Mission Research Corporation

Copy No. 4

MRC-R-1517

Final Technical Report

IMPROVED TESTS FOR GLOBAL WARMING TREND EXTRACTION IN OCEAN ACOUSTIC TRAVEL-TIME DATA

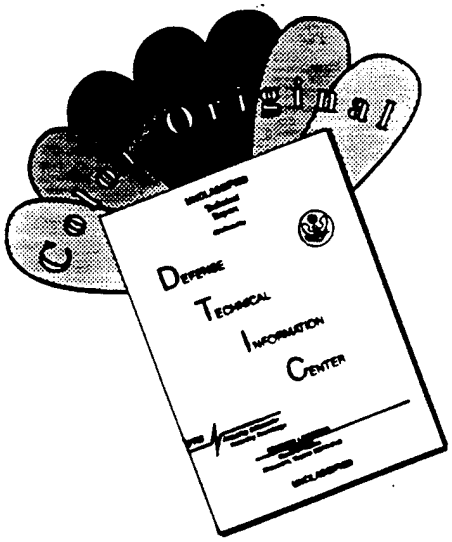
Steven Bottone
Henry L. Gray
Wayne A. Woodward

April 1996

Submitted by: MISSION RESEARCH CORPORATION
735 State Street, P. O. Drawer 719
Santa Barbara, CA 93102-0719

Sponsored by: Advanced Research Projects Agency
NMRO
ARPA Order No. 9376 Program Code No. 3716
Issued by ARPA/CMO under Contract
#MDA972-93-C-0021

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF COLOR PAGES WHICH DO NOT REPRODUCE LEGIBLY ON BLACK AND WHITE MICROFICHE.

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704.0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1 Apr 1996	Final Technical Report

4. TITLE AND SUBTITLE
Improved Tests For Global Warming Trend Extraction In Ocean Acoustic Travel-Time Data

5. FUNDING NUMBERS
Contract MDA972-93-C-0021

6. AUTHOR(s)
Steven Bottone
Henry L. Gray* Wayne A. Woodward*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Mission Research Corporation
735 State Street, P. O. Drawer 719
Santa Barbara, CA 93102-0719

8. PERFORMING ORGANIZATION REPORT NUMBER
MRC-R-1517

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714
Attn: Dr. Ralph Alewine/NMRO

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
* Southern Methodist University, Dallas, TX

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public distribution;
distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A possible indication of the existence of global climate warming is the presence of a trend in the travel time of an acoustic signal along several ocean paths over a period of many years. This report describes new, improved tests for testing for linear trend in time series data with correlated residuals. Previous tests that adjust for the correlation in the residuals tend to have the problem that observed significance levels are higher than nominal for small to moderate realization lengths whenever the residuals are correlated. We introduce a bootstrap-based procedure to test for trend in this setting which is better adapted to controlling the significance levels. The procedure is applied to acoustic travel-time data generated by the MASIG ocean model. It is shown how to generalize the improved method to multivariate, or vector, time series, which, in the ocean acoustics setting, corresponds to travel-time data on many ocean paths. An appendix describes the **TRENDS** software, which enables the user to perform these calculations using a graphical user interface (GUI).

14. SUBJECT TERMS		15. NUMBER OF PAGES
Global Warming	Time Series Analysis	
Ocean Acoustics	Statistical Models	16. PRICE CODE
Trend Extraction	Climate Change	

17. Security CLASSIFICATION OF REPORT	18. Security CLASSIFICATION OF THIS PAGE	19. Security CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18
298-102

SECURITY CLASSIFICATION OF THIS PAGE

CLASSIFIED BY:

DECLASSIFY ON:

TABLE OF CONTENTS

SECTION **PAGE**

1. INTRODUCTION 1

2. TESTING FOR TREND IN UNIVARIATE TIME SERIES DATA 4

   2.1. Ordinary Linear Regression 4

   2.2. A New Testing Procedure 6

   2.3. Results 8

3. TESTING FOR TREND IN MULTIVARIATE TIME SERIES DATA 12

   3.1. Ordinary Multivariate Linear Regression 12

   3.2. Extending the New Testing Procedure to the Multivariate Case 15

4. REFERENCES 18

APPENDIX

A. IMPROVED TESTS FOR TREND IN TIME SERIES DATA A-1

B. THE **TRENDS** SOFTWARE B-1

   B.1. Introduction B-1

   B.2. Getting Started B-1

   B.3. Trend Detection B-1

   B.4. Power of the Test (Probability of detecting a trend) B-2

   B.5. Trend Stability (Will the trend continue?) B-4

   B.6. Classification Statistics B-4

   B.7. Example B-6

   B.8. CD ROM B-7

LIST OF FIGURES

FIGURE **PAGE**

2-1 Time series for acoustic travel-time anomaly from the MASIG model. 9

A-1 Hansen and Lebedeff Series..... A-26

A-2 Simulation from MASIG Model: San Diego to Hawaii..... A-27

B-1 Main window of **TRENDS** show input time series.....B-2

B-2 Main window of **TRENDS** after complete run.....B-7

B-3 Cover for the CD ROM multimedia tutorial.....B-8

LIST OF TABLES

TABLE **PAGE**

2-1 Observed significance levels and powers for AR(10) model fit to monthly MASIG data 10

A-1 Burg estimates of ϕ_1 before and after removing the least squares line A-21

A-2 Observed 97.5th percentiles of $\hat{t}$ values calculated as in 5(d) based on 1000 realizations from the model $Y_t = a + bt + Z_t$ where $(1 - \phi_1 B)Z_t = a_t$ and $b = 0$ A-21

A-3 Observed significance levels associated with tests for $b = 0$ based on the model $Y_t = a + bt + Z_t$ where $(1 - \phi_1 B)Z_t = a_t$ A-22

A-4 Observed power associated with TB and TBA for $n = 100$ and various values of b where the model is $Y_t = a + bt + Z_t$ where $(1 - \phi_1 B)Z_t = a_t$ A-23

A-5 Observed significance levels and powers for "no line" and "line" models fit to the Hansen and Lebedeff temperature data A-24

A-6 Observed significance levels and powers for AR(10) model fit to monthly MASIG data A-25

1. INTRODUCTION

A possible indication of the existence of global climate warming is the presence of a trend in the travel time of an acoustic signal along several ocean paths over a period of many years. This report presents improved techniques for testing for trend in such time series data. For background on the use of statistical time series and trend extraction methods for ocean acoustic global warming studies the reader is referred to our previous report [Bottone, Gray, and Woodward, 1995].

The specific problem we address in this report is that of testing for trend using the model

$$X_t = a + bt + E_t, \tag{1}$$

where E_t may be highly correlated stationary noise. In the ocean acoustics problem, X_t represents the acoustic travel time along a fixed path at time t, however, all of the techniques discussed in this report apply to any time series which can be represented by the model given by equation (1). Testing for trend in a given set of time series data assumed to be modeled by equation (1) amounts to testing the hypothesis $b = 0$. If the value of b estimated from the data is significantly (at some given significance level) different from zero, a trend is said to exist in the data. If the distribution of the noise, E_t, is completely known, this problem is relatively simple. If, however, as is usually the case in practice, the distribution of E_t, or, at least, its parameters, must the estimated from the data, the problem is somewhat more difficult. In particular, when the correlation in the E_t is high, as appears to be the case with ocean acoustic data, the problem becomes extremely difficult due to the difficulty in obtaining reasonable parameter estimates without exceedingly long samples lengths.

Previously existing methodology for testing for trend in the model given by equation (1) is generally invalid when the correlation in the E_t is high, unless the sample size is unacceptably large. These methods are unable to differentiate between trends which are due to $b \neq 0$ and "temporary trends" that are due to high correlation in E_t. This inability to differentiate between a true trend and high correlation manifests itself in inflated significance levels. That is, when b is equal to zero in equation (1) and the E_t are highly correlated and existing methods are used to compute the significance level using a nominal 5% level, the percentage of realizations from (1) for which a nonzero b is (incorrectly) detected can be as high as 25%–50% for all existing tests. The impact is serious since Woodward and Gray [1993] show for atmospheric temperature data that when existing

methods for testing $b = 0$ in equation (1) are applied to the data one concludes that b is significantly greater than zero and, hence, warming is present. On the other hand, they also model the data as an ARIMA($p,1,q$), which is a plausible "correlation model", and the inference is the opposite, i.e., the current warming trend is temporary and should not be predicted to continue. The source of the conflicting results can be traced to the invalidity of existing tests for $b = 0$ in this setting.

As a result of this difficulty, Woodward and Gray stressed the importance of selecting the statistical model used to represent the data with care. Consequently, they developed a method to let the data select the model [Woodward and Gray, 1995]. In that paper, their procedure selected the "correlation model" as appropriate for the atmospheric temperature data, which implied that warming should not be predicted to continue. This method was also applied to simulated ocean acoustic travel time data in our previous report, where it was shown that it would take well over 20 years to reliably distinguish whether such data was best classified as coming from a line plus noise model as in equation (1) or a correlation model (ARIMA).

What is needed, however, is a more robust test for trend in equation (1) than currently exists. That is, we need a test for $b = 0$ that is valid for much higher correlated data than previous tests allow. By developing such a test we will have a model which is more robust to high correlation and, hence, should be more compatible with correlation models. It will not be possible to make such models completely compatible since the assumption that the E_t in equation (1) is stationary may not be reasonable. However, it is possible to dramatically improve the test for $b = 0$ so that it is compatible with correlation models in most cases. In our current work, described in this report, we have developed a new test for testing $b = 0$ in equation (1) that is valid for correlations as high as .95. That is, the new test appropriately maintains the specified significance level while continuing to have good detection capability when a trend is actually present. It should be remarked that although equation (1) models a linear trend, the test is sensitive to detecting any general increase, or decrease, in the data. It is interesting to note that when this new test is applied to the atmospheric temperature data we get agreement with the previous result of Woodward and Gray, that is, the new test determines that b is not significantly different from zero, which is compatible with the classification of the data as more likely coming from a correlation model that would predict no warming. This new method is described in some detail in section 2, with some of the more difficult technical discussion found in appendix A, which is a copy of a paper on this topic by the authors submitted for publication in the *Journal of Time Series Analysis.*

This new method for testing for trend has been developed in such a way that it can be directly extended to the multivariate, or vector, case. In the ocean acoustics problem, this corresponds to having a set of data consisting of travel times on several paths and testing for trend on all of the paths simultaneously. This generalization will be presented in detail in section 3.

Appendix B contains a description of the **TRENDS** software, which allows the user to perform the new test for trend on a selected set of time series data. It also contains routines which allow the data to select between the model of equation (1) or an ARIMA(p,1,0) correlation model. A given set of data can be modeled in various ways, the noise structure can be approximated, and the power of the test for trend can be computed. Questions such as how long will it take to detect a trend in data similar to a given data set and how large would the trend have to be in a given data set to be significant can be answered with simple applications of the software.

2. TESTING FOR TREND IN UNIVARIATE TIME SERIES DATA

In this section we develop a new method to test the hypothesis H_0: $b = 0$ in equation (1) against either the two-sided alternative H_1: $b \neq 0$ or the one-sided alternative H_1: $b < 0$ (or H_1: $b > 0$). The one-sided alternative $b < 0$ is more appropriate to the ocean acoustics problem since the travel time of any acoustic signal is expected to decrease with time on most paths if there is warming. We will concentrate here, however, on the two-sided alternative, $b \neq 0$, since it is most readily generalizable to the multivariate case.

In equation (1) X_t is assumed to be a random variable given by the discrete stochastic process $\{X_t; t = 0,\pm 1,\ldots\}$. In the ocean acoustics problem X_t represents travel time as a function of time, t. A realization of length n of the time series, X_t, is a set of real-valued outcomes which will be denoted $\{x_t; t = 1,\ldots,n\}$. Loosely speaking, a set of data, x_t, will be considered a realization from the time series, X_t. The noise component in equation (1), E_t, will be assumed to be a stationary autoregressive (AR) process of order p satisfying [Box and Jenkins, 1976; Gray et al., 1996]

$$\phi(B)E_t = a_t, \tag{2}$$

where B is the backward shift operator given by $B^k X_t = X_{t-k}$,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{3}$$

and a_t is discrete white (Gaussian) noise with zero mean and variance σ_a^2, i.e., $E[a_t] = 0$ and $E\left[a_t^2\right] = \sigma_a^2$. To motivate the new method, we first discuss testing for trend under the assumption that the noise in equation (1) is white, i.e., $\phi(B) \equiv 1$.

