수질 자료에 대한 ARIMA 모형 적용

ARIMA Modeling for Monthly Oxygen Demand Data

허 용 구* · 박 승 우(서울대)
Hou, Yong Ku · Park, Seung Woo

Abstract

A multiplicative ARIMA model was tested and applied to analyze the periodicity and trends of 168 monthly oxygen demand data from the Noryanggin water quality gauging station in the downstream Han River. ARIMA model was identified to fit to the data using ACF and PACF tests, and the parameters estimated using an unconditional least square method. The residuals between the observed and forecasted data were acceptable with the Porte-Manteau test. A forecast of DO changes was made for its applications.

Ⅰ. 서 론

하천의 수질은 물의 물리적 거동으로 나타나는 유량의 변화와 밀접한 관계를 가지면서 시간에 따라 끊임없이 변화하고 있다. 수질은 수문량과 함께 하나의 추계학적 현상으로 볼 수 있는데, 수질은 수문량보다 인위적인 생활환경의 영향을 많이 받고 있기 때문에 수질의 변화특성은 보다 복잡한 형태를 띄고 있다. 이러한 수질자료에 대한 현재의 평가와 장래의 예측은 하천 수질 관리 대책의 수립에 있어 가장 기본적인 것으로써, 수치모형이 수질 및 환경관리를 위한 예측 수단으로 많이 사용되고 있다.

하천 수질 예측을 위한 수학적 모형에는 확정론적 모형과 통계적 모형이 있으며, 그 동안 국내의 수질예측 및 수질관리 계획에서는 수역에 유출입되는 오염물의 경계값으로부터 수질을 계산하는 QUAL2E, WASP5 등의 확정론적 방법이 많이 사용되어 왔다. 확정론적 모형은, 모형을 수립하는 과정에서 자료들이 가지는 추세 또는 변동의 원인을 알 수 있고, 자료의 변화 기작을 설명해주는 장점을 가지고 있다. 그러나 확정론적 모형은 물리적 변화에 따른 매개변수 추정이 어렵고 실제 유역에서 변수의 측정이 어려울 경우 적용에는 한계가 있다. 이에 반해 평활법, 분해법, ARIMA모형 등의 통계적 방법 또는 모형들은 자료로 쓰이는 시계열자료가 보여 주

2000년도 한국농공학회 학술발표회 논문집(2000년 10월 14일)

는 형태를 단순히 기술적으로 설명하려는 모형들로서, 추세나 계절적 변동성분의 이유를 설명해 주는 것이 아니라 단지 그러한 성분이 존재한다는 사실만을 확인해 준다. 통계적 모형은 필요로 하는 신뢰도를 만족하는 수의 자료만 확보된다면, 과거 자료의 통계학적 특성변수를 매개변수로 하여 통계적으로 정형화된 방법을 통해, 비교적 적은 노력으로 모형의 구축과 장래의 예측을 행할 수 있다. 그러나 시계열은 여러 가지 형태로 나타날 수 있으므로, 각 경우에 가장 적합한 이론적 모형을 개발한다는 것은 쉬운 일이 아니다. 그리고 우리 나라의 경우는 하천 및 호소의 수질 측정 자료의 부족으로 통계적 모형의 적용에 근본적인 문제점을 내포하고 있다. 이러한 통계적 모형은 수문 분야에서 복잡한 수자원 시스템 계획, 설계 및 운영조작을 위한 각종 해석의 입력자료 제공을 위해 모의발생기법으로서 사용되기도 하고 국내에서도 상당한 연구가 진행되었으나, 국내에서 수질 분야에의 활용은 자료의 부족 등의 이유로 아직 미비한 실정이다. 이홍균(1982)은 한강 하류부의 수질변동에 대한 추계학적 특성을 고찰했고, 류병로와 한양수(1998) 등은 ARIMA 모형이 월별 수질계열의 예측에 접합한지를 검토한바 있다. 국외에서는 F. Worral과 T.P. Burt(1999)는 영국에 있는 네 개의 하천의 수질에 대해 ARIMA모형을 구축하고 각 하천수질의 상관성에 대해 분석했다. 수질자료 이외의 시계열의 분석에서 ARIMA 모형의 적용은 광범위한 영역에서 활발히 행해지고 있다. 김광진, 이상훈, 정용(1988)은 ARIMA모형으로 서울시 일부지역의 SO_2 오염도의 월변화를 분석했고. 박무종과 윤용남(1989)은 다중 ARIMA모형에 의해 월유량을 예측한바 있다. 또한 심순보, 김만식, 한재석(1992)은 저수지 유입량 예측을, 이경훈, 문병석, 박성천(1997)은 ARIMA모형으로 상수도 일별 합수량 결정에 관해 연구한바 있다.

