ISSN 1226-9999 (print)
ISSN 2287-7851 (online)

Korean J. Environ. Biol. 34(4) : 365~373 (2016) https://doi.org/10.11626/KJEB.2016.34.4.365

〈Review paper〉

환경생태 자료 분석을 위한 시계열 분석 방법 연구

모형호 · 조기종[1] · 신기일[2,*]

고려대학교 생명자원연구소, [1]고려대학교 환경생태공학부, [2]한국외국어대학교 통계학과

A Review of Time Series Analysis for Environmental and Ecological Data

Hyoung-ho Mo, Kijong Cho[1] and Key-Il Shin[2,*]

Institute of Life Science and Natural Resources, Korea University, Seoul 02841, Republic of Korea
[1]Division of Environmental Science and Ecological Engineering, Korea University, Seoul 02841, Republic of Korea
[2]Department of Statistics, Hankuk University of Foreign Studies, Yongin 17035, Republic of Korea

Abstract - Much of the data used in the analysis of environmental ecological data is being obtained over time. If the number of time points is small, the data will not be given enough information, so repeated measurements or multiple survey points data should be used to perform a comprehensive analysis. The method used for that case is longitudinal data analysis or mixed model analysis. However, if the amount of information is sufficient due to the large number of time points, repetitive data are not needed and these data are analyzed using time series analysis technique. In particular, with a large number of data points in the current situation, when we want to predict how each variable affects each other, or what trends will be expected in the future, we should analyze the data using time series analysis techniques. In this study, we introduce univariate time series analysis, intervention time series model, transfer function model, and multivariate time series model and review research papers studied in Korea. We also introduce an error correction model, which can be used to analyze environmental ecological data.

Key words : ARIMA, intervention model, multivariate time series, error correction

서 론

시계열 분석은 시간의 흐름에 따라 얻어진 시계열 자료를 분석하는 방법이다. 시간에 영향을 받는 자료를 분석하는 기법은 다양하지만 충분히 많은 시점에서 자료가 얻어졌다면 시계열 분석 기법을 사용해야 한다. 자료가 시간의 흐름에 따라 얻어지기 때문에 각 자료는 종속적인 관계를 갖게 되며 이 종속적인 관계를 분석에 사용함으로써 자료 분석의 정확성 향상에 많은 도움을 준다. 또한 현재 자료를 분석함으로써 미래에 발생되는 시계열 자료 값을 예측할 수 있다. 그러나 종속적인 관계 규명은 분석을 복잡하게 하고, 결과 해석을 어렵게 하는 요인이 되기도 한다. Boero *et al.* (2015)은 생태학 등에서 시계열 자료 분석의 필요성을 설명하고 있다. 이미 국내외에서 시계열 자료 분석이 수행되고 있으나 다른 분야, 특히 경영, 경제 분야에 비해 시계열 모형을 이용한 환경생태 자료 분석은 국내를 포함하여 국외에서도 미미한 상태이다. 본 논문에서는 환경생태 자료 분석에 도움을 줄 수 있는 다양한 시계열 분석 기법을 설명하였다.

먼저 가장 간단하면서도 중요한 분석 기법은 단변량 시계

* Corresponding author: Key-Il Shin, Tel. 031-330-4718, Fax. 031-330-4719, E-mail. keyshin@hufs.ac.kr

