수문학적 시계열모형의 동정 및 매개변수의 추정
ESTIMATION AND IDENTIFICATION OF HYDROLOGIC TIME SERIES

연세대학교 토목공학과 교수 이 원환

연세대학교 토목공학과 조교수 조 원철

연세대학교 토목공학과 박사 이 재준

연세대학교 토목공학과 석사과정 조 재원

1. 서 론

근래 들어 수문학적 시계열을 모형화하기 위하여 여러가지 확률모형들이 이용되어 왔는데, 특히 하천의 유량, 증발량과 강우량 등이 그 대표적인 예라 할 수 있다. 여기에 이용 되어질 수 있는 모형은 여러가지가 있을 수 있지만, 그중 어느 것도 실제 수문사상의 시계열을 정확히 모형화할 수는 없는 것이다. 그 까닭은 여러가지 요인에 기인하는데, 그중 하나는 적용하려는 시계열을 어떠한 모형에 적용할 것인가하는 문제이고, 또 하나는 적용하려는 모형을 선택하였다고 했을 때 그 모형의 적정차수와 최적계수들을 어떻게 정확히 구해내는 가 하는 문제이다. 이것 또한 정확한 결정방법은 없으나 일반적으로 관측값과 모형에서 추정된 값과의 오차를 최소화함으로써 구할 수 있다.

본 연구에서는 수문학적 시계열에 흔히 이용되는 AR(autoregressive)모형과 ARMA(autoregressive moving average)모형의 각종 경합 모형에 관해서 MAICE(minimization Akaike information criterion estimate), MFPEE(minimization final prediction error estimate) 등의 방법을 이용하여 적정차수를 구하고 이를 이용하여 최적의 모형을 선정한다.

또한, 최적차수결정과정에서 모형에 필요한 AR 계수는 Levinson-Durbin알

고리즘을 ARMA 모형에서는 연세 알고리즘을 이용하여 풀기 때문에 모형화도 동시에 가능하게 한다.

이러한 알고리즘은 종래의 Box-Jenkins 의 방법에 비해 명확하고, 수치적으로 비교하면서 최적모형을 결정할 수 있는 장점이 있다. 또한 ARMA 모형의 매개변수의 결정에서 이용된 연세 알고리즘은 비선형 Yule-Walker 방정식을 정확하고 빠르게 풀 수 있는 알고리즘이다.

본 연구에서는 태양 흑점자료의 최적차수 및 모형의 계수를 결정하고, 서울, 부산과 울릉도의 강수량을 해석하여, 최적차수와 선정된 모형에 대해서 적절한 계수를 구하는 연구이다.

2. AR(p) 모형의 차수 및 계수

2.1 AR(p) 모형 (autoregressive model)

시계열 $\{y_t \mid t=..,-1,0,1,..\}$이 $y_t=\Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + v_t$ 를 만족하는 확률변수일 때, $\{y_t\}$ 를 차수가 p 인 자기회귀계열이라 하며, 그 계수는 $\Phi_1,...,\Phi_p$ 이다. 이를 간략하게 표시하면 식 (2-1)과 같다.

$$\Phi(B)\, y_t = v_t \qquad \text{----------(2-1)}$$

여기서,

$$\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - ..... - \Phi_p B^p, \qquad (2\text{-}2)$$

식 (2-2)에서 B 는 후진 연산자이다.

AR모형에서는 $\Phi(z) = 0$ 의 근이 1 보다 클 경우에만 시계열 $\{y_t\}$의 정상성 (stationarity)이 만족된다.

시계열의 정상성이 만족될 때 식(2-1)은 다음 식 (2-3)과 같이 표시될 수 있다.

$$y_t = \sum_{i=0}^{\infty} \Phi_i^{-1}(B)\, v_{t-i} \qquad (2\text{-}3)$$

식 (2-1)과 식 (2-3) 를 곱하여 기대값을 구하면, 다음과 같은 Yule-Walker 방정식을 구할 수 있다.

$$\sigma(0) - \Phi_1 \sigma(-1) - ..... - \Phi_p \sigma(-p) = \sigma_v^2$$

$$\sigma(k) = \Phi_1 \sigma(k-1) - .... - \Phi_p \sigma(k-p) \quad \text{for } k \geq 1 \qquad (2\text{-}4)$$

여기서, $\sigma(k)$ 는 자기공분산함수이며, $\sigma(k)/\sigma(0)$ 를 r_k 라두고 이를 자기상관함수(ACF)라 한다.