본 연구에서는 계절 ARIMA모형을 노량진지점의 월 평균 DO 자료에 적용하여 모형의 선정, 매개변수의 산정, 적합성 판정, 모의발생 및 예측 등에 관해 고찰하였다.

Ⅱ. 기본이론

1. Simple ARIMA 모형

일반적인 Simple ARIMA(p,d,q) 모형(Autoregressive Intergrated Moving Average Model)은 다음과 같이 정의된다.

$$Y_t = \sum_{j=0}^{p} \phi_j Y_{t-j} + \varepsilon_t - \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \quad \theta_0 = -1 \tag{1}$$

여기서 Y_t는 시계열 X_t의 d차 differencing에 의한 값이며, ϕ_j 및 θ_j는 각각 Autoregressive 및 Moving Average 계수이다.

2. Multiplicative ARIMA 모형

Multiplicative ARIMA 모형은 일반적으로 계차 d차로 표현되는 Simple ARIMA(p,d,q) 모형과 주기가 w인 계절적 계차를 이용한 Periodic ARIMA(P,D,Q)

모형의 결합으로 표현될 수 있다. 즉, 주기가 w인 D차 차분이 고려된 ARMA(P,Q) 모형(ARIMA(P,D,Q))의 forward form은 다음과 같다.

$$\Phi(B^w)\phi(B)(1-B^w)^D(1-B)^d X_t = \Theta(B^w)\theta(B)\varepsilon_t \tag{2}$$

여기서 B는 backward operator이다.

3. Multiplicative ARIMA 모형의 매개변수 산정

모형의 매개변수는 적률법(method of moment), 최우추정법(maximum likelihood estimation), 조건최소제곱추정법(conditional least square estimation), 비조건최소제곱추정법(unconditional least square estimation) 등이 있다. 여기서 비조건최소제곱추정법은 조건최소제곱추정과는 달리 초기조건을 주는 대신에 미지의 과거의 시계열값 X_*과 ε_*를 식(3)과 같은 backward form에 의해 추적하여 계산한 후, 비선형 반복법에 의해 최소제곱추정량 $S(\phi,\theta,\mu)=\sum_{t=-\infty}^{n}[\varepsilon_t]^2$ 을 최소로 하는 비교적 정확한 최소제곱추정치를 구하게 된다.

$$\Phi(F^w)\phi(F)(1-F^w)^D(1-F)^d X_t = \Theta(F^w)\theta(F)\alpha_t \tag{3}$$

여기서, F는 forward operator이다.

4. 모형의 적합성 판정

Multiplicative ARIMA 모형의 가정에 있어서 잔차항 ε_t는 정규분포를 가지며 서로 독립적인 무작위 계열이어야 하므로 모형의 매개변수가 일단 추정되고 나면 실측자료 각각에 해당하는 잔차를 계산하고, 이 잔차의 독립성과 정규성을 검사함으로써 모형의 적합성을 검정하게 된다. 독립성은 잔차 시계열의 자기상관 상관도를 구하여 Porte Manteau Test 등에 의한다.

$$Q = N\sum_{k=1}^{L}\gamma_k^2(\varepsilon_t) \tag{4}$$

이다. 여기서 $\gamma_t(\varepsilon_t)$는 추정된 매개변수를 이용하여 계산한 잔차계열 ε_t의 Lag k 자기상관계수이고, N은 자료의 수, L은 고려하는 Lag의 수로써 통상 N의 10~20%를 취한다. 식(4)로 계산한 Q값이 신뢰도 $(1-\alpha)$, 자유도 (L-p-q)인 Chi-Square 매개변수값 $\chi^2_{1-\alpha,1-p,q}$보다 작으면 잔차계열은 독립 시계열로 볼 수 있다. 그러나, Lung과 Box는 n이 작을 때 Porte Manteau Test 통계량 Q는 χ^2분포를 따르지 않음을 보였고, 이를 수정하여 식(5)와 같은 수정된 Box-Pierce통계량을 제시하였다.

$$Q^* = N(N+2)\sum_{k=1}^{L}\frac{\gamma_k^2(\varepsilon_t)}{n-k} \tag{5}$$

5. 예측 및 예측 오차

예측(forecasting)이란 과거 관측치의 거동을 조건으로하여 미래의 발생가능한 계열을 계산하는 것이다. 본 논문에서 채택된 ARIMA(1,0,0)×(1,1,1)$_{12}$ 의 예측에 대해 보면 다음 식과 같다.