열 분석 기법이다. 이 분석 기법 중에서 중요하게 사용하는 방법이 Box-Jenkins 분석 방법이다. 이 방법은 ARIMA(p,d,q) 모형을 기반으로 분석한다. 주어진 자료에 가장 타당한 ARIMA(p,d,q) 모형을 식별하고, 즉 차수 p, d, q를 결정하고, 모형에 포함된 모수를 추정하며 이를 기반으로 예측을 수행한다. 그러나 변수가 하나이기 때문에 얻어진 모형을 설명하는 것이 큰 의미가 없을 수 있으며 예측 결과가 매우 단순할 수 있다. 그러나 단변량 ARIMA 모형 분석은 향후 복잡한 모형의 기본적인 아이디어와 이론을 제공하기 때문에 매우 중요하다. 또한 단변량 시계열 분석에서 쉽게 사용할 수 있는 분석 모형이 자기회귀오차 모형(autoregressive error model)이다. 이 방법은 일반적으로 사용하는 선형회귀모형의 오차가 시계열적으로 독립이 아니라는 특징을 이용하여 분석한다. 만약 오차가 독립이라고 판단되면 이 모형은 흔히 사용하는 선형회귀모형과 동일해진다. 물론 자기회귀오차 모형은 다음에 설명할 전이함수 모형(transfer function model)의 특별한 경우이다.

단변량 시계열에서 특정 시점에 개입(intervention)이 일어나 시계열에 왜곡이 발생할 수 있다. 이렇게 특정 시점에 특정 사건이 일어난 이유가 알려진 경우의 시계열 분석은 개입 모형을 이용하여 분석하게 된다. 개입 분석은 특정 사건이 시계열에 얼마나 영향을 주는지 파악하고, 향후 비슷한 사건이 다시 일어날 경우 분석 결과를 이용하여 예측에 이를 반영하게 된다. 이를 위해 특정 시점을 설명할 수 있는 시간 변수를 생성한 후 이 변수를 독립변수로 사용하여 분석한다. 생성되는 시간 변수의 형태는 개입이 발생한 한 시점만 영향을 줄 경우에 사용하는 pulse 변수, 그리고 개입이 발생한 시점 이후 계속적으로 영향을 줄 경우에 사용하는 step 변수를 사용할 수 있다.

만약 독립변수가 개입 변수가 아니라 시계열 변수인 경우에는 개입 모형을 사용하지 않고, 개입 모형의 확장 모형인 전이함수 모형(transfer function model)을 사용하여 분석한다. 예를 들면 바람의 세기가 풍력 발전의 발전량에 영향을 주는 상황을 살펴보자. 이 경우 바람의 세기는 시간에 따라 변하게 되므로 시계열 자료이며 이 자료는 풍력 발전량의 시계열 분석에 독립변수로 사용될 수 있다. 그러나 풍력 발전에서 얻어진 발전량은 바람의 세기에 영향을 주지 않는다. 따라서 시계열 자료인 발전량을 종속변수로, 역시 시계열 자료인 바람의 세기를 독립변수로 하는 모형을 만들 수 있다. 이러한 모형을 전이함수 모형이라 하며 이 모형에서 독립변수로 사용하는 시계열 자료를 입력 계열(input series), 종속변수로 사용하는 시계열 자료를 출력 계열(output series)이라 한다. 전이함수 모형은 입력 계열이 얼마나 출력 계열에 영향을 주는지 모형을 만들며, 입력 계열의 예측값이 얻어지게 되면 이 값을 이용하여 출력 계열값을 예측할 수 있게 된다.

최종적으로 두 시계열 자료가 서로 영향을 주게 되어 각각이 종속변수로 그리고 독립변수로 사용될 수 있다. 예를 들면 predator-prey 모형의 경우이다. 시점별로 포식자 수는 피식자 수에 영향을 주고 또한 피식자 수는 포식자 수에 영향을 준다. 이와 같이 두 시계열 자료가 서로 영향을 줄 경우에는 다변량 시계열 모형(multivariate time series model)을 이용하여 분석할 수 있다. 다변량 시계열 모형은 단변량 시계열 모형 분석 방법과 매우 유사하지만 모형에 미지의 모수가 매우 많아지게 되기 때문에 시계열 분석의 세부 분석 방법인 모형 식별, 모수 추정 및 모형 검진이 매우 복잡하게 된다. 이로 인해 다변량 시계열 모형의 특별한 경우인 다변량 자기회귀 모형(MAR: multivariate autoregressive model)이 흔히 사용된다. 2007년 San Jose에서 열린 MAR Work ESA 2007을 생각한다면 향후 MAR 모형을 사용한 자료 분석은 매우 중요하게 사용될 것으로 판단된다. 또한 MAR 모형 중에서 향후 사용 가능성이 매우 높은 오차수정 모형(error correction model)이 있다. 이 분석 방법은 기존의 다변량 ARIMA 모형을 확장한 개념이다.