2.2 AR 모형의 최적차수 추정을 위한 Levinson-Durbin 알고리즘

앞 절에서 구한 식 (2-4), 즉 Yule-Walker 방정식을 풀면 AR 계수 혹은 편자기상관함수를 구할 수가 있으며 이 방정식을 푸는 데는 Gauss 소거법, Cholesky 방법 그리고 Gauss-Seidal 반복해법 등이 있다. 하지만 본 연구에서는 m번째 계수$\{\Phi^m\}$을 구할 때 m-1 번째의 $\{\Phi^{m-1}\}$을 이용하는 Levinson-Durbin알고리즘을 이용하기로 하며 이 알고리즘은 컴퓨터시간도 줄일 수 있는 동시에 절차가 간단하다는 장점이 있다.

먼저 다음과 같은 정의를 두기로 한다.

$$\begin{aligned}
\vec{\Phi}_k &= (\Phi_{k,1} \quad \Phi_{k,2} \quad \cdots \quad \Phi_{k,k})^T \\
\vec{\Phi}_{k+1} &= (\Phi_{k+1,1} \quad \cdots \quad \Phi_{k+1,k})^T \\
\vec{r}_k &= (r_1 \quad r_2 \quad r_3 \quad \cdots \quad r_k)^T \\
\overleftarrow{r}_k &= (r_k \quad r_{k-1} \quad \cdots \quad r_1)^T
\end{aligned} \tag{2-5}$$

이 정의를 이용하면 Yule-Walker 방정식은 다음과 같이 된다.

$$\begin{bmatrix} \Sigma_m & \overleftarrow{r(k)} \\ \overleftarrow{r(k)}^T & 1 \end{bmatrix} \begin{bmatrix} \vec{\Phi}_{k+1} \\ \Phi_{k+1,k+1} \end{bmatrix} = \begin{bmatrix} \vec{r(k)} \\ r(k+1) \end{bmatrix} \tag{2-6}$$

이것을 풀면, 식(2-7)과 같다.

$$\begin{aligned}
\phi_{k+1,k+1} &= \frac{r(k+1) - \overleftarrow{r(k)}^T \vec{r(k)}}{1 - \vec{\phi}_k^T \vec{r(k)}}, \\
\phi_{k+1,j} &= \phi_{k,j} - \phi_{k+1,k+1}\, \phi_{k,k+1-j}, \quad j=1,2,\ldots,k.
\end{aligned} \tag{2-7}$$

이 된다.

2.3 . AR(p) 차수결정을 위한 정보기준량

2.3.1 FPE(final prediction error)

시계열 $\{y_t\}$와 같은 통계적 성질을 가지지만 $\{y_t\}$ 와는 독립적인 AR(p)과정인 시계열 $\{x_t\}$를 생각하면, 일반적 산출값 x_{T+1}^*은

$$x_{T+1}^{*} = \phi_{p,1} x_T + \cdots + \phi_{p,p} x_{T+p-1}$$

과 같이 된다.

T 가 충분히 큰 경우 산출값의 평균 제곱오차는

$$E(x_{T+1}^{*} - x_{T+1})^2 = \sigma^2 (1 + p/T)$$

과 같이 된다.

여기서, σ^2 값은 Yule-Walker 방정식에서 구한 잔차계열의 분산값이다. 여기에 근거하여 Akaike 는 FPE 값을 다음과 같이 제안하였다.

$$FPE(p) = (1 + 2p/T)\,\sigma^2 \quad \text{---------- (2-8)}$$

이 FPE(p)값을 최소화하는 모형이 오차를 최소화하는 최적의 모형인 것이다.

