7. Segmentation of Non-Stationary Time Series

7.1. INTRODUCTION

In many practical situations, a model of piecewise-stationary time series with successive stationary segments is assumed for nonstationary series (Fig. 7.1). In order to establish this piecewise-stationary time series model, segmentation algorithms have been developed to detect segment boundaries and estimate the parameters characterizing each segment. Techniques of segmentation without requiring *a priori* spectral information have been developed based on autoregressive (AR) models. It is assumed in these techniques that statistical properties described by a set of AR parameters remain the same in each segment. If these algorithms yield a single series – the original series – then the series is stationary.

Several tests have been proposed for use in segmentation algorithms. de Souza and Thomson (1982) derived a test based on a likelihood ratio statistic, which is insensitive to changes in the signal energy but sensitive to changes in spectral shape. This test statistic has an asymptotic chi-square distribution with p degrees of freedom under the null hypothesis of no changes in the AR parameters. Appel and Brandt (1983) transformed a likelihood ratio to a measure of the information loss caused by the assumption that the null hypothesis is true. Under the null hypothesis of no changes in the AR parameters and variance, the threshold is empirically adjusted according to the particular application for which the segmentation procedure is designed. Because the threshold of the distance is empirically determined, it is difficult to analyze it analytically. Lovell and Boashash (1987) adopted the segmentation procedure of Appel

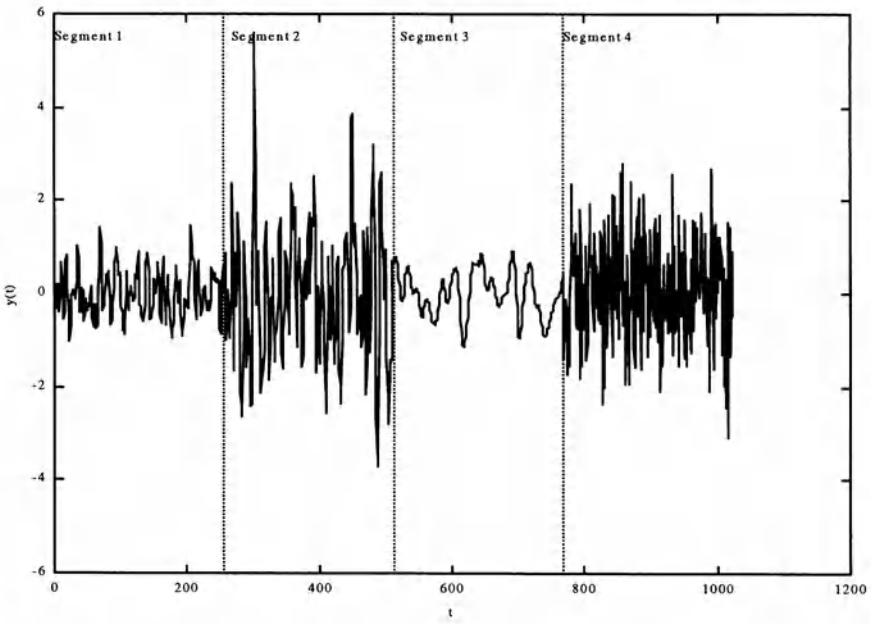

Figure 7.1. A piecewise-stationary time series with successive stationary segments.

A. Ramachandra Rao et al., *Nonstationarities in Hydrologic and Environmental Time Series*

and Brandt (1983), but replaced the test with the one of de Souza and Thomson (1982) instead. Since they maintain that variations in the signal energy are largely due to the measuring instrument, the segmentation algorithm of Lovell and Boashash (1987) is designed to determine segments according to changes in the spectral shape alone. Because the test derived by de Souza and Thomson (1982) is insensitive to changes in the signal energy, it is questionable whether the resulting segments from Lovell and Boashash's (1987) method are still stationary.

The algorithm used in Appel and Brandt's (1983) segmentation procedure consists of two stages: boundary detection and boundary optimization. In the first stage, a preliminary boundary is detected after the inherent change point; and the optimal boundary is then searched in the second stage. A disadvantage of this algorithm is that both boundary detection and optimization procedures are not symmetric with respect to time, which may give different results from the same time series in reverse order. Imberger and Ivey (1991) defined a distance derivative, a simple and symmetric segmentation technique suggested in part by personal communication with de Souza. The distance derivative is computed by taking the derivative of the test statistic derived by de Souza and Thomson (1982). A threshold is required for the segmentation algorithm so that stationary segments are determined where values of the distance derivative are below this threshold. However, the value of threshold is empirically adjusted rather than statistically determined. Besides, the distance derivative is based on the test statistic of de Souza and Thomson (1982) which is insensitive to changes in the signal energy.