2.1. Ordinary Linear Regression

Given a set of time series data, $\{x_t; t = 1,\ldots,n\}$, we wish to test the hypothesis $b = 0$ in equation (1) under the assumption that $E_t = a_t$. If the hypothesis is rejected at an appropriate significance level (usually 5%), then it is generally accepted that a trend is present. In ordinary linear regression , the least squares estimators for b and a in equation (1) are

$$\hat{b} = \frac{\sum_{t=1}^{n} (t - \bar{t}) X_t}{\sum_{t=1}^{n} (t - \bar{t})^2}, \tag{4}$$

$$\hat{a} = \bar{X} - \hat{b}\bar{t}, \tag{5}$$

where

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t, \tag{6}$$

$$\bar{t} = \frac{1}{n} \sum_{t=1}^{n} t = \frac{n+1}{2}. \tag{7}$$

Sample estimates of these quantities are obtained by substituting the sample data x_t into the appropriate equation. Under the usual regression assumptions that the residuals are independent and normally distributed with mean zero and variance σ_a^2, the estimated standard error of $\hat{b}$ is given by

$$\hat{SE}(\hat{b}) = \left[\frac{12 \sum_{t=1}^{n} (X_t - \hat{a} - \hat{b}t)^2}{(n-2)n(n^2-1)} \right]^{1/2}. \tag{8}$$

Under these assumptions, the test statistic $\hat{t} = \hat{b} / \hat{SE}(\hat{b})$ is distributed as Student's t with $n-2$ degrees of freedom when the null hypothesis H_0: $b = 0$ is true. For a given realization, $|\hat{t}|$ can then be compared with the 2.5% critical value of the Student's t distribution with $n-2$ degrees of freedom, for the two-sided test. The null hypothesis is rejected, and the trend is said to be significant (at the 5% level), if $|\hat{t}| > t_{.975}(n-2)$, which is the critical region for the test. ($t_{.975}(n-2)$ is equal to 2.01 for $n = 50$, 1.98 for $n = 100$ and asymptotically equal to 1.96 for large n, i.e., it is asymptotically normal). For the one-sided (negative) test, the trend will be significant (at the 5% level) if $\hat{t} < -t_{.95}(n-2)$. A thorough discussion of linear regression theory can be found in many textbooks on mathematical statistics, such as Robinson and Silva [1979].

2.2. A New Testing Procedure

In this subsection we outline the new method to test H_0: $b = 0$ in equation (1). Full details can be found in appendix A. Notice that if $\phi(B)$ in (1) were known, then we could rewrite (1) as

$$\begin{aligned} \phi(B)X_t &= \phi(1)a + \left(\sum_{i=1}^{p} i\phi_i\right)b + \phi(1)bt + \phi(B)E_t \\ &= c + dt + a_t, \end{aligned} \tag{9}$$

where $c = \phi(1)a + \left(\sum_{i=1}^{p} i\phi_i\right)b$, $d = \phi(1)b$ and a_t is white noise. If E_t is stationary, which implies $\phi(1) > 0$, then $d = 0$ if and only if $b = 0$, and d and b have the same sign when $b \neq 0$. To test $b = 0$ in equation (1) we simply test $d = 0$ in equation (9), in which case we are able to use the usual regression-based standard errors as given in section 3.1 since the residuals are white.

In practice $\phi(B)$ is not known and must be estimated from the sample data. To estimate $\phi(B)$, we subtract the least-squares estimates of a and b (denoted $\hat{a}$ and $\hat{b}$) from the data to obtain the residuals

$$\hat{E}_t = X_t - \hat{a} - \hat{b}t. \tag{10}$$

These residuals do not follow the same model as E_t and in general are not stationary since

$$\begin{aligned} \hat{E}_t &= a + bt + E_t - \hat{a} - \hat{b}t \\ &= (a - \hat{a}) + (b - \hat{b})t + E_t, \end{aligned} \tag{11}$$

which does not have constant mean unless $b = \hat{b}$. However, in most cases we find it reasonable to assume these residuals are approximately AR(p), and we let $\hat{\phi}(B)$ denote the estimated autoregressive operator. We transform the data using $\hat{\phi}(B)$ to obtain

$$\begin{aligned} W_t &= \hat{\phi}(B)X_t \\ &= \hat{\phi}(1)a + \left(\sum_{i=1}^{p} i\hat{\phi}_i\right)b + \hat{\phi}(1)bt + g_t \\ &= c' + d't + g_t, \end{aligned} \tag{12}$$

where $c' = \hat{\phi}(1)a + \left(\sum_{i=1}^{p} i\hat{\phi}_i\right)b$, $d' = \hat{\phi}(1)b$, and $g_t = \hat{\phi}(B)\phi^{-1}(B)a_t$, which will not be white noise but should be a reasonably close approximation to it.

A straightforward application of the procedure (assuming g_t is white) is to use standard regression procedures to test for the significance of $\hat{d}'$, which should be a good estimate of d if g_t is close to being white. This estimation procedure is summarized as follows:

1. Estimate a and b using least squares.
2. Calculate $\hat{E}_t$ as in (3).
3. Find Burg estimates of $\phi(B)$ where $\phi(B)\hat{E}_t = a_t$. Call this estimate $\hat{\phi}(B)$.
4. Transform the data to obtain $\hat{\phi}(B)X_t = c' + d't + g_t$ where g_t is nearly white.
5. Calculate $\hat{t} = \hat{d}' / \hat{SE}(\hat{d}')$, where $\hat{d}'$ and its standard error are the usual least squares-based quantities assuming uncorrelated residuals. Compare $\hat{t}$ with $\hat{t}(n-p-2)$ critical values based on Student's t since $\hat{\phi}(B)X_t$ is of length $n-p$.

As is shown in appendix A, because of bias in the estimate $\hat{\phi}(B)$, the distribution of the test statistic, $\hat{t}$, defined above, is not close to Student's t distribution when the residuals are highly correlated, i.e., $\phi(1) \approx 0$, and the series length is small to moderate, $n \approx 100$. For example, for the model in equation (1) with $\phi(B) = 1-.95B$ and $b = 0$, this test had a significance level of approximately 25% for $n = 100$ using the usual critical regions based on Student's t, instead of the nominal 5%. We see that this procedure suffers from the same problem of excessive actual significance levels that occurs with all existing tests.

2.2.1. A bootstrap approach

This deficiency in maintaining the true significance level when $b = 0$ can be remedied by finding the true distribution and critical regions of the test statistic $\hat{t}$, which we propose to estimate by employing a bootstrap procedure. The advantages of using the test statistic $\hat{t}$ defined above is that it is quickly calculated so a bootstrap will not consume a prohibitive amount of computer time and it is readily generalizable to the multivariate case, as will be seen in section 3. In the bootstrap procedure we obtain $\hat{t}$ for a sample data set using steps 1–5 above. To simulate realizations under H_0 we estimate $\phi(B)$ in (1) assuming $b = 0$ and

denote the estimate by $\hat{\phi}^{(0)}(B)$. That is, under H_0 we assume that any trending behavior in the series is due to correlation structure alone. We then obtain by simulation B realizations from the autoregressive model with AR operator given by $\hat{\phi}^{(0)}(B)$. For the bth realization, $b = 1,\ldots,B$, we calculate $\hat{t}_b^*$ as in step 5. For the two-sided test, the null hypothesis is rejected at the α level of significance if $\hat{t} > t^*_{1-\alpha/2}$ or $\hat{t} < t^*_{\alpha/2}$ where t^*_β is the βth empirical quantile of $\left\{\hat{t}_b^*\right\}_{b=1}^{B}$. Because of the symmetric nature of $\hat{t}$, in practice we accomplish this test by rejecting H_0 if $|\hat{t}|>|t|^*_{1-\alpha}$, where $|t|^*_{1-\alpha}$ is the $(1-\alpha)$th empirical quantile of $\left\{|\hat{t}_b^*|\right\}_{b=1}^{B}$. Since the probability that a randomly selected member from the population is greater than or equal to the jth largest value is $j/(B+1)$, then by setting $\alpha = j/(B+1)$ it follows that $|t|^*_{1-\alpha}$ is the jth largest value of $\left\{|\hat{t}_b^*|\right\}_{b=1}^{B}$, e.g., if $\alpha = 0.05$ and $B = 399$ then $|t|^*_{1-\alpha}$ is the 20th largest value of $\left\{|\hat{t}_b^*|\right\}_{b=1}^{399}$. For a one-sided test, the α-level critical value is the $(1-\alpha)$th or αth empirical quantile of $\left\{\hat{t}_b^*\right\}_{b=1}^{B}$ depending on whether the alternative is H_1: $b > 0$ or H_1: $b < 0$, respectively.

2.2.2. A second application of the bootstrap

As is shown in appendix A, if $\phi(1)$ is near zero, that is, there is a root of the characteristic polynomial close to unity, the significance levels are still high after the bootstrap has been used to approximate the critical region. This phenomenon is caused by the bias in estimating $\phi(B)$. Appendix A describes a procedure, which relies on a second application of the bootstrap, which adjusts the test statistic, $\hat{t}$, by a factor, $\hat{C}$, which is less than one when $\phi(1)$ is near zero, to yield an adjusted test statistic, $\hat{t}_{\text{adj}} = \hat{C}\hat{t}$. This adjusted test statistic is then compared to the critical values of the distributions, such as $\left\{|\hat{t}_b^*|\right\}_{b=1}^{B}$, described above.

2.3. Results

Section 3 of appendix A shows in great detail the results of simulation studies designed to examine the performance of the new testing procedures. There it is shown (see table 3 in appendix A), that for a variety of noise models with high correlation, the observed significance levels are near the nominal 5% even for series lengths as small as $n = 50$. It is also shown there that the new tests also have substantial power in detecting trends in the time series studied.

2.3.1. Analysis of time series data from the MASIG model

As an example of the use of the new testing procedures, we analyze a time series produced by the MASIG model. The MASIG (Mesoscale Air-Sea Interaction Group) model is a reduced gravity ocean model driven by COADS (Comprehensive Ocean-Atmosphere Data Set) winds, coupled to an equatorial model at its southern boundary [Pares-Sierra and O'Brien, 1989]. The acoustic travel-time anomaly for a path from Hawaii to San Diego for a 20 year period is plotted in figure 2-1. The time axis is given in years, with data plotted every month, giving 240 points in the time series. As can be seen in the figure, the travel-time anomalies are between ± 2 seconds. This time series, which represents the model years 1970–1990, has no trend (the slope of the best fit straight line is close to zero). If there were warming occurring during this period the time series would look similar to that shown in figure 2-1, except that there would be added to it a warming trend. In this context, a warming trend would be given by a negative slope (the best fit straight line through the data would have a negative slope). A slope of –0.10 seconds/year, a change of –2 seconds over 20 years, would correspond to an approximate increase in temperature along the path of 0.01 degrees Celsius per year, or 0.2° C increase over 20 years (a slope of –0.05 sec/yr would give half these values).

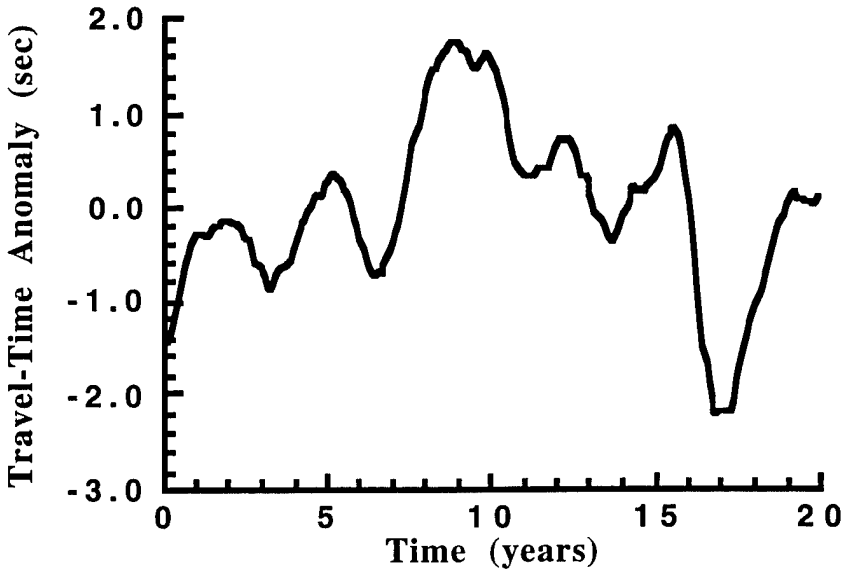


Figure 2-1. Time series for acoustic travel-time anomaly from the MASIG model.

We modeled the last 160 points of the MASIG data as an AR(10). This model satisfies

$$\begin{aligned}\big(1-2.1765B+1.6926B^2-.9443B^3+.7730B^4-.6793B^5+.4348B^6\\ -.0608B^7+.2369B^8-.4745B^9+.2024B^{10}\big)E_t = a_t,\end{aligned} \tag{13}$$

with white noise variance $\sigma_a^2 = .001196$. Our simulations consisted of generating realizations from the model given by equation (1) with E_t given by equation (13) with two warming scenarios: lines with slopes of $b = -0.05$ and $b = -0.10$ seconds/year. In table 2-1 we give results using the bootstrap test procedure (labeled TB in the table) and the adjusted bootstrap test procedure (TBA) assuming 5 to 30 years of monthly data. The table also contains the case of no warming, $b = 0$. There it can be seen that when $b = 0$ both tests produce actual significance levels which do not differ significantly from the nominal 5% level, with the only exception being TB at 5 years.

	Significance Level		Power –0.05 s/y slope		Power –0.10 s/y slope	
Years	TB	TBA	TB	TBA	TB	TBA
5	7.5	4.3	9.5	5.3	9.8	4.9
10	6.0	4.3	10.5	7.5	14.4	8.7
15	5.3	5.1	13.7	11.6	26.8	22.7
20	4.6	4.2	21.7	18.4	45.8	39.0
25	5.2	5.0	31.8	28.6	70.5	62.0
30	6.0	5.7	49.5	45.7	88.0	83.7

Table 2-1. Observed significance levels and powers for AR(10) model fit to monthly MASIG data.