식(2)에서 p=1, d=0, q=0, P=1, D=1, Q=1, w=12로 하여 X에 관해 전개하면,

$$X_t = \varepsilon_t + X_{t-12} + \phi_1 X_{t-1} - \phi_1 X_{t-13} + \Phi_1 X_{t-12} - \Phi_1 X_{t-24} - \phi_1 \Phi_1 X_{t-13} + \phi_1 \Phi_1 X_{t-25} - \Theta_1 \varepsilon_{t-12} \quad (6)$$

이고, 시각 t에서 Lead Time L (L =1, 2, ...)을 가지고 예측하면, L = 12일 때,

$$\hat{X}_t(12) = X_t + \phi_1 \hat{X}_t(11) - \phi_1 X_{t-1} + \Phi_1 X_t - \Phi_1 X_{t-12} - \phi_1 \Phi_1 X_t + \phi_1 \Phi_1 X_{t-13} - \Theta_1 \varepsilon_t \quad (7)$$

이 된다.

여기서 $\hat{X}_t(L)$은 예측시점 t에서의 Lead Time L 이후의 기대치(expectation)를 의미한다. 그리고, 시각 t+L에서의 자료는 시각 t에서의 L 선행예측치 $X_t(L)$과 예측오차 $\varepsilon_t(L)$의 합이라 할 수 있으므로 다음과 같다.

$$X_{t+L} = X_t(L) + \varepsilon_t(L) \quad (8)$$

또한, 예측 오차 및 예측오차를 계산하기 위한 가중치 Ψ_j는 다음 식(9), 식(10)으로부터 구할 수 있다.

$$\varepsilon_t(L) = \sum_{j=0}^{L-1} \Psi_j \varepsilon_{t+L-j} \quad (9)$$

$$\Phi(B^\omega)\phi(B)(1-B^\omega)(1-B)^d \Psi(B) = \Theta(B^\omega)\theta(B) \quad (10)$$

여기서, $\Psi(B) = \Psi_0 + \Psi_1 B + \Psi_2 B^2 + \Psi_3 B^3 + \ldots$, $\Psi_0 = 1$

그리고, 자료의 실시간 예측을 위한 식은 다음과 같이 표현된다.

$$X_{t+1}(L) = X_t(L) + \Psi_L [X_{t+1} - X_t(1)] \quad (11)$$

Ⅲ. 자료의 분석 및 결과

수질자료 예측을 위해 선택된 수질자료는 환경연감(환경부, 1985~1999)에 수록된 노량진 지점의 월 평균 DO 측정 자료이다. 시계열 분석에 있어서는 보다 많은 연속적인 자료를 확보하는 것이 모형에 의한 예측의 정확도를 높이는 방법이나, 우리나라의 수질측정의 역사가 길지 않고, 수록된 자료도 연속성이 결여된 시계열이 많은 것이 모형의 수립에 가장 큰 장애로 생각된다.

1. 수질 자료의 변동

노량진 지점의 DO 측정 자료의 분포형 시간에 따른 분산이 일정하고 정규분포를 띄고 있기 때문에 분산안정화 변환은 필요하지 않다.

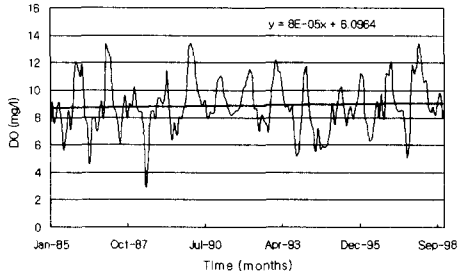

Fig. 3.1 Monthly mean DO

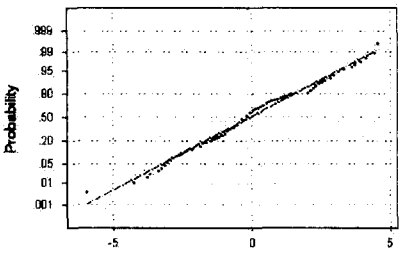

Fig. 3.2 Normality of DO

2. 모형의 차수 추정

노량진 지점에서의 DO 측정 자료는 추세를 무시할 수 있을 정도로 작기 때문에 비계절성 차분은 실시하지 않고, 추정될 모형에서 상수항을 제거하기 위해 평균을 일정하게 보아(μ=8.9) 원 자료에 평균을 빼준다. 즉, $X_t = Z_t - \mu$ (t=1, 2, 3, ...)이다. 차수를 선정하기 위해 원 자료 및 필요한 처리 후의 자료에 대한 표본자기상관함수(이하 SACF)를 나타내보면 다음 그림(Fig. 3.3~Fig. 3.6)과 같다.