본 논문에서는 2절에 시계열 분석 방법의 전반적인 이론이 설명되었다. 먼저 단변량 시계열 분석을 위해 사용되는 ARIMA 모형과 자기회귀오차 모형이 설명되었으며 단변량 모형의 확장인 개입 모형과 전이함수 모형이 설명되었다. 또한 다변량 시계열 모형 중에서 MAR 모형이 설명되었고 향후 매우 중요하게 사용될 가능성이 높은 오차수정 모형을 설명하였다. 3절에서는 발표된 논문을 중심으로 각 분석 방법을 설명하였다. 4절에 토론 및 결론이 있다.

본 론

1. 시계열 분석 방법 개요

1) 단변량 시계열(univariate time series model: ARIMA model)

시계열 자료는 시간의 흐름에 따라 얻어진 자료이다. 여기서 시간은 일정 간격을 두고 자료가 얻어져야 한다. 즉 매일 자료가 얻어진다면 매일 같은 시각에 자료가 얻어져야 하며 월별 자료의 경우에는 같은 일자에 자료가 얻어져야 한다. 이는 같은 시간 간격을 유지하게 됨으로써 효과적인 분석이 가능하기 때문이다. 다음으로 시계열 자료는 크게 정상 시계열(stationary time series)과 비정상 시계열(nonstationary time series)로 나누어진다. 정상 시계열은 각 시점별로 평균

과 분산이 일정하며 공분산이 시점에 무관하고 시점 간의 차이인 시차(lag)에만 관계가 있는 시계열이다. 반면 정상 시계열을 만족하지 못하는 시계열을 비정상 시계열이라 한다. 모든 비정상 시계열은 변환(transformation)과 차분(difference)을 이용하여 정상 시계열로 전환될 수 있다. Box-Jenkins 분석 방법은 정상 시계열만이 분석 가능하며 정상 시계열인 ARMA(p,q) 모형을 이용하여 분석하게 된다. 또한 단변량 시계열 분석에서 회귀분석과 시계열 모형이 결합된 자기회귀오차 모형(autoregressive model)을 이용하여 자료를 분석할 수 있으며 이 경우 특히 장기 추세를 살펴볼 수 있다. 이 방법은 흔히 사용하는 회귀 모형에서 오차가 시계열적인 관계가 있을 때 사용하는 방법이다. 환경/생태 자료에서 장기 추세 유무를 결정하는 데 유효하게 사용될 수 있다.

(1) Box-Jenkins 분석 3단계 방법

ARIMA(p,d,q) 모형을 이용한 Box-Jenkins의 분석 방법은 모형 식별, 모수 추정, 모형 검진의 세 단계로 나누어진다. 그러나 이 분석 단계는 정상 시계열 자료일 때 분석하는 단계이므로 자료 분석 전에 사전 분석 단계를 거쳐야 한다. 즉 분산이 일정하지 않은 비정상 시계열인 경우에는 변환을 이용해야 한다. 변환의 경우 흔히 사용하는 방법이 Box-Cox (1964) 변환이다. 만약 자료에 음수가 있다면 양수 c를 더하여 모든 자료를 양수로 만든 후 사용한다. 공식은 다음과 같다.