2.3.2. AIC(Akaike information criterion).

Kullback(1951) 은 통계적으로 상이성을 검토하기 위하여, 확률밀도함수 p.d.f. g(x) 와 f(x)의 상이성을 나타내는 다음과 같은 수를 제안하였다.

$$D(f;g) = \int f(x) \ln \{f(x)/g(x)\}\, dx$$

D 를 g에 대한 f 의 Kullback-Leiber information number 라고 한다. AIC 는 D 를 최소화시키는 데 이론적인 근거를 두며, 식(2- 9)와 같이 유도하였다.

$$AIC^{*}(k) = -2 \sum_{i=1}^{T} \ln f(x_i/\Theta_k) + 2k \quad \text{---------- (2-9)}$$

식 (2-12)을 Akaike 의 Information criterion 이라고 한다. 이 AIC 값은 시계열, 인자분석(factor analysis) 및 회귀분석 등에 적용되어 최적모형을 선택하는 데 사용되고 있으며, 식 (2- 9)를 최소화시키는 p 차수는 또한 다음 식도 최소화시킨다.

$$AIC(p) = \ln v(p) + 2p/T \quad \text{---------- (2-10)}$$

이 값은 ARMA 모형에 대해서는 식 (2-11)로 표현된다.

$$AIC(p,q) = \ln v(p,q) + 2\,(p+q)\,/T \quad \text{---------- (2-11)}$$

2.3.3 BIC 와 HQ

앞 절에서 언급된 FPE 와 AIC를 최소화시키는 차수 p를 구하는 방법을

각각 MFPEE(Minimizing final predection error estimator), MAICE 라고 한다.

이 두가지 방법은 실제의 차수 보다 크게 산정할 확률이 있기 때문에 불일치성이 존재하는 것으로 알려져 있다. 이 단점을 보완하는 방법이 BIC (bayesian information criterion)와 HQ(Hannan-Quinn)가 제안되었으며, 그 기본식은 다음과 같다.

$$BIC(k) = \ln \sigma_k^2 + (\ln T/T)\, k \quad \text{---------- (2-12)}$$

$$HQ(k) = \ln \sigma_k^2 + 2(k)\ln(\ln T)/T \quad \text{---------- (2-13)}$$

여기서, σ_k^2 은 잔차계열의 분산이며, T는 시계열의 총 수를 의미한다.

BIC, HQ 를 이용하는 방법은 일치성 조건은 어느 정도 만족하지만, 실제 차수보다 작게 산정할 우려가 있는 단점이 있다.

3. ARMA(p,q) 모형의 차수 및 계수결정

3.1 확장된 Yule-Walker 방정식 및 연세 알고리즘

차수가 (p,q) 이고 정상상태인 ARMA 시계열은 다음과 같다.

$$\phi(B)\, x_t = \Theta(B)\, v_t \quad \text{---------- (3-1)}$$

여기서 $\Gamma(z)$를 $\Theta(z)/\phi(z)$로 정의한 후 AR 에서와 마찬가지로 식 (3-1)에 변변 $y_{t-k} = \Sigma\, \Gamma_i\, v_{t-k-i}$ 를 곱한 후 기대값을 구하면, 다음과 같은 확장된 Yule-Walker 방정식이 얻어진다. 단, $\{v_t\}$ 는 i.i.d.(identically independent distributed)조건을 만족한다.

$$\phi_0\sigma(k)+\phi_1\sigma(k-1)+\ldots+\phi_p\sigma(k-p)$$

$$= \begin{cases} \sigma^2\{\Theta_k\Gamma_0+\Theta_{k+1}\Gamma_1+\ldots+\Theta_q\Gamma_{q-k}\}, & 0\le k\le q \quad \text{-------- (3-2a)} \\ 0, & k > q \quad \text{--------- (3-2b)} \end{cases}$$

식 (3-2a)와 식 (3-2b)를 확장된 Yule-Walker 방정식이라고 한다.