The power in table 2-1 is computed by generating 1000 realizations from the model of equation (1) with noise from equation (13) and the number of bootstraps replications $B = 399$. The power is estimated by the percentage of realizations which have significant slope at the 5% nominal level using the one-sided tests describes above. As can be seen in the table, the TB test has more power than the TBA test, which one would be justified in using for those cases where the significance level is near the 5% nominal level, which is

usually the case here (except, perhaps, for 5 years). Even using the TB test, for warming to be detected at least 50% of the time in the simulations, over 20 years of data would be needed if the slope is –0.10 seconds/year and at least 30 years if the slope is –0.05 seconds/year.

3. TESTING FOR TREND IN MULTIVARIATE TIME SERIES DATA

In this section we address the problem of testing for trend in the multivariate, or vector, model

$$\mathbf{X}_t = \mathbf{a} + \mathbf{b}t + \mathbf{E}_t, \tag{14}$$

where the vector random variable $\mathbf{E}_t = [E_{t1}, E_{t2}, \ldots, E_{tm}]'$ may be highly correlated (in time) stationary noise. In the ocean acoustics case, $\mathbf{X}_t = [X_{t1}, X_{t2}, \ldots, X_{tm}]'$, where X_{ti} represents the travel time on path i, $i = 1, \ldots, m$ (the total number of paths is m), at time t. A sample set of data can be arranged in an $n \times m$ data matrix, $\mathbf{X}$, where $\{\mathbf{X}\}_{ti} = x_{ti}$, and x_{ti} is the measured travel time on path i, $i = 1, \ldots, m$, at time t, $t = 1, \ldots, n$. For a given data set, $\mathbf{X}$, we wish to test the hypothesis H_0: $\mathbf{b} = \mathbf{0}$ against the alternative hypothesis H_1: $\mathbf{b} \neq \mathbf{0}$. Other alternative hypotheses will not be dealt with here for two reasons: 1) it is not clear in the ocean acoustics problem what the alternative hypothesis should be since it is expected under the global warming hypothesis that some paths will warm (decrease in travel time) while others may cool (increase in travel time) and 2) the mathematics necessary to treat other alternative hypotheses is quite formidable and beyond the approach used here.

We assume that the noise in equation (14), $\mathbf{E}_t$, is given by a multivariate autoregressive process of order p satisfying

$$\mathbf{E}_t = \Phi_1 \mathbf{E}_{t-1} + \Phi_2 \mathbf{E}_{t-2} + \ldots + \Phi_p \mathbf{E}_{t-p} + \mathbf{U}_t, \tag{15}$$

where Φ_1, Φ_2,..., Φ_p, are real $m \times m$ matrices and $\mathbf{U}_t$ is a multivariate white noise vector such that $E[\mathbf{U}_t] = \mathbf{0}$, $E[\mathbf{U}_t \mathbf{U}_t'] = \Sigma$, and $E[\mathbf{U}_t \mathbf{U}_{t+k}'] = \mathbf{0}$, $k \neq 0$. To motivate the generalization of the new method for testing for trend from the univariate to the multivariate case, we first discuss testing for trend in the multivariate setting under the assumption that the noise in equation (14) is white, i.e., $\mathbf{E}_t = \mathbf{U}_t$.

3.1. Ordinary Multivariate Linear Regression

Let us rewrite equation (14) for n data vectors as

$$\mathbf{X} = \mathbf{HB} + \mathbf{E}, \tag{16}$$

where $\mathbf{X}$ is the $n \times m$ data matrix defined above

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}, \tag{17}$$

H, the "design matrix", is an $n \times 2$ matrix given by

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}, \tag{18}$$

B is a $2 \times m$ matrix defined by

$$\mathbf{B} = \begin{bmatrix} \mathbf{a}' \\ \hline \mathbf{b}' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{bmatrix}, \tag{19}$$

and the error matrix, $n \times m$, is given by

$$\mathbf{E} = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1m} \\ E_{21} & E_{22} & \cdots & E_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ E_{n1} & E_{n2} & \cdots & E_{nm} \end{bmatrix}. \tag{20}$$

The assumption in ordinary linear regression is that the m observations at time t have covariance matrix Σ, but observations from different times are uncorrelated, i.e., $E[E_{ti}E_{sj}] = \delta_{ts}\Sigma_{ij}$, $i,j = 1,\ldots,m$, $t,s = 1,\ldots,n$, where δ_{ts} is the Kronecker delta.

The least-squares estimate of **B**, denoted $\hat{\mathbf{B}}$, is found by minimizing $\mathrm{tr}[(\mathbf{X} - \mathbf{HB})'(\mathbf{X} - \mathbf{HB})]$. The result is [Johnson and Wichern, 1988]

$$\hat{\mathbf{B}} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{X}. \tag{21}$$

It can be shown that $\hat{\mathbf{B}}$ is an unbiased estimator for **B**, i.e., $E[\hat{\mathbf{B}}] = \mathbf{B}$ and that the covariance of $\hat{\mathbf{B}}$ is given by [Anderson, 1984]

$$\mathrm{cov}[\hat{\mathbf{B}}] = \Sigma \otimes (\mathbf{H}'\mathbf{H})^{-1}, \tag{22}$$

where $\Sigma \otimes (\mathbf{H}'\mathbf{H})^{-1}$ is the direct product of the matrices Σ and $(\mathbf{H}'\mathbf{H})^{-1}$. Focusing on $\hat{\mathbf{b}}$, the second row of $\hat{\mathbf{B}}$, we have

$$\text{cov}[\hat{\mathbf{b}}] = \lambda^{-1}\Sigma, \tag{23}$$

where λ^{-1} is the 22 element of $(\mathbf{H}'\mathbf{H})^{-1}$. It can easily be shown that

$$\lambda^{-1} = \left\{(\mathbf{H}'\mathbf{H})^{-1}\right\}_{22} = \left(\sum_{t=1}^{n}(t-\bar{t})^2\right)^{-1} = \frac{12}{n(n^2-1)}. \tag{24}$$

To establish distributional results the further assumption is made that the noise vector, $\mathbf{E}_t$, is m-variate normal. Under this assumption, it follows that the estimator, $\hat{\mathbf{b}}$, which is some linear combination of the $\mathbf{E}_t$, is also m-variate normal with mean $\mathbf{b}$ and covariance given by equation (24), i.e., $\hat{\mathbf{b}} \sim N(\mathbf{b}, \lambda^{-1}\Sigma)$. The residual matrix, $\hat{\mathbf{E}}$, is defined by

$$\hat{\mathbf{E}} = \mathbf{X} - \mathbf{H}\hat{\mathbf{B}} = \left[\mathbf{I} - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right]\mathbf{X}, \tag{25}$$

and the residual sum of squares, $\hat{\mathbf{E}}'\hat{\mathbf{E}}$, can be shown to be independent of $\hat{\mathbf{b}}$ and to be distributed as a Wishart distribution with $n-2$ degrees of freedom, i.e., $\hat{\mathbf{E}}'\hat{\mathbf{E}} \sim W_m(n-2,\Sigma)$ [Johnson and Wichern, 1988]. An unbiased estimator of Σ is given by

$$\mathbf{S} = \frac{1}{n-2}\hat{\mathbf{E}}'\hat{\mathbf{E}} = \frac{1}{n-2}(\mathbf{X} - \mathbf{H}\hat{\mathbf{B}})'(\mathbf{X} - \mathbf{H}\hat{\mathbf{B}}). \tag{26}$$

We now define the test statistic, T^2, by

$$T^2 = \lambda(n-2)(\hat{\mathbf{b}} - \mathbf{b})'(\hat{\mathbf{E}}'\hat{\mathbf{E}})^{-1}(\hat{\mathbf{b}} - \mathbf{b}) = \lambda(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{S}^{-1}(\hat{\mathbf{b}} - \mathbf{b}). \tag{27}$$

It can be proved [Seber, 1984] that in our case, i.e., $\hat{\mathbf{b}} \sim N(\mathbf{b}, \lambda^{-1}\Sigma)$, $\hat{\mathbf{E}}'\hat{\mathbf{E}} \sim W_m(n-2,\Sigma)$ and $\hat{\mathbf{b}}$ independent of $\hat{\mathbf{E}}'\hat{\mathbf{E}}$, then

$$\frac{n-2-m+1}{m}\frac{T^2}{n-2} \sim F(m, n-2-m+1), \tag{28}$$

where $F(m,n-2-m+1)$ is the F-distribution with m and $n-2-m+1$ degrees of freedom.

To test the hypothesis H_0: $\mathbf{b} = \mathbf{0}$ against the alternative hypothesis H_1: $\mathbf{b} \neq \mathbf{0}$, we assume $\mathbf{b} = \mathbf{0}$ in equation (27) and compare

$$\hat{F} = \frac{n-2-m+1}{m}\frac{T^2}{n-2} \tag{29}$$

with the critical value at the 5% significance level of the F-distribution, $F_{.95}(m, n-2-m+1)$. The null hypothesis is rejected and the trend is said to be significant (at the 5% level) if $\hat{F} > F_{.95}(m, n-2-m+1)$.

3.2. Extending the New Testing Procedure to the Multivariate Case

In this subsection we generalize to new testing procedure introduced in section 2.2 to test the hypothesis H_0: $\mathbf{b} = \mathbf{0}$ in the vector line plus noise model of equation (14). The noise in equation (14) is given by the AR(p) process satisfying equation (15) which we rewrite

$$\Phi(B)\mathbf{E}_t = \mathbf{U}_t, \tag{30}$$

where the $m \times m$ matrix operator $\Phi(B)$ is given by

$$\Phi(B) = \mathbf{I} - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p, \tag{31}$$

with $\mathbf{I}$ the $m \times m$ identity matrix. If $\Phi(B)$ were known, equation (14) could be rewritten

$$\begin{aligned} \Phi(B)\mathbf{X}_t &= \Phi(1)\mathbf{a} + \left(\sum_{i=1}^{p} i\Phi_i\right)\mathbf{b} + \Phi(1)\mathbf{b}t + \Phi(B)\mathbf{E}_t \\ &= \mathbf{c} + \mathbf{d}t + \mathbf{U}_t, \end{aligned} \tag{32}$$

where $\mathbf{c} = \Phi(1)\mathbf{a} + \left(\sum_{i=1}^{p} i\Phi_i\right)\mathbf{b}$, $\mathbf{d} = \Phi(1)\mathbf{b}$, and $\mathbf{U}_t$ is vector white noise. If $\mathbf{E}_t$ is a stationary operator, which implies $\det(\Phi(1)) > 0$, then $\mathbf{d} = \mathbf{0}$ if and only if $\mathbf{b} = \mathbf{0}$. To test $\mathbf{b} = \mathbf{0}$ in equation (14) we test $\mathbf{d} = \mathbf{0}$ in equation (32) using the ordinary multivariate linear regression-based test described in section 3.1.

As in the univariate case, $\Phi(B)$ is not known and must be estimated from the sample data. After computing the least-squares estimates, $\hat{\mathbf{B}} = \left[\hat{\mathbf{a}}' \,\vdots\, \hat{\mathbf{b}}'\right]'$, as in equation (21) compute the residuals

$$\hat{\mathbf{E}}_t = \mathbf{X}_t - \hat{\mathbf{a}} - \hat{\mathbf{b}}t. \tag{33}$$

These residuals do not follow the same model as $\mathbf{E}_t$ and in general are not stationary since

$$\begin{aligned} \hat{\mathbf{E}}_t &= \mathbf{a} + \mathbf{b}t + \mathbf{E}_t - \hat{\mathbf{a}} - \hat{\mathbf{b}}t \\ &= (\mathbf{a} - \hat{\mathbf{a}}) + (\mathbf{b} - \hat{\mathbf{b}})t + \mathbf{E}_t, \end{aligned} \tag{34}$$

which does not have constant mean unless $\mathbf{b} = \hat{\mathbf{b}}$. However, in most cases we find it reasonable to assume these residuals are approximately multivariate AR(p), and we let $\hat{\Phi}(B)$ denote the estimated autoregressive operator. We transform the data using $\hat{\Phi}(B)$ to obtain

$$\begin{aligned} \mathbf{W}_t &= \hat{\Phi}(B)\mathbf{X}_t \\ &= \hat{\Phi}(1)\mathbf{a} + \left(\sum_{i=1}^{p} i\hat{\Phi}_i \right)\mathbf{b} + \hat{\Phi}(1)\mathbf{b}t + \mathbf{g}_t \\ &= \mathbf{c}' + \mathbf{d}'t + \mathbf{g}_t, \end{aligned} \tag{35}$$

where $\mathbf{c}' = \hat{\Phi}(1)\mathbf{a} + \left(\sum_{i=1}^{p} i\hat{\Phi}_i \right)\mathbf{b}$, $\mathbf{d}' = \hat{\Phi}(1)\mathbf{b}$, and $\mathbf{g}_t = \hat{\Phi}(B)\Phi^{-1}(B)\mathbf{U}_t$, which will not be white noise but should be a reasonably close approximation to it.

We proceed as in the univariate case (assuming $\mathbf{g}_t$ is white) and begin to use standard regression procedures to test for the significance of $\hat{\mathbf{d}}'$, which should be a good estimate of $\mathbf{d}$ if $\mathbf{g}_t$ is close to being white. This estimation procedure is summarized as follows:

1. Estimate $\mathbf{a}$ and $\mathbf{b}$ using least squares.
2. Calculate $\hat{\mathbf{E}}_t$ as in (33).
3. Find estimates of $\Phi(B)$ where $\Phi(B)\hat{\mathbf{E}}_t = \mathbf{U}_t$. Call this estimate $\hat{\Phi}(B)$.
4. Transform the data to obtain $\mathbf{W}_t = \hat{\Phi}(B)\mathbf{X}_t = \mathbf{c}' + \mathbf{d}'t + \mathbf{g}_t$ where $\mathbf{g}_t$ is nearly white.
5. Calculate $\hat{F}$ as in section 3.1 using the vector series $\mathbf{W}_t$ and assuming uncorrelated residuals. Compare $\hat{F}$ with $F_{.95}(m, n-p-2-m+1)$, the critical value based on F-distribution, since $\hat{\Phi}(B)\mathbf{X}_t$ is of length $n-p$.

