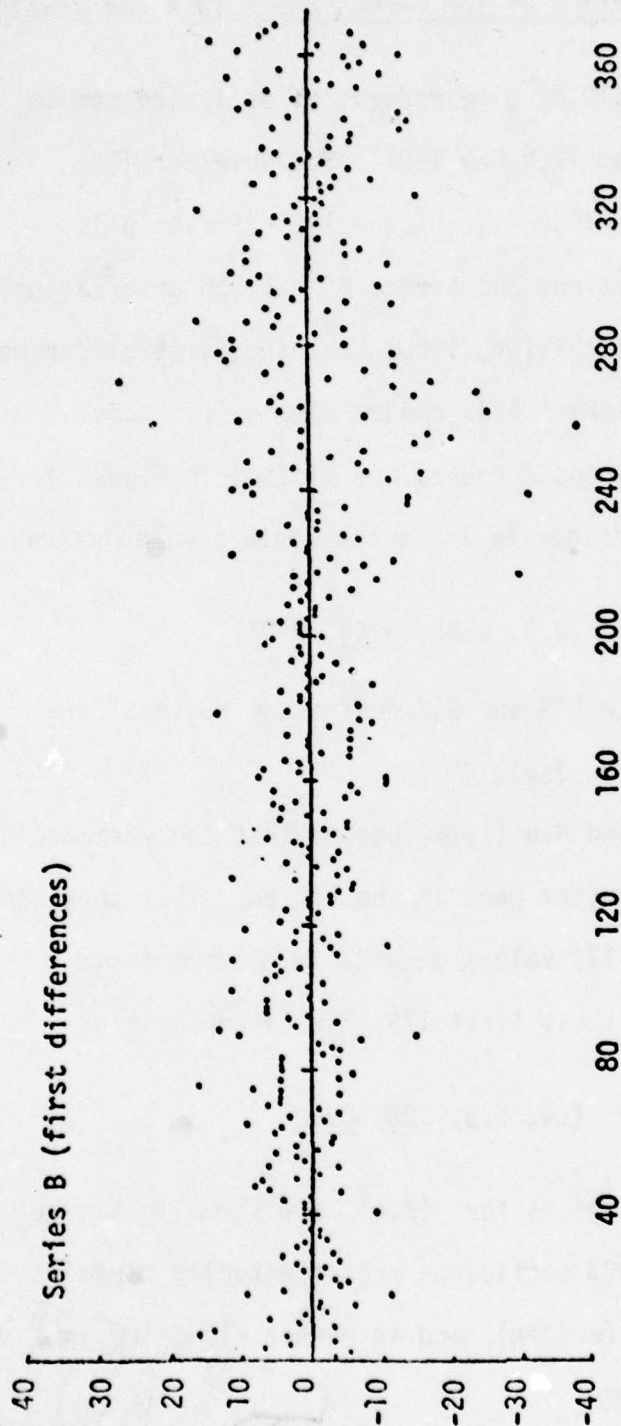


Figure 1. First Differences for Series B and Series B'

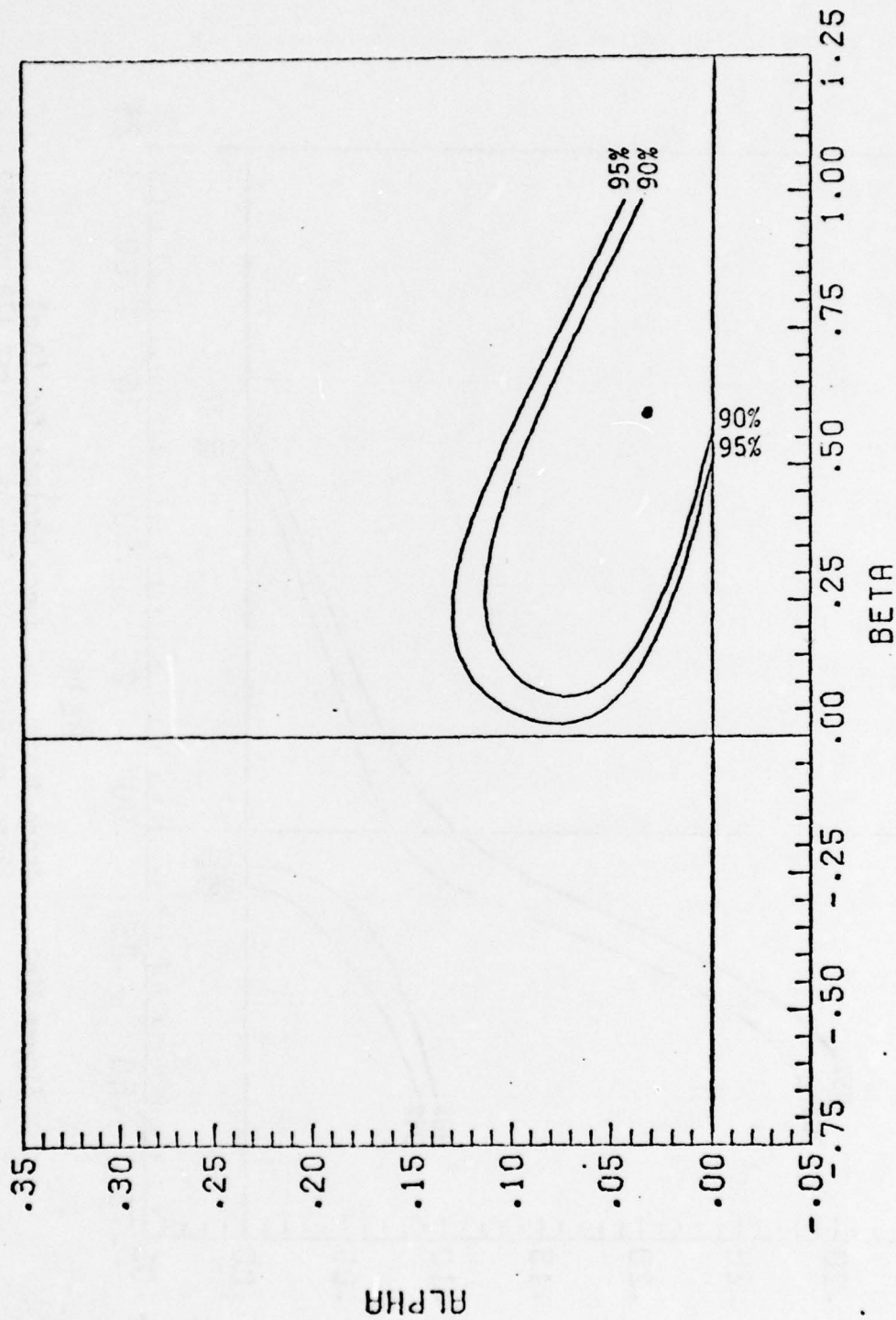


Figure 2(a). Approximate confidence regions for (β, α)
(First differences of Series B)

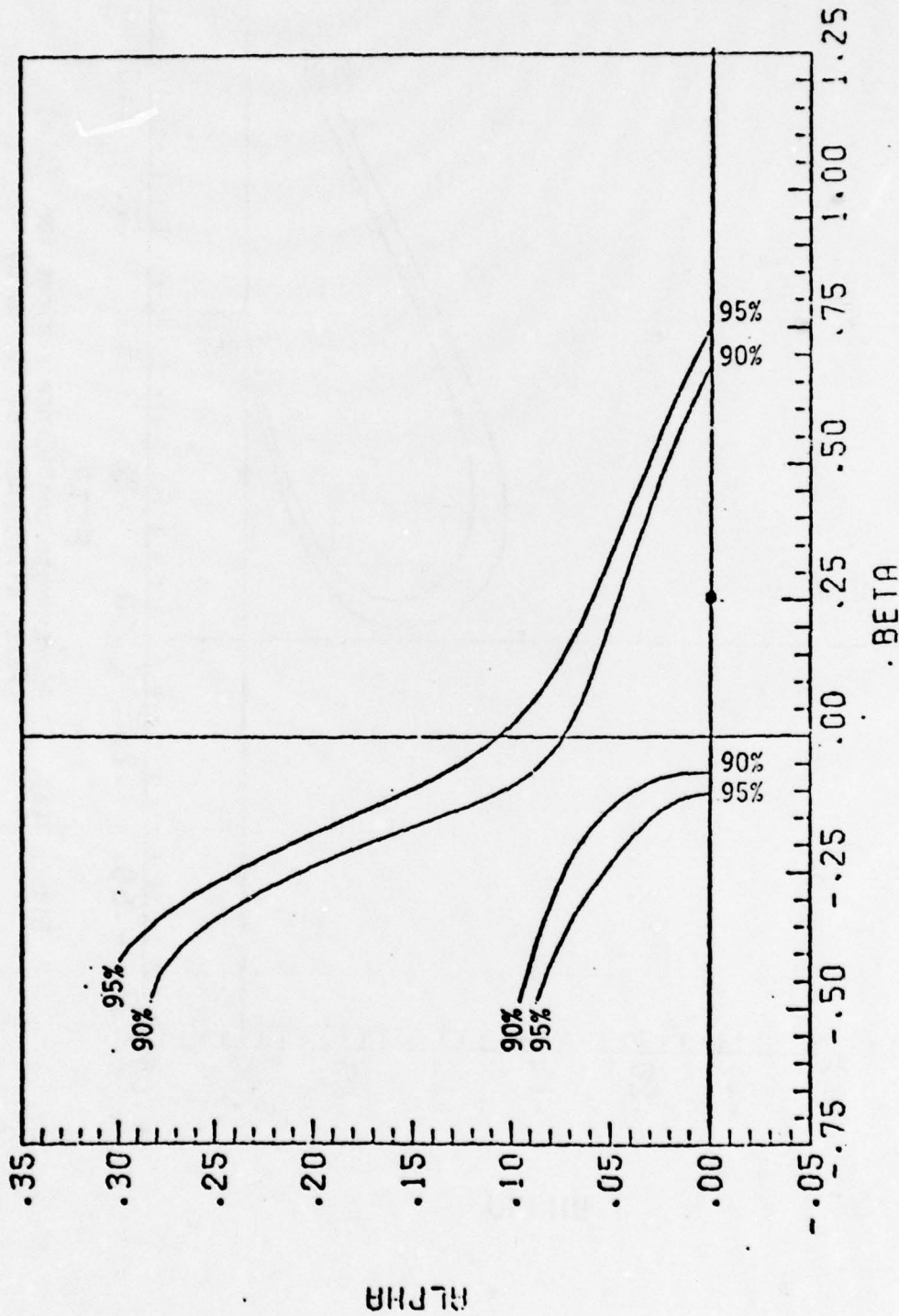


Figure 2(b). Approximate confidence regions for (β, α)
 (First differences of Series B, first 179 values)

The example shows that for these data, tail heaviness of the whole sample arises because of changes in variance in the latter part of the series. It seems that for the first 179 observations the parent distribution could very well be represented by a normal, or slightly contaminated, normal distribution. As for Series B', maximum likelihood estimation was carried out for the first differences of the series to obtain

$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (0.0, 3.54, \underline{0.20}, \underline{0.16})$ and the approximate confidence regions shown in Figure 2(c). The confidence regions are elongated as expected showing the trade off between β and α . A contaminated normal distribution with $\alpha \approx .25$ could well represent the data.

4. Measures of Velocity of Light

Measurement of the Velocity of Light in a Partial Vacuum by
Michelson, Pease and Pearson (Michelson, Pease and Pearson (1935))

During the period September 1929 to March 1933, the velocity of light was measured at the Irvine Ranch near Santa Ana, California by Michelson, Pease and Pearson. The observations were made by the rotating-mirror method. In all 233 series were recorded, the averages for which are plotted in Figure 3. Every series represents two to eight sets of observations taken during over about an hour's time with each set of observations usually furnishing three values of velocity of light.

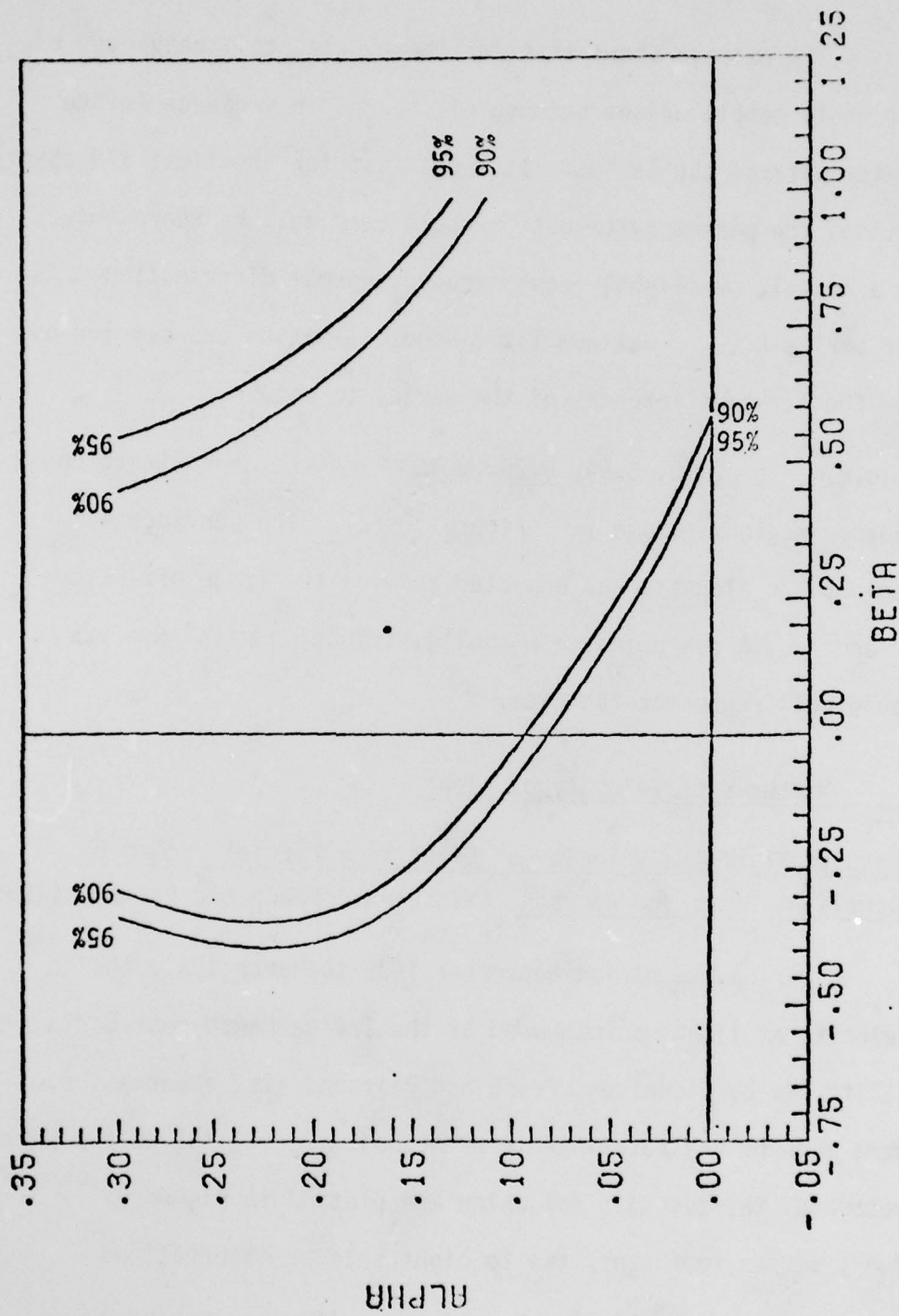


Figure 2(c). Approximate confidence regions for (β, α)
(First differences of Series B')

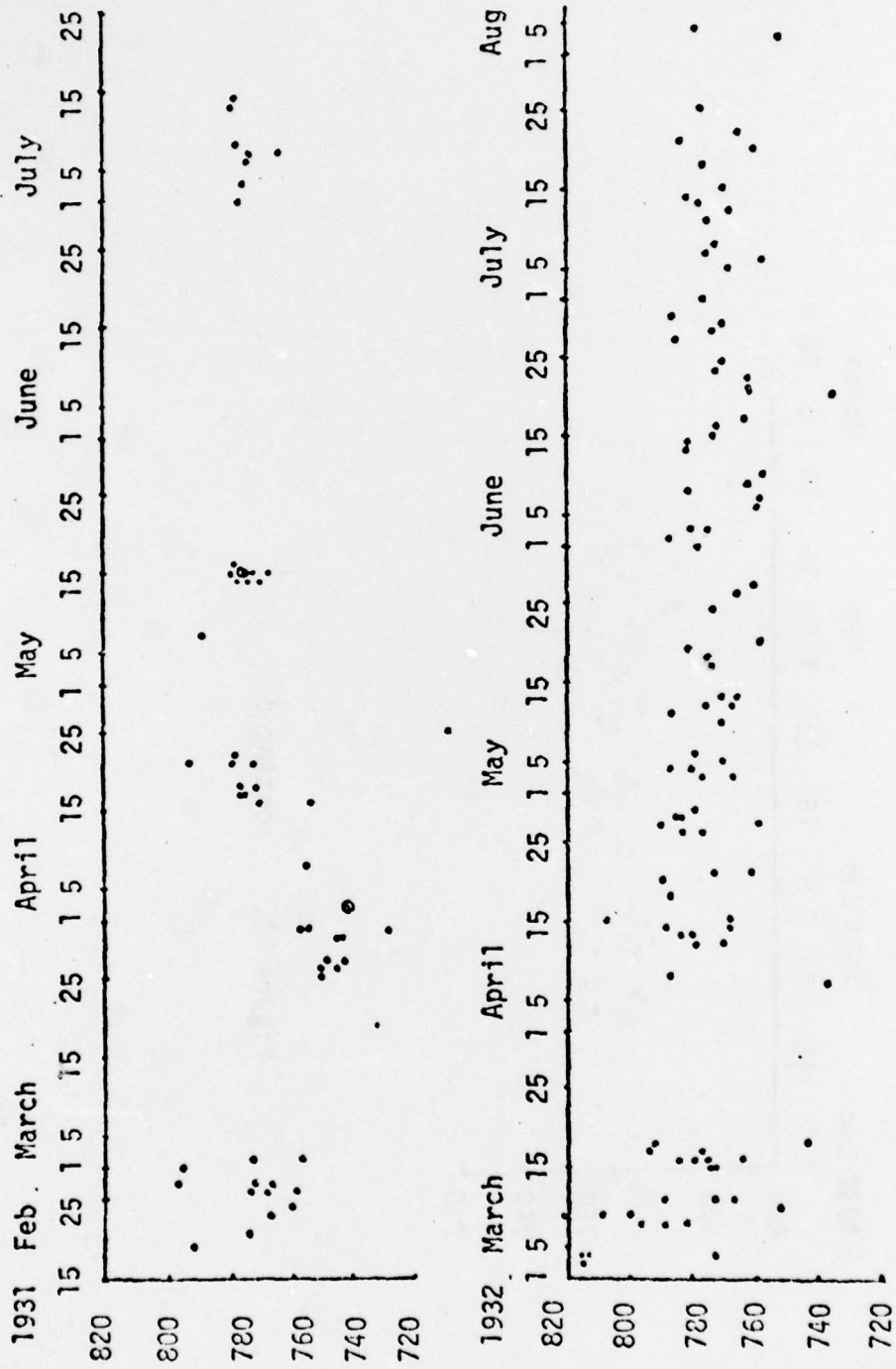


Figure 3. Michelson, Pease and Pearson's data
 (Mean velocity - 299.000) measured between Feb.1931-March 1933
 (continued on next page)

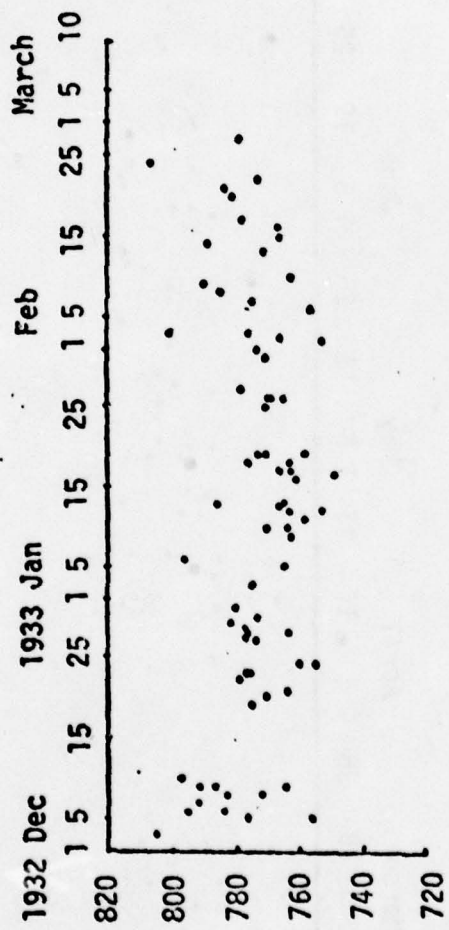


Figure 3. continued

All these observations were actually made between February 1931 and March 1933. In 1931, observations were not made continuously over the year; instead, a group of observations were made in February, another group in late March and early April and so on. The means and variances of these groups fluctuate considerably. The 155 observations made after April 1932 were, however, much more homogeneous. Behavior of this kind is common where a new method of scientific measurement takes some time to "debug" and to settle down to produce homogeneous results.