식 (3-2a)로 부터 MA 계수인 Θ_i를 구하게 되고, 식 (3-2b)로 부터 AR 계수 ϕ_i 를 구하게 된다.

먼저 ϕ_i, i =1,2,..,p 를 구하기 위해서는 식(3-2b)에서 k=q+1,..,q+p를 적용하여, Gauss 소거법을 통하여 $\{\phi_i\}$를 구한다. 이같이 구한 $\{\phi_i\}$는 식 (3-2a)에 대입되어 $\{\Theta_i\}$를 구하는데 이용된다.

정의에 의해서 $\Theta_j = \phi_0\Gamma_j + \phi_1\Gamma_{j-1} + \cdots + \phi_j\Gamma_0$ $(0 \le j \le q)$ (3-3)

이다. 다시 행렬로 표시하면, 식 (3-4)와 같다.

$$[\Gamma]\ \vec{\phi} = \vec{\Theta} \quad \text{---------- (3-4)}$$

식 (3-2a)를 k=0, 1, 2, ..., q 에 대해 적용하여 행렬로 표시하면

$$\sigma^2 \cdot [\Gamma_q]\ \vec{\Theta} = \Sigma\ \vec{\phi} \quad \text{---------- (3-5)}$$

식 (3-5)에 식 (3-4)를 대입하면,

$$(\sigma\ \Gamma_q)\ (\sigma\ \Gamma)\ \vec{\phi} = \Sigma\ \vec{\phi} \quad \text{---------- (3-6)}$$

가 된다. 다시 $\sigma\ \Gamma = H$ 로 표시하면, 다음과 같이 된다.

$$H_q \cdot H\ \vec{\phi} = \Sigma\ \vec{\phi} \quad \text{---------- (3-7)}$$

혹은

$$H\ \vec{\phi} = H_q^{-1}\Sigma\ \vec{\phi} \quad \text{---------- (3-8)}$$

수렴성을 증가시키기 위해서 양변에 $H\ \vec{\phi}$ 를 더하여 정리하면,

$$H^m\vec{\phi} = \{ (H^{m-1}\ H_q^{-1m-1}\ \Sigma) \}/2 \cdot \vec{\phi} \quad \text{---------- (3-9)}$$

가 된다.

(3-9)식과 같은 비선형 방정식을 이용해서 σ^2 과 MA계수를 구할 수 있다. 즉, 행렬 [H]의 요소들을 구하면 Γ_i 를 구할 수 있고, 이 값을 식(3-3) 에 대입하면 Θ_i를 구할 수 있다. 이와 같은 알고리즘을 연세 알고리즘이라고 한다.

3.2 ARMA(p, q) 모델의 차수 결정방법

AR(p) 모델의 경우와 마찬가지의 개념으로, 오차를 최소화하는 FPE 와 Kullback-number 를 최소화하는 AIC, FPE 와 AIC 의 불일치성을 보완하는 BIC 및 HQ값을 (p, q) 차수에 대해 정의하면 다음 식과 같다.

$$FPE(p,q) = (1 + 2(p+q)/T)\ \sigma_{p,q}^2 \quad \text{---------- (3-10)}$$

$$AIC(p,q) = \ln v(p,q) + 2(p+q)/T \quad \text{---------- (3-12)}$$

$$BIC(p,q) = \ln v(p,q) + (p+q)\ \ln T/T \quad \text{---------- (3-13)}$$

$$HQ(p,q) = \ln v(p,q) + 2(p+q)\ \ln \ln T/T \quad \text{---------- (3-14)}$$

윗 식의 모든 오른쪽 항은 벌칙함수이다. 이 값들을 최소화하는 차수(p, q)가 가장 적합한 ARMA(p, q) 차수가 되는 것이다.