There are three important parameters in these segmentation algorithms based on AR models: (1) the threshold for the test used in the algorithm; (2) the minimum starting segment length; and (3) the AR order p. For selecting the threshold, a tradeoff has to be considered between true and false detection rates. The minimum starting segment length is suggested by Appel and Brandt (1983) to be less than 70% of the length of the smallest segment and greater than $(p/3)^2$, in order to obtain small false detection rates. Imberger and Ivey (1991) compared estimated Batchelor spectra (defined in Section 2.2) with the AR spectra of order p = 4, 8, 12, 16 and 19. The results favored rather large lags; however, the overall spectral shape was adequately developed by AR models of order p = 4 (Imberger and Ivey, 1991).

Several tests based on AR models have been proposed for detecting spectral changes in time series, although the performances of these tests have not been evaluated. In this chapter, the results of comparison of performances of four tests based on AR models is presented first. Synthetic series are used for the comparison. The tests include the first test which is sensitive to changes in the spectral shape (de Souza and Thomson, 1982), the second test which is a derivative of the first test statistic (Imberger and Ivey, 1991), the third test which is designed to detect changes in the AR parameters (Davis, Huang and Yao, 1995), and the fourth test which is modified from Tsay (1988) to detect changes in the AR variance.

Tests based on AR models require specifying the AR order p which is unknown in observed data. In order to avoid the task of determining the AR order, a non-parametric test based on wavelet analysis is proposed. Based on the orthogonality and the property that at a given scale the correlation among wavelet coefficients is negligible (Wornell, 1996), the proposed test is derived for detecting changes in the wavelet variance. The

performance of this non-parametric test based on wavelet analysis is compared with those based on AR models and the results are presented.

A modified segmentation algorithm consisting of three stages, including boundary detection and two stages of boundary optimization is presented. The first stage of this algorithm is a modified form of the boundary detection procedures of Appel and Brandt (1983). In the two stages of boundary optimization, a symmetric approach is modified from the segmentation algorithm of Imberger and Ivey (1991). In general, a preliminary boundary is detected in the first stage, the optimal ending boundary is determined in the second stage (boundary optimization one), and the optimal starting boundary is determined in the third stage (boundary optimization two).

This chapter is organized as follows: four tests based on AR models, including test 1 by de Souza and Thomson (1982), test 2 by Imberger and Ivey (1991), test 3 by Davis, Huang and Yao (1995) and test 4 modified from Tsay (1987), are defined in Section 7.2; a test based on wavelet analysis (test 5) is proposed in Section 7.3; a modified segmentation algorithm is presented in Section 7.4; two types of synthetic series are used in Section 7.5 to examine variations of the test statistics of tests 1, 3 and 4 based on AR models with respect to the AR order p; the performances of tests 1-5 are compared in Section 7.6 by using synthetic series consisting of multiple stationary segments and a comparison is made in Section 7.7 between the algorithms with and without boundary optimization, using synthetic series composed of multiple stationary segments and non-stationary segments.

7.2. TESTS BASED ON AR MODELS

7.2.1. Test 1 (de Souza and Thomson, 1982). Consider a series of random variables $\{Y_t\}$ satisfying the AR(p) model in Eq. 7.1:

$$Y_t = \varepsilon_t - a_0 - \sum_{i=1}^{p} a_i Y_{t-i}, \; t = 1,2,\ldots,N. \tag{7.1}$$

Assume $a_0 = 0$, and $\{\varepsilon_t\}$ are independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance σ^2. Let **a** be the p-dimensional column vector of AR(p) parameters: $[a_1, a_2,\ldots, a_p]'$, the least square estimates of the parameters of the AR model are

$$\hat{\mathbf{a}} = -\mathbf{D}^{-1}\mathbf{d}, \tag{7.2}$$

and

$$\hat{\sigma}^2 = \hat{\mathbf{a}}'\mathbf{C}\hat{\mathbf{a}}, \tag{7.3}$$

where **D** is a p by p submatrix of covariance matrix **C** obtained by deleting row and column zero, and **d** is the p-dimensional column vector equivalent to the first column of **D**.

Let $\{Y_{Rt}\}$ and $\{Y_{Tt}\}$ denote two series of length N_R and N_T respectively. Let us assume that $\{Y_{Rt}\}$ is generated by an AR(p) model with $\mathbf{a} = \mathbf{a}_R$, $\sigma^2 = \sigma_R^2$, and $\{Y_{Tt}\}$, independent of $\{Y_{Rt}\}$, is generated by an AR(p) model with $\mathbf{a} = \mathbf{a}_T$, $\sigma^2 = \sigma_T^2$.