We shall, therefore, again perform two analyses—the first for all the data and the second for the last 155 observations.

(i) For the whole data set

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (774.0, 14.6, \underline{.80}, \underline{0.0})$$

Approximate confidence regions for (β, α) are shown in Figure 4(a).

(ii) For the 155 observations between April 1932 and February 1933

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (773.1, 9.47, \underline{-.35}, \underline{0.10})$$

Approximate confidence regions for (β, α) are drawn in Figure 4(b).

Again, notice that the maximum likelihood estimates $(\hat{\beta}, \hat{\alpha})$ move from $(.80, 0.0)$ to $(-.35, 0.10)$ as we leave out the inhomogeneous part of the data, and the confidence regions, which originally cover

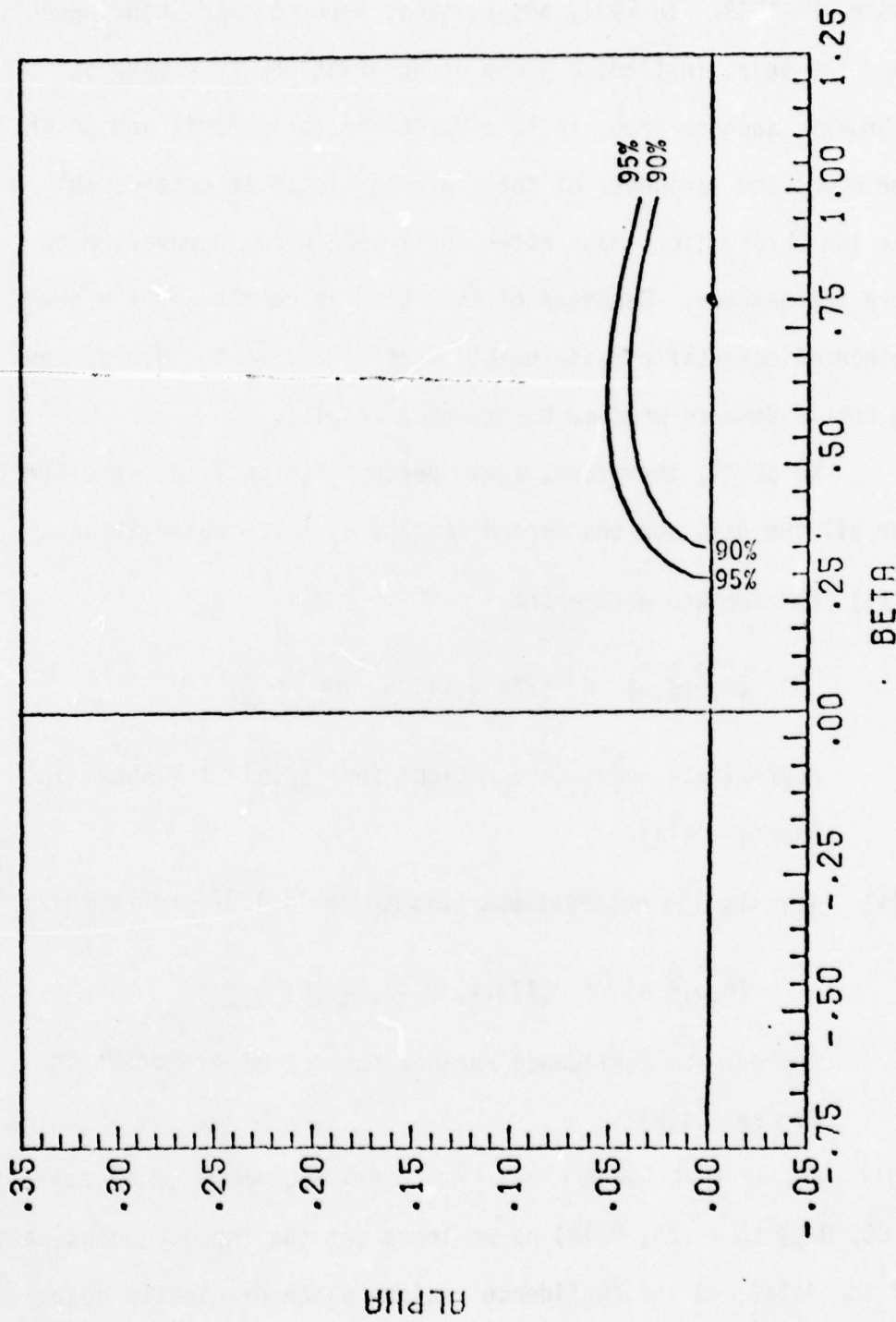


Figure 4(a). Confidence regions for (β, α)
 (Measurements of the velocity of light by
 Michelson, Pease, Pearson)

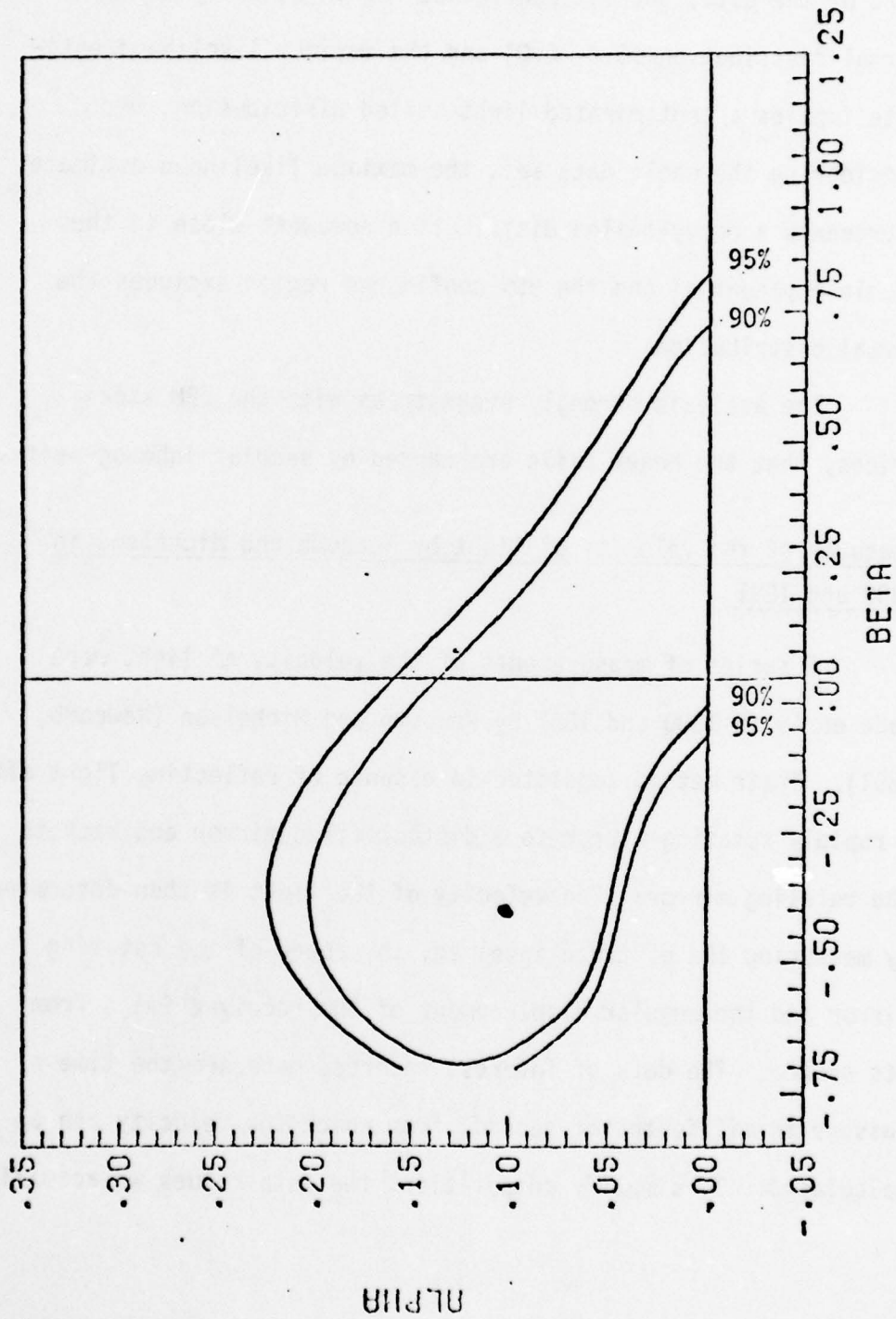


Figure 4(b). Confidence regions for (β, α)
 (155 measurements of the velocity of light made
 between April 1932 and February 1933 by Michelson,
 Pease and Pearson)

the lower right corner, now cover the central part and some upper left part of the plane. Also note that while, for the homogeneous part of the data, the 90% confidence region actually contains the normal distribution (0.0, 0.0) and the maximum likelihood estimate implies a contaminated light-tailed distribution, when considering the whole data set, the maximum likelihood estimate represents a heavy-tailed distribution somewhat close to the double exponential and the 95% confidence region excludes the normal distribution.

The analysis strongly suggests, as with the IBM stock prices, that the heavy tails are caused by secular inhomogeneity.

Measures of the velocity of light by Newcomb and Michelson in 1880 and 1881

A series of measurements of the velocity of light were made early in 1880 and 1881 by Newcomb and Michelson (Newcomb, 1891). Their method consisted in essence of reflecting light off a rapidly rotating mirror to a distant fixed mirror and back to the rotating mirror. The velocity of the light is then determined by measuring the distance involved, the speed of the rotating mirror and the angular displacement of the received image from its source. The data of interest recorded here are the time of passage in millionths of second, from which the velocity can be calculated. To simplify computation, the data values we actually

considered are (recorded time of passage - 17,000). The observations made on each day are plotted in Figure 5. We can see from the plot that there are appreciable day to day variations. This was to be expected because of factors such as weather conditions. Some of the factors which might have affected the results were recorded. For example, it was noticed that it was cloudy on August 25, and that images were faint for some observations so that on that day we might expect a larger variance. Again on September 15, the instrument was adjusted, and this might be expected to influence both the mean and variance on that day. With these in mind, any assumption of constant mean and variance is dubious. One way to get rid of day to day variation is to standardize the observations within each day, i.e., to calculate the mean and standard deviation for data on each day, and use $\frac{\text{observation} - \text{mean}}{\text{standard deviation}}$ as new observations.

What we have done with this set of data is, first, analyze the whole set of the original data; second, leave out the three obvious outliers as shown by an arrow in Figure 5; and finally, analyze the data set after standardizing within each day. Note that in standardizing the data, the three obvious outliers were discarded, and when only a single observation was made in a day, this observation was also excluded since no standard deviation was available for it. The results are

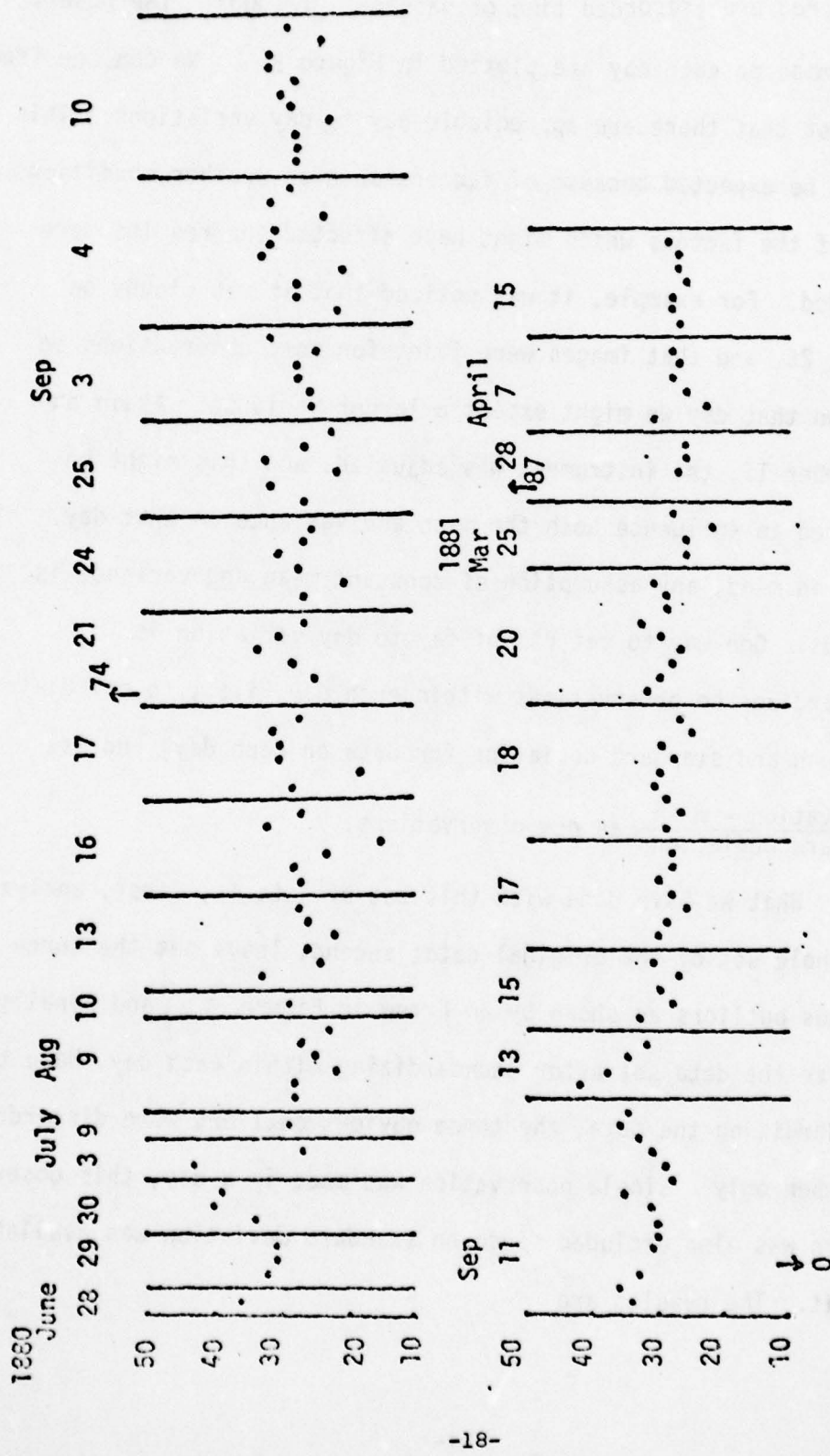


Figure 5. Measurements of the velocity of light by Newcomb and Michelson in 1880 and 1881.

- (i) If we take the whole set of data as a random sample
($n = 150$), then

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (28.0, 4.16, \underline{1.0}, \underline{.11})$$

Approximate confidence regions for (β, α) are given in Figure 6(a), which show that the parent distribution have very heavy tail (with both large β and α values) if the data set is really a random sample.

- (ii) If we throw out three obvious wild observations and regard the rest as a random sample ($n = 147$), then

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (28.0, 4.3, \underline{1.0}, \underline{0.0})$$

and the approximate confidence regions for (β, α) are given in Figure 6(b). These confidence regions have quite different shapes from the others, the 90% confidence region has two parts, located at the lower right and upper left plane, respectively. This came as no surprise because we know there is compensation between β and α . A distribution with heavy tail and small contamination could be somewhat similar to a distribution with light or moderate tail but high contamination, and for this particular set of data, both areas in the 90% confidence region contain distributions which can adequately describe the data.

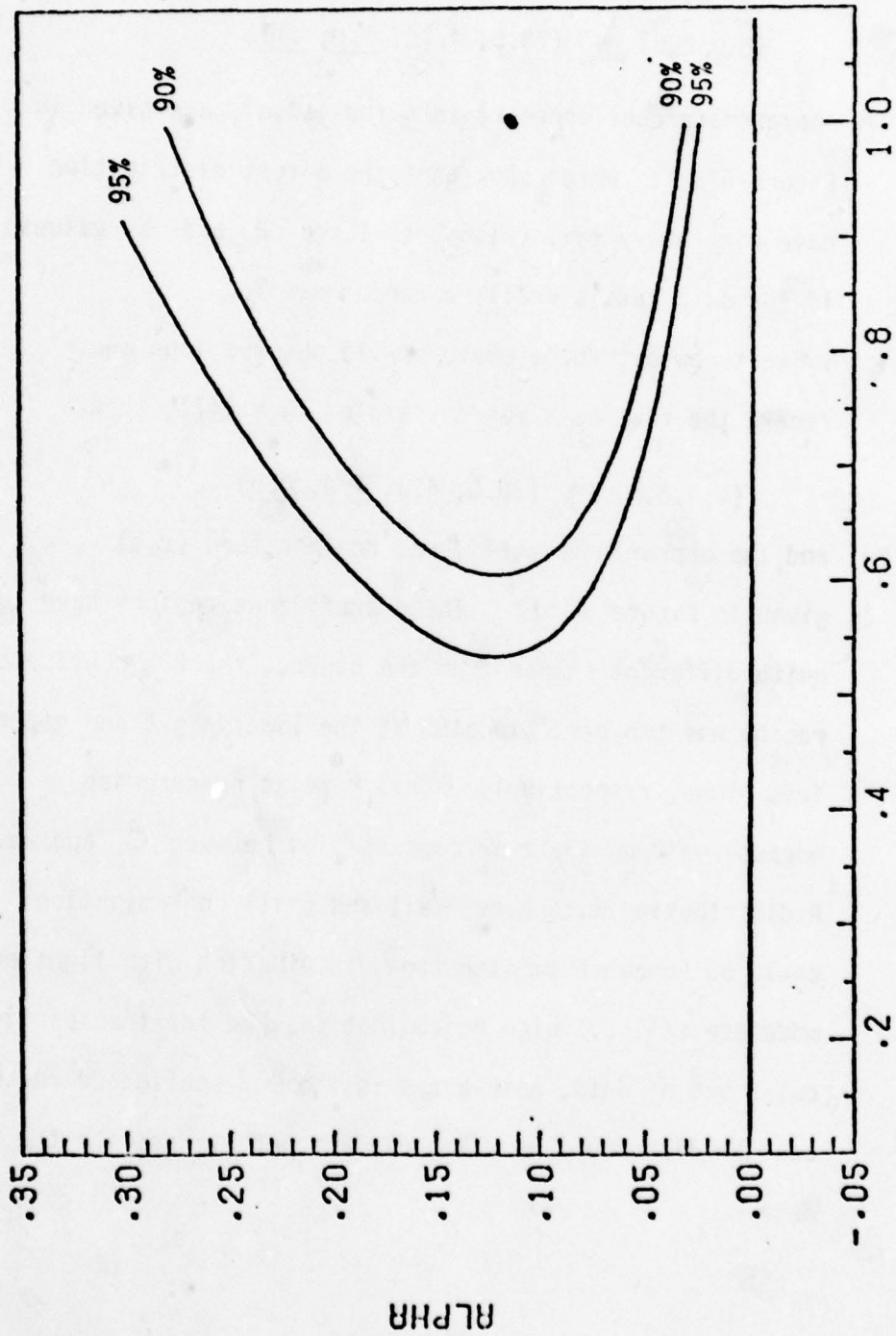


Figure 6(a) Approximate confidence regions for (β, α)
 (Measurements of the velocity of light by Newcomb
 and Michelson in 1880 and 1881)

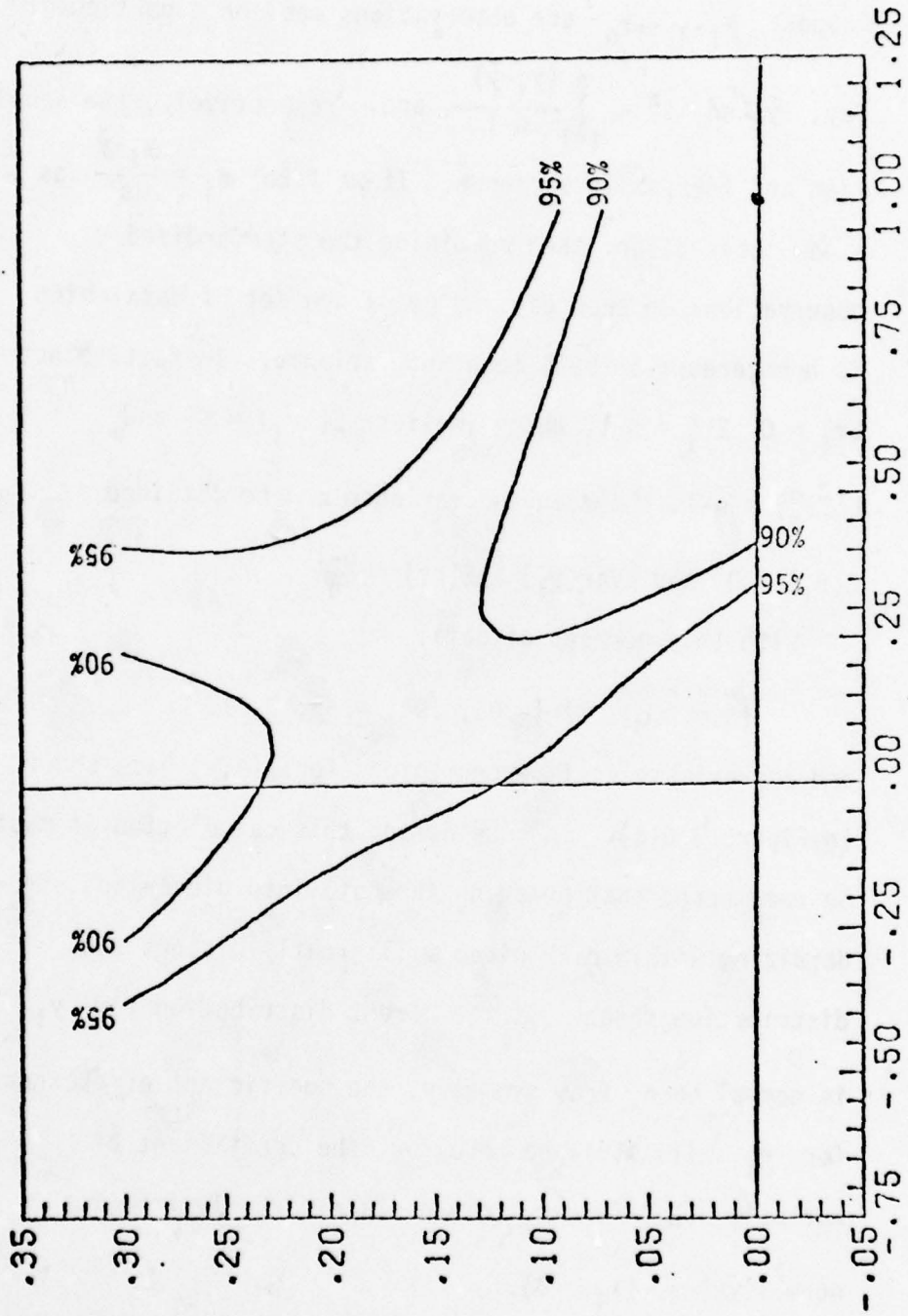


Figure 6(b) Approximate confidence regions for (β, α)
 (Newcomb and Michelson data after throwing
 out 3 outliers)

(iii) We standardize the observations made on each day.