$$Z=(X+c)^{\lambda},\ \lambda>0$$

$$Z=\log(X+c),\ \lambda=0$$

사용되는 모수 값 λ 중에서 주로 사용하는 변환은 $\lambda=0.5$인 제곱근 변환과 log 변환이다. 변환된 자료를 이용하여 실제 분석이 이루어지며 분석에서 얻어진 예측은 재변환(back transformation)을 이용하여 원자료와 같은 scale의 예측값을 얻게 된다. 또한 확률 보행 과정(random walk process)이 시계열에 있다고 판단되는 경우는 차분(difference)을 통하여 정상 시계열로 바꿀 수 있다. 차분이 필요하다고 의심되는 경우에는 단위근 검정(unit root test)을 실시하여 차분을 결정하게 된다. 이상 변환과 차분을 이용하게 되면 비정상 시계열은 정상 시계열로 바뀌게 된다. 이렇게 정상 시계열로 바뀐 자료를 Box-Jenkins 방법을 이용하여 분석하게 된다.

(2) Box-Jenkins 분석에서의 모형 식별

시계열 분석에서는 다양한 모형이 사용된다. 이 중에서 가장 흔하면서도 중요하게 사용되는 모형이 ARIMA(p,d,q) (autoregressive integrated moving average) 모형이다. 이 모형은 비정상 시계열을 고려하는 모형으로 자료를 d번 차분하게 되면 정상 시계열인 ARMA(p,q) 모형이 된다. 즉 ARMA (p,q) 모형은 정상 시계열 자료에 사용하는 모형이다. 이 모형은 자료들 간의 선형 결합으로 이루어진 AR(p) (p-th order autoregressive) 모형과 백색잡음으로 불리는 오차의 선형 결합으로 이루어진 MA(q) (q-th order moving average) 모형이 결합된 모형이다. ARMA(p,q) 모형의 정의는 다음과 같다.

$$\begin{aligned}&(1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p)(Z_t-\mu)\\&=(1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q)a_t\end{aligned}\qquad(1.1)$$

여기서 B는 후향연산자(Backward shift operator)로 $BX_t=X_{t-1}$을 의미하고 AR(p) 모형의 특성 방정식 $\phi(B)=(1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p)$와 MA(q) 모형의 특성 방정식 $\theta(B)=(1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q)$는 공통근이 존재하지 않으며 백색잡음 과정 a_t는 평균이 "0"이고 분산이 σ_a^2을 따르는 오차이다. 또한 $E(Z_t)=\mu$이다. 모형 식별 과정은 모형(1.1)에서 차수 p와 차수 q를 결정하는 과정이다. 이때 사용하는 통계량은 대표적으로 자기상관함수(ACF: autoregressive correlation function)와 부분(편)자기상관함수(PACF: partial autoregressive correlation function)이다. 이 통계량은 대표적으로 사용되는 통계 패키지인 SAS에서 자동으로 출력되므로 이 결과를 이용하여 차수를 결정하면 된다.

(3) Box-Jenkins 분석에서의 모수 추정

모수 추정은 모형(1.1)에 포함된 모수를 추정하는 과정이다. 즉 모형에 포함된 $\phi_i\theta_j$와 평균 μ 그리고 오차의 분산인 σ_a^2을 추정한다. 다양한 방법으로 모수를 추정할 수 있으며 가장 흔하게 사용하는 방법이 조건부 최소제곱추정법(conditional least squares estimation)과 최대우도추정법(maximum likelihood estimation)이다. 시계열에서의 모수 추정법은 매우 복잡하여 이론적으로 파악하기 어려우나 대부분의 통계 패키지를 이용하면 쉽게 모수 추정이 가능하다.