As in the univariate case, the distribution of the test statistic, $\hat{F}$, defined in step 5, is not close to an F-distribution, yielding excessive significance levels when $\mathbf{b} = \mathbf{0}$, when the residuals are highly correlated and the series length is small to moderate. The bootstrap approach used in the univariate case may be used in the same way to lower the significance levels.

3.2.1. A bootstrap approach

To estimate the actual distribution and critical regions of the test statistic $\hat{F}$ when $\mathbf{b} = \mathbf{0}$ in model (14), we employ a similar bootstrap procedure as in the univariate case. In the bootstrap procedure we obtain $\hat{F}$ for a sample data set using steps 1–5 above. To simulate realizations under H_0 we estimate $\Phi(B)$ in (14) assuming $\mathbf{b} = \mathbf{0}$ and denote this estimate by $\hat{\Phi}^{(0)}(B)$. That is, under H_0 we assume that any trending behavior in the series is due to correlation structure alone. We then obtain by simulation B realizations from the autoregressive model with AR operator given by $\hat{\Phi}^{(0)}(B)$. For the bth realization, $b = 1,\ldots,B$, we calculate $\hat{F}_b^*$ as in step 5. The null hypothesis is rejected at the α level of significance if $\hat{F} > F_{1-\alpha}^*$ where F_β^* is the βth empirical quantile of $\left\{\hat{F}_b^*\right\}_{b=1}^{B}$. Since the probability that a randomly selected member from the population is greater than or equal to the jth largest value is $j/(B+1)$, then by setting $\alpha = j/(B+1)$ it follows that $F_{1-\alpha}^*$ is the jth largest value of $\left\{\hat{F}_b^*\right\}_{b=1}^{B}$, e.g., if $\alpha = 0.05$ and $B = 399$ then $F_{1-\alpha}^*$ is the 20th largest value of $\left\{\hat{F}_b^*\right\}_{b=1}^{399}$.

This multivariate testing procedure is basically equivalent to the univariate procedure when $m = 1$, because the distribution of $\hat{t}^2$ with $n-2$ degrees of freedom in the univariate case with white residuals is distributed as $F(1,n-2)$, which is the same as equation (28) with $m = 1$. As in the univariate case, when the correlation (in time) in the noise is high, i.e., $\det(\Phi(1))$ is near zero, the significance levels are still high after the bootstrap has been used to approximate the critical region. The actual significance levels determined by the test increase as m increases for cases where there is no correlation between paths (components of $\mathbf{X}_t$). It is reasonable to believe that an adjustment procedure similar to that described in section 2.2.2 could be used to help correct the inflated significance levels. A difficulty arises in choosing the "median" model to be employed in the second application of the bootstrap, as described in appendix A. This difficult issue must be left for the future.

4. REFERENCES

Anderson, T. W., 1984: *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, New York, 675 pages, second edition.

Bottone, S., H. L. Gray, and W. A. Woodward, January 1995: *Stochastic Modeling and Global Warming Trend Extraction for Ocean Acoustic Travel Times.* MRC-R-1488, Mission Research Corporation, 39 pages.

Box, G. E. P. and G. M. Jenkins, 1976: *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco, 575 pages, revised edition.

Gray, H. L., W. A. Woodward, and N. F. Zhang, 1996: *Time Series Analysis.* In preparation.

Johnson, R. J. and D. E. Wichern, 1988: *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, New Jersey, second edition.

Manabe, S., R. Stouffer, M. Spellman, and K. Bryan, 1991: Transient responses of a coupled ocean-atmosphere model to gradual changes of atmospheric CO_2. Part I: Annual mean response. *Journal of Climate*, **4,** 785–818.

Pares-Sierra, A. and J. O'Brien, 1989: The seasonal and internal variability of the California current system: A numerical model. *Journal of Geophysical Research*, **94,** 3159–3180.

Robinson, E. A. and M. T. Silva, 1979: *Digital Foundations of Time Series Analysis: Volume 1 — The Box-Jenkins Approach.* Holden-Day, San Francisco, 451 pages.

Seber, G. A. F., 1984: *Multivariate Observations.* John Wiley and Sons, New York.

Woodward, W. A. and H. L. Gray, 1993: Global warming and the problem of testing for trend in time series data. *Journal of Climate*, **6**, 953–962.

Woodward, W. A. and H. L. Gray, 1995: Selecting a model for detecting the presence of a trend. *Journal of Climate*, **8**, 1929–1937.

APPENDIX A

IMPROVED TESTS FOR TREND IN TIME SERIES DATA

This appendix contains a paper by Wayne A. Woodward, Steven Bottone, and H. L. Gray entitled "Improved Tests for Trend in Time Series Data", which has been submitted for publication to the *Journal of Time Series Analysis*.

IMPROVED TESTS FOR TREND IN TIME SERIES DATA

Woodward, Wayne A.*, Bottone, Steven**, and Gray, H.L.*

* Southern Methodist University

** Mission Research Corporation

ABSTRACT

The difficult problem of testing for linear trend in the presence of correlated residuals is addressed. Because of the correlated residuals, tests for trend based on the classical least-squares regression techniques are inappropriate. Even procedures in the literature that adjust for the correlation in the residuals tend to have the problem that the observed significance levels are higher than nominal levels for small to moderate realization lengths whenever the residuals are highly correlated. We introduce a bootstrap-based procedure to test for trend in this setting which is better adapted to controlling the significance levels, and this testing procedure is studied via simulation results. The procedure is then applied to the problem of testing for trend in global atmospheric temperature data and in data from models for ocean acoustic travel time along a path.

Keywords: regression with correlated residuals, bootstrap, global warming, hypothesis testing

The research was partially supported by DOE Environmental Sciences Division Grant DE-FG03-93ER61645 and ARPA contract MDA972-93-C-0021.

1. INTRODUCTION

The challenging problem of testing for trend in time series data is one which arises in many areas of application. This problem has been addressed by many authors including recent discussions by Brillinger (1989, 1994), Woodward and Gray (1993, 1995), Bloomfield and Nychka (1992), Bloomfield (1992), and Harvey (1989).

The specific problem we address in this paper is that of testing for trend using the model

$$Y_t = a + bt + Z_t \tag{1}$$

where Z_t is a stationary autoregressive process of order p satisfying $\phi(B)Z_t = a_t$, where a_t is white noise and $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, where B is the backshift operator defined by $B^k Y_t = Y_{t-k}$. In Figure A-1 we show the global temperature series for the years 1880-1987 as obtained by Hansen and Lebedeff (1987, 1988). It is clear that temperatures have had a tendency to rise over this time span. When viewing this series, it is difficult to ascertain whether the trend in the series is due to some deterministic component such as the bt term in (1) or is simply due to wandering or random trending behavior caused by roots of $\phi(r) = 0$ near (or equal to) one. Given a model such as (1) and a series such as that given in Figure A-1, it is clear that estimation procedures would be expected to have difficulty "knowing" whether to attribute apparent trends to deterministic components (i.e. nonzero b) or to high autocorrelation in Z_t (i.e., roots of $\phi(B) = 0$ near one). When roots of $\phi(r) = 0$ are near unity the estimation procedures currently in use tend to be biased in favor of attributing random trending behavior to the existence of a nonzero slope. Specifically, when $\phi(r) = 0$ has a root or roots near unity, tests for trend assuming model (1) have a tendency to inflate the significance level over

nominal amounts, i.e., to attribute the trend behavior to the term bt. Park and Mitchell (1980) studied the case in which the autoregressive order $p = 1$ and used simulations to examine a number of estimation procedures including a maximum likelihood method due to Beach and MacKinnon (1978). They concluded that all the test procedures studied had a tendency for the slope to appear more significant than it really is when $\phi(r) = 0$ has a root near one, i.e., when ϕ_1 is near one. The SAS AUTOREG procedure uses the Beach and MacKinnon ML estimates and quotes p-values associated with the test $H_o : b = 0$. These p-values are based on the assumption that the test statistic calculated is distributed as t with $n - 2$ degrees of freedom. Park and Mitchell (1980) indicate that these types of results can be very misleading and that when $\phi(r) = 0$ has roots near one, the user should consider using lower nominal significance levels to adjust for the fact that actual significance levels are higher than the nominal levels. A discussion of these problems is also given in the SAS/ETS User's Manual (1993). Woodward and Gray (1993) found that in simulated realizations of length $n = 100$ from (1) where $\phi(B) = 1 - .95B$ and $b = 0$, a significant trend is found about 35% of the time using the Bloomfield and Nychka (1993) test (using a 2-sided test at the nominal 5% level) which adjusts the standard error of the least squares estimator of b to account for the correlation structure. Additionally, Brillinger's (1989) test for a monotonic trend in a time series found a significant trend in about 50% of the realizations in this case. Use of SAS to obtain ML estimates in this same setting resulted in significant results about 25% of the time.

Based on the results discussed above, we see that the existing procedures can have such high true significance levels that the finding of a significant trend using these results must be viewed with extreme caution. One might even conclude that it is often simply not possible to examine a series of short to moderate length and make an intelligent decision concerning whether an apparent trend is deterministic or random. In this paper we help make such a decision more plausible by introducing a new procedure for testing the hypothesis $H_o : b = 0$ in (1) which is more effective in controlling the actual significance

level even in the presence of highly autocorrelated errors. In Section 2 we briefly discuss tests for trend and introduce our new procedure. In Section 3 we discuss simulation studies that demonstrate the fact that the new procedure does a much better job of maintaining actual significance levels near nominal levels. Finally, in Section 4 we apply the test proposed here to some actual data.

2. A NEW TESTING PROCEDURE

In this section we specifically address the problem of testing $H_o : b = 0$ in (1) against one and two-sided alternatives. Some techniques proposed in the literature (e.g. Bloomfield and Nychka, 1992) estimate b using usual least squares estimators and adjust the standard error of the least squares estimator to account for the correlation structure. Others such as the Beach and MacKinnon (1978) ML technique involve iterative procedures for simultaneously estimating a, b, and the coefficients in $\phi(B)$. The technique we propose uses the usual least squares estimates of a and b. Notice that if $\phi(B)$ in (1) were known, then we could rewrite (1) as follows:

$$\begin{aligned} \phi(B)Y_t &= \phi(1)a + \left(\sum_{i=1}^{p} i\phi_i\right)b + \phi(1)bt + \phi(B)Z_t \\ &= c + dt + a_t \end{aligned} \tag{2}$$

where $c = \phi(1)\,a + \left(\sum_{i=1}^{p} i\phi_i\right)b$ and $d = \phi(1)\,b$ and where a_t is white noise.

Note that if Z_t is stationary, which implies $\phi(1) > 0$, then $d = 0$ if and only if $b = 0$, and d and b have the same sign when $b \neq 0$. To test $b = 0$ in (1) we simply test $d = 0$ in (2) in which case we are able to use the usual regression-based standard errors since the residuals are white.

In practice $\phi(B)$ is not known so, for now, we consider the following approach. The least squares estimates of a and b (denoted $\hat{a}$ and $\hat{b}$) are obtained and the residuals from the regression line are calculated using

$$\hat{Z}_t = Y_t - \hat{a} - \hat{b}t. \tag{3}$$

These residuals do not follow the same model as Z_t and are in general not stationary since

$$\begin{aligned}\hat{Z}_t &= a + bt + Z_t - \hat{a} - \hat{b}t \\ &= (a - \hat{a}) + (b - \hat{b})t + Z_t\end{aligned}$$

which does not have constant mean unless $\hat{b} = b$. However, in most cases we find it reasonable to assume that $\hat{Z}_t$ is approximately AR(p), and we let $\hat{\phi}(B)$ denote the estimated autoregressive operator. We transform the data using $\hat{\phi}(B)$ to obtain

$$\begin{aligned}W_t &= \hat{\phi}(B)Y_t \\ &= \hat{\phi}(1)a + \left(\sum_{i=1}^{p} i\hat{\phi}_i\right)b + \hat{\phi}(1)bt + g_t \\ &= c' + d't + g_t\end{aligned} \tag{4}$$

where $\hat{c} = \hat{\phi}(1)\,a + \left(\sum_{i=1}^{p} i\hat{\phi}_i\right)b$, $\hat{d} = \hat{\phi}(1)b$, and where $g_t = \hat{\phi}(B)\phi^{-1}(B)a_t$ which will not be white noise but should be a reasonably close approximation to it.

A straightforward application of the procedure (assuming g_t is white) is to use standard regression procedures to test for the significance of $\hat{d}$. This estimation procedure is summarized in the following:

(i) Estimate a and b using least squares.

(ii) Calculate $\hat{Z}_t$ as in (3).

(iii) Find Burg estimates of $\phi(B)$ where $\phi(B)\widehat{Z}_t = a_t$ (see Burg, 1967 and Marple, 1987). Call this estimate $\widehat{\phi}(B)$.

(iv) Transform the data to obtain $\widehat{\phi}(B)Y_t = c' + d't + g_t$ (5)

where g_t is nearly white noise.