Suppose y_1, \dots, y_n are observations made on a particular day, \bar{y} and $s^2 = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1}$ are, respectively, the sample mean and the sample variance. If we take $r_i = \frac{y_i - \bar{y}}{s}$ as a new observation, then combining the standardized observations on each day, we get a new set of data which is homogeneous in both mean and variance. In fact, since $\sum r_i = 0$ $\sum r_i^2 = n-1$, which implies $E(\sum r_i) = 0$ and $E(\sum r_i^2) = n-1$, the mean and variance can be obtained as $E(r_i) = 0$ and $\text{Var}(r_i) = E(r_i^2) = \frac{n-1}{n}$.

With this new set of data,

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (-.05, .94, \underline{-.65}, \underline{0.0})$$

and approximate confidence regions for (β, α) are shown in Figure 3.6(c). In considering this calculation it must be remembered that breaking the data into pieces and standardizing within each piece will greatly distort the distribution shape. If the parent distribution for y_i is normal then, from symmetry, the coefficient of skewness for r_i will still be zero, but the coefficient of kurtosis (λ_4) for r_i can deviate markedly from its normal value ($\lambda_4 = 0$).

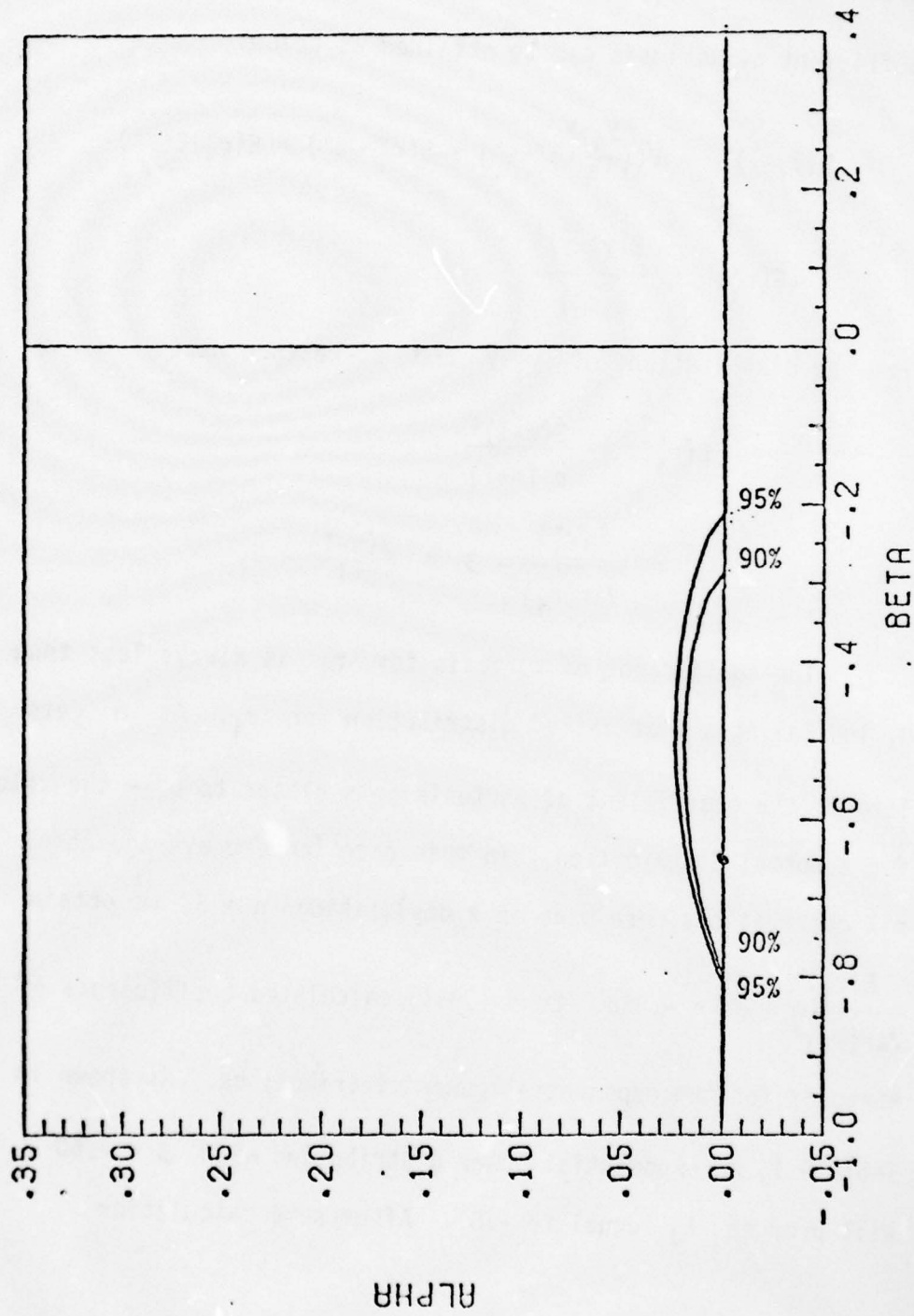


Figure 6(c) Approximate confidence regions for (β, α)
 (Newcomb and Michelson data after standardization)

Since r_i is independent of σ and s is a sufficient statistic for σ , r_i is independent of s . It follows that the coefficient of kurtosis can be obtained as below:

$$E(y_i - \bar{y})^4 = E\left(\left(\frac{y_i - \bar{y}}{s}\right)^4 \cdot s^4\right) = E(r_i^4 \cdot s^4) = E(r_i^4)E(s^4)$$

$$E(r_i^4) = \frac{E(y_i - \bar{y})^4}{E(s^4)}$$

From the distributions of $y_i - \bar{y}$ and s , we then have

$$E(r_i^4) = \frac{3(n-1)^3}{n(n+1)}$$

and

$$\lambda_4 = \frac{E(r_i^4)}{\text{Var}(r_i)^2} - 3 = \frac{3(n-1)}{n+1} - 3.$$

The coefficient of kurtosis for r_i is always less than 0, indicating a flat tailed distribution for r_i . As n gets larger, the coefficient of kurtosis gets closer to 0 — the value for a normal distribution. In this case (on the average about six observations were made in a day), taking $n = 6$, we obtain

$$\frac{E(r_i^4)}{\text{Var}(r_i)^2} - 3 = -.858. \text{ Lund (1967) calculated coefficients of}$$

kurtosis for the exponential power distributions. As shown in Table 3.1, an exponential power distribution with $\beta = -.50$ will produce λ_4 equal to $-.81$. After some calculation

β	-1.0	- .75	-.50	-.25	0.0	.25	.50	.75	1.00
λ_4	-1.20	-1.07	-.81	-.45	0.00	.55	1.22	2.03	3.00

Table 1

one can also show that for $\beta = -.55$, $\lambda_4 = -.87$. Therefore, we concluded that if the parent distribution for y_i is normal, one would expect that the distribution for r_i (when n is approximately 6) would have tail behavior similar to that of an exponential power distribution with $\beta = -.55$ which is very close to the maximum likelihood estimate of $-.65$. Thus it appears that the normal distribution alone, or maybe with some slight contamination, could adequately describe the data if there had not been daily changes in levels and variances.

5. Some Data Sets Investigated by Tocher (1928)

In 1928, Tocher published the paper "An Investigation of the Milk Yield of Dairy Cows," in which a large amount of data were given. These are the records about milk yields of dairy cows supplied by "The Scottish Milk Records Association." Various "characters" are investigated for each cow, e.g., estimated total quantity of milk in gallons for the entire lactation period, total quantity of butter fat, age of the cow, ..., etc. Tocher's investigation had several purposes. In one interesting

analysis he fitted Pearson curves to a number of large data sets.

Three of these data sets which we will consider in our study are

- (1) Total yield of butter fat in pounds for 449 cows at age 7.
- (2) Total yield of butter fat in pounds for 251 cows at age 9.
- (3) Quantity of milk in gallons for 240 cows with percentages of butter fat between 3.00 and 3.25.

The exact data values are not available. Tocher's paper gives frequency distributions for the data, with a grouping interval of 50 gallons for quantity of milk and 20 pounds for yield of butter fat. The frequency distributions are shown in Figure 7.

Following the procedure stated before, we have

- (i) For total yield of butter fat for cows at age 7

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (249.75, 52.15, \underline{-.20}, \underline{.01})$$

Approximate confidence regions are shown in Figure 8(a).

- (ii) For total yield of butter fat for cows at age 9

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (265.8, 60.15, \underline{0.35}, \underline{0.0})$$

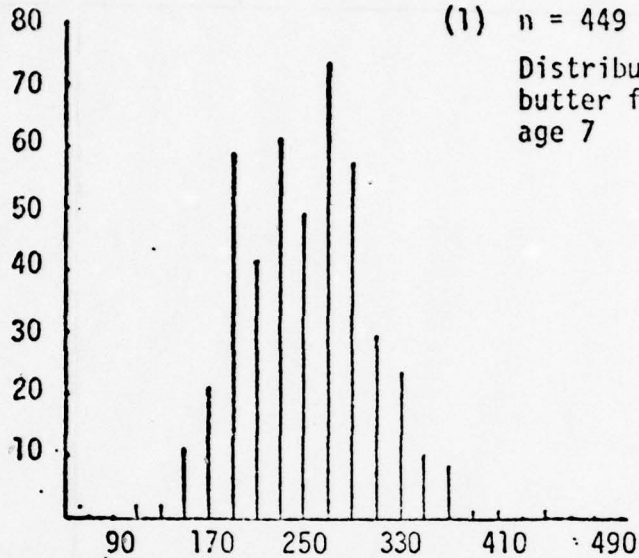
Approximate confidence regions are given in Figure 8(b).

- (iii) For quantity of milk in gallons for cows with percentages of butter fat between 3.00 and 3.25.

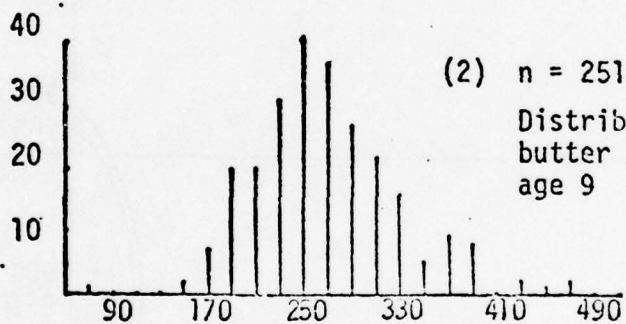
$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (664.0, 176.0, \underline{0.25}, \underline{0.0})$$

Approximate confidence regions are given in Figure 8(c).

Yield of butter fat



Yield of butter fat



Quantity of milk in gallons

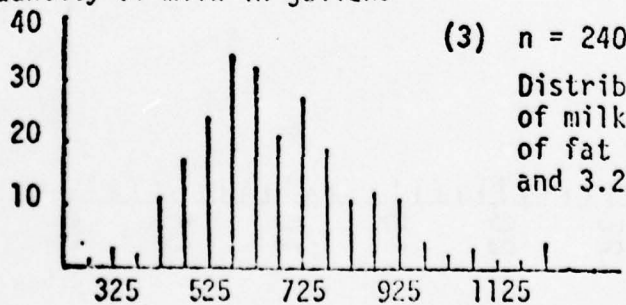


Figure 7 Frequency distributions for Techer's data.

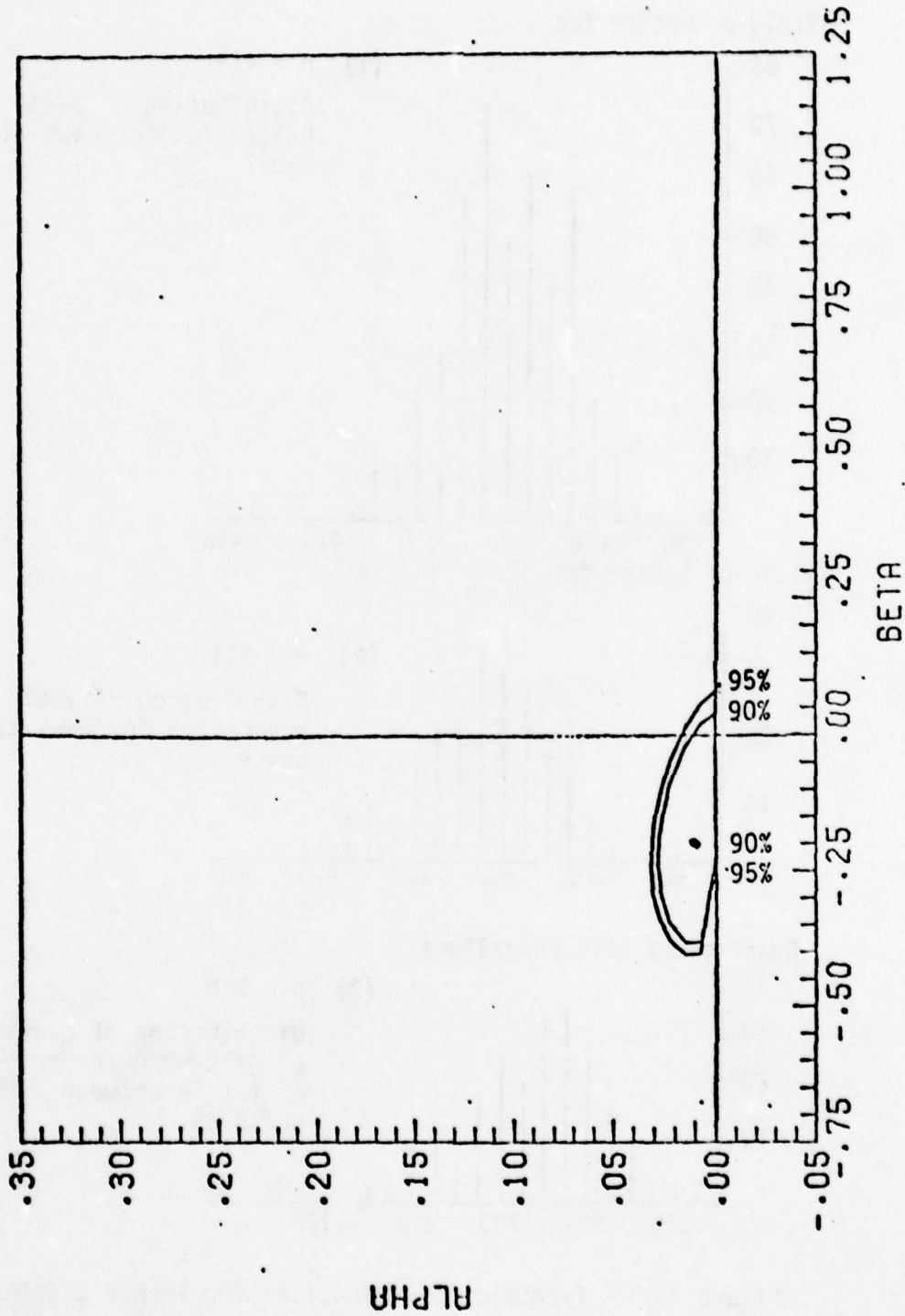


Figure 8(a) Approximate confidence regions for (β, α)
 (Yield of butter fat for cows at age 7)

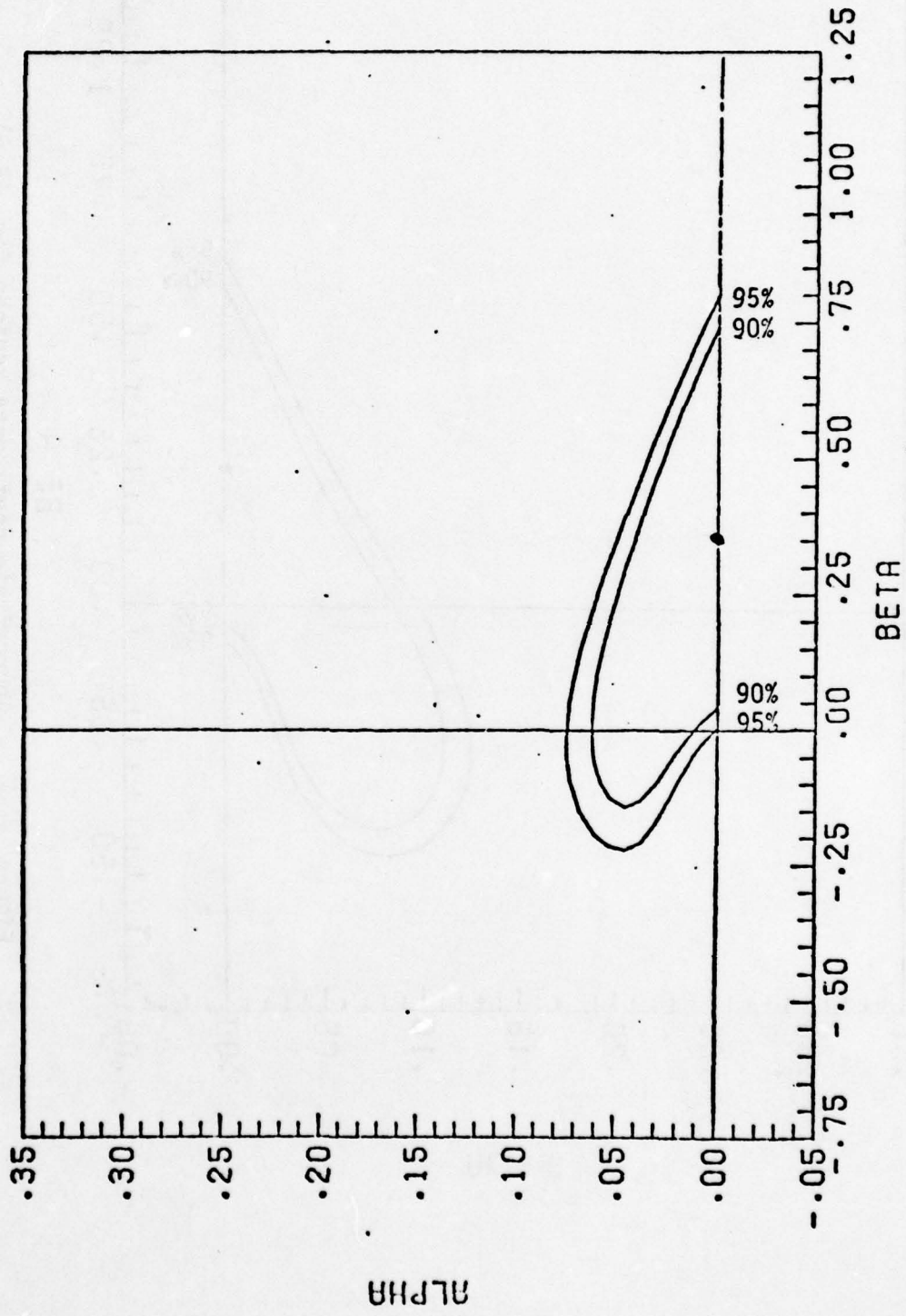


Figure 8(b) Approximate confidence regions for (β, α)
(Yield of butter fat for cows at age 9)