If σ_R^2 and σ_T^2, not necessarily the same, are known as *apriori*, the null hypothesis of no changes in the AR parameters is

$$\mathbf{H_0}: \mathbf{a}_R = \mathbf{a}_T.$$

The relevant likelihood ratio is

$$\exp\left[-\frac{1}{2}\left(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T\right)'\left\{\sigma_R^2\left(N_R'\mathbf{D}_R\right)^{-1} + \sigma_T^2\left(N_T'\mathbf{D}_T\right)^{-1}\right\}^{-1}\left(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T\right)\right], \tag{7.4}$$

where $N_R' = N_R - p$, $N_T' = N_T - p$, and $\mathbf{D}_R$ and $\mathbf{D}_T$ denote the submatrices of $\mathbf{C}_R$ and $\mathbf{C}_T$ obtained from $\{Y_{Rt}\}$ and $\{Y_{Tt}\}$, respectively. Under the null hypothesis, this statistic

$$\left(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T\right)'\left\{\sigma_R^2\left(N_R'\mathbf{D}_R\right)^{-1} + \sigma_T^2\left(N_T'\mathbf{D}_T\right)^{-1}\right\}^{-1}\left(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T\right) \tag{7.5}$$

has an asymptotic chi-squared distribution with p degrees of freedom. However, this test is of limited use in practice since the variances σ_R^2 and σ_T^2 are typically unknown, and the computation of the quadratic form

$$\left\{\sigma_R^2\left(N_R'\mathbf{D}_R\right)^{-1} + \sigma_T^2\left(N_T'\mathbf{D}_T\right)^{-1}\right\}^{-1} \tag{7.6}$$

may be infeasible, especially when singular matrices are encountered.

Two modifications are made to Eq. 7.5. The first modification is to replace σ_R^2 and σ_T^2 with the unbiased estimates

$$\sigma_R^{*2} = \frac{N_R'\hat{\sigma}_R^2}{N_R' - p}, \tag{7.7}$$

and

$$\sigma_T^{*2} = \frac{N_T'\hat{\sigma}_T^2}{N_T' - p}, \tag{7.8}$$

where $\hat{\sigma}_R^2$ and $\hat{\sigma}_T^2$ are the least squares estimates of σ_R^2 and σ_T^2. The second modification is to replace the quadratic form (Eq. 7.6) with

$$\frac{N'_R N'_T}{(N'_R + N'_T)^2}\left(\frac{N'_T \mathbf{D}_R}{\sigma_R^{*2}} + \frac{N'_R \mathbf{D}_T}{\sigma_T^{*2}}\right), \tag{7.9}$$

in the sense that the harmonic mean is replaced with the arithmetic mean of $\frac{\mathbf{D}_R}{\sigma_R^{*2}}$ and $\frac{\mathbf{D}_T}{\sigma_T^{*2}}$. These modifications yield

$$d_s = \frac{N'_R N'_T}{(N'_R + N'_T)^2}(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T)'\left(\frac{N'_T \mathbf{D}_R}{\sigma_R^{*2}} + \frac{N'_R \mathbf{D}_T}{\sigma_T^{*2}}\right)(\hat{\mathbf{a}}_R - \hat{\mathbf{a}}_T), \tag{7.10}$$

where d_s has an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis. Consequently, the decision rule is

$$\begin{cases} d_s < \chi_\alpha^2(p), \text{ accept Ho} \\ d_s \geq \chi_\alpha^2(p), \text{ do not accept Ho} \end{cases}.$$

The threshold $\chi_\alpha^2(p)$ is chosen corresponding to a given level of significance α.

7.2.2. Test 2 (Imberger and Ivey, 1991). If $\{Y_{Rt}\}$ and $\{Y_{Tt}\}$ are two halves of a series; then, a simple alternative is defined as distance derivative by following the definition of test statistic d_s in Eq. 7.10

$$dd/dt = \frac{2d_s}{N_R \Delta}, \textit{if } N_R = N_T, \tag{7.11}$$

where Δ is the sampling time interval. A minimum value is specified for the distance derivative below which the record is assumed stationary. Unlike test 1, the threshold for the distance derivative is empirically selected.