(v) Calculate $\widehat{t} = \widehat{d}/\widehat{\mathrm{SE}}(\widehat{d})$ where $\widehat{d}$ and its standard error are the usual least squares-based quantities assuming uncorrelated residuals. Compare $\widehat{t}$ with $t(n-p-2)$ critical values based on Student's t since $\widehat{\phi}(B)Y_t$ is of length $n-p$.

We considered a 2-sided version of this procedure on simulated realizations of length 100 from the model in (1) with $\phi(B) = 1 - .95B$ and $b = 0$. We found that a significant trend is typically found over 25% of the time using usual critical regions based on Student's t, and so we see that this procedure suffers from the same problem of excessive actual significance levels which occurred with the previously mentioned tests.

It seems that the primary reason for the excessive significance levels in this case is that when ϕ_1 in $1 - \phi_1 B$ is close to 1, $\widehat{\phi}_1$ tends to be less than ϕ_1 so that $1 - \widehat{\phi}_1$ tends to be larger than $1 - \phi_1$. This happens in general, i.e., when $\phi(B)$ has a factor close to $1 - B$ then $\widehat{\phi}(1)$ tends to be larger than $\phi(1)$. In the AR(1) case, for example, it is well known that usual estimators (i.e.,OLS, ML and Burg) for ϕ_1 exhibit a bias toward zero, i.e., away from the nonstationary region (see Kang, 1992). This bias is demonstrated in Table A-1 where we show the average of the Burg estimates of ϕ_1 over 250 replications from a variety of realization lengths and values of ϕ_1. For each configuration we show the average of ϕ_1 estimates before and after removing the least squares line. The bias that has been addressed by Kang is that found before the line is removed. There it can be seen that for $n = 50$ and $n = 100$ this bias can be substantial. However, it should also be noticed that after removal of the line, the estimates are even more biased away from the nonstationary region. This is not surprising since the effect of removing the line will be to

use the line to account for some of the behavior associated with the correlation structure in the original series. Thus, in the $n = 100$ case with $\phi = 0.95$, we see that $\widehat{\phi}(1) = 1 - \widehat{\phi}_1 \approx 0.124$ which is substantially larger than $\phi(1) = 0.05$. Our simulations show that transforming the series as in (2) with the true value of $\phi(B)$ and testing $H_0 : d = 0$ using standard regression-based test statistics, leads to a test that appears to have the appropriate significance level, while as we have seen, transforming by $\widehat{\phi}(B)$ instead of $\phi(B)$, which we of course must do in practice, leads to a test with inflated significance levels.

Examination of the calculated $\widehat{t}$ values using this procedure indicates that the variability of $\widehat{t}$ is substantially larger than that which would be expected based on the Student's t with $n - p - 2$ degrees of freedom. In Table A-2 we show empirical 97.5th percentiles obtained from 100 simulated realizations from the model in (1) with $b = 0$ and $p = 1$ for several values of n and ϕ_1. In the table we see that for ϕ_1 near unity, these percentiles are substantially larger than the percentiles of a Student's t with $n - p - 2$ degrees of freedom which in the cases considered here are slightly larger than 1.96. Thus, in these cases, the distribution of $\widehat{t}$ is symmetric about zero but does not follow a Student's t with $n - p - 2$ degrees of freedom. It is clear that for a given n, the variability of $\widehat{t}$ increases as the true value of ϕ_1 approaches +1.

(a) A bootstrap approach

In order to estimate the actual distribution of $\widehat{t}$ when Y_t follows the model in (1) with $b = 0$, we propose a bootstrap procedure. In this procedure we obtain $\widehat{t}$ as described previously in (5). To simulate realizations under H_0 we estimate $\phi(B)$ in (1) assuming $b = 0$. We denote this estimate by $\widehat{\phi}^{(0)}(B)$. That is, under H_0 we assume that any trending behavior in the series is due to the correlation structure alone. We then obtain B realizations from the autoregressive model with AR operator given by $\widehat{\phi}^{(0)}(B)$. For the bth realization, $b = 1, \ldots, B$ we calculate $\widehat{t}^{\,*}_b$ as in (v). It can be shown that $\widehat{t}^{\,*}_b$

does not depend on the white noise variance, which thus can be chosen arbitrarily for the bootstrap replications. For the 2-sided test, the null hypothesis is rejected at the α level of significance if $\hat{t} > t^*_{1-\alpha/2}$ or $\hat{t} < t^*_{\alpha/2}$ where t^*_β is the βth empirical quantile of $\{\hat{t}^*_b\}_{b=1}^B$. Because of the symmetric nature of $\hat{t}$, in practice we accomplish this test by rejecting H_0 if $|\hat{t}| > |t|^*_{1-\alpha}$ where $|t|^*_{1-\alpha}$ is the $(1-\alpha)$th empirical quantile of $\{|\hat{t}^*_b|\}_{b=1}^B$. Since the probability a randomly selected member from the population is greater than or equal to the jth largest value is $j/(B+1)$, then by setting $\alpha = j/(B+1)$ it follows that $|t|^*_{1-\alpha}$ is the jth largest value of $\{|\hat{t}^*_b|\}_{b=1}^B$, i.e., if $\alpha = 0.05$ and $B = 399$ then $|t|^*_{1-\alpha}$ is the 20th largest values of $\{|\hat{t}^*_b|\}_{b=1}^{399}$. For a 1-sided test, the α-level critical value is the $(1-\alpha)$th or αth empirical quantile of $\{\hat{t}^*_b\}_{b=1}^B$ depending on whether the alternative is $H_1: b > 0$ or $H_1: b < 0$ respectively.

(b) A second application of the bootstrap

It should be noted that for ϕ_1 near one, the observed significance levels are high. This phenomenon is caused by the bias shown in Table A-1. That is, if $\hat{t}$ is calculated from data associated with $\phi_1 = 0.99$, $b = 0$, and $n = 100$, the estimate $\hat{\phi}^{(0)}_1$ is likely to be about 0.95. Thus, the bootstrap realizations are generated from the AR(1) model $(1 - \hat{\phi}^{(0)}_1)X_t = a_t$ with $\hat{\phi}^{(0)}_1 \approx 0.95$, and this will result in $\hat{t}^*_b$ values that are not as variable as those for the original model with $\phi_1 = 0.99$ as can be seen in Table A-2. An intuitively appealing procedure would be to scale the original $\hat{t}$ so that it has variance comparable to that of the bootstrap distribution for $\hat{t}^*_b$, i.e., we obtain $\hat{t}_{adj} = C\,\hat{t}$ where $C = \sigma_{\hat{t}^*_b}/\sigma_{\hat{t}}$ and where C would be less than one when ϕ_1 is near one. Clearly, $\sigma_{\hat{t}^*_b}$ can be estimated from $\hat{t}^*_b$, $b = 1, \ldots, B$. However, no comparable estimate of $\sigma_{\hat{t}}$ is available. It is clear from Table A-1 that the estimates of the autoregressive coefficient of the bootstrap realizations from $(1 - \hat{\phi}^{(0)}_1)X_t = a_t$ will in general underestimate $\hat{\phi}^{(0)}_1$ in much the same way that $\hat{\phi}^{(0)}_1$ tends to underestimate ϕ_1. Let $\hat{\phi}^*_{1(b)}$, $b = 1, \ldots, B$ denote the coefficient estimates of $\hat{\phi}^{(0)}_1$ from the B bootstrap realizations, and let $\hat{\phi}^*_1(m)$ denote the

median of the coefficient estimates obtained from these realizations. We generate a second set of bootstrap realizations from the AR(1) model with coefficient $\hat{\phi}_1^*(m)$, and we denote the $\hat{t}$ values calculated from this second set of bootstrap realizations as $\hat{\hat{t}}_b^*$, $b = 1, \ldots, B$. Since $\sigma_{\hat{t}_b^*}/\sigma_{\hat{t}} \approx \sigma_{\hat{\hat{t}}_b^*}/\sigma_{\hat{t}_b^*}$ we calculate $\hat{t}_{adj} = \hat{C}\,\hat{t}$ where $\hat{C} = \hat{\sigma}_{\hat{\hat{t}}_b^*}/\hat{\sigma}_{\hat{t}_b^*}$ and the $\hat{\sigma}$'s are the sample standard deviations from the B values of $\hat{\hat{t}}_b^*$ and $\hat{t}_b^*$.

The concept of the "median" model is not quite so clear-cut when an AR(p) model with $p > 1$ is used. Simply selecting a model whose coefficients are the corresponding median values $\hat{\phi}_j^*(m)$, $j = 1, \ldots, p$ may result in an AR(p) model that does not have the desired properties and may even have roots inside the unit circle. Since the phenomenon of excessive significance levels is caused by roots of $\phi(r) = 0$ close to one, we take $\hat{\phi}(1)$ to be our measure of the extent to which a fitted model has a root near +1. In the general AR(p) case, the "median" model is then selected as the model from the first set of bootstrap realizations associated with the median value of $\{\hat{\phi}_{(b)}^*(1)\}_{b=1}^{B}$ where $\hat{\phi}_{(b)}^*(1) = 1 - \hat{\phi}_{1(b)}^* - \cdots - \hat{\phi}_{p(b)}^*$. This procedure is equivalent to the approach described in the preceding paragraph when $p = 1$.

3. SIMULATION STUDIES

In this section we discuss the results of simulation studies designed to examine the performance of the testing procedures discussed in the previous sections. In Table A-3 we show the observed significance levels from testing $H_0 : b = 0$ vs. $H_1 : b \neq 0$ in (1) with $p = 1$ based on simulated realizations for a variety of values of n and ϕ_1. The testing procedures used were the Beach and MacKinnon (MLE) (1978) procedure using SAS, the Bloomfield and Nychka (BN) (1992) procedure, the transformation (T) procedure discussed in (5) using Student's t-based critical values, the bootstrapped version (TB) of the transformation procedure, and the "adjusted" bootstrap approach (TBA) using the second bootstrap application. The tabled values are the percentage of simulated realizations for which a trend was detected. For MLE, BN and T the results shown are

based on 250 replications while those for TB and TBA are based on 1000 replications. In the table it can be seen that for small to moderate realization lengths, the observed significance levels for MLE, BN, and T is substantially higher than the nominal 5% level especially when ϕ_1 is near +1. Only for $n \geq 500$ with $\phi_1 = 0.8$ did the actual observed significance levels attain the nominal levels. For $\phi_1 = 0.95$ it seems that these levels are approaching 5% as n increases, but the 5% level was not attained by $n = 1000$. Thus, based on the evidence presented here it is seen that the MLE, BN, and T techniques do not behave well for highly correlated residuals and small to moderate realization lengths. However, even in the case of highly correlated residuals these tests appear to behave properly asymptotically.

Results for TB and TBA are given for $\phi_1 = 0.8$, 0.95, 0.99 and -0.95. The significance levels for TB are much closer to the nominal levels, being somewhat too large for $n \leq 100$ when $\phi_1 = 0.95$ and for $n \leq 500$ when $\phi_1 = 0.99$. The significance levels of 6.4% in the table are borderline significantly too high (about 2 SE's above 5%). As expected, the use of TBA improved the significance levels and only in the cases of $n = 50$ and 500 with $\phi_1 = 0.99$ were the observed significance levels significantly larger than 5%. It seems that as ϕ_1 approaches -1, no corresponding significance level problems arise. This is reasonable since the inflated significance levels are attributable to apparent trends in Z_t due to roots near +1. It is clear from the table that for values of ϕ_1 well removed from +1 and for larger realization lengths, significance levels for TB are acceptable and the adjustment does not have a substantial effect.

In Table A-4 we show power results for a variety of values of ϕ_1 and slopes when $n = 100$ and $p = 1$. In order to standardize the variance of the residual series, Z_t, in all cases the white noise variance of the Z_t series is selected so that $\sigma_Z^2 = \sigma_a^2/(1 - \phi_1^2) = 1$. It can be seen that in the cases considered these tests do have substantial power, especially for lower values of ϕ_1. In fact the trend is detected over 90% of the time with TB and at least 82% of the time with TBA for $\phi_1 \leq 0.8$ for the slopes considered. The use of TBA,

which does a better job of controlling the significance level, causes a dramatic reduction in power when ϕ_1 is near one and the realization lengths are not large. It is not surprising that smaller power is encountered for small to moderate slopes in the case in which ϕ_1 is near +1 due to the confusion between random and deterministic trends. However, if slight inflation of significance level (say to about 10%) can be tolerated, then one might consider the use of TB because of its substantially higher power than TBA in these cases. It should be noted that when nominal significance levels for TB were dropped below 5% in order to obtain actual significance levels of about 5% in these cases, the power for TB was comparable to that of TBA shown in the table.

It should be noted that all simulations discussed in this section are concerned with the case in which Z_t is AR(1). We have found these results to be representative of the case $p > 1$ if $\phi(B)$ has at most a single factor near $1 - B$. In the next section, in conjunction with the analysis of actual data series, we will examine simulation results for $p > 1$.

4. APPLICATION TO TREND TESTING IN ACTUAL DATA

In this section we consider the application of the tests proposed here to actual data sets of interest. We consider the Hansen and Lebedeff (1987, 1988) annual atmospheric temperature series for 1880-1987 shown in Figure A-1 as well as acoustic travel time signals. Because of the small to moderate lengths of these series the bootstrap tests recommended here will be applied.