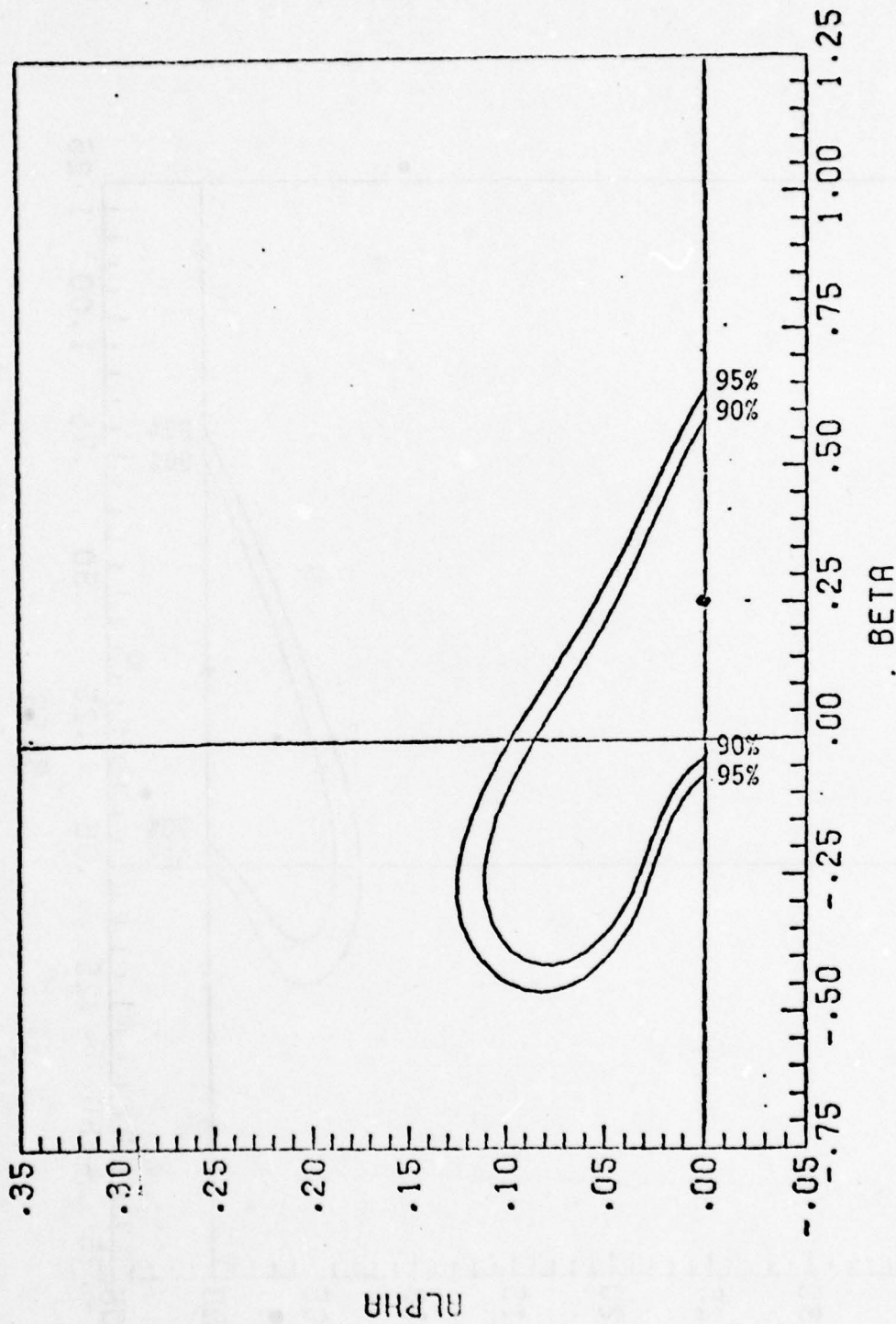


Figure 8(c) Approximate confidence regions for (β, α)
 (Quantity of milk in gallons for cows with
 percentages of butter fat between 3.00 and 3.25)

All the confidence regions in Figure 8 are reasonably close to the center of the plane, suggesting that the data sets could be random samples from distributions not too far away from the normal. It is interesting to notice that the analysis suggests that the distribution of the yields of the butter fat for cows at age 7 is a short-tailed distribution. This runs contrary to the common belief that most data sets have heavy-tailed distributions. Cox (Discussion of Stigler's paper, 1977), however, has said that according to his experience, short-tailed distributions could arise as well as heavy-tailed distributions.

These three sets of data are different in the nature from those discussed before, all three sets of data being measurements made on members of a particular population at approximately the same time, while the sets discussed previously are measurements or observations made sequentially over a long period of time. The latter type suffers from the effect of secular inhomogeneity but the former does not. This can partly explain why Tocher's first three sets of data are quite close to a normal distribution.

Due to the high cost involved in collecting the data, the Scottish Milk Record Association was not able to make daily observations of milk for each cow. The morning and evening milk of each cow was tested at intervals ranging from 14 to 28 days. The operation was repeated after an interval of about fourteen days. The quantity of milk and the percentage of butter fat yielded

by each cow in the interval were assumed to be those given on the date of the previous test. So the yield of milk determined by the Scottish Milk Record Association is not the actual yield but an estimated one. During the period of 1911-13, some daily records were available from private sources. Tocher proposed to compare these actual yields with those estimated by the Association to see if the estimate is any good. The data are available on 107 cows, and we will take the difference of yields of milk (Association - Private) as observations and examine the distributional property of these 107 observations. A plot of these data is given in Figure 9. $(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha})$ for these 107 data are (.10, 11.09, .20, .11), approximate confidence regions are shown in Figure 10. The confidence region is very large because the sample size is relatively small, and it is elongated, as expected.

6. Wiebe's Agricultural Data

Wiebe (1935) gives the grain yield of 1,500 rows of Federation wheat in a uniformity trial. He studied the variation and correlation in grain yield among the 1,500 wheat nursery plots to seek for criteria for better seeding arrangement in future field trials. Barbacki and Fisher (1936) used his data to examine the relative precision of systematic and randomized experimental arrangements. In field trials of this kind it is well known that results from plots close together are inevitably

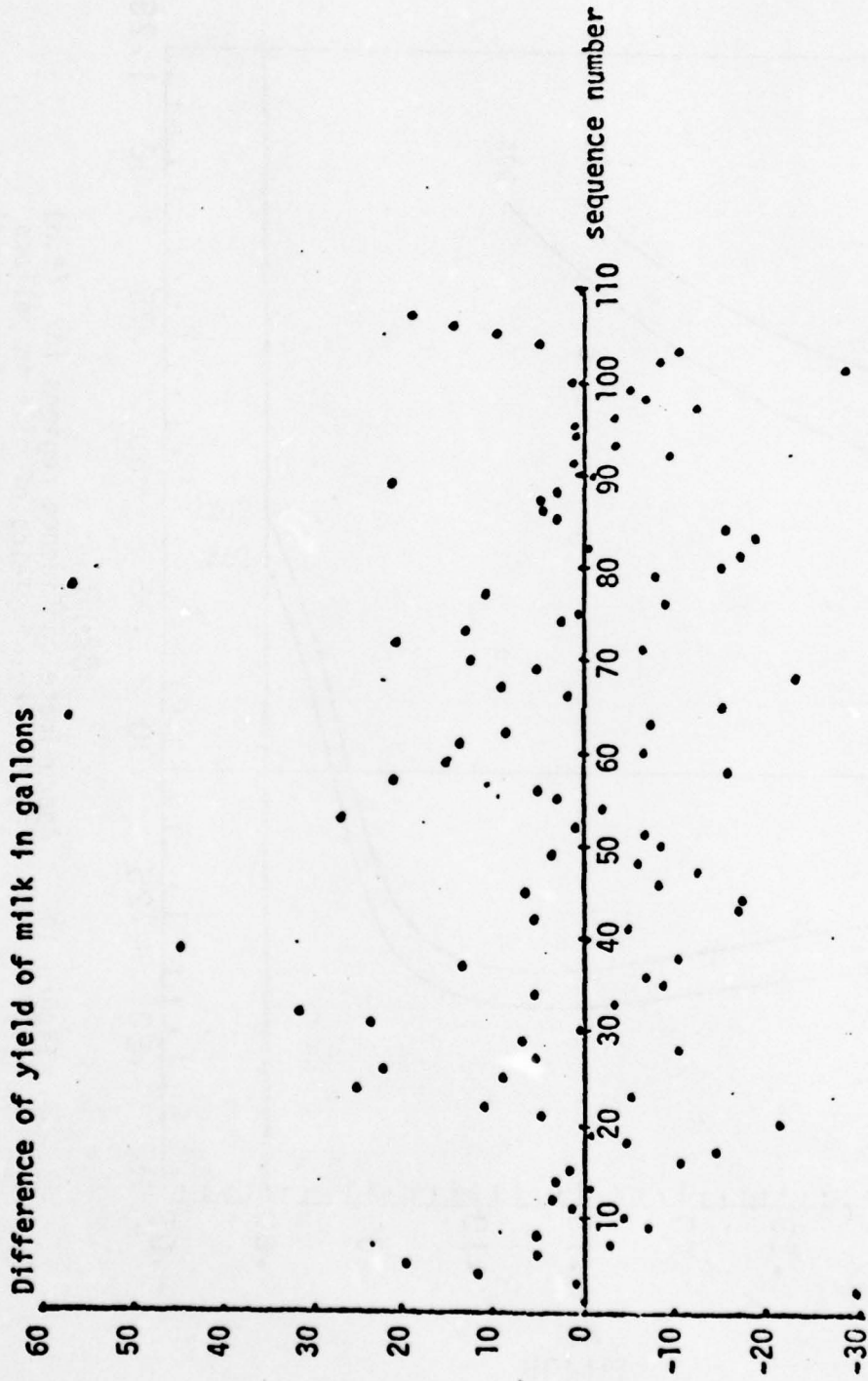


Figure 9 Difference of the yield of milk between Association and Private records.

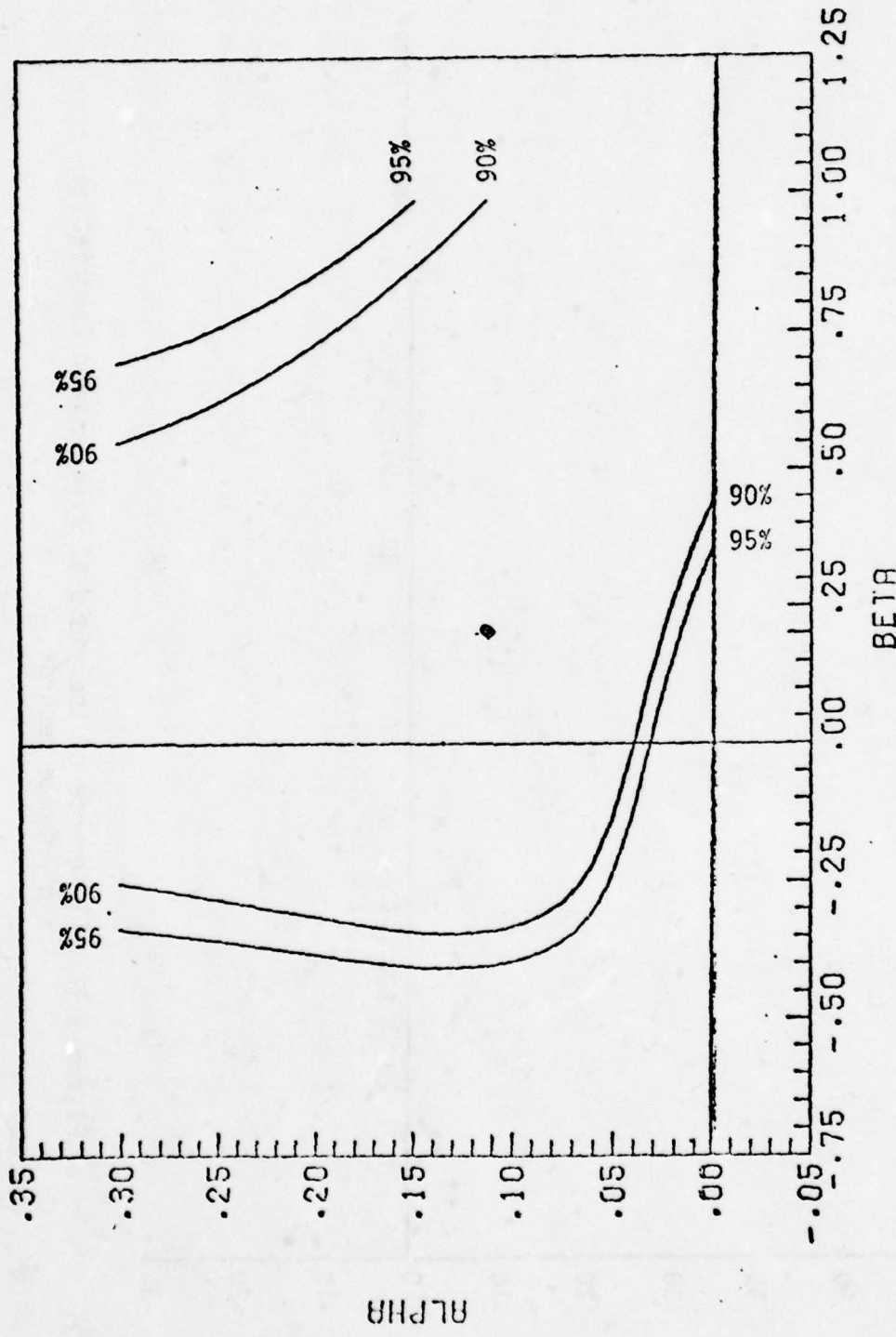


Figure 10 Approximate confidence regions for (β, α)
 (Difference of yields of milk in gallons
 between Association and private records)

correlated, and blocking and randomization were introduced to ensure proper isolation of the real effects of treatments in these circumstances.

This 1,500 rows of wheat were grown in the summer of 1927 on the Aberdeen Substation, Aberdeen, Idaho. The 1,500 rows were grown in 12 series, each having 125 rows. Figure 11 shows the general arrangement of the crop. The rows were 12 inches apart and 15 feet long. The plot was seeded on April 18 with a grain drill that sowed eight rows at a time. The individual rows were harvested in August and threshed with a small nursery thresher. The grain yields, recorded in grams per row are shown in Table . A contour map of the yield obtained in the plot given by Wiebe is shown in Figure 12 (The contour map is not obvious from the data; Wiebe must have used some kind of smoothing process which is not given in his paper). Sequence plots of all twelve series are shown in Figure 13. If we ignore the obvious serial effects and treat these series as random samples drawn from some distribution, then for any data set we can estimate θ , σ , β and α using the procedure described on page 4. Thus for series 1, we find $(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (651.6, 79.7, 0.45, 0.0)$ with the approximate confidence regions shown in Figure 14. Although the maximum likelihood estimates suggest an approximate uncontaminated somewhat heavy-tailed parent distribution, the confidence regions for (β, α) still include the normal distribution.

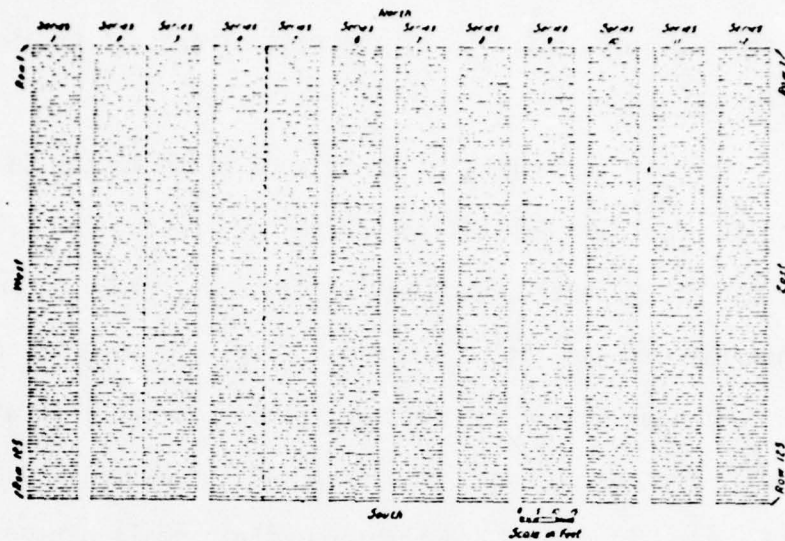


Figure 11 Field plan of uniformity experiments showing general arrangement of rows. Series 2 and 3, 4 and 5, were grown as 30 foot rows but were cut in 15 foot lengths at harvest time.

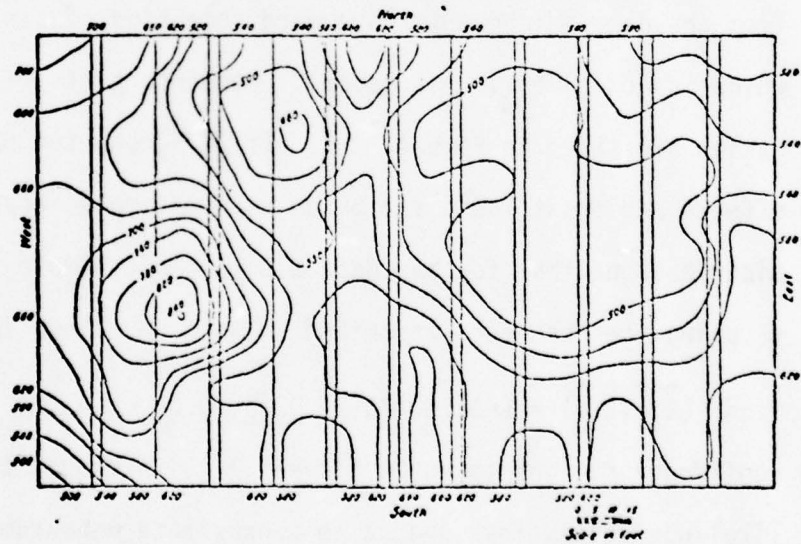


Figure 12 Wiebe's contour map of the grain yields obtained. The lines are at 40-g intervals.

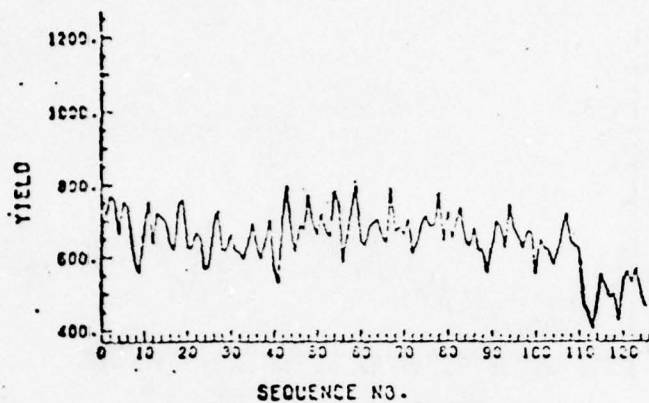
Table 2 Yield of grain from each of 1,500 15-foot rows of wheat, 12 inches apart, grouped in 12 series of 125 rows each, Wiebe's data.