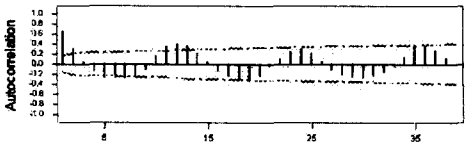

Fig. 3.3 ACF of DO

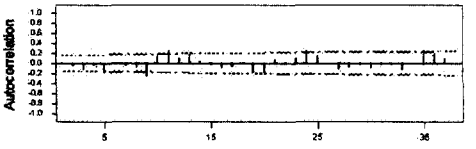

Fig. 3.4 ACF after 1th differencing for DO

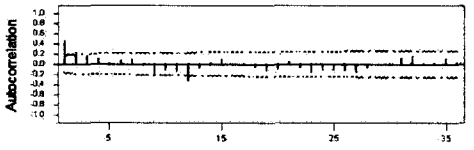

Fig. 3.5 ACF after 1th seasonal differencing of DO

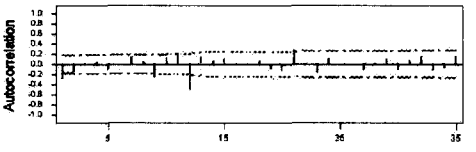

Fig. 3.6 ACF after 1th differencing and 1th seasonal differencing of DO

Fig. 3.3에서 보는 바와 같이, 원 자료의 SACF는 지수적으로 감소하면서 진동폭이 sine곡선이다. 이는 AR항의 특징이다. 1차 비계절성 차분 후 SACF(Fig. 3.4)는 주

기가 12인 계절성을 나타내고 있어, 계절성 차분이 필요함을 나타낸다. 1차 계절성 차분 후 SACF(Fig. 3.5)는 1차 비계절성 차분 후 1차 계절성 차분의 SACF(Fig. 3.6)에 비해 유의하지 않으므로, 비계절성 차수는 p=1, d=0, q=0로 한다. 그리고, 원 자료의 SACF는 주기 12를 중심으로 앞뒤의 11과 13에서 높은 상관관계를 보이므로 계절성 차수는 P=1, D=1, Q=1로 하여, ARIMA$(1,0,0)\times(1,1,1)_{12}$ 을 모형으로 선정한다.

3. 매개변수의 산정

미지의 과거의 시계열값 X_*과 ε_*를 추정하기 위해서 후방예측을 실시한다. 이를 위해서 식(3)의 backward form을 이용하며, ε_n에서 ε_1까지 역으로 계산한다. 그 후 X_*를 구하기 위해서 같은 식을 이용해서 X_0에서 X_Q까지 추정한다. 여기서 Q는 충분히 작은 수 η(<<1)에 대해 $|[X_i]-[X_{i+1}]|<\eta$ 일 때, Q = i (i < 0) 이다. X_*를 추정한 후에 마지막으로 ε_t를 계산하게 되는데, 이는 식(2)의 forward form을 이용하여, ε_Q 에서 ε_n 까지 계산한다. 이러한 과정을 통해 계산되는 비조건부최소제곱합 $S(\hat{\phi},\hat{\theta},\hat{\mu})=\sum_{t=-Q}^{n}[\varepsilon_t]^2$ 을 최소로 만드는 $\hat{\phi}$, $\hat{\theta}$, $\hat{\mu}$가 비조건최소제곱추정량이 된다. 이는 비선형반복법(nonlinear iterative method)에 의해 수행되는데, 이는 $\hat{\phi}_1$, $\hat{\Phi}_1$, $\hat{\theta}_1$의 초기치를 가정하고, $S(\hat{\phi},\hat{\theta},\hat{\mu})$를 계산한 다음, 보다 작은 제곱합을 주는 방향으로 매개변수의 초기값을 계속하여 갱신해 나가며 어떤 만족할 만한 수렴기준에 도달할 때 까지 반복적으로 매개변수를 추정해 나가는 방법이다.