7.2.3. Test 3 (Davis, Huang and Yao, 1995). Consider a series $\{Y_t\}$ specified in Eq. 7.1, and $\{\varepsilon_t\}$ is a fourth-order white noise series, that is, for $i \leq j \leq k \leq l$,

$$E(\varepsilon_i) = 0, \tag{7.12}$$

$$E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, \textit{if } i = j, \\ 0, \textit{if } i < j, \end{cases} \tag{7.13}$$

$$E(\varepsilon_i \varepsilon_j \varepsilon_k) = \begin{cases} \mu_3, \textit{if } i = j = k, \\ 0, \textit{otherwise}, \end{cases} \tag{7.14}$$

$$E(\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l)=\begin{cases}\mu_4, \textit{if } i=j=k=l,\\ \sigma^4, \textit{if } i=j<k=l,\\ 0, \text{otherwise.}\end{cases} \tag{7.15}$$

If $\{\varepsilon_t\}$ satisfies Eqs. 7.12 to 7.15 with $\sup_t E|\varepsilon_t|^{4+\delta} < \infty$ for some $0 < \delta \le 1$, and $\{Y_t\}$ is strongly mixing with the mixing function $\rho(N) << N^{-(1+\varepsilon)(1+4/\delta)}$ for some $\varepsilon > 0$, the null hypothesis of no changes in the AR parameters after $t = k$ is

H_0: no change occurred in the AR parameters after $t = k$.

The log likelihood ratio, conditional on the first p observations, is given by

$$\Lambda_N(k)=-2\ln\left(\frac{L(N)}{L(k)L(N-k)}\right)=\min_a\sum_{t=p+1}^{N}\varepsilon_t^2-\min_a\sum_{t=p+1}^{k}\varepsilon_t^2-\min_a\sum_{t=k+1}^{N}\varepsilon_t^2=Q_1-Q_2-Q_3, \tag{7.16}$$

where

$$Q_1=Y_N'Y_N-Y_N'M_N(M_N'M_N)^{-1}M_N'Y_N', \tag{7.17}$$

$$Q_2=Y_k'Y_k-Y_k'M_k(M_k'M_k)^{-1}M_k'Y_k', \tag{7.18}$$

$$Q_3=\tilde{Y}_k'\tilde{Y}_k-\tilde{Y}_k'\tilde{M}_k(\tilde{M}_k'\tilde{M}_k)^{-1}\tilde{M}_k'\tilde{Y}_k', \tag{7.19}$$

$$Y_k=[Y_{p+1},\ldots,Y_k]', \tag{7.20}$$

$$\tilde{Y}_k=[Y_{k+1},\ldots,Y_N]', \tag{7.21}$$

$$M_k=\begin{bmatrix}1 & Y_p & Y_{p-1} & \cdots & Y_1\\ 1 & Y_{p+1} & Y_p & \cdots & Y_2\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & Y_{k-1} & Y_{k-2} & \cdots & Y_{k-p}\end{bmatrix}, \tag{7.22}$$

and

$$\tilde{M}_k=\begin{bmatrix}1 & Y_k & Y_{k-1} & \cdots & Y_{k-p+1}\\ 1 & Y_{k+1} & Y_k & \cdots & Y_{k-p+2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & Y_{N-1} & Y_{N-2} & \cdots & Y_{N-p}\end{bmatrix}. \tag{7.23}$$

Under the null hypothesis, Eq. 7.1 yields

$$Y_k = M_k a_{p+1} + E_k , \tag{7.24}$$

$$\tilde{Y}_k = \tilde{M}_k a_{p+1} + \tilde{E}_k , \tag{7.25}$$

where

$$a_{p+1} = [a_0, a_1, \ldots, a_p]' , \tag{7.26}$$

$$E_k = [\varepsilon_{p+1}, \varepsilon_{p+2}, \ldots, \varepsilon_k]' , \tag{7.27}$$

and

$$\tilde{E}_k = [\varepsilon_{k+1}, \varepsilon_{k+2}, \ldots, \varepsilon_N]' . \tag{7.28}$$

By replacing Y_N, Y_k *and* $\tilde{Y}_k$ in Eqs. 7.17 to 7.19, the log likelihood ratio is rearranged as

$$\Lambda_N(k) = -S_N' P_N S_N + S_k' P_k S_k + (S_N - S_k)' \tilde{P}_k (S_N - S_k), \tag{7.29}$$

where

$$S_k = M_k' E_k , \tag{7.30}$$

$$P_k = (M_k' M_k)^{-1}, \tag{7.31}$$

and

$$\tilde{P}_k = (\tilde{M}_k' \tilde{M}_k)^{-1} . \tag{7.32}$$

Under the null hypothesis, the normalized log likelihood ratio in Eq. 7.33

$$\gamma_n = \frac{\sigma^{-2}\Lambda_N - b_N(p+1)}{a_N(p+1)} , \tag{7.33}$$

has the following distribution (Theorem 2.2 in Davis, Huang and Yao (1995)):