Series 1	Series 2	Series 3	Series 4	Series 5	Series 6	Series 7	Series 8	Series 9	Series 10	Series 11	Series 12
715	595	580	580	615	610	540	515	557	675	570	612
770	210	655	675	710	670	565	565	550	574	511	618
760	670	670	690	655	725	675	640	625	755	616	705
665	615	685	555	585	630	570	570	550	616	573	570
755	730	670	560	545	620	580	525	475	565	579	612
745	670	585	560	550	710	580	545	538	587	670	674
645	670	570	520	470	670	555	535	520	455	611	578
585	425	455	470	445	555	550	430	420	481	531	559
560	540	470	500	525	525	545	500	478	538	453	670
655	730	670	580	555	645	615	535	534	583	616	638
755	810	675	570	525	715	650	550	613	610	742	657
640	635	585	475	470	615	550	515	379	493	635	567
725	635	520	455	475	590	580	470	435	520	519	555
715	775	615	545	530	675	575	510	527	561	557	577
700	705	555	440	455	670	575	425	474	448	532	577
640	655	495	435	485	575	475	435	450	516	458	559
620	635	425	445	455	620	505	455	419	591	545	537
750	630	535	455	510	575	550	470	427	545	552	582
760	675	610	540	570	675	580	460	513	599	555	675
630	645	490	445	440	585	490	420	460	542	474	563
625	540	495	415	445	500	440	430	455	445	461	523
670	670	550	485	445	570	525	410	416	570	533	623
655	630	540	465	460	560	465	435	476	454	458	523
570	535	525	435	415	510	460	420	370	442	425	446
575	495	540	480	525	555	440	437	437	541	525	514
690	645	595	515	470	570	515	445	528	538	517	533
730	730	710	565	475	640	565	550	494	521	554	597
620	635	555	485	435	530	470	470	465	521	554	537
620	600	530	435	435	545	515	430	462	486	525	578
665	685	625	520	490	525	530	465	497	594	563	634
620	615	530	440	450	570	470	445	510	591	478	514
615	525	610	515	525	570	435	520	485	534	448	533
595	625	640	440	435	510	515	525	576	522	468	514
640	580	630	445	480	610	525	515	514	525	458	518
635	650	695	550	515	685	570	525	613	540	532	525
630	610	545	475	450	555	500	575	458	477	480	567
595	555	540	505	475	540	490	467	477	479	444	537
645	635	650	560	525	645	515	520	477	510	465	533
705	660	595	495	465	595	490	505	477	489	529	533
575	530	630	460	410	500	460	505	459	431	438	578
530	555	570	450	425	490	450	495	425	443	449	469
715	620	685	575	535	620	495	500	474	491	495	533
795	715	740	590	565	585	535	570	515	513	541	545
650	635	620	495	510	585	520	475	444	405	579	469
615	580	525	490	505	565	450	470	458	442	474	484
690	645	630	490	565	645	570	550	434	474	522	543
675	615	640	505	550	615	535	500	465	456	513	563
770	725	685	570	575	685	540	475	534	456	467	545
695	700	770	540	515	685	530	495	538	454	422	670
650	660	705	570	545	625	570	540	494	570	532	687
720	730	825	665	545	645	520	540	584	546	577	630
670	640	710	590	535	575	495	435	512	447	463	600
655	590	675	575	515	570	490	470	477	477	472	589
780	790	855	625	575	655	530	540	575	572	523	582
750	705	770	575	560	565	510	435	486	470	502	570
585	620	685	620	550	620	435	470	429	485	514	619
715	760	790	645	625	595	560	485	466	487	476	645
795	795	850	700	655	645	540	470	521	582	584	750
645	885	720	610	520	550	460	440	425	425	462	694
630	695	665	630	510	575	445	430	443	423	450	627
625	840	875	640	575	625	570	490	514	529	486	717
695	740	790	575	520	630	545	470	455	476	488	619
705	745	840	710	640	610	500	570	448	479	434	574
640	840	795	640	610	585	560	450	460	562	450	570
790	765	830	770	580	615	585	460	451	472	504	630
670	690	785	645	510	680	575	490	445	473	529	552
645	790	770	645	510	560	535	435	474	455	458	555
660	825	900	645	620	545	535	390	389	451	439	615
705	835	860	750	660	680	620	510	531	450	534	694
610	720	705	635	540	650	535	435	479	457	431	649
640	735	845	645	540	625	595	390	379	483	522	638
690	855	905	700	615	605	580	470	477	441	495	638
830	765	945	695	645	650	615	495	531	444	453	645
830	765	945	695	645	650	615	495	531	444	453	645

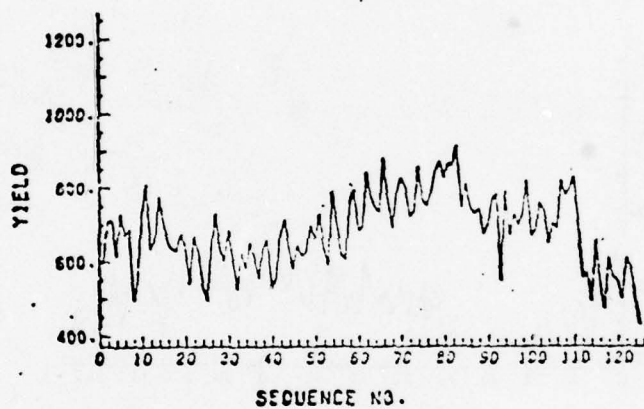
Table 2 continued

Series 1	Series 2	Series 3	Series 4	Series 5	Series 6	Series 7	Series 8	Series 9	Series 10	Series 11	Series 12
645	725	825	715	550	650	655	595	414	451	542	642
645	729	720	615	535	625	510	525	425	471	478	514
725	845	965	820	655	675	640	545	474	513	559	527
645	870	840	740	615	655	575	490	433	473	500	567
725	825	920	640	640	620	575	525	451	484	510	567
650	855	935	640	645	640	590	520	476	507	525	563
695	860	840	725	615	710	620	525	438	520	646	623
735	910	875	775	680	700	770	615	546	526	514	638
645	745	835	700	635	615	605	550	455	534	533	589
640	810	700	625	535	650	735	535	458	554	535	615
640	745	840	730	645	650	775	840	557	561	565	619
670	730	775	640	610	610	660	515	537	525	553	570
670	745	690	565	520	635	610	450	476	520	562	575
560	675	640	635	525	635	635	515	494	516	558	559
625	700	725	645	645	580	615	640	516	618	615	604
700	765	725	615	640	705	710	540	619	627	620	645
645	785	655	690	570	615	670	575	546	602	540	600
625	550	540	540	635	535	590	595	579	629	571	658
745	790	675	690	625	685	725	635	686	678	568	672
640	670	630	640	645	670	645	610	627	579	537	675
655	730	615	650	640	645	655	620	611	642	685	630
625	790	675	720	635	680	720	615	595	556	648	578
670	735	645	620	705	675	690	580	623	541	590	630
670	820	685	665	715	715	740	660	658	689	725	675
555	670	560	540	590	625	640	605	553	580	635	638
650	685	505	525	555	590	590	575	592	559	584	518
625	760	625	545	625	635	670	625	672	656	577	645
670	740	575	565	565	610	650	595	515	641	533	623
590	650	570	575	450	445	670	645	587	634	534	574
615	705	590	515	550	485	610	570	565	622	518	623
675	670	615	610	590	625	745	540	584	616	535	668
720	820	720	650	840	615	770	645	615	755	625	713
660	780	640	620	645	540	695	590	615	679	540	570
630	800	715	710	625	575	490	485	594	625	645	582
625	830	745	755	610	630	775	630	614	760	730	619
470	685	640	735	590	580	670	610	556	539	488	638
445	555	635	685	540	615	610	525	546	639	594	739
490	470	555	655	475	615	725	545	640	740	671	795
480	495	610	690	535	640	760	655	488	621	578	615
555	650	715	690	625	655	775	625	622	720	689	728
620	530	620	695	480	545	585	600	551	670	622	623
490	475	595	660	475	545	635	580	524	538	625	552
600	810	540	680	555	625	670	630	591	613	583	585
425	560	570	570	455	605	695	530	544	645	612	537
535	550	540	630	480	565	745	515	467	646	593	510
560	590	560	575	540	475	610	525	492	541	532	490
610	610	600	620	480	585	650	570	577	612	684	630
670	585	635	765	590	625	765	620	608	705	677	660
505	530	540	655	470	555	570	555	537	585	549	619
465	430	510	640	460	600	670	615	620	594	616	744

SERIES 1



SERIES 2



SERIES 3

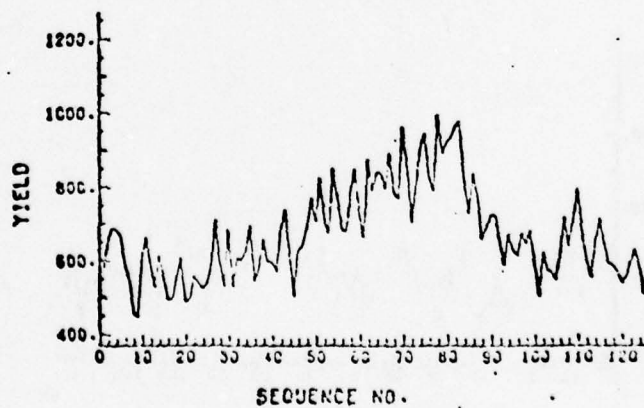
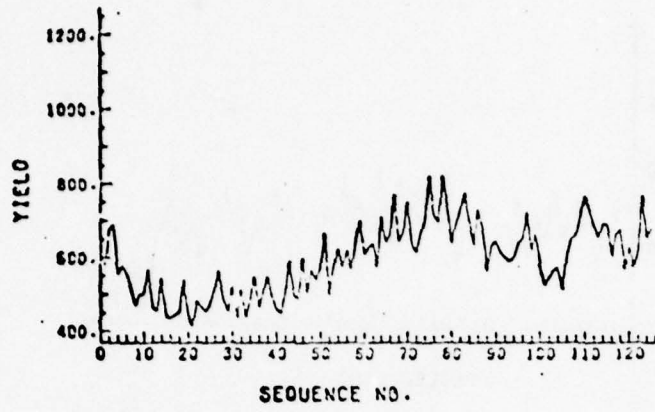
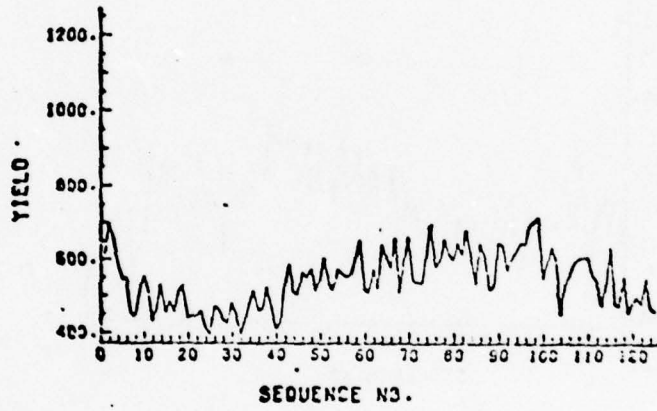


Figure 13 Sequence plot of Wiebe's 12 series given in Table 3.2.

SERIES 4



SERIES 5



SERIES 6

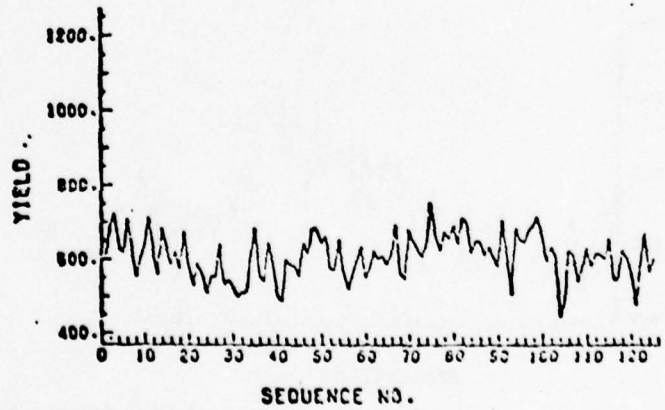
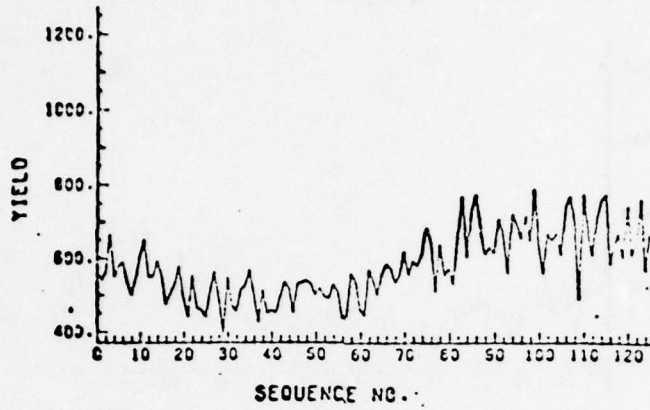
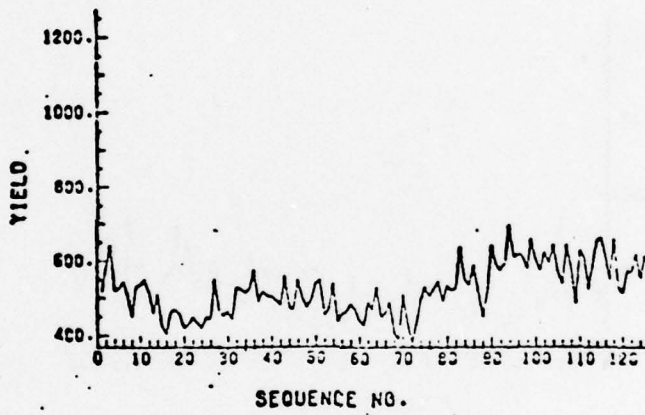


Figure 13 continued

SERIES 7



SERIES 8



SERIES 9

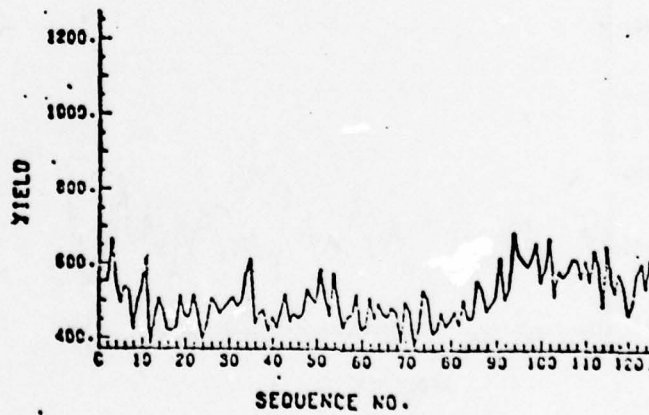
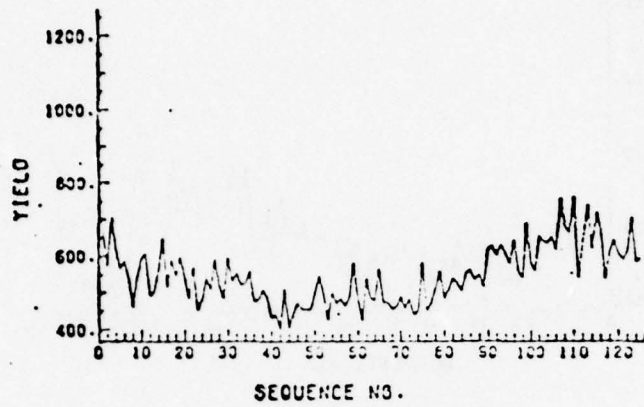
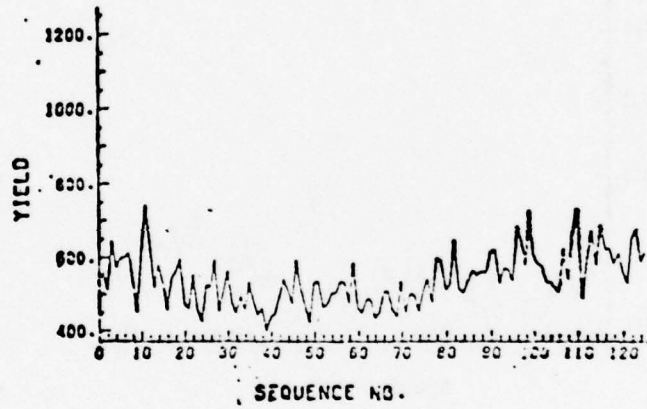


Figure 13 continued

SERIES 10



SERIES 11



SERIES 12

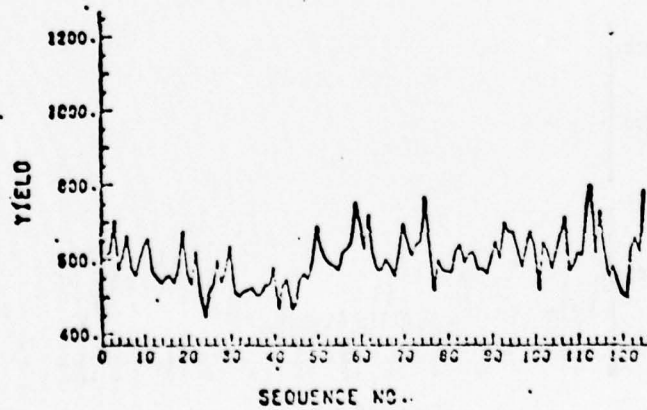


Figure 13 continued

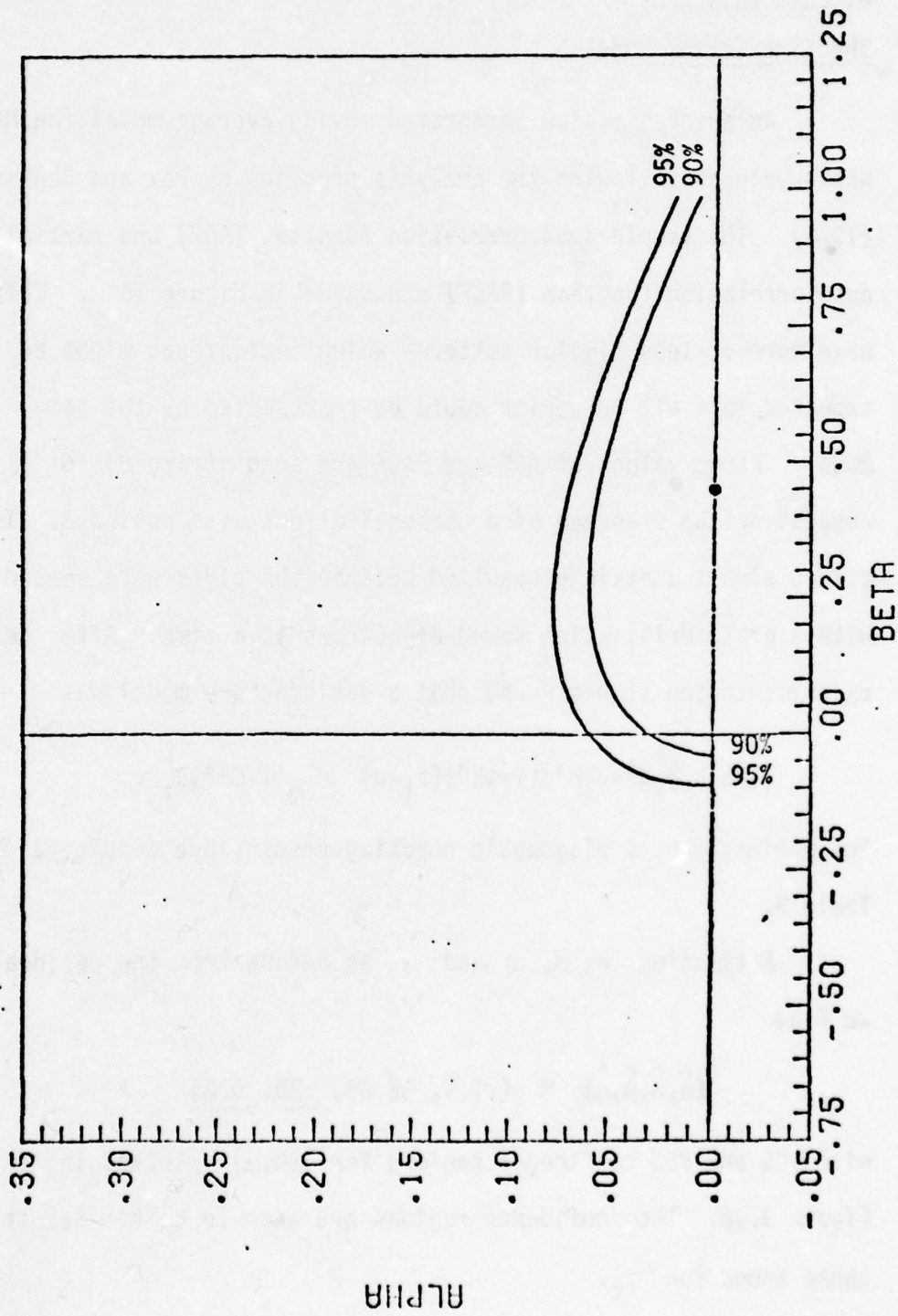


Figure 14 Approximate confidence regions for (β, α)
 (Wiebe's agricultural data, series 1)

More complicated models which take into account the serial correlations were also fitted to this set of data. We discuss this now.