4. 자료 및 결과분석

4.1 기본 자료

본 연구에서는 Box-Jenkins 가 제안한 자료A와 태양흑점자료 및 서울, 부산 그리고 울릉도 우량관측소의 년강수량과 월강수량에 대해서 해석을 하였다. 태양흑점자료의 경우에는 상관함수를 구해본 결과 11년주기가 명확하게 나타났으며, 월강수량의 해석시에는 시계열자료의 계절성을 계차화함으로써 제거하여 해석하였다. 그리고 각 강수량 자료에 대해서 비정상성을 정상화하기 위하여 평균에 대해서는 계차화를 하며, 분산의 비정상성에 대해서는 Log 변환을 하여 정상화하였다.

4.2 동정

본 연구에서의 동정절차는 MAICE, MFPEE, MBICE와 MBICE 방법을 이용하여 적정차수를 각 자료에 대하여 구한다. 각 자료에 대해서 동정을 각각의 방법을 이용하여 구한 후에 정상성이 만족하지 않는 차수는 고려하지 않고, 정상성이 만족되는 차수중에서 가장 최소인 차수를 결정한다. 각 자료에 대한 AR, ARMA 및 MA에 대해 구한 차수결정값을 그림 1- 그림 3 에 나타내었다.

4.3 매개변수 추정

AR 모형에 대해서 매개변수의 추정을 하기위해서 Levinson-Durbin 알고리즘을 사용하며 ARMA 모형과 MA 모형에 대해서는 연세 알고리즘을 이용한다. AR의 경우에는 차수 p 에 대해서는 반복없이 한번만에 구하지만, 연세 알고리즘은 비선형방정식을 풀어야 하기 때문에 반복해법을 사용한다. 이때 반복이 수렴을 만족해야 차수를 구할 수 있다.

동정의 과정에서 결정한 최적차수에 대해서 모형에 맞는 알고리즘을 이용하여 매개변수를 추정한 후 모형식을 구하면 표 1 과 같다.