(4) Box-Jenkins 분석에서의 모형 검진

최종적으로 모형이 식별되고, 모수 추정이 완성되면 모형 검진을 실시한다. 모형 검진은 주로 잔차의 독립성 검정에 초점이 맞추어지고 있다. 잔차의 ACF와 잔차의 PACF를 통하여 모형 검진이 수행된다. 그러나 잔차의 ACF와 PACF는 모든 잔차의 독립성을 동시에 검정하지는 못하기 때문에 시계열 분석에서 Ljung-Box 통계량 또는 퍼트멘토 검정 통계량이라 부르는 다음의 통계량을 이용하여 독립성을 검정한다.

$$Q=n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{(n-k)}\qquad(1.2)$$

이 통계량은 자유도 h를 갖는 카이제곱 분포를 따르게 된

다. 여기서 $\hat{\rho}_k$은 표본에서 구해진 ACF이다. 만약 통계량 Q가 0.05보다 크다면 잔차가 독립이라는 가설을 기각할 수 없게 되고 이를 기초로 분석은 일단락되게 된다.

(5) Box-Jenkins 분석에서의 예측(forecast)

시계열 분석의 중요한 목적 중의 하나는 얻어진 모형식을 이용하여 미래를 예측하는 것이다. 즉 현재 n 시점까지 n개의 자료가 얻어졌을 때 $n+l$ 시점의 자료 값을 예측하는 것이다. 이때 사용하는 예측 방법은 최소평균제곱오차예측(minimum mean square error forecast) 방법이다. 이 방법은 다른 통계 분야 분석에서 사용하는 BLUP (best linear unbiased prediction) 방법과 동일한 방법이며 공식은 다음과 같다.

$$\hat{Z}_n(l)=E(Z_{n+l}|Z_n,Z_{n-1},\cdots,Z_1) \tag{1.3}$$

즉 과거의 자료 $Z_1,Z_2,\cdots,Z_n$이 주어진 경우, Z_{n+l} 시점의 기댓값을 예측값으로 사용한다.

(6) 자기회귀오차 모형(autoregressive error model)

자기회귀오차 모형은 다양한 독립변수를 이용하여 분석할 수 있다. 그러나 회귀 모형과 달리 오차는 서로 독립이 아니고 종속적인 관계가 있으므로 이 종속적인 관계를 모형에 추가함으로써 정확한 모수 추정이 가능하다. 특히 독립변수에 시간 변수를 추가하게 되면 시간이 흐르면서 다양한 형태의 추세를 확인할 수 있다. 흔히 사용하는 모형은 다음과 같다.

$$\begin{aligned} &Z_t=\beta_0+\beta_1X_{1t}+\beta_2X_{2t}+\cdots\beta_kX_{kt}+\epsilon_t,\ t=1,\cdots,n\\ &\epsilon_t=\phi_1\epsilon_{t-1}+\cdots+\phi_p\epsilon_t+a_t \end{aligned} \tag{1.4}$$

이 분석은 SAS 등의 통계 패키지를 사용하면 매우 쉽게 분석을 수행할 수 있다.

2) 개입 모형(intervention model)

시간에 따라 얻어진 시계열 자료는 특정 시점 T에서 발생한 특정 사건으로 인해 시계열이 크게 영향을 받을 수 있다. 먼저 국내로 들어오는 외국인 여행자 수는 시계열 분석에서 대표적으로 사용되는 자료이다. 그러나 2015년 5월 메르스(중동호흡기증후군: MERS) 사태가 국내에서 발생하였을 때 국내로 들어오는 외국인 여행자 수는 급격히 감소하였다. 이후 메르스 사태가 끝난 후 점차적으로 국내에 들어오는 외국인 수는 정상 수준으로 되돌아가게 되었다. 이렇게 관심 시계열 자료에 특정 시점에서 구체적인 사건이 발생하게 되고 이로 인해 관심 시계열에 큰 변화가 발생한다면 이 사건의 효과를 시계열 분석에 추가하여 분석해야 한다. 만약 특정 사건이 구체적으로 알려져 있지 않다면 이상점으로 파악될 수 있으나 특정 사건이 구체적으로 알려지고, 이 사건으로 시계열 자료에 변화가 발생한다면 이 사건의 영향력을 분석할 수 있다. 이를 위해 특정 사건, 즉 시계열 자료의 급격한 변화에 영향을 주는 사건을 변수로 만들며 이 변수를 개입 변수 $I_t^{(T)}$라 한다. 이 경우는 T 시점에서 사건이 발생하였을 때 사용하는 표시 방법으로 흔히 사용하는 개입 변수의 형태는 step 함수와 pulse 함수로 먼저 step 함수는 $S_t^{(T)}=\begin{cases}0, & t<T\\ 1, & t\geq T\end{cases}$ 이고 pulse 함수는 $P_t^{(T)}=\begin{cases}1, & t=T\\ 0, & t\neq T\end{cases}$이다. 따라서 step 함수는 T 시점 이후 계속 영향을 주는 반면 pulse 함수는 T 시점에만 영향을 준다. 물론 기본적인 개입 변수는 이 두 형태이지만 다양한 방법을 이용하여 시계열에 다양한 형태의 영향력을 줄 수 있다. k개의 개입 변수 $I_{jt}^{(T_k)}$를 독립변수로 하는 개입 모형의 형태는 다음과 같다.