(a) Atmospheric Temperature Series

Woodward and Gray (1993, 1995) considered model (1) where Z_t is modeled as an AR(8) as a possible model for the Hansen and Lebedeff data. The model they obtain, using least squares estimates of a and b and Burg estimates of the autoregressive parameters has $\hat{a} = -0.410$, $\hat{b} = 0.0055$, with Z_t modeled as

$$Z_t = .4943Z_{t-1} - .0688Z_{t-2} + .0333Z_{t-3} + .1253Z_{t-4} - .1035Z_{t-5} + .2973Z_{t-6} - .2305Z_{t-7} + .1970Z_{t-8} + a_t \quad (6)$$

where $\widehat{\sigma}_a^2 = .013064$. Woodward and Gray (1995) display the factors of the eighth order characteristic polynomial associated with the model in (6) and show that it has a factor of $(1 - .916B)$ indicating a root near but not extremely close to +1. Bloomfield and Nychka (1992) considered the test statistic given by $\widehat{b}/\widehat{\mathrm{SE}}(\widehat{b})$ where $\widehat{\mathrm{SE}}(\widehat{b})$ takes the correlation in the residuals into account and is given by

$$\widehat{\mathrm{SE}}(\widehat{b}) = \left[2\int_0^{.5} W(f)\widehat{S}(f)df\right]^{1/2}, \quad (7)$$

where

$$W(f) = \left|\sum_{t=1}^{n} b_t e^{-2\pi i f t}\right|^2,$$

with

$$b_t = \frac{t - \overline{t}}{\sum_{t=1}^{n}(t - \overline{t})^2},$$

i.e., $\widehat{b} = \sum_{t=1}^{n} b_t Y_t$. For a given realization, Z_t is estimated by $\widehat{Z}_t = Y_t - \widehat{a} - \widehat{b}t$ and $\widehat{S}(f)$ is an appropriate estimate of the spectrum of $\widehat{Z}_t$.

Using this procedure a test statistic of 4.70 was obtained which when compared to usual normal or Student's t critical values is highly significant indicating the presence of a trend. These results were consistent with Bloomfield and Nychka's (1992) analysis of the temperature series. However, as pointed out here and by Woodward and Gray (1993), the true significance levels of these tests can be sufficiently large to make the finding of significance meaningless. We used TB and TBA to perform a 1-sided test against the

alternative that $b > 0$. Use of TB resulted in a test statistic of 2.73. The 1-sided critical value $t^*_{.95}$ was obtained from $B = 399$ bootstrap samples to be 4.50, i.e., the test does not indicate a significant slope. The bootstrap-based p-value is 14.3%. For TBA we obtained $t_{adj} = 1.95$ which should be compared with the same 95% critical value of 4.50 and this again is not significant at the 5% level. The bootstrap-based p-value in this case is 22.3%.

In order to examine the observed significance level in this case we fit a stationary AR(8) model to the temperature data without first removing the line. The model obtained is the AR(8) model

$$Y_t = .5547Y_{t-1} - .0421Y_{t-2} + .0669Y_{t-3} + .1569Y_{t-4} - .0919Y_{t-5} + .3221Y_{t-6} - .2321Y_{t-7} + .2017Y_{t-8} + a_t \quad (8)$$

with $\sigma_a^2 = .0138$. In this case the characteristic polynomial has a factor of $1 - .98B$ indicating that the model is very nearly nonstationary with a root near +1. Thus, the trending behavior is accounted for in this model via this near nonstationarity. Realizations of length $n = 100$, 150 and 200 were generated from the model in (8) and the percentage of realizations for which the tests found significant trends are shown in Table A-5(a) where it can be seen that the significance levels for TB are around 8% while those for TBA are close to the nominal level of 5%. Consistent with the simulation results of the previous section, the significance levels using BN were excessively high, with the significance levels in this case being about 25%. The power figures for TB and TBA are shown in Table A-5(b) based on realizations generated from (6) which actually contains a line. If this were the true model and 100 years of data were available, then it can be seen that the trend will be detected about 76% of the time with TB and about 50% of the time with TBA. If 150 or 200 years of observations were available, then it can be seen that both tests would have a strong chance of detecting the trend. These results certainly do not indicate a trend in the Hansen and Lebedeff data, and even TB (which has an inflated

significance level) does not yield significance while it has 76% power against alternatives such as (6) for $n = 100$.

A final comment in this section concerns the ARIMA(9,1,0) model fit by Woodward and Gray (1993) to the temperature series. Woodward and Gray showed that about 67% of realizations of length $n = 108$ from this model were seen to have a significant trend using a 2-sided version of Bloomfield and Nychka's (1992) test at the 5% nominal significance level. For comparison with the results in Table A-5, it should be noted that this figure reduces to approximately 35% when a 1-sided test is used. In Table A-5(c) we show the percentage of realizations from this ARIMA(9,1,0) model for which TB and TBA found significant slopes. There it can be seen that TB and TBA found a significant slope at the nominal 5% level about 13% and 7% of the time respectively for the realization lengths considered. It should be realized that the ARIMA(9,1,0) is not a special case of (1) with $b = 0$ since in (1) Z_t is assumed to be stationary. Thus, the ARIMA(9,1,0) model is simply an alternative to (1), and we would not expect the percentage of realizations for which a significant slope is (incorrectly) found to be at the nominal 5% level. This is the case because of our procedure of generating the bootstrap realizations from a stationary AR(p) model fit to the data. However, we see that these percentages are much closer to 5% than with the Bloomfield and Nychka test, and thus there is less likelihood of confusion between a series with real trend and one with ARIMA-type random trends using the new tests. If we included a check for unit root and generated the bootstrap realizations from a model including a unit root model if this was indicated, then the percentage of realizations from the ARIMA(9,1,0) model for which a significant trend is detected should be approximately the nominal level.

(b) Ocean Acoustic Travel Times

Since the speed of sound in the ocean increases as ocean temperature increases, an indication of the presence of global warming is a negative trend for the travel time of an

acoustic pulse along a long fixed path. The MASIG (Mesoscale Air-Sea Interaction Group) model is a reduced gravity ocean model driven by COADS (Comprehensive Ocean-Atmosphere Data Set) winds, coupled to an equatorial model at its southern boundary (Pares-Sierra and O'Brien, 1989). A 20 year simulation from this model, assuming no warming, for acoustic travel-time anomaly along a path from Hawaii to San Diego is plotted in Figure A-2. A question of interest concerns the length of time before a warming (if it existed) would be detected based on the test for trend described in Section 2. The data plotted in Figure A-2 consist of monthly values for the 20 years simulated. We modeled the last 160 points of the MASIG data as an AR(10). This model is given by

$$Z_t = 2.1765Z_{t-1} - 1.6926Z_{t-2} + .9443Z_{t-3} - .7730Z_{t-4} + .6793Z_{t-5} - .4383Z_{t-6} + .0608Z_{t-7} - .2369Z_{t-8} + .4745Z_{t-9} - .2024Z_{t-10} + a_t \quad (9)$$

with $\sigma_a^2 = .001196$. There are no positive real roots in the characteristic equation but there is a pair of complex roots with small complex component (i.e., associated with a frequency near zero). Our simulation consisted of generating realizations from (1) with Z_t given in (9) based on two warming scenarios: lines with slopes of -0.05 and -0.1 seconds/year. These slopes correspond to increases in ocean temperature along the path of 0.005 and 0.01 Celsius per year. In Table A-6 we give the results using TB and TBA assuming 5 to 30 years of monthly data. Again, the other tests would not be appropriate because of the small realization lengths required in this application. In the table we also consider the case in which there is no warming, i.e., $b = 0$. There it can be seen that when $b = 0$ both tests produce actual significance levels which did not differ significantly from the nominal 5% level, with the only exception being TB at 5 years.

For the two warming scenarios considered, it is clear that warming will not be able to be detected with high probability unless data are collected for a sufficiently long period. Based on the significance level results, one would probably be willing to use TB because

of its higher power. Even in this case, for warming to be detected at least 50% of the time in the simulations, at least 25 years of data are needed if the slope is -0.1/year, and at least 30 years of data are needed if the slope is -0.05/year.

6. CONCLUDING REMARKS

The results presented here are encouraging and indicate that it is possible to construct a test for trend with appropriate significance levels in the presence of highly autocorrelated noise when the realization length is not large. The results presented here also indicate that the test has reasonable power.

The bootstrap procedure described here could be used analogously to obtain tests based on the test statistic $\widehat{b}/\widehat{\mathrm{SE}}(\widehat{b})$ where $\widehat{b}$ is the least-squares estimate of b and where $\widehat{\mathrm{SE}}(\widehat{b})$ is given in (7). Alternatively, the procedure could be based on a test statistic $\widehat{t}_{\mathrm{ML}}$ obtained using the Beach and MacKinnon (1978) ML procedure. In both cases, however, the procedures are relatively computationally intensive, making bootstrapping less practical than in the current implementation.

It should be noted that in this paper we have assumed that the order p of $\widehat{\phi}(B)$, the autoregressive model fit to the estimated residuals $\widehat{Z}_t$, is the same as the order of $\widehat{\phi}^{(0)}(B)$ obtained by assuming $b = 0$, i.e., the autoregressive order of the model fit to the Y_t series itself. In fact, this need not be the case and the procedure described here could be modified to allow for different orders. In fact, as indicated in Section 5, the model may be allowed to have one or more unit roots if these are indicated as appropriate in the modeling procedure. In this case the problem with high significance levels when roots are on or very near the unit circle may be alleviated without the adjustment technique described here.

The procedure described here is similar to that given by Woodward and Gray(1995). In that paper, a bootstrap-based classification analysis technique was presented for ascertaining which of two competing models produced realizations most like

the observed data. The results obtained there do not deal with the issue of testing the hypothesis $H_0 : b = 0$ and controlling the probability of a type one error. Their findings related to the temperature data were consistent with those presented here in that a trend component was not indicated for the temperature series.

REFERENCES

Beach, C.M. and MacKinnon, J.G. (1978). "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors," *Econometrica* 46, 51-58.

Bloomfield, P. (1992). "Trends in Global Temperature," *Climatic Change* 21, 1-16.

Bloomfield, P. and Nychka, D. W. (1992). "Climate Spectra and Detecting Climate Change," *Climatic Change* 21, 275-287.

Brillinger, D.R. (1989). "Consistent Detection of a Monotonic Trend Superimposed on a Stationary Time Series," *Biometrika* 76, 23-30.

Brillinger, D.R. (1994). "Trend Analysis: Time Series and Point Process Problems," *Environmetrics* 5, 1-19.

Burg, J.P. (1967). "Maximum Entropy Spectral Analysis," *Proceedings of the 37th Meeting of the Society of Exploration Geophysicists.*

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Hansen, J. and Lebedeff, S. (1987). "Global Trends of Measured Surface Air Temperature," *Journal of Geophysical Research* 92, 13345-13372.

Hansen, J. and Lebedeff, S. (1988). "Global Surface Air Temperatures: Update through 1987," *Geophysical Research Letters* 15, 323-326.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Kang, H. (1992). "Bias Correction of Autoregressive Parameters and Unit Roots Tests," American Statistical Association 1992 Proceedings of the Business and Economic Statistics Section, 45-50

Marple, S.L. (1987). *Digital Spectral Analysis with Applications*, Prentice Hall, Inc., Englewood Cliffs, N.J.

Pares-Sierra, A. and O'Brien, J. (1989). "The Seasonal and Internal Variability of the California Current System: A Numerical Model," *Journal of Geophysical Research* 94, 3159-3180.

Park, R.E. and Mitchell, B.M. (1980). "Estimating the Autocorrelated Error Model with Trended Data," *Journal of Econometrics* 13, 185-201.

SAS/ETS User's Guide, Version 6, Second Edition (1993), SAS Institute, Cary, NC.

Woodward, W.A. and Gray, H.L. (1993). "Global Warming and the Problem of Testing for Trend in Time Series Data," *Journal of Climate* 6, 953-962.

Woodward, W.A. and Gray, H.L. (1995). "Selecting a Model for Detecting the Presence of a Trend," *Journal of Climate* 8, 1929-1937.