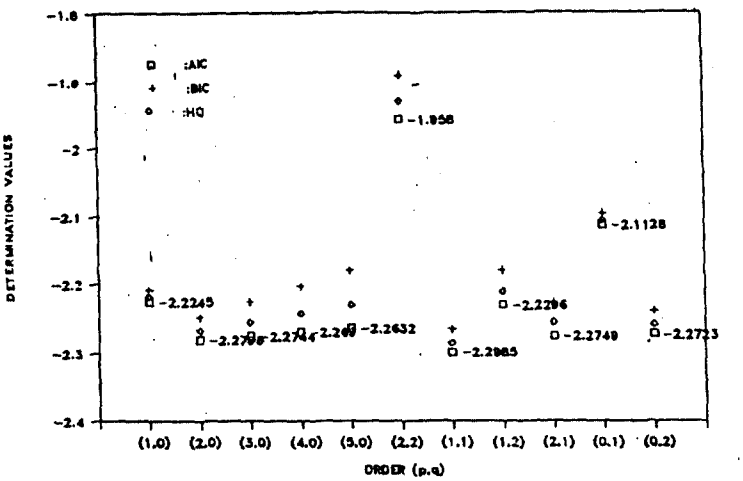

그림. 1 Box-Jenkins가 제안한 자료 A

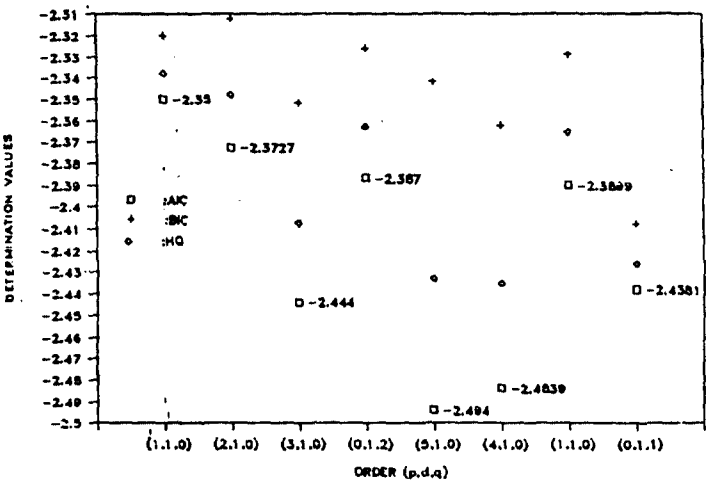

그림. 2 서울지방의 계차화된 년강수량

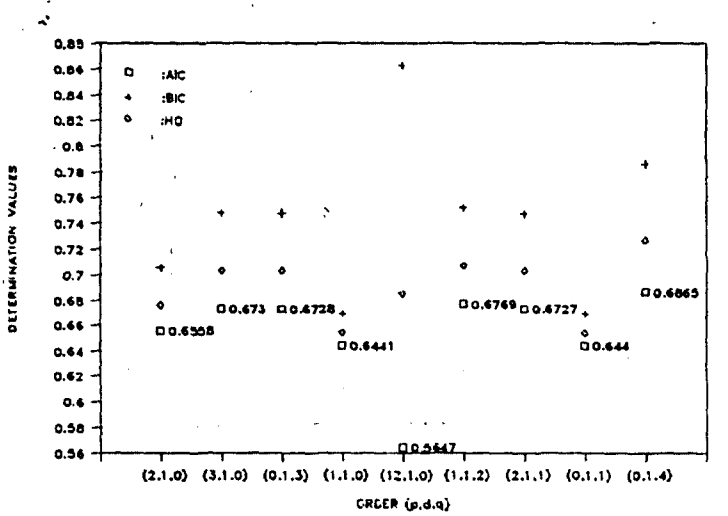

그림. 3 부산지방의 계차화된 월강수량

표.1 각 자료에 대한 모형

자 료	AR, MA, ARMA 모형
Box-Jenkins 자료	$(1-.868B)y_t = (1-.481B)e_t$
태양흑점 자료	$(1+1.318B-.634B^2)y_t = e_t$
서울 계수화된 년강수량	$(1-.725B-.562B^2-.545B^3-.381B^4-.188B^5)(y_t-y_{t-1})=e_t$ $(1+.0394B)(y_t-y_{t-1})=(1-.528B)e_t$
울릉도 년강수량	$(1-.575B-.290B^2)y_t = (1-.478B)e_t$
부산 월강수량	$(1+.294B+.114B^2-.043B^3-.110B^4-.144B^5-.192B^6)y_t=e_t$
부산 계차화된 년강수량	$(1+.067B)(y_t-y_{t-12})=e_t$ $(1-.534B)(y_t-y_{t-12})=(1-.475B-.013B^2)e_t$
부산의 계차화된 년강수량	$(1-.831B-.59B^2-.435B^3-.242B^4)(y_t-y_{t-1})=e_t$ $(1+.131B)(y_t-y_{t-1})=(1-.661B)e_t$

5. 결 론

본 연구에서 이용한 MAICE, MFPEE, MBICE, MHQE 의 방법은 통계학적 성질을 만족할 뿐만 아니라, 수치적으로 비교하면서 최적차수를 결정할 수 있기 때문에 명백한 차수를 밝힐 수 있다는 장점이 있다. 또한 AR 모형에서는 Levinson-Durbin 알고리즘을 ARMA 모형에서는 연세 알고리즘을 이용하여 각 모형의 매개변수를 추정할 수 있기 때문에 본 연구의 분석결과로 주어진 자료에 대해서 모형화가 가능하였다.

특히 연세 알고리즘에서는 Toeplitz행렬의 역행렬성질을 이용하고 수렴성을 좋게 하기 위해서 Newton-Raphson 반복기법을 사용하기 때문에, 기존의 방법보다 훨씬 정확하고 빠른 속도로 확장된 Yule - Walker 비선형방정식을 풀 수 있었다.

그러므로, 본 연구에서 제시하는 알고리즘은 시계열의 차수를 분명하게 제시하여 줄 뿐만 아니라, AR 계수는 물론 MA 계수를 구할 수 있기 때문에 예측을 위한 AR(p), ARMA(p,q) 모형의 모형화도 동시에 완성할 수 있다.