The Time Series Model

An autoregressive integrated moving average model (ARIMA) was developed following the analysis proposed by Box and Jenkins (1976). The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) are shown in Figure 15 . They have more or less similar patterns which indicate, as might be expected, that all 12 series could be represented by the same model. Large values of ACF and PACF are seen at lag 8, 16 suggesting the presence of a seasonal effect with period 8. This effect almost certainly resulted because the plots were seeded with a grain drill which sowed eight rows at a time. After some experimentation it was found that a satisfactory model was

$$(1-\phi_1B-\phi_2B^2-\phi_3B^3)(1-\phi B^8)(z_t-u) = (1-\theta B^8)a_t .$$

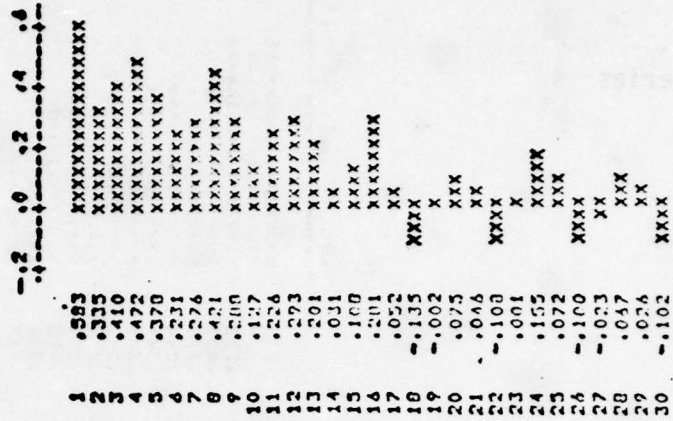
The estimation and diagnostic checking results are displayed in Table 3.

Estimating θ , σ , β and α as before from the residuals, we find

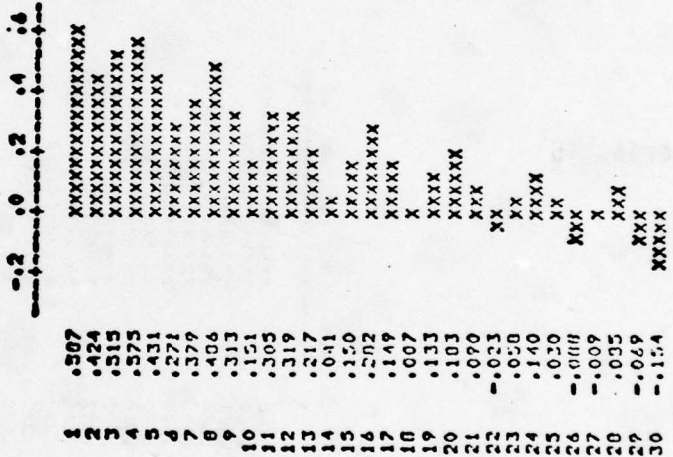
$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (-1.1, 56.45, \underline{.25}, \underline{0.0})$$

with 90% and 95% confidence regions for (β, α) plotted in Figure 3.16. The confidence regions are seen to be similar to those found for z_t .

Series 1



Series 2



Series 3

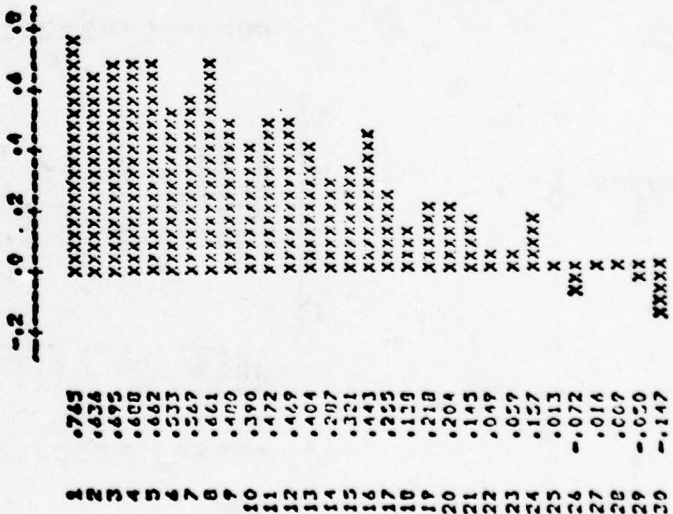
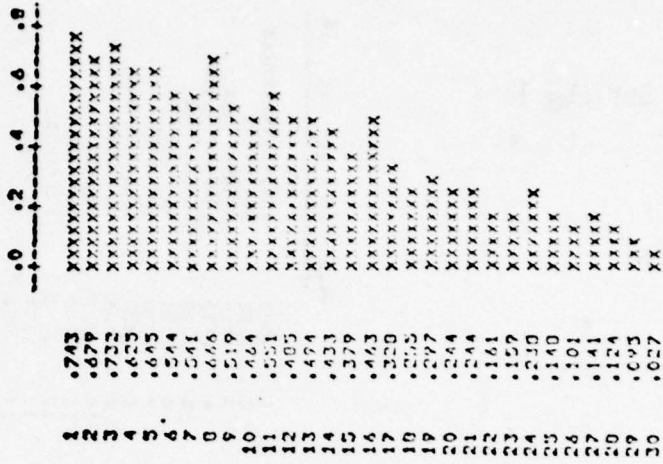
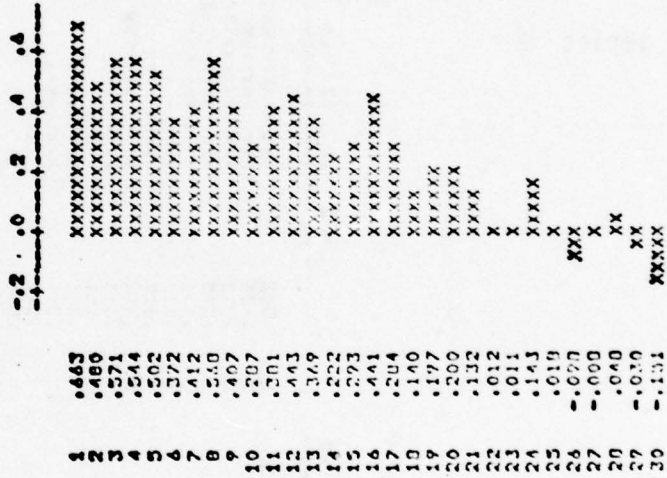


Figure 15(a) ACF for Wiebe's 12 series given in Table 2.

Series 4



Series 5



Series 6

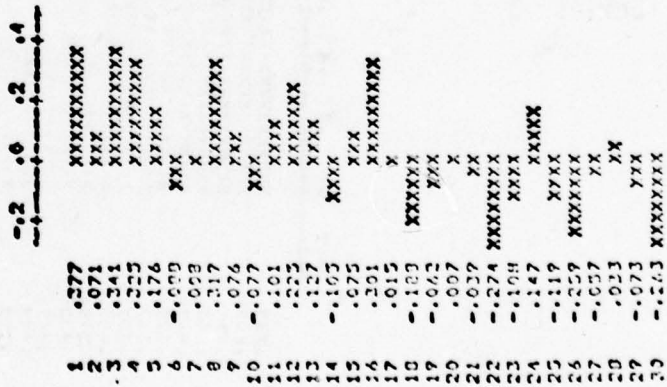


Figure 15(a) continued

Series 7

	.0	.2	.4	.6	.8
1	.597	XXXXXXXXXXXXXXXXXX			
2	.527	XXXXXXXXXXXXXXXXXX			
3	.639	XXXXXXXXXXXXXXXXXX			
4	.412	XXXXXXXXXXXXXXXXXX			
5	.577	XXXXXXXXXXXXXXXXXX			
6	.444	XXXXXXXXXXXXXXXXXX			
7	.570	XXXXXXXXXXXXXXXXXX			
8	.700	XXXXXXXXXXXXXXXXXX			
9	.534	XXXXXXXXXXXXXXXXXX			
10	.466	XXXXXXXXXXXXXXXXXX			
11	.534	XXXXXXXXXXXXXXXXXX			
12	.541	XXXXXXXXXXXXXXXXXX			
13	.503	XXXXXXXXXXXXXXXXXX			
14	.304	XXXXXXXXXXXXXXXXXX			
15	.427	XXXXXXXXXXXXXXXXXX			
16	.568	XXXXXXXXXXXXXXXXXX			
17	.400	XXXXXXXXXXXXXXXXXX			
18	.294	XXXXXXXXXXXXXXXXXX			
19	.303	XXXXXXXXXXXXXXXXXX			
20	.303	XXXXXXXXXXXXXXXXXX			
21	.397	XXXXXXXXXXXXXXXXXX			
22	.240	XXXXXXXXXXXXXXXXXX			
23	.300	XXXXXXXXXXXXXXXXXX			
24	.302	XXXXXXXXXXXXXXXXXX			
25	.241	XXXXXXXXXXXXXXXXXX			
26	.136	XXXX			
27	.192	XXXXXX			
28	.250	XXXXXXXXXX			
29	.177	XXXXXX			
30	.002	XXX			

Series 8

	.0	.2	.4	.6
1	.638	XXXXXXXXXXXXXXXXXX		
2	.527	XXXXXXXXXXXXXXXXXX		
3	.643	XXXXXXXXXXXXXXXXXX		
4	.577	XXXXXXXXXXXXXXXXXX		
5	.474	XXXXXXXXXXXXXXXXXX		
6	.441	XXXXXXXXXXXXXXXXXX		
7	.520	XXXXXXXXXXXXXXXXXX		
8	.544	XXXXXXXXXXXXXXXXXX		
9	.412	XXXXXXXXXXXXXXXXXX		
10	.394	XXXXXXXXXXXXXXXXXX		
11	.470	XXXXXXXXXXXXXXXXXX		
12	.310	XXXXXXXXXXXXXXXXXX		
13	.337	XXXXXXXXXXXXXXXXXX		
14	.292	XXXXXXXXXXXXXXXXXX		
15	.247	XXXXXXXXXXXXXXXXXX		
16	.336	XXXXXXXXXXXXXXXXXX		
17	.243	XXXXXX		
18	.169	XXXXXX		
19	.191	XXXXXX		
20	.160	XXXXXX		
21	.171	XXXXXX		
22	.051	XX		
23	.042	XX		
24	.171	XXXXXX		
25	.014	X		
26	-.027	XX		
27	.022	XX		
28	-.003	X		
29	-.020	XX		
30	-.023	XX		

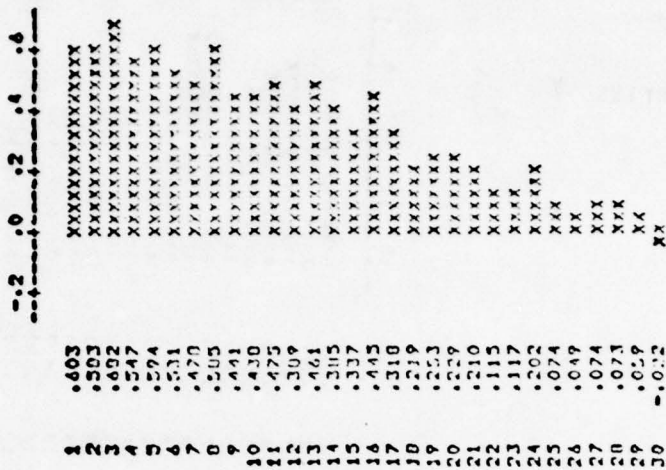
Series 9

	-.2	.0	.2	.4	.6
1	.305	XXXXXXXXXXXXXXXXXX			
2	.363	XXXXXXXXXXXXXXXXXX			
3	.559	XXXXXXXXXXXXXXXXXX			
4	.477	XXXXXXXXXXXXXXXXXX			
5	.407	XXXXXXXXXXXXXXXXXX			
6	.327	XXXXXXXXXXXXXXXXXX			
7	.369	XXXXXXXXXXXXXXXXXX			
8	.503	XXXXXXXXXXXXXXXXXX			
9	.259	XXXXXXXXXXXXXXXXXX			
10	.230	XXXXXXXXXXXXXXXXXX			
11	.304	XXXXXXXXXXXXXXXXXX			
12	.271	XXXXXXXXXXXXXXXXXX			
13	.270	XXXXXXXXXXXXXXXXXX			
14	.127	XXXX			
15	.151	XXXXXX			
16	.123	XXXXXXXXXXXXXXXXXX			
17	.122	XXXXXX			
18	.016	X			
19	.120	XXXXXX			
20	.124	XXXXXX			
21	.097	XXXX			
22	-.028	XX			
23	.009	X			
24	.129	XXXXXX			
25	-.040	XX			
26	-.074	XX			
27	-.023	XX			
28	-.003	X			
29	-.014	X			
30	-.113	XXXX			

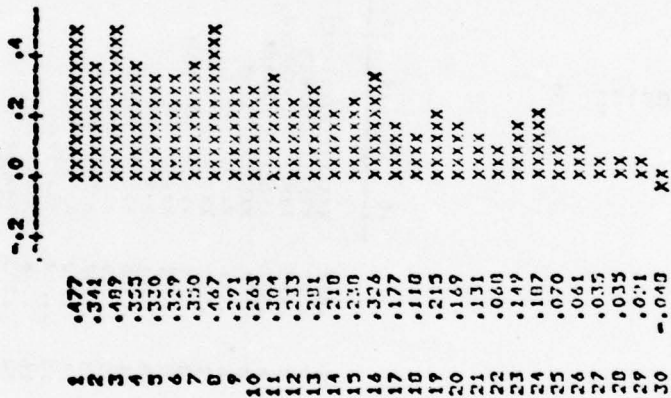
Figure 15(a)

continued

Series 10



Series 11



Series 12

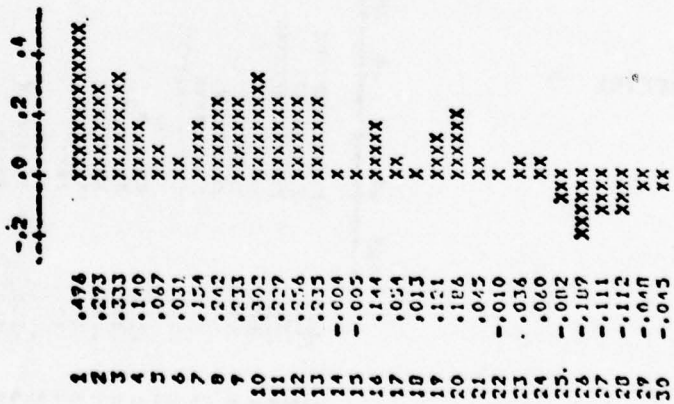
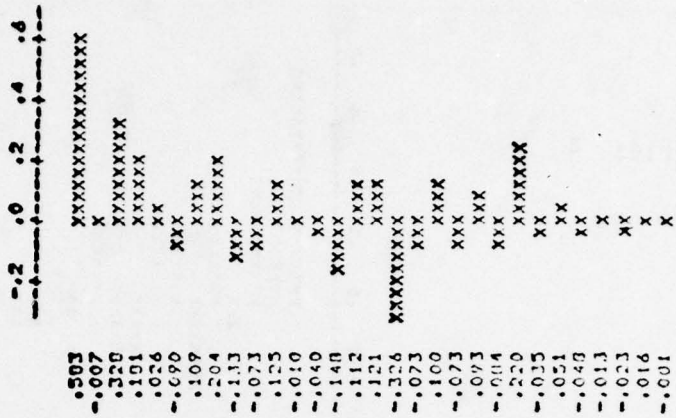
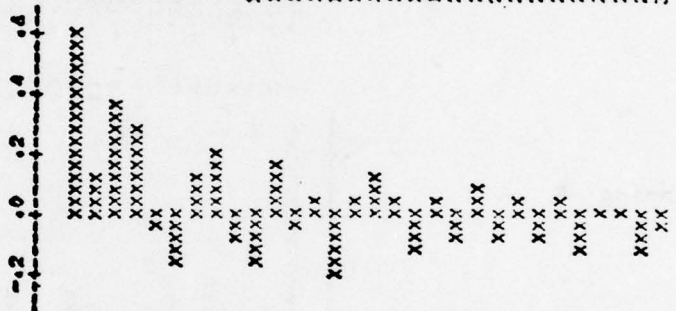


Figure 15(a) continued

Series 1



Series 2



Series 3

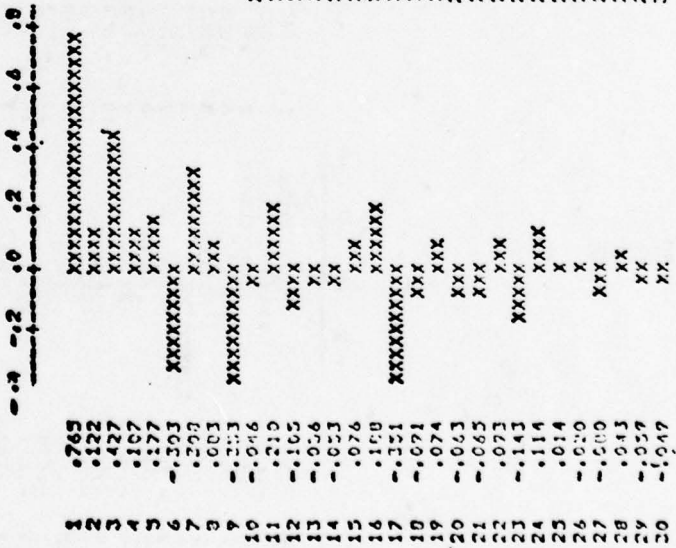
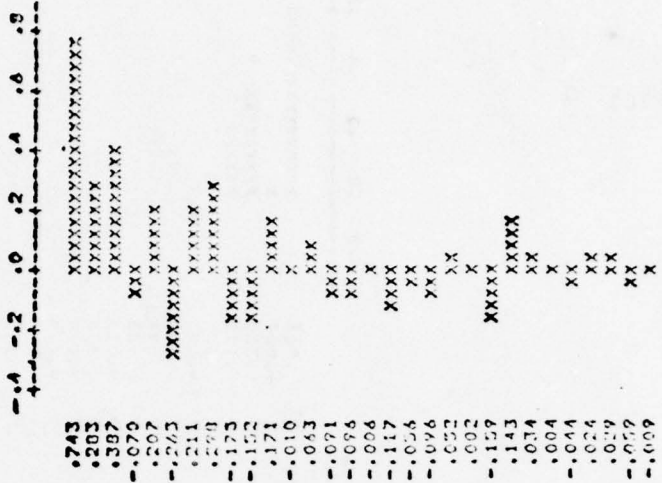
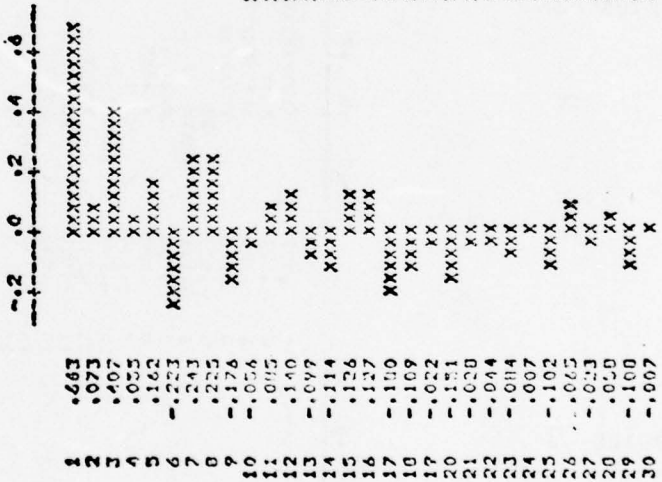


Figure 15(b) PACF for Wiebe's 12 series given in Table 2.