Table A-1. Burg Estimates of ϕ_1 Before and After Removing the Least Squares Line

(250 replicates)

			True ϕ_1				
			.5	.8	.9	.95	.99
n	50	Before	.455	.718	.818	.857	.890
		After	.421	.666	.757	.784	.818
	100	Before	.473	.765	.860	.908	.943
		After	.458	.745	.833	.876	.902
	500	Before	.499	.792	.893	.942	.981
		After	.495	.788	.889	.938	.974

Table A-2. Observed 97.5th percentiles of $\hat{t}$ values calculated as in 5(d) based on 1000 realizations from the model

$$Y_t = a + bt + Z_t$$

where $(1 - \phi_1 B)Z_t = a_t$ **and** $b = 0$

		ϕ_1				
		0.50	0.80	0.95	0.99	$t_{.975}(n-p-2)$
n	50	2.40	3.01	5.12	7.21	2.01
	100	2.15	2.65	4.18	6.98	1.98
	150	2.00	2.26	2.93	5.03	1.98

Table A-3. Observed significance levels associated with tests for $b = 0$ based on the model

$$Y_t = a + bt + Z_t$$

where $(1 - \phi_1 B)Z_t = a_t$

Nominal level = 5%

2-sided tests

1000 replications with $B = 399$ for TB and TBA

(250 replications for MLE, BN and T)

		$\phi_1 = .8$					$\phi_1 = .95$				
		MLE	BN	T	TB	TBA	MLE	BN	T	TB	TBA
	50	16.2	22.0	18.4	5.9	5.3	36.2	36.0	37.2	10.0	6.2
	100	12.0	16.8	16.0	6.4	5.9	25.2	40.0	28.4	7.6	4.3
n	250	7.6	12.4	8.0	3.8	4.1	15.6	16.8	17.6	6.4	5.7
	500	6.2	4.8	5.6	4.6	5.1	10.8	12.8	15.2	6.4	6.2
	1000	5.4	4.4	4.8	4.1	4.0	8.4	6.0	9.6	6.1	6.2
	SE	1.4			0.7		1.4			0.7	

		$\phi_1 = .99$		$\phi_1 = -.95$	
		TB	TBA	TB	TBA
	50	14.9	8.2	4.5	4.5
	100	13.6	6.1	5.0	5.1
n	250	10.5	6.3	4.3	4.9
	500	8.3	6.7	6.0	6.1
	1000	5.7	4.9	5.4	5.6
	SE	0.7			

Table A-4. Observed power associated with TB and TBA for $n = 100$ and various values of b where the model is

$$Y_t = a + bt + Z_t$$

where $(1 - \phi_1 B)Z_t = a_t$

Nominal level = 5%
2-sided tests
1000 replications with $B = 399$

		b = 0.05		0.10		0.15	
		TB	TBA	TB	TBA	TB	TBA
ϕ_1	0.95	47.3	26.8	80.3	48.3	93.5	71.7
	0.90	58.4	38.3	86.4	58.3	97.2	78.7
	0.80	90.4	82.0	99.2	92.5	100.0	97.0
	0.00	100.0	100.0	100.0	100.0	100.0	100.0
	SE	1.6					

Table A-5: Observed significance levels and powers for "no line" and "line" models fit to the Hansen and Lebedeff temperature data

1000 replications for TB and TBA, $B = 399$
250 replications for BN
Nominal level = 5%
1-sided tests

(a) Observed Significance Level

AR(8)

n	BN	TB	TBA
100	24.4	8.0	5.9
150	23.6	9.1	6.6
200	27.6	7.6	5.5
SE	1.4	0.7	

(b) Observed Power

AR(8) + line

TB	TBA
76.2	50.2
93.3	77.6
99.9	94.8
1.6	

(c) Alternative Nonstationary Model

ARIMA(9,1,0)

n	TB	TBA
100	11.2	6.0
150	14.4	8.9
200	12.1	7.0
SE	1.6	

Table A-6: Observed significance levels and powers for AR(10) model fit to monthly MASIG data

1000 replications, $B = 399$

Nominal level = 5%

1-sided tests

		Significance Level		Power			
				− 0.05/yr slope		− 0.1/yr slope	
		TB	TBA	TB	TBA	TB	TBA
Years	5	7.5	4.3	9.5	5.3	9.8	4.9
	10	6.0	4.3	10.5	7.5	14.4	8.7
	15	5.3	5.1	13.7	11.6	26.8	22.7
	20	4.6	4.2	21.7	18.4	45.8	39.0
	25	5.2	5.0	31.8	28.6	70.5	62.0
	30	6.0	5.7	49.5	45.7	88.0	83.7
	SE	0.7		1.6		1.6	

Figure A-1. Global Temperature Series

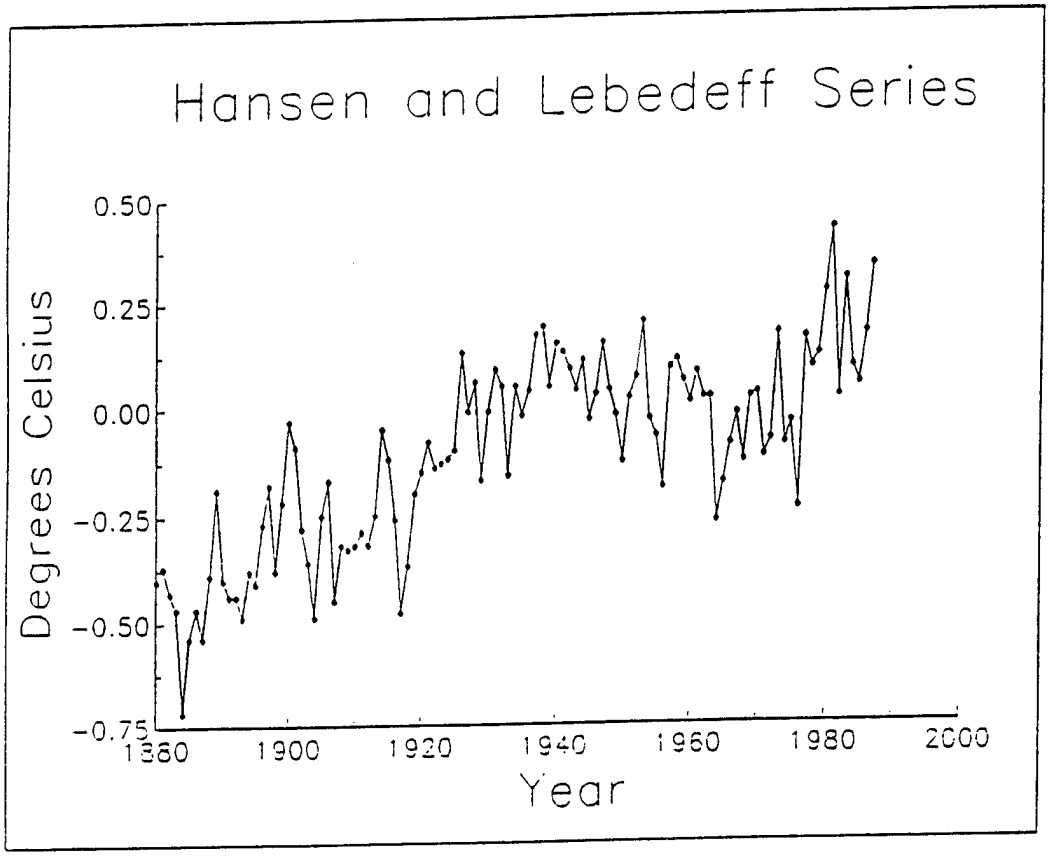


Figure A-2. Simulation from MASIG Model: Hawaii to San Diego

APPENDIX B

THE TRENDS SOFTWARE

B.1. Introduction

The TRENDS software performs four basic computations related to detecting a trend in time series data: 1) Trend Detection, 2) Power of the Test (probability of detecting a trend), 3) Trend Stability (will the trend continue?), and 4) Classification Statistics. The trend detection software runs in one of two modes: automatic (novice) and manual (expert). At any stage in the analysis the user may turn off or on the customization button. With the customization button off the program will perform the calculation with no intermediate input from the user. The customization on mode allows the user to check the calculations at various intermediate stages and to change inputs if desired.

B.2. Getting Started

To get started a set of time series data must be loaded. The time series data is assumed to be equally spaced in time. The data should be in a file with the first entry an integer giving the number of data points, n, followed by a column of n real data values. Once loaded, a plot of the data appears in the main window (see figure B-1). A slider bar allows the user to choose any contiguous subset of the data on which to perform the subsequent analysis.

B.3. Trend Detection

This option answers the question: is there a trend in the selected time series (or subset)? For the selected time series, the test described in section 2.2 is performed (the default number of bootstrap replications is 399). If the test statistic, $\hat{t}$, falls in the critical region the trend is significant, at the 5% level, and is said to have been detected. When the customization button is on, the user is shown the trend program input dialog before it is executed. At this point the user has the option of changing any of the input parameters, which have their default values. The program chooses an order for the autoregressive process used to model the noise. The user has the option of changing this parameter at this time.

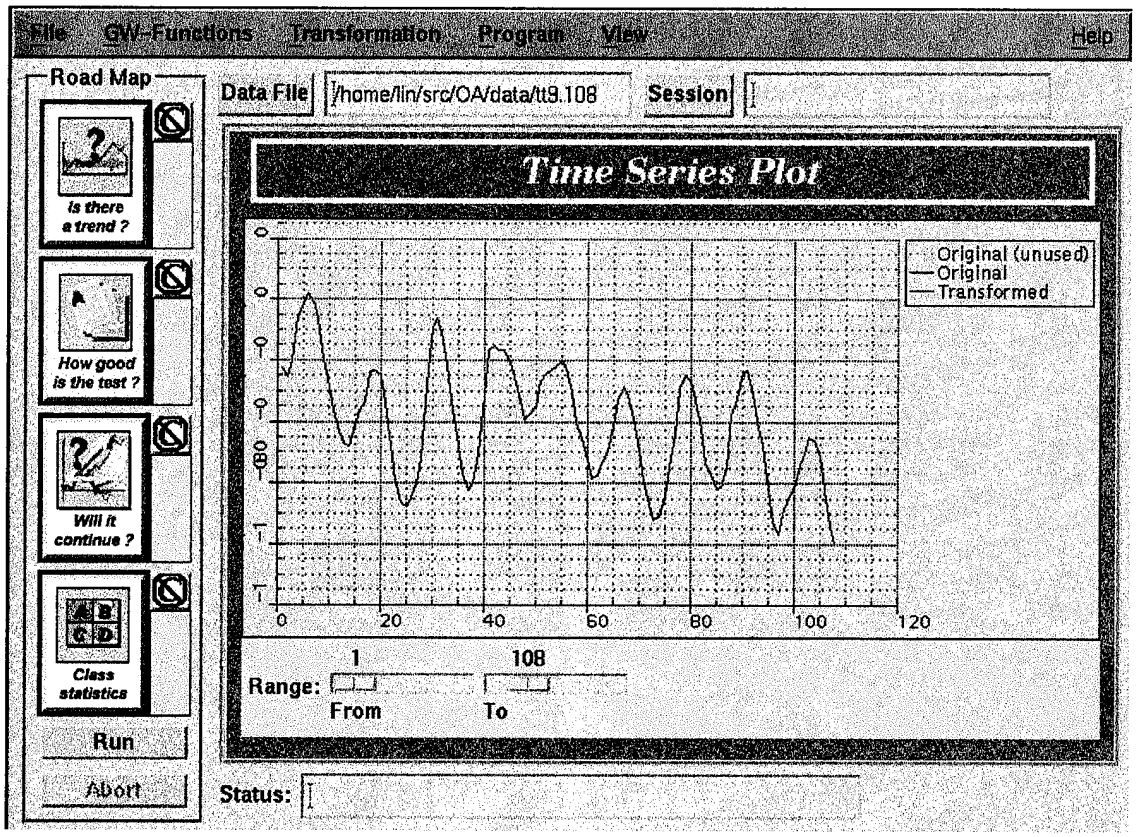


Figure B-1. Main window of TRENDS show input time series.

The user may wish to carry out this entire process with no automation. After having selected the data (or subset), select "Pre-Processor" on the Transformation menu. Remove best fit line by selecting "Remove from Data" under Linear Trend. The transformed data with the best fit line removed is plotted along with the original data. (If the plot is not on the screen it can be selected from the View menu.) Estimate the order of an autoregressive process (AR) to model this transformed series by selecting "AIC Estimate" from the GW-Functions menu and executing. To run the trend program select "Trend Program" from the program menu and choose "Single Data Set" and "Original" buttons. One- or two-sided testing may be chosen and the AR order for the transformed data may be changed. The final result can be viewed by selecting "Current Session" from the View menu. Detailed output may be viewed by selecting "Reports" from the view menu.

B.4. Power of the Test (Probability of detecting a trend)

This option computes an estimate of the power of the test of significance of trend, i.e., an estimate of the probability that the trend will be judged as significant. This probability is computed by generating many realizations (the default is 100) from a line plus (AR) noise

model fit to the data and calculating the percentage of those realizations with significant trend using the test procedure described in section 2.2.

When the customization button is on, the user is shown the trend program input dialog before it is executed. At this point the user has the option of changing any of the input parameters, which have their default values (the default value for the number of realizations is 100). The model parameters may be viewed, and changed if desired, by selecting the "Model Parameters" button in the Trend dialog or by selecting "Parameters" from the View menu. The series length for the data to be generated and the number of realizations to be performed (the larger the number of realizations, the better the statistics but the longer the execution time) are selected using the slider bars. With this option one can obtain answers to questions such as how long will it take to detect a trend in data similar to a given data set and how large would the trend have to be in a given data set to be significant. For example, to answer the question how long will it take to detect a trend in data similar to a given data set one increases the length of the realizations in the dialog box until the desired probability of detection is reached. One can easily determine the probability of detection as a function of series length by running the program with sequentially increasing values of the series length.

The user may wish to carry out this entire process with no automation. After having selected the data (or subset), select "Pre-Processor" on the Transformation menu. Remove the best fit line by selecting "Remove from Data" under Linear Trend. The transformed data with the best fit line removed is plotted along with the original data. (If the plot is not on the screen it can be selected from the View menu.) To model this transformed series as an autoregressive process (AR) first estimate the order by selecting "AIC Estimate" from the GW-Functions menu and executing. Next estimate the AR coefficients by selecting "Estimate Coeffs" from the GW-Functions menu. To run the trend program select "Trend Program" from the program menu and choose "Generate Data from Model". Choose the realization length and the number of realizations using the slider bar. The entire set of model parameters may be viewed, and changed if desired, by selecting "Parameters" from the View menu. One- or two-sided testing may be chosen and the AR order for the transformed data may be changed (it should be equal to the model order of the data to be generated). The final result can be viewed by selecting "Current Session" from the View menu. Detailed output may be viewed by selecting "Reports" from the view menu.

B.5. Trend Stability (Will the trend continue?)