$$Z_t=\sum_{j=1}^{k}\frac{\omega_j^{s(j)}(B)B^{b_j}}{\delta_j^{r(j)}(B)}I_{jt}+\frac{\theta(B)}{\phi(B)}a_t \tag{1.5}$$

개입 모형 분석은 개입 변수 I_t의 특성 방정식에 해당되는 $\omega_j^{s(j)}(B)=\omega_{0(j)}-\omega_{1(j)}B-\cdots-\omega_{s(j)}B^{s(j)}$와 $\delta_j^{(r)}(B)=1-\delta_{1(j)}B-\cdots-\delta_{r(j)}B^{r(j)}$의 차수를 결정하고 결정된 차수별로 부여된 모수를 추정하는 것이다. 또한 단변량 시계열 모형에서 설명한 것처럼 AR 모형의 특성방정식과 MA 모형의 특성방정식인 $\theta(B)$와 $\phi(B)$의 차수와 모수를 추정하는 것이다. 이에 추가하여 b_j는 지체모수(delay effect parameter)로 일정 시간 이후에 지체하여 영향을 주는 효과를 파악하기 위한 모수이다. 즉 개입이 일어났지만 시차를 두고 그 영향력이 발생하는 경우에 사용하는 모수이다. 이후의 분석 방법은 단변량 시계열과 동일하다.

3) 전이함수 모형(transfer function model)

개입 모형에서 사용한 개입 변수는 '0'과 '1'의 값만 갖는 자료이다. 그러나 지속적으로 변하는 시계열 자료가 관심 시계열 자료인 분석 변수에 영향을 줄 수 있다. 지속적으로 영향을 주는 시계열 자료를 입력 계열(input series)이라 하며 관심 변수인 분석 변수를 출력 계열(output series)이라고 한다. 전이함수 모형은 개입 모형을 확장한 모형이라 생각하면 된다. 모형을 간단히 표시하기 위하여 입력 계열이 하나인 모형은 다음과 같다.

$$Z_t=\frac{\omega_s(B)B^b}{\delta_r(B)}X_t+\frac{\theta(B)}{\phi(B)}a_t \tag{1.6}$$

모형(1.5)와 (1.6)을 비교하면 (1.6)은 하나의 입력 계열이 있는 경우이며, 개입 변수 I_t 대신 입력 계열 X_t가 사용되었다는 차이가 있다. 물론 여러 개의 입력 계열이 있다면 이를 모형에 추가하면 된다.