Series 4



Series 5



Series 6

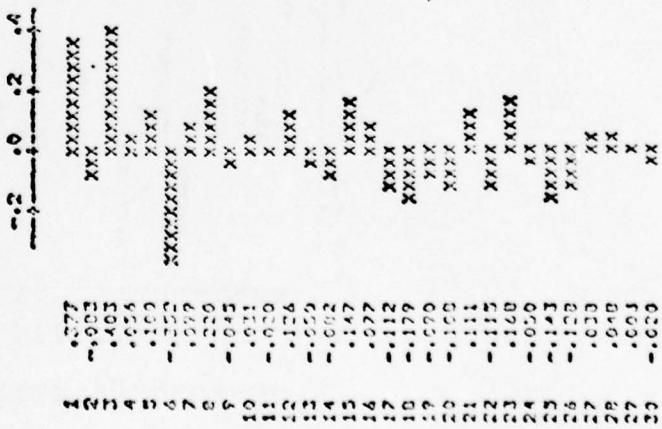


Figure 15(b) continued

Series 7

	-.2	.0	.2	.4	.6
1	.597	XXXXXX	XXXXXX	XXXXXX	XXXXXX
2	.267	XXXXXX	XXXXXX	XXXXXX	XXXXXX
3	.413	XXXXXX	XXXXXX	XXXXXX	XXXXXX
4	.224	XXXXXX	XXXXXX	XXXXXX	XXXXXX
5	.184	XXXXXX	XXXXXX	XXXXXX	XXXXXX
6	-.181	XXXXXX	XXXXXX	XXXXXX	XXXXXX
7	.199	XXXXXX	XXXXXX	XXXXXX	XXXXXX
8	.186	XXXXXX	XXXXXX	XXXXXX	XXXXXX
9	.044	XXXXXX	XXXXXX	XXXXXX	XXXXXX
10	-.134	XXXXXX	XXXXXX	XXXXXX	XXXXXX
11	-.092	XXXXXX	XXXXXX	XXXXXX	XXXXXX
12	-.030	XXXXXX	XXXXXX	XXXXXX	XXXXXX
13	.103	XXXXXX	XXXXXX	XXXXXX	XXXXXX
14	-.057	XXXXXX	XXXXXX	XXXXXX	XXXXXX
15	-.157	XXXXXX	XXXXXX	XXXXXX	XXXXXX
16	.001	XXXXXX	XXXXXX	XXXXXX	XXXXXX
17	-.023	XXXXXX	XXXXXX	XXXXXX	XXXXXX
18	-.109	XXXXXX	XXXXXX	XXXXXX	XXXXXX
19	-.047	XXXXXX	XXXXXX	XXXXXX	XXXXXX
20	-.049	XXXXXX	XXXXXX	XXXXXX	XXXXXX
21	.168	XXXXXX	XXXXXX	XXXXXX	XXXXXX
22	-.053	XXXXXX	XXXXXX	XXXXXX	XXXXXX
23	-.043	XXXXXX	XXXXXX	XXXXXX	XXXXXX
24	-.120	XXXXXX	XXXXXX	XXXXXX	XXXXXX
25	-.029	XXXXXX	XXXXXX	XXXXXX	XXXXXX
26	-.120	XXXXXX	XXXXXX	XXXXXX	XXXXXX
27	.026	XXXXXX	XXXXXX	XXXXXX	XXXXXX
28	.084	XXXXXX	XXXXXX	XXXXXX	XXXXXX
29	-.003	XXXXXX	XXXXXX	XXXXXX	XXXXXX
30	-.020	XXXXXX	XXXXXX	XXXXXX	XXXXXX

Series 8

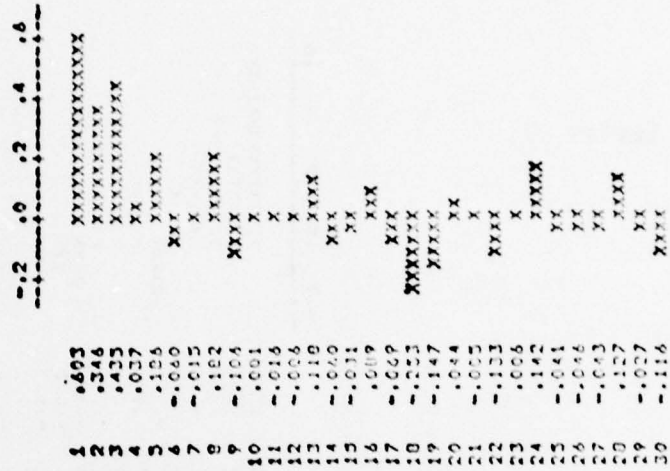
	-.2	.0	.2	.4	.6
1	.638	XXXXXX	XXXXXX	XXXXXX	XXXXXX
2	.202	XXXXXX	XXXXXX	XXXXXX	XXXXXX
3	.433	XXXXXX	XXXXXX	XXXXXX	XXXXXX
4	.092	XXXXXX	XXXXXX	XXXXXX	XXXXXX
5	.044	XXXXXX	XXXXXX	XXXXXX	XXXXXX
6	-.105	XXXXXX	XXXXXX	XXXXXX	XXXXXX
7	.220	XXXXXX	XXXXXX	XXXXXX	XXXXXX
8	.186	XXXXXX	XXXXXX	XXXXXX	XXXXXX
9	-.100	XXXXXX	XXXXXX	XXXXXX	XXXXXX
10	-.072	XXXXXX	XXXXXX	XXXXXX	XXXXXX
11	-.077	XXXXXX	XXXXXX	XXXXXX	XXXXXX
12	-.143	XXXXXX	XXXXXX	XXXXXX	XXXXXX
13	.163	XXXXXX	XXXXXX	XXXXXX	XXXXXX
14	-.073	XXXXXX	XXXXXX	XXXXXX	XXXXXX
15	-.074	XXXXXX	XXXXXX	XXXXXX	XXXXXX
16	.131	XXXXXX	XXXXXX	XXXXXX	XXXXXX
17	-.129	XXXXXX	XXXXXX	XXXXXX	XXXXXX
18	-.046	XXXXXX	XXXXXX	XXXXXX	XXXXXX
19	-.010	XXXXXX	XXXXXX	XXXXXX	XXXXXX
20	.040	XXXXXX	XXXXXX	XXXXXX	XXXXXX
21	.049	XXXXXX	XXXXXX	XXXXXX	XXXXXX
22	-.165	XXXXXX	XXXXXX	XXXXXX	XXXXXX
23	-.077	XXXXXX	XXXXXX	XXXXXX	XXXXXX
24	.139	XXXXXX	XXXXXX	XXXXXX	XXXXXX
25	-.005	XXXXXX	XXXXXX	XXXXXX	XXXXXX
26	.043	XXXXXX	XXXXXX	XXXXXX	XXXXXX
27	-.120	XXXXXX	XXXXXX	XXXXXX	XXXXXX
28	.021	XXXXXX	XXXXXX	XXXXXX	XXXXXX
29	.179	XXXXXX	XXXXXX	XXXXXX	XXXXXX
30	.087	XXXXXX	XXXXXX	XXXXXX	XXXXXX

Series 9

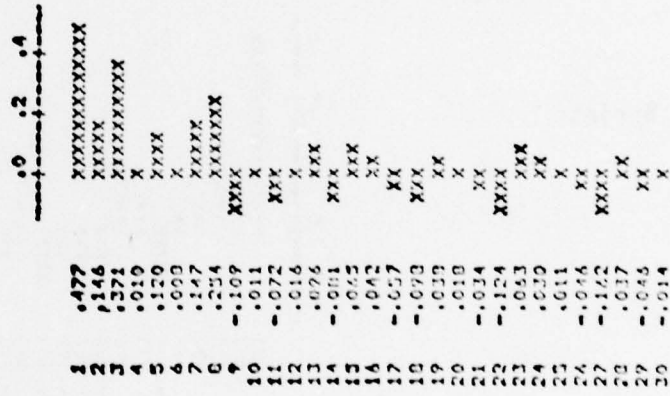
	-.2	.0	.2	.4	.6
1	.503	XXXXXX	XXXXXX	XXXXXX	XXXXXX
2	.172	XXXXXX	XXXXXX	XXXXXX	XXXXXX
3	.431	XXXXXX	XXXXXX	XXXXXX	XXXXXX
4	.352	XXXXXX	XXXXXX	XXXXXX	XXXXXX
5	.097	XXXXXX	XXXXXX	XXXXXX	XXXXXX
6	-.112	XXXXXX	XXXXXX	XXXXXX	XXXXXX
7	.039	XXXXXX	XXXXXX	XXXXXX	XXXXXX
8	.264	XXXXXX	XXXXXX	XXXXXX	XXXXXX
9	-.185	XXXXXX	XXXXXX	XXXXXX	XXXXXX
10	.093	XXXXXX	XXXXXX	XXXXXX	XXXXXX
11	-.117	XXXXXX	XXXXXX	XXXXXX	XXXXXX
12	.052	XXXXXX	XXXXXX	XXXXXX	XXXXXX
13	.100	XXXXXX	XXXXXX	XXXXXX	XXXXXX
14	-.169	XXXXXX	XXXXXX	XXXXXX	XXXXXX
15	-.022	XXXXXX	XXXXXX	XXXXXX	XXXXXX
16	.140	XXXXXX	XXXXXX	XXXXXX	XXXXXX
17	.023	XXXXXX	XXXXXX	XXXXXX	XXXXXX
18	-.103	XXXXXX	XXXXXX	XXXXXX	XXXXXX
19	-.039	XXXXXX	XXXXXX	XXXXXX	XXXXXX
20	-.042	XXXXXX	XXXXXX	XXXXXX	XXXXXX
21	.077	XXXXXX	XXXXXX	XXXXXX	XXXXXX
22	-.125	XXXXXX	XXXXXX	XXXXXX	XXXXXX
23	.010	XXXXXX	XXXXXX	XXXXXX	XXXXXX
24	.030	XXXXXX	XXXXXX	XXXXXX	XXXXXX
25	-.046	XXXXXX	XXXXXX	XXXXXX	XXXXXX
26	.044	XXXXXX	XXXXXX	XXXXXX	XXXXXX
27	-.120	XXXXXX	XXXXXX	XXXXXX	XXXXXX
28	.093	XXXXXX	XXXXXX	XXXXXX	XXXXXX
29	-.001	XXXXXX	XXXXXX	XXXXXX	XXXXXX
30	.010	XXXXXX	XXXXXX	XXXXXX	XXXXXX

Figure 15(b) continued

Series 10



Series 11



Series 12

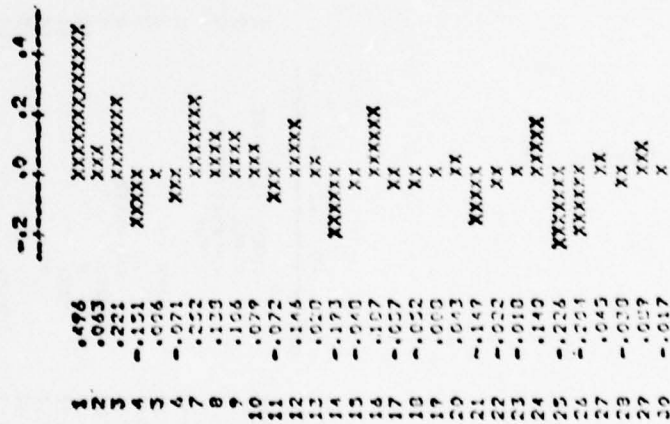


Figure 15(b) continued

Table 3. Estimation results for model $(1-\phi_1B-\phi_2B^2-\phi_3B^3)(1-\phi_8B^8)(z_t-u) = (1-\theta_8B^8)a_t$ for Wiebe's twelve series.

Series	estimated parameters								M.S.E. ($\times 10^{-2}$) 108 d.f.	Diagnostic checking	
	ϕ_1	ϕ_2	ϕ_3	ϕ_8	u	θ_8	$X^2_{.05}(14) = 23.69$	ACF, PACF		lack of fit	
1	.55	.04	.17	.89	577.63	.73	26.89	21.67	not very good	boundary	
2	.36	.27	.20	.79	630.28	.54	46.06	11.72	fair		
3	.68	.04	.14	.87	687.81	.48	34.09	26.88		yes	
4	.37	.22	.20	.95	946.15	.82	20.84	26.01		yes	
5	.52	.01	.25	.90	610.60	.65	19.68	13.15	random		
6	.51	.03	.29	.95	595.65	.87	18.15	13.91	fair		
7	.22	.27	.28	.98	1060.70	.76	29.04	10.79	fair		
8	.41	.00	.39	.95	726.38	.76	16.38	13.24	random		
9	.34	.04	.44	.87	571.52	.70	19.73	8.43	random		
10	.29	.13	.41	.87	615.28	.68	22.41	17.87	fair		
11	.31	.00	.37	.47	534.67	.19	27.61	11.52	fair		
12	.55	.00	.17	.93	643.77	.90	27.41	45.34		yes	

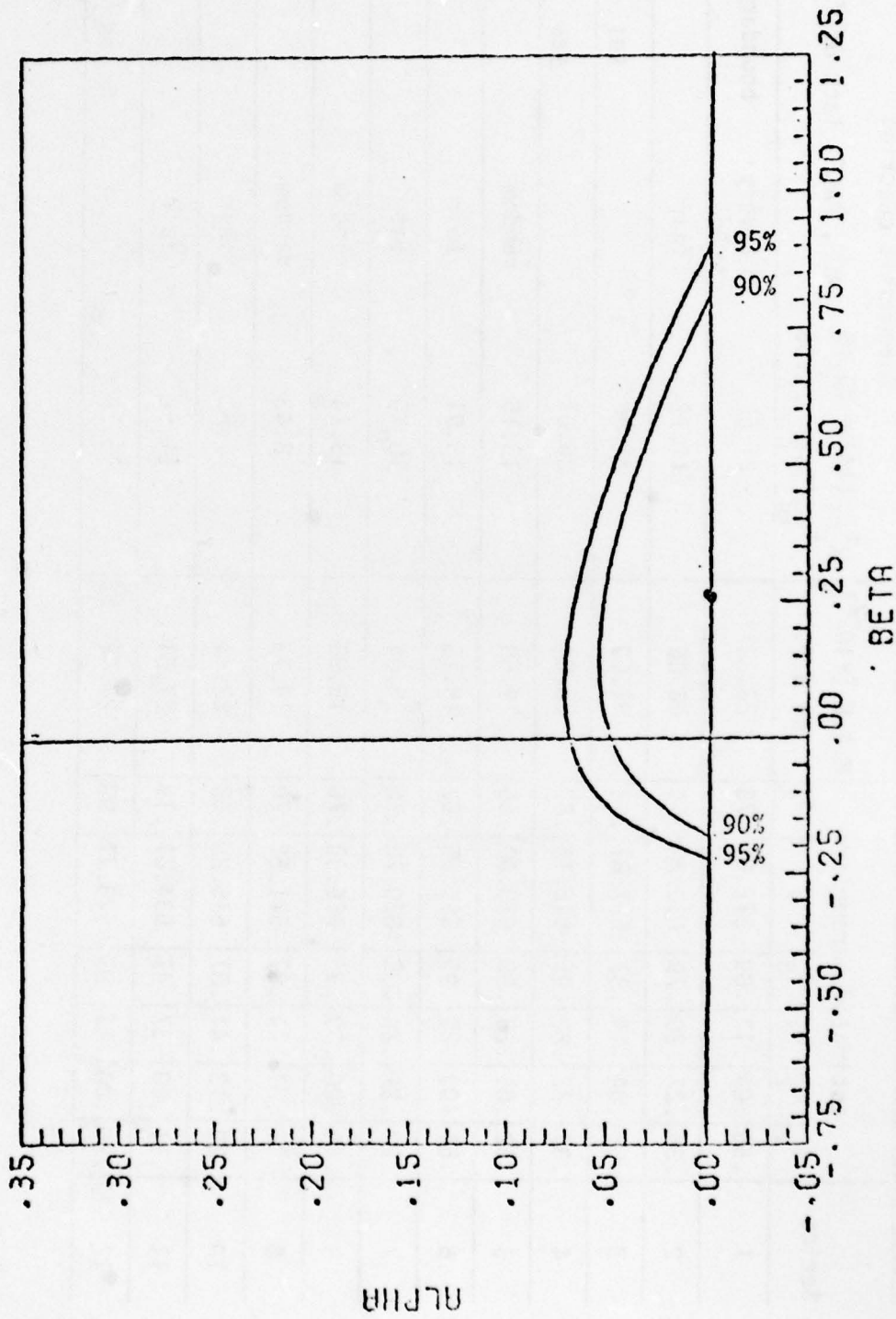


Figure 16. Approximate confidence regions for (β, α)
 (Wiebe's data: residuals from time series model)

In general, the relationship between the kurtosis for z_t and a_t is readily established for a stationary process $\phi(B)z_t = \theta(B)a_t$.

Consider first a moving average process

$$z_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t \quad \theta_i \text{'s not all zero}$$

$$= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}.$$

If $a_t, a_{t-1}, \dots, a_{t-q}$ are independently and identically distributed with cumulants $k_i(a_t)$ and coefficient of kurtosis $\lambda_4(a_t)$, then the i -th cumulant for z_t will be

$$k_i(a_t) + \theta_1^i k_i(a_t) + \theta_2^i k_i(a_t) + \dots + \theta_q^i k_i(a_t) = (1 + \theta_1^i + \dots + \theta_q^i) k_i(a_t).$$

The coefficient of kurtosis $\lambda_4(z_t)$ is then

$$\lambda_4(z_t) = \frac{k_4(z_t)}{k_2^2(z_t)} = \frac{(1 + \theta_1^4 + \dots + \theta_q^4) k_4(a_t)}{(1 + \theta_1^2 + \dots + \theta_q^2)^2 k_2^2(a_t)} = \frac{1 + \theta_1^4 + \dots + \theta_q^4}{(1 + \theta_1^2 + \dots + \theta_q^2)^2} \lambda_4(a_t).$$

It is easy to see since $\frac{1 + \theta_1^4 + \dots + \theta_q^4}{(1 + \theta_1^2 + \dots + \theta_q^2)^2} < 1$, that

z_t is more normal than a_t and in particular, z_t is normally distributed provided that a_t is so distributed.

Any stationary model $\phi(B)z_t = \theta(B)a_t$ can be written in the form

$$z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \quad \text{where} \quad \sum_{j=0}^{\infty} \psi_j \quad \text{converges.}$$

It then follows that in general

$$\lambda_4(z_t) = \frac{\left(\sum_{i=0}^{\infty} \psi_i^4\right) k_4(a_t)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2 k_2^2(a_t)} = \frac{\left(\sum_{i=0}^{\infty} \psi_i^4\right)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2} \lambda_4(a_t)$$

and $\frac{\left(\sum_{i=0}^{\infty} \psi_i^4\right)}{\left(\sum_{i=0}^{\infty} \psi_i^2\right)^2} < 1$. The previous conclusions for the moving

average, therefore, hold for any stationary series.

What we have is evidently a central limit effect which makes z_t more normal than a_t . In the important special case of a first order autoregressive process

$$\begin{aligned} (1-\phi B)z_t &= a_t & |\phi| < 1 \\ z_t &= (1+\phi B+\phi^2 B^2+\phi^3 B^3+\dots)a_t \\ &= a_t+\phi a_{t-1}+\phi^2 a_{t-2}+\phi^3 a_{t-3}+\dots \end{aligned}$$

$$\text{and } \lambda_4(z_t) = \frac{\sum_{i=0}^{\infty} (\phi_i)^4}{\left(\sum_{i=0}^{\infty} (\phi_i)^2\right)^2} \lambda_4(a_t) = \frac{1}{\left(\frac{1}{1-\phi^2}\right)^2} \lambda_4(a_t) = \frac{1-\phi^2}{1+\phi^2} \lambda_4(a_t).$$

The larger ϕ is, the smaller $\frac{1-\phi^2}{1+\phi^2}$ will be and the stronger will be the central limit effect.

The situation is different for a nonstationary process. This may be demonstrated by a simulation study using a random walk model.

(i) Generate n random normal deviates a_1, \dots, a_n .

(ii) Calculate $z_t = \sum_{i=1}^t a_i$.

(iii) Calculate $\frac{n \sum_{i=1}^n (z_i - \bar{z})^4}{\left(\sum_{i=1}^n (z_i - \bar{z})^2 \right)^2} - 3$ which is an estimate of the

coefficient of kurtosis λ_4 .

With $n = 100$, five such simulations gave estimates of λ_4 of $-.44, -.61, -.56, -.28, -.61$. With $n = 200$, three simulations gave estimates of λ_4 of $-1.14, -.95$ and -2.0 .

For Wiebe's data, nonstationary models were tried but did not fit as well as the stationary model. The observed result for this data that the tail behavior for z_t and a_t is similar will be expected when a_t is nearly normally distributed and a stationary model is appropriate.

An Alternative Approach: the Block Model

Traditionally, the kind of design that has been used for agricultural data employs small blocks. Since the drill machine

sowed eight rows at a time, every eight rows can conveniently be viewed as a block and associated with a block effect B_j . A treatment effect D_i could be associated with the position of the drill.

A plausible model is then

$$z_t = \mu + D_{i_t} + B_{j_t} + \epsilon_t \quad t = 1, \dots, 125, \quad 1 \leq i_t \leq 8, \\ 1 \leq j_t \leq 16$$

where $i_t = t_{\text{mod } 8} = t - [(t-1)/8] * 8$

$$j_t = [(t-1)/8] + 1$$

and $\sum D_{i_t} = 0 \quad \sum B_{j_t} = 0.$

The "design" thus produced is systematic and not randomized. However, we might allow for serial correlation in the errors by letting ϵ_t be an autoregressive process

$$(1 - \phi B)\epsilon_t = a_t \quad \text{where } a_t \text{'s are i.i.d. } N(0, \sigma^2).$$

The model now becomes

$$z_t = \mu + D_{i_t} + B_{j_t} + \epsilon_t \quad \text{with } (1 - \phi B)\epsilon_t = a_t \quad \text{Model (1)} \\ a_t \text{ i.i.d. } N(0, \sigma^2).$$

The parameters can now be estimated using nonlinear least squares.