이러한 방법을 통하여 추정된 매개변수 $\hat{\phi}_1$, $\hat{\Phi}_1$, $\hat{\theta}_1$는 각각 0.595, -0.245, 0.871이다.

4. 모형의 검정

ARIMA$(1,0,0)\times(1,1,1)_{12}$ 모형의 잔차 ε_t는 평균 0.0192, 표준편차 1.302로 나타났으며, 잔차의 정규성을 검토하기 위해 정규분포에 적합시켜보면, 모형의 잔차는 정규분포를 나타냄을 알 수 있다. 이로서 잔차 ε_t가 0에 가까운 평균치를 가지고 정규분포를 따름에 따라, 잔차의 정상성이 성립하여 잔차의 가정을 만족한다.

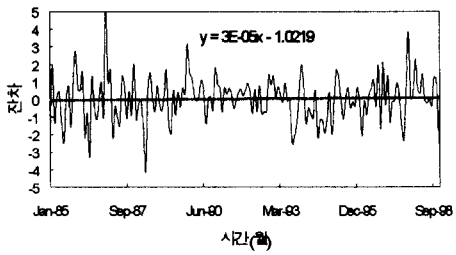

Fig. 3.7 Residual of ARIMA(1,0,0)×(1,1,1)$_{12}$ model applied to DO

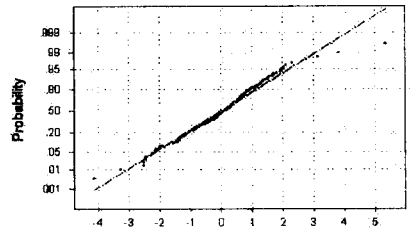

Fig. 3.8 Normality of the residual of ARIMA(1,0,0)×(1,1,1)$_{12}$ model applied to DO

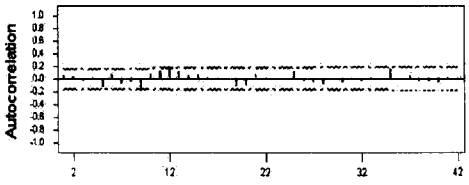

Fig. 3.9 SACF of the residual of ARIMA(1,0,0)×(1,1,1)$_{12}$ model applied to DO

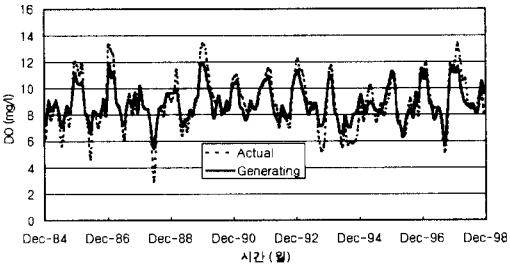

Fig. 3.10 Actual and generating DO from ARIMA(1,0,0)×(1,1,1)$_{12}$ model applied to DO

그리고 잔차의 독립성을 검사하기 위해 수정 Porte Manteau Test (Box-Pierce통계량)를 시행하면, 처음 36개 자료에 대해 Q=42.15 (자유도 33 = 36 - 1 - 1 - 1) 이고 5% 유의수준에서 $\chi^2_{1-\alpha,\nu} = \chi_{1-0.05,33} = 47.4$ 이므로, $Q < \chi_{1-0.05,34}$ 이 되어 잔차는 서로 독립적이라고 할 수 있고 잔차의 자기상관도도 난수계열임을 입증되었다.

5. 예측 및 예측오차

예측은 과거의 자료를 이용하므로 forward form을 이용하여 식(6)과 같이 표현한 뒤 ARIMA(1,0,0)×(1,1,1)$_{12}$에 대하여 기점 t = 168부터 예측을 실시하면 다음과 같다. (Fig. 3.11)

$$\widehat{X_{168}}(1) = X_{157} + \phi_1 X_{168} - \phi_1 X_{156} + \Phi_1 X_{157} - \Phi_1 X_{145}\phi_1\Phi_1 X_{156} + \phi_1\Phi_1 X_{144} - \Theta_1\varepsilon_{157} = 0.727$$

$$\widehat{X_{168}}(12) = X_{168} + \phi_1 X_{179} - \phi_1 X_{167} + \Phi_1 X_{168} - \Phi_1 X_{156}\phi_1\Phi_1 X_{167} + \phi_1\Phi_1 X_{155} - \Theta_1\varepsilon_{168} = 1.164$$

ARIMA$(1,0,0)\times(1,1,1)_{12}$ 모형을 이용한 DO 예측의 오차는 식(9)에 의해서 계산되고, 가중치 Ψ는 다음 표와 같다.