This option determines whether the trend detected as significant in option 1 is predicted to continue by determining if the selected series is best classified as a line plus noise model or an ARIMA model using the approach described in Bottone, Gray, and Woodward [1995]. The answer is either "the trend will continue" (line plus noise) or "the trend will not continue" (ARIMA).

When the customization button is on, the user is shown the bootstrap program input dialog before it is executed. At this point the user has the option of changing any of the input parameters, which have their default values (the default number of bootstrap replications is 399). The program chooses an order for the autoregressive process for the noise in a line plus noise model and the order of the ARIMA process used in the bootstrap classification. The user has the option of changing these parameters at this time. The user may also choose between parametric and non-parametric bootstrapping and choose a white noise variance classification or an Anderson classification.

The user may wish to carry out this entire process with no automation. After having selected the data (or subset), select "Pre-Processor" on the Transformation menu. Remove best fit line by selecting "Remove from Data" under Linear Trend. The transformed data with the best fit line removed is plotted along with the original data. (If the plot is not on the screen it can be selected from the View menu.) Estimate the order of an autoregressive process (AR) to model this transformed series by selecting "AIC Estimate" from the GW-Functions menu and executing. To run the bootstrap program select "Boot Program" from the program menu and choose "Single Data Set" and "Original" buttons. Parametric or non-parametric bootstrapping may be chosen and the white noise variance (WNV) test or the Anderson test may be chosen. The AR order for the line plus noise model and the orders for the ARIMA model may be changed. The final result can be viewed by selecting "Current Session" from the View menu. Detailed output may be viewed by selecting "Reports" from the view menu.

B.6. Classification Statistics

This option computes the classification statistics showing how well the classification procedure of option 3 works. The result is a 2 x 2 matrix giving an estimate of the probability that the data will be classified as line plus noise given that it is line plus noise in the upper left hand corner. The lower right hand entry is the probability that the data will be classified as ARIMA given that it is ARIMA. The off-diagonal entries give the

probability that line plus noise is chosen when the data actually comes from an ARIMA model and vice versa.

When the customization button is on, the user is shown the boot program input dialog before both the first stage (A) and second stage (B) are executed. The first stage computes the probabilities that realizations generated from a line plus noise model fit to the data is classified as line plus noise or ARIMA. The second stage computes the probabilities that realizations generated from an ARIMA model fit to the data is classified as line plus noise or ARIMA. At this point the user has the option of changing any of the input parameters, which have their default values (the default for the number of realizations is 100). The model parameters may be viewed, and changed if desired, by selecting the "Model Parameters" button in the Boot dialog or by selecting "Parameters" from the View menu. The series length for the data to be generated and the number of realizations to be performed (the larger the number of realizations, the better the statistics but the longer the execution time) are selected using the slider bars. The number of bootstrap replications can also be changed using the slider bar.

The user may wish to carry out this entire process with no automation:

A. After having selected the data (or subset), select "Pre-Processor" on the Transformation menu. Remove best fit line by selecting "Remove from Data" under Linear Trend. The transformed data with the best fit line removed is plotted along with the original data. (If the plot is not on the screen it can be selected from the View menu.) To model this transformed series as an autoregressive process (AR) first estimate the order by selecting "AIC Estimate" from the GW-Functions menu and execute. Next estimate the AR coefficients by selecting "Estimate Coeffs" from the GW-Functions menu. Run the bootstrap program by selecting "Boot Program" from the program menu and choose "Generate Data from Model". Choose the realization length, the number of realizations and the number of replications using the slider bar. The entire set of model parameters may be viewed, and changed if desired, by selecting "Parameters" from the View menu. Parametric or non-parametric bootstrapping may be chosen and the white noise variance (WNV) test or the Anderson test may be chosen. The AR orders for the line plus noise model and the ARIMA model fit to the realizations may be changed. The final result (left hand entries of the classification matrix) can be viewed by selecting "Current Session" from the View menu. Detailed output may be viewed by selecting "Reports" from the view menu.

B. Re-enter the data. Select "operator" from the Transformation menu and apply the operator (1-B) (which is the default) on the series. The transformed data is plotted along with the original data. (If the plot is not on the screen it can be selected from the View menu.) To model this transformed series as an autoregressive process (AR) first estimate the order by selecting "AIC Estimate" from the GW-Functions menu and executing. Next estimate the AR coefficients by selecting "Estimate Coeffs" from the GW-Functions menu. Choose "MULT" from the GW-Functions menu and load the AR(p) operator, multiply by (1-B) (the default) and apply. Run the bootstrap program by selecting "Boot Program" from the program menu and choose "Generate Data from Model". Choose the realization length, the number of realizations and the number of replications using the slider bar. The entire set of model parameters may be viewed, and changed if desired, by selecting "Parameters" from the View menu. Parametric or non-parametric bootstrapping may be chosen and the white noise variance (WNV) test or the Anderson test may be chosen. The AR orders for the line plus noise model and the ARIMA model fit to the realizations may be changed. The final result (right hand entries of the classification matrix) can be viewed by selecting "Current Session" from the View menu. Detailed output may be viewed by selecting "Reports" from the view menu.

B.7. Example

Figure B-2 shows the main window after a complete run. The time series analyzed, shown in figure B-2 is 108 monthly values (9 years) of simulated data from the GFDL model [Manabe et al., 1991] representing the travel-time anomaly along a path from Hawaii to San Diego. The upper left hand corner shows the result of step 1 that a trend is detected in the selected series. The power of the test, which is an estimate of the probability that a trend will be detected in data similar to the input data, appears as a pie chart in the upper right hand corner of the main window, which in this case is 96%. In other words, one expects to detect a trend in data similar to the input data 96% of the time.

The lower left hand corner shows that the best eventual forecast is for the trend to continue by classifying the input time series as line plus noise as opposed to ARIMA. The classification statistics appear as a 2x2 matrix in the lower right hand corner of the main window. For this case, the probability of classifying a time series similar to the input series as line plus noise when it really is line plus noise is 95%. The probability of classifying the time series as ARIMA when it really is ARIMA is 66%. The off-diagonal entries give the probability of classifying the time series as ARIMA when it is actually line plus noise and vice versa.

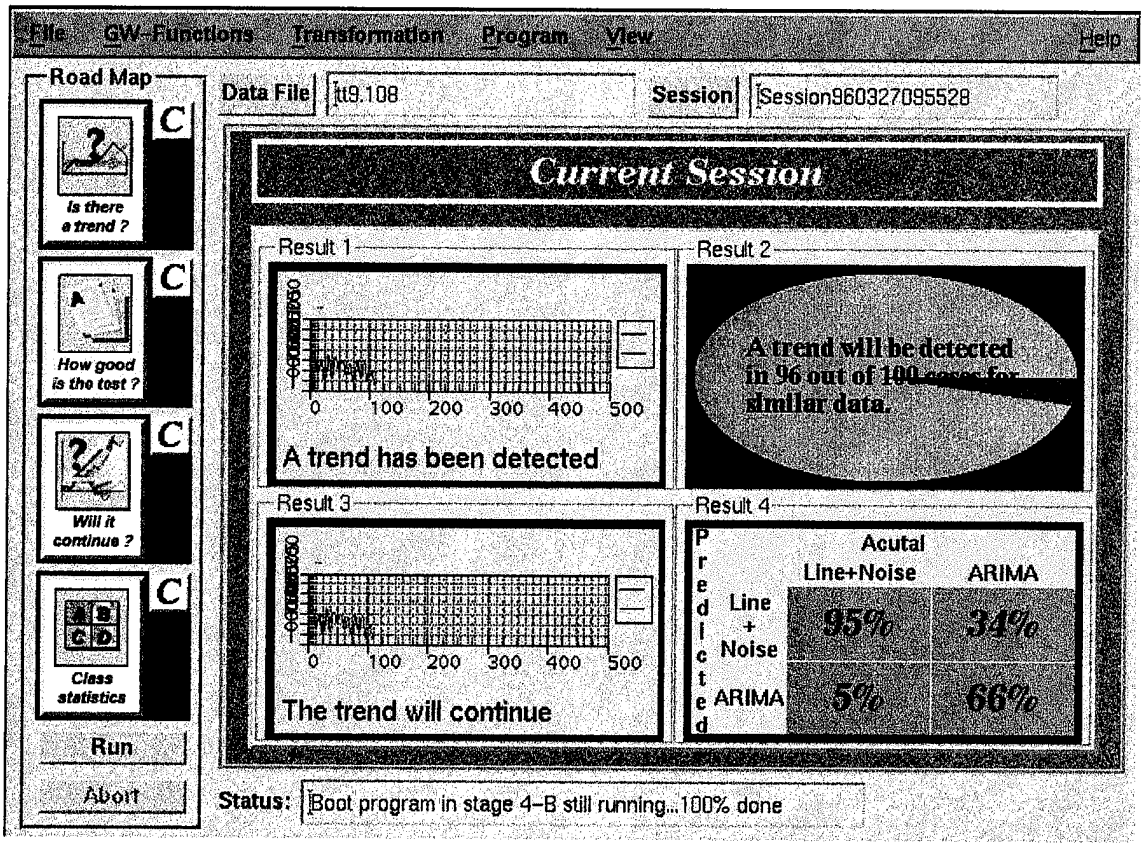


Figure B-2. Main window of TRENDS after complete run.

B.8. CD ROM

A multimedia tutorial about the use of statistical methods in the ARPA Acoustic Monitoring of Global Ocean Climate (AMGOC) program has been developed by Mission Research Corporation. It is supplied on CD ROM and runs with a supplied viewer on both Macintosh and PC Compatible computers. It incorporates advanced user navigation techniques to allow exploration of information in visual form by providing access to several layers of material at successive levels of detail. As such, it is intended to guide the user through an introduction to the background of the overall program, an elementary understanding of the statistical methods employed, their importance to the problem of extraction of warming trends from ocean acoustic data and finally an example of application of the **TRENDS** software. The cover for the CD ROM for the tutorial is shown in figure B-3. The implementation of this multimedia software includes access to a navigation Help section, a full Index of the entire tutorial and a Map which allows the user to view the tree of choices and also to move instantly to any chosen page. Details about

this CD ROM tutorial can be obtained at the internet address: http://chapman.mrcsb.com/OA.html.

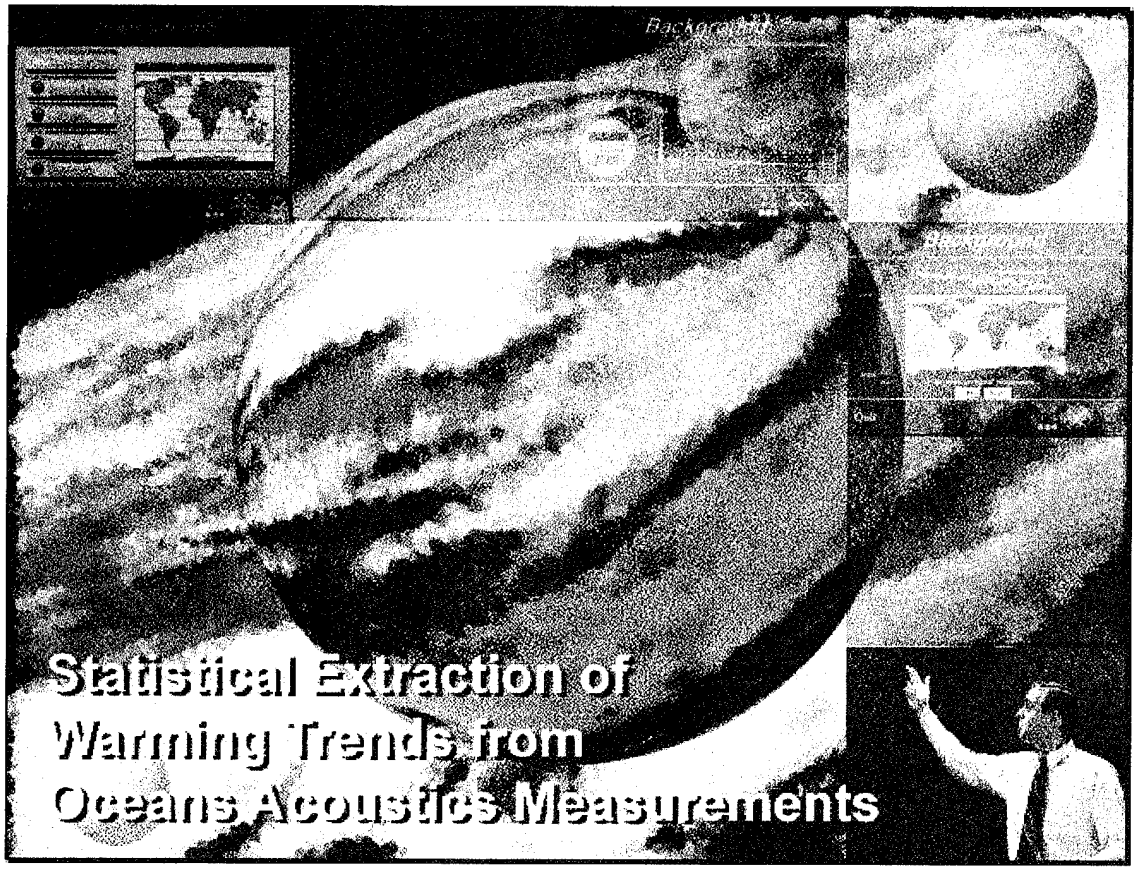


Figure B-3. Cover for the CD ROM multimedia tutorial.