Alternatively we can apply linear least squares estimation in the model

* $[x]$ means the integer part of x .

$$Y_t = (1-\phi)\mu + \nabla_i i_t + \beta_j j_t + a_t \quad \text{where}$$

$$Y_t = z_t - \phi z_{t-1}, \quad \nabla_i i_t = D_i - \phi D_{i,t-1}, \quad \beta_j j_t = B_j - \phi B_{j,t-1}$$

for various ϕ values, and plot the residual mean square against ϕ . The estimate of $\hat{\phi}$ is obtained at the minimum of the curve and all the other parameters are estimated by linear least squares with this $\hat{\phi}$ value.

Figure 17 shows this estimation procedure for Series 1 which yields $\hat{\phi} = 0.26$. The estimates of all the other parameters are given in Table 4. The ACF and PACF for the residuals are given in Figure 18. No significant pattern is noticeable.

If we now look back to the time series models we have fitted, we see that even the best fitted model for Series 1 has a residual mean square of over 2600 which is considerably larger than the residual mean square for the present model (2078). Even if we did not allow for serial correlation for the errors and made the assumption that the errors were i.i.d. $N(0, \sigma^2)$, i.e., $\phi = 0$, the residual mean square 2216 is still relatively small although ACF and PACF for the residuals (Figure 19) do now show some pattern. It is clear, however, that the block model does very well in explaining the data.

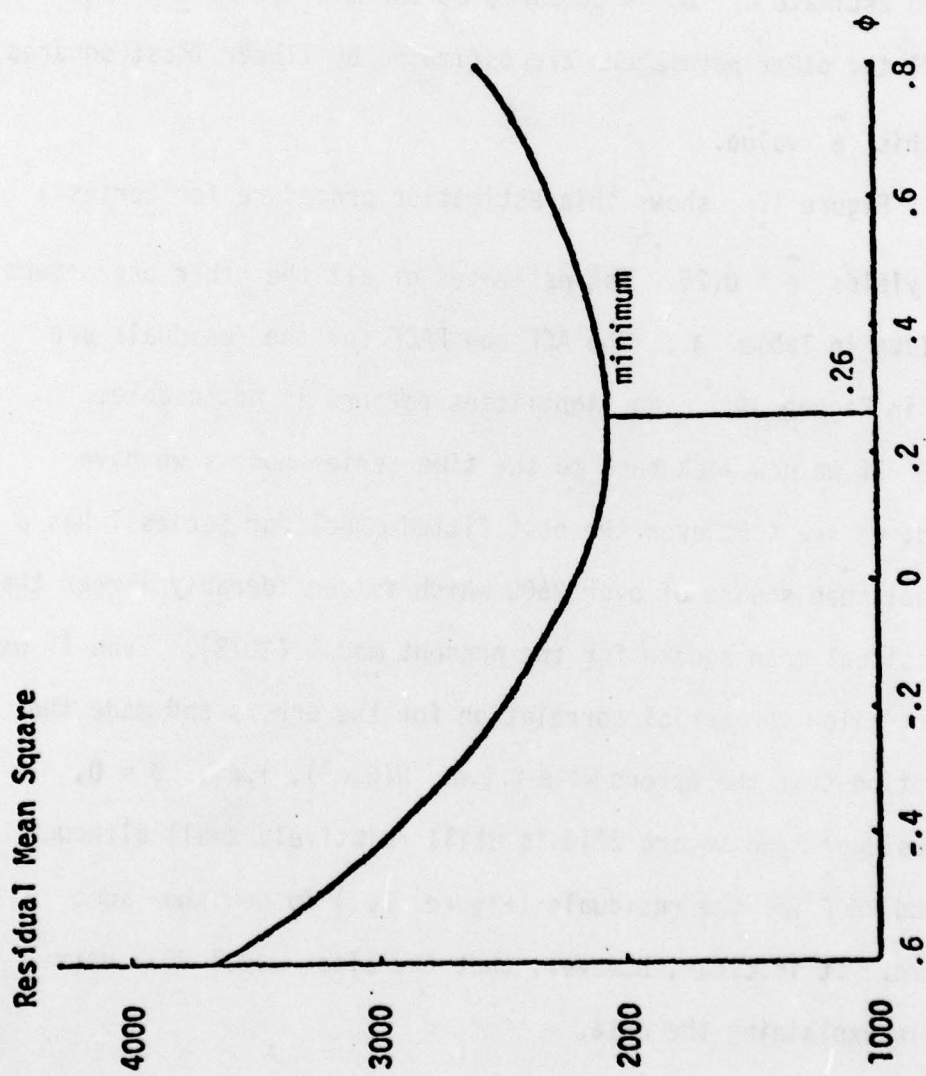
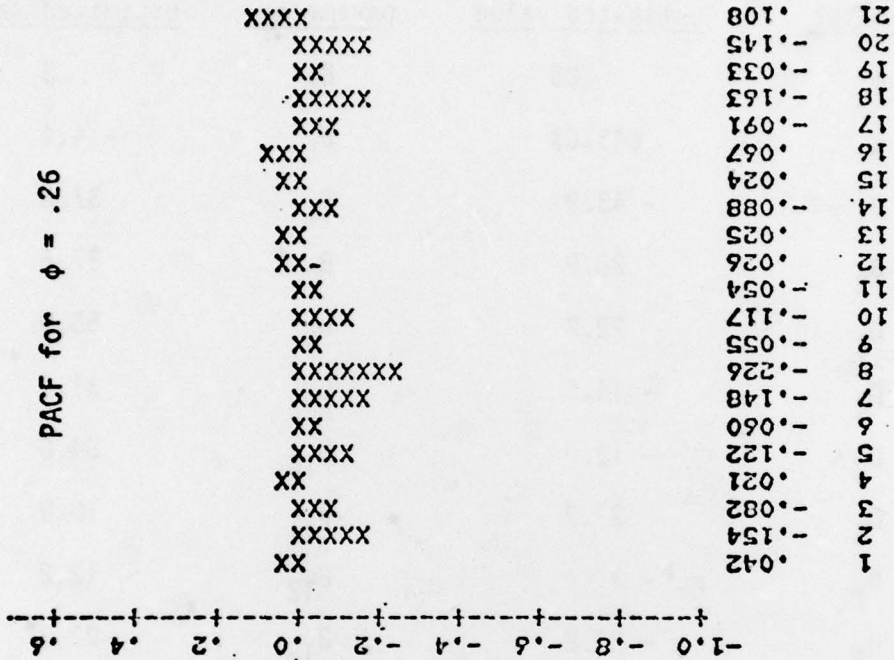


Figure 17 Residual mean square curve with the block model for Series 1.

Table 4 Estimates of the parameters in model (1)

<u>parameter</u>	<u>estimated value</u>	<u>parameter</u>	<u>estimated value</u>
ϕ	.26	B_4	.3
μ	643.63	B_5	- 4.4
D_1	- 43.9	B_6	37.0
D_2	20.9	B_7	37.6
D_3	72.7	B_8	55.9
D_4	- 14.6	B_9	31.3
D_5	- 13.6	B_{10}	54.0
D_6	27.7	B_{11}	16.8
D_7	- 11.4	B_{12}	12.2
D_8	- 37.8	B_{13}	-22.2
B_1	50.5	B_{14}	-41.7
B_2	33.5	B_{15}	-151.5
B_3	15.1	B_{16}	-124.4

PACF for $\phi = .26$



ACF for $\phi = .26$

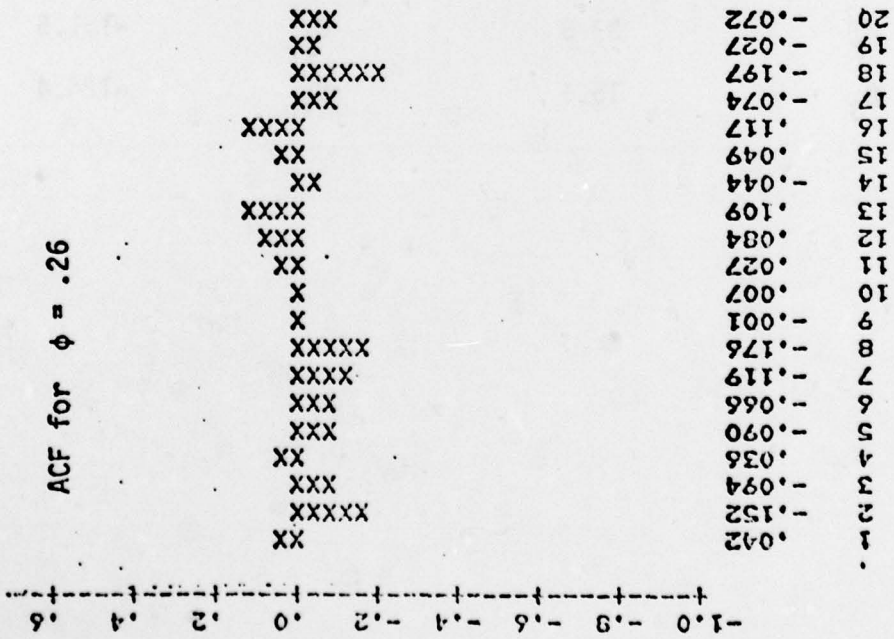


Figure 18 ACF and PACF for residuals obtained from fitting Series 1 to model (1).

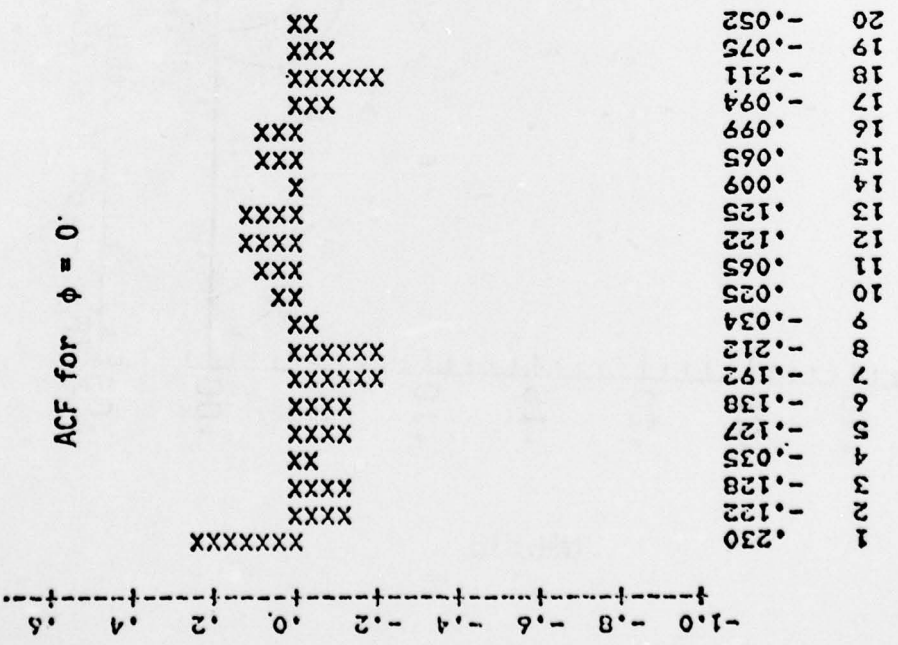
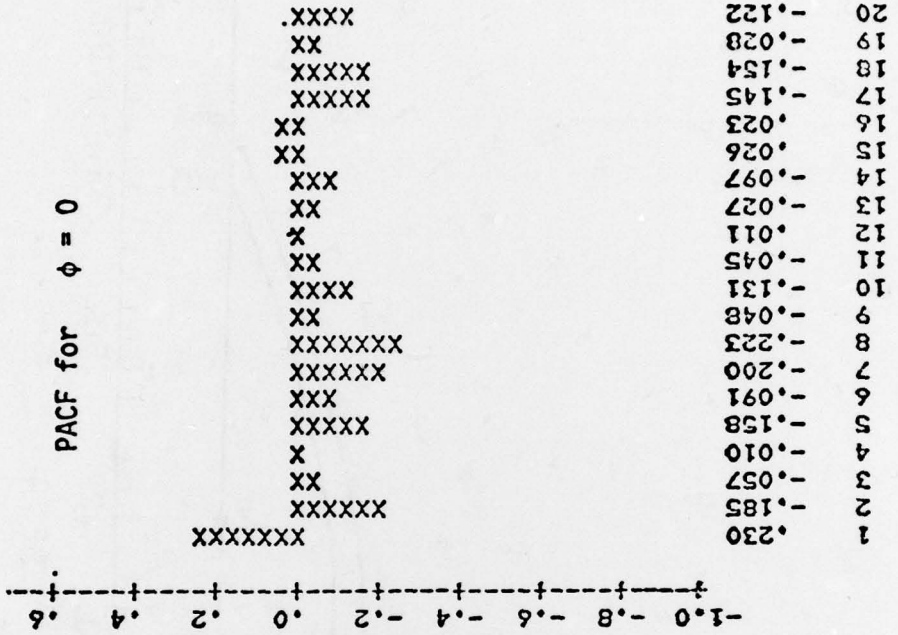


Figure 19 ACF and PACF for residuals obtained from fitting Series 1 to model (1).

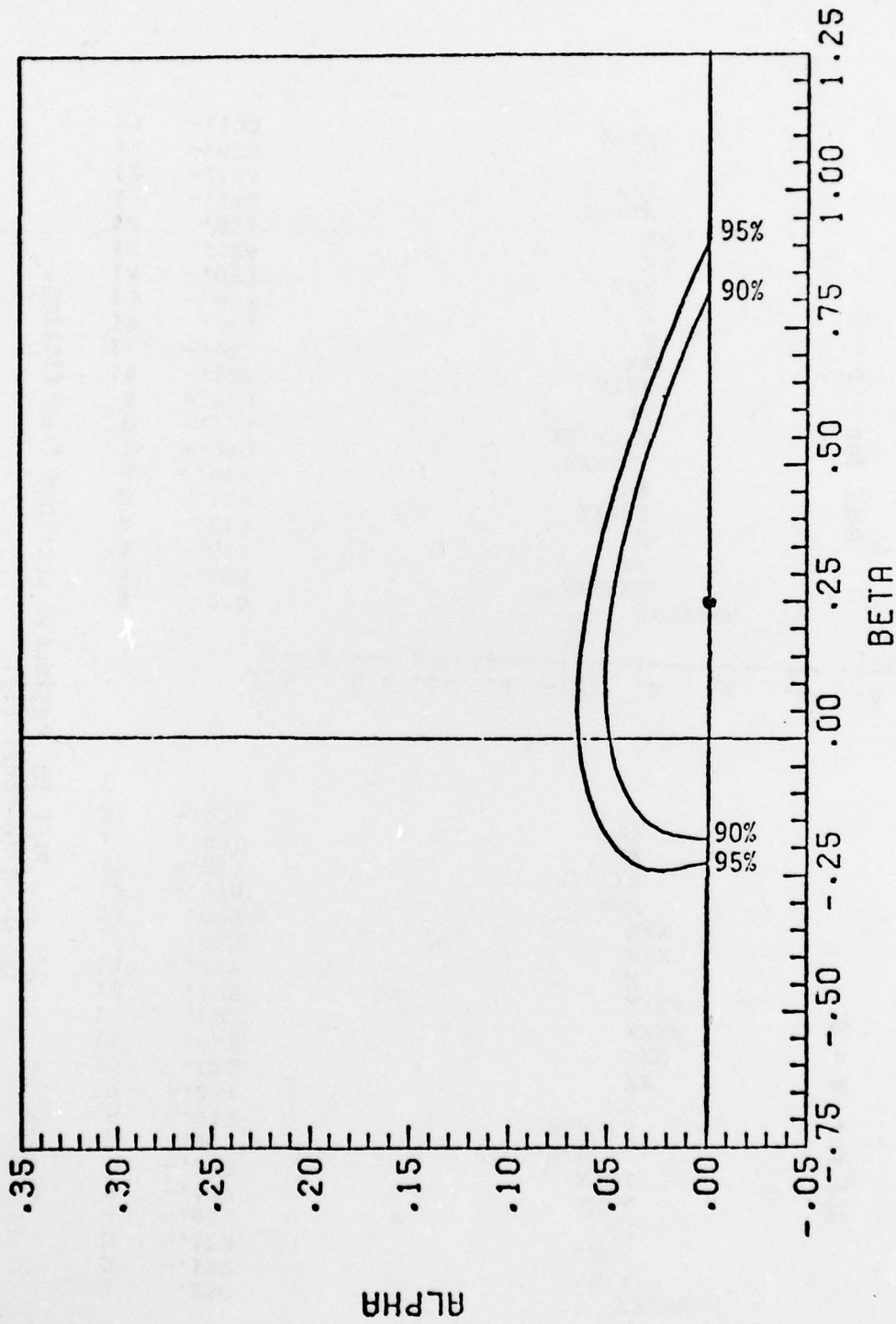


Figure 20 Approximate 90% and 95% confidence regions for (β, α)
 (Data: Residuals obtained from fitting Series 1 to
 model (1) with $\phi = 0.26$)

References

1. Bachelier, L. (1900), "Theorie de la speculation," Ann. Sci. Ec. norm. sup., Paris, Series 3, 17, 21.
2. Barbacki, S. and R. A. Fisher (1936), "A test of the supposed precision of systematic arrangements," Annals of Eugenics, 7, p 189-193.
3. Box, G. E. P. and G. M. Jenkins (1976), Time Series Analysis: Forecasting and Control, Holden-Day.
4. Lund, D. R. (1967), Parameter Estimation in a Class of Power Distributions, Ph.D. thesis, the University Wisconsin-Madison.
5. Michelson, A. A., Pease, F. G. and F. Pearson (1935), "Measurement of the velocity of light in a partial vacuum," Astrophysical Journal, 82, p 26-61.
6. Newcomb, S. (1891), "Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880-1882," Astronomical Papers, 2, p 107-203, U. S. Nautical Almanac Office.
7. Stigler, S. M. (1977), "Do robust estimators work with real data?" The Annals of Statistics, 5, p 1055-1098.
8. Tocher, J. F. (1928), "An investigation of the milk yield of dairy cows," Biometrika, 208, Part II, p 105-244.
9. Wichern, D. W., Miller, R. B. and D. Hsu (1976), "Changes of variance in first-order autoregressive time series models - with an application," Applied Statistics, 25, p 248-256.
10. Wiebe, G. A. (1935), "Variation and correlation in grain yield among 1,500 wheat nursery plots," Journal of Agricultural Research, 50, p 331-357.

If after fitting the block model we apply our analysis to the residuals, we obtain

$$(\hat{\theta}, \hat{\sigma}, \hat{\beta}, \hat{\alpha}) = (-.023, 1.0, .25, 0.0)$$

and approximate confidence regions for (β, α) are shown in Figure 20.

Again comparison with Figure 14 shows that the confidence region is remarkably little changed.

7. Conclusions

From this study of real data sets, we conclude the following:

- 1) Very heavy-tailed distributions do occur in practice. However, they are often caused by inhomogeneity in levels and variances. Serial correlation in a stationary process makes the data more normal than the generating shocks. However, serial correlation in a finite sequence from a non-stationary process generated from normal errors produces an apparent light-tailed distribution for z_t .
- 2) Data collected over a long period of time usually suffer from secular inhomogeneity and would, therefore, not be expected to be normally distributed. However, much of the inhomogeneity of this kind would be eliminated or reduced when appropriate designs and blocking were adopted, and appropriate input variables were included in the model.

- 3) For the data we have studied which was reasonably homogeneous, the distributions are most frequently slightly heavy-tailed and/or contaminated. However, light-tailed and contaminated light-tailed distributions seem also to occur.
- 4) After obvious inhomogeneities in a series (for example, in most cases, due to start-up phenomena) have been allowed for, the trade off between α and β causes the confidence regions to be elongated usually and to include suitably contaminated normal distributions. We believe, therefore, that most data sets of this kind could be adequately described by contaminated normal distributions with suitably chosen α and k .

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2002	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER 9 Technical
4. TITLE (and Subtitle) 6 A STUDY OF REAL DATA		5. TYPE OF REPORT & PERIOD COVERED Summary Report, no specific reporting period
7. AUTHOR(s) 10 Gina Chen George E. P. Box		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706		8. CONTRACT OR GRANT NUMBER(s) 15 DAAG29-75-C-0024
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, North Carolina 27709		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 4 - Probability, Statistics and Combinatorics
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 12 70		13. REPORT DATE 11 October 1979
		14. NUMBER OF PAGES 67
		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. 14 MRC-MSR-2002		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) real data sets, contaminated exponential power distribution, maximum likelihood estimation, secular inhomogeneity, serial correlation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Nine sets of real data are analyzed. The distributions within the contaminated exponential power family which best describe these data sets are obtained by maximum likelihood. It appears that heavy tailed distributions are often produced by secular inhomogeneity in mean and variance.		

221200

B