Table 3.1 Weights used in calculating confidence limits and updating forecasts.

J	0	1	2	3	4	5
Ψ(J)	1.0000	0.5950	0.3540	0.2106	0.1253	0.0746

J	6	7	8	9	10	11
Ψ(J)	0.0444	0.0264	0.0157	0.0093	0.0056	0.0033

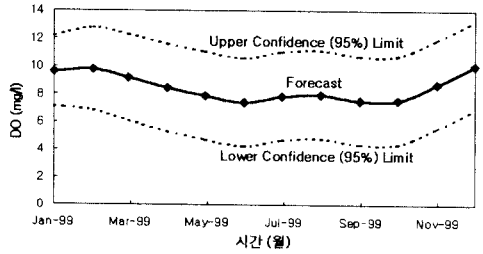

Fig. 3.11 Forecasting and 95% confidence limit of DO from ARIMA$(1,0,0)\times(1,1,1)_{12}$ model applied to DO

Ⅳ. 결 론

본 연구에서는 노량진 지점의 14년간의 DO 측정 자료를 Multiplicative ARIMA 모형에 적용하기 위한 모형의 선정, 매개변수 추정, 모형의 검증 및 예측을 행하였으며 결론은 다음과 같다.

1. ARIMA 모형의 매개변수 및 차수 결정은 계열 자기상관함수와 부분 계열 자기상관함수를 이용하여 행하였고, 잔차와 잔차의 계열 자기상관함수를 이용, 수정 Porte Manteau Test를 실시하여 잔차의 독립성을 검증하였다.

2. ARIMA 모형의 매개변수 추정은 비조건최소제곱법(unconditional least square estimation)에 의한 잔차의 제곱합을 최소로 하는 매개변수를 찾기 위해 비선형추적법(nonlinear iterative estimation)을 이용하여 행하였고, 그 결과 추정치 $\hat{\phi}_1$, $\hat{\Phi}_1$, $\hat{\theta}_1$는 각각 0.595, -0.245, 0.871 였다.

3. 계산된 매개변수의 ARIMA$(1,0,0)\times(1,1,1)_{12}$ 모형을 이용하여 DO자료를 예측하였으나, 그 결과는 자료의 부재로 비교되지 못하였고, 더불어 예측수정에 의한 실시간 예측도 불가능하였다.

참 고 문 헌

1. 박무종, 윤용남, 1989, Multiplicative ARIMA 모형에 의한 월유량의 추계학적 모형 예측, 한국수자원학회지, 제22권, pp.331-338
2. 이경훈, 분병석, 박성천, 1997, ARIMA 모델에 의한 상수도 일별 합수량 결정에 관한 연구, 한국수자원학회지, 제30권, pp.45-54
3. 류병로, 한양수, 1998, ARIMA 모형에 의한 하천수질 예측, 한국환경학회지, 제7권, pp.433-440
4. 민병준, 1995, 돼지의 사육농가호수와 두수의 변동 예측, 한국동물과학회지, 제37권, pp.558-566
5. 진영민, 박승우, 강문성, 1996, 추계학적 기상모의발생을 위한 매개변수 추정, 한국농공학회 학술발표회 논문집, pp.67-72
6. K. J. Hollenbeck and K. H. Jensen, 1998, Maximum-likelihood estimation of unsaturated hydraulic parameters, Journal of Hydrology, 210, pp.192-205
7. F. Worrall and T. P. Burt, 1999, A univariate model of river water nitrate time series, Journal of Hydrology, 214, pp.74-90
8. 김원경, 1999, 시계열 분석의 이해, 교우사, pp.1-209
9. 조신섭, 손영숙, 1999, 시계열 분석, 율곡출판사, pp.11-343
10. George E. P. Box, Gwilym, M, Jenkins and Gregory C. Reinsel, 1994, Time Series Analysis, Prentice Hall, New Jersey, pp.181-366
11. A. Ian McLeod and Keith W. Hipel, 1994, Time Series Modeling of Water Resources and Environmental Systems, ELSEVIER, London, pp.171-255
12. E. Lukacs and Z. W. Birnbaum, 1996, Spectral Analysis and Time Series, ACADEMIC PRESS, San Diego, pp.816-877
13. J. Keith Ord and Maurice Kendall, 1993, Time Series, Edward Arnold, New York, pp.27-120