$$P[\gamma_n \le z] \to \exp\left(-2\exp\left(-\frac{z}{2}\right)\right), \tag{7.34}$$

where *P* denotes the accumulative probability,

$$b_N(p) = \frac{2\ln\ln(N) + \left(\frac{p}{2}\right)\ln\ln\ln(N) - \ln\Gamma\left(\frac{p}{2}\right)^2}{2\ln\ln(N)}, \tag{7.35}$$

$$a_N(p) = \sqrt{\frac{b_N(p)}{2\ln\ln(N)}}, \tag{7.36}$$

and Γ is the gamma function. The decision rule is

$$\begin{cases} z_{1-\frac{\alpha}{2}} < \gamma_n < z_{\frac{\alpha}{2}}, \text{accept } H_0 \\ \gamma_n \le z_{1-\frac{\alpha}{2}} \text{ or } \gamma_n \ge z_{\frac{\alpha}{2}}, \text{do not accept } H_0 \end{cases}.$$

The threshold is chosen corresponding to a given level of significance α.

It should be noted that this likelihood ratio test statistic is applied to detection of a change point in a single series.

7.2.4. Test 4. Tsay (1988) proposed a test for detecting changes in the AR variance. Assume a change occurs after time $t = k$ in the AR variance of a series $\{Y_t\}$ specified in Eq. 7.1 with $a_0 = 0$, such that

$$\varepsilon_t = \begin{cases} e_t, \text{for } t < k \\ e_t(1+\omega_v), \text{for } t \ge k \end{cases} \tag{7.37}$$

where $\{e_t\}$ are i.i.d. Gaussian with zero mean and variance σ^2, and ω_v is a constant. If **a** and k are known, the null hypothesis that there are no changes in the AR variance after $t = k$ is

$\mathbf{H_0}$: $\omega_v = 0$.

The variance ratio of ε_t before and after the time k is given by r_v:

$$r_v = \frac{(k-1)\sum_{t=k}^{N}\varepsilon_t^2}{(N-k+1)\sum_{t=1}^{k-1}\varepsilon_t^2}. \tag{7.38}$$

The likelihood ratio test statistic r_v under the assumption of Gaussianity has a central F-distribution with degrees of freedom $(N + k - 1)$ and $(k - 1)$.

In order to apply the variance ratio test to two independent series, a simple modification is made in this study. Let $\{Y_{Rt}\}$ and $\{Y_{Tt}\}$ denote two independent series of length N_R and N_T, respectively. Assume that both series are generated from two AR(p) models with the same **a**, but different white-noise series $\{\varepsilon_{Rt}\}$ and $\{\varepsilon_{Tt}\}$ which are Gaussian distributed with zero mean and variance σ_R^2 and σ_T^2, respectively. Assuming **a** is known *a priori*, a new null hypothesis of no changes in the AR variance is:

$$\mathbf{H_{0n}}: \sigma_R^2 = \sigma_T^2.$$

The variance ratio is redefined as

$$r_v' = \frac{(N_R - p)\sum_{t=p+1}^{N_T} \varepsilon_{Tt}^2}{(N_T - p)\sum_{t=1}^{k-1} \varepsilon_{Rt}^2}. \tag{7.39}$$

Since **a** is unknown in practice, the least square estimates of **a** corresponding to $\{Y_{Rt}\}$ is used for both series under the null hypothesis and the same **a** assumption. Thus, the modified variance ratio is equivalent to

$$r_V' = \frac{\tilde{\hat{\sigma}}_T^2}{\hat{\sigma}_R^2}, \tag{7.40}$$

where

$$\hat{\sigma}_R^2 = \hat{a}'_R C_R \hat{a}_R, \tag{7.41}$$

$$\tilde{\hat{\sigma}}_T^2 = \hat{a}'_R C_T \hat{a}_R, \tag{7.42}$$

$$\hat{a}_R = -D_R^{-1} d_R. \tag{7.43}$$

This modified variance ratio r'_v has a central F-distribution with degrees of freedom $(N_T - p)$ and $(N_R - p)$. The decision rule is

$$\begin{cases} F_{1-\frac{\alpha}{2}}(N_T - p, N_R - p) < r_v' < F_{\frac{\alpha}{2}}(N_T - p, N_R - p), \text{ accept } H_{0n} \\ r_v' \le F_{1-\frac{\alpha}{2}}(N_T - p, N_R - p) \text{ or } r_v' \ge F_{\frac{\alpha}{2}}(N_T - p, N_R - p), \text{ do not accept } H_{0n} \end{cases}.$$

The threshold is chosen corresponding to a given level of significance α.