분석은 개입 모형 분석과 같이 입력 계열에 주어진 $\omega_s(B)$와 $\delta_r(B)$인 특성 방정식의 모수를 추정하는 것이며 AR 부분과 MA 부분의 특성 방정식인 $\theta(B)$와 $\phi(B)$에 포함된 모수를 추정하는 것이다. 그러나 개입 분석과 달리 본 분석의 모형 식별을 위해서는 사전백색화작업(pre-whitening)이 필요하며 이 결과를 기초로 얻어지는 교차상관함수(CCF: cross correlation function)를 이용하여 $\omega_s(B)$와 $\delta_r(B)$의 차수를 구하게 된다. 또한 향후 출력 계열 Z_t의 예측을 위해서는 입력 계열 X_t의 예측값이 필요하기 때문에 전이함수 모형 분석을 위해서는 입력 계열 X_t의 모형 분석 및 예측이 선행되어야 한다. 이후 특성 방정식인 $\theta(B)$와 $\phi(B)$의 모수 추정 방법과 최종 모형 검진 방법은 단변량 시계열 분석과 같은 방법을 사용하면 된다.

4) 다변량 시계열 모형(multivariate time series model)

다변량 시계열 모형 분석은 단변량 시계열 분석 방법과 매우 유사하게 분석이 수행된다. 다만 자료가 m-차원의 다변량 자료로 각각의 시계열이 서로 영향을 주고받게 되기 때문에 모형 분석은 매우 복잡하게 이루어진다. 흔히 사용하는 다변량 시계열 모형 분석법은 Box-Jenkins가 제안한 방법으로 모든 시계열 자료를 정상 시계열로 바꾼 후 분석하는 방법과 상대적으로 최근에 개발된 오차수정 모형은 비정상 시계열을 직접 사용하여 분석하는 방법이 있다.

(1) MAR 모형(multivariate autoregressive model)

다변량 시계열 모형도 자기회귀 모형(AR: autoregressive model)과 이동 평균 모형(MA: moving average model)이 모두 포함된 MARIMA (multivariate autoregressive integrated moving average) 모형이 정립되어 있다. 물론 실제 분석에서는 차분을 통하여 모든 시계열 자료가 정상 시계열이 되도록 만든다. 또한 현실적으로는 MA 부분이 없는 MAR 모형이 주로 사용된다. 일부 다른 학문 분야에서는 특히 경영, 경제 분야에서는 다변량을 의미하는 MAR 대신 VAR (vector autoregressive model)라 부른다. 다변량 시계열에는 많은 수의 모수가 포함되어 있어 시계열의 길이가 길지 않으면 다변량 시계열 모형 분석이 불가능하게 된다. 흔히 사용하는 MAR(p) 모형은 다음과 같다.

$$Z_t-\mu=\Phi_1(Z_{t-1}-\mu)+\cdots+\Phi_p(Z_{t-p}-\mu)+a_t \qquad (1.7)$$

여기서 $Z_t=(Z_{1t},Z_{2t},\cdots,Z_{mt})'$, $\mu=E(Z_t)=E(Z_{1t},Z_{2t},\cdots,Z_{mt})'$이고, $E(a_t)=0_m$이며 $Var(a_t)=\Sigma=\begin{pmatrix}\sigma_1^2, & \sigma_{12}, & \cdots,\sigma_{1m}\\ \sigma_{21}, & \sigma_2^2, & \cdots,\sigma_{2m}\\ & \vdots & \\ \sigma_{m1}, & \sigma_{m2}, & \cdots,\sigma_m^2\end{pmatrix}$, m은 시계열 변수 수 또는 차원이다. 또한 Φ_k는 $m\times m$행렬로 $\Phi_k=\begin{pmatrix}\phi_{k,11}, & \phi_{k,12}, & \cdots, & \phi_{k,1m}\\ \phi_{k,21}, & \phi_{k,22}, & \cdots, & \phi_{k,2m}\\ & & \vdots, & \\ \phi_{k,m1}, & \phi_{k,m2}, & \cdots, & \phi_{k,mm}\end{pmatrix}$이다.

본 모형의 모형 식별, 즉 차수 p를 결정하는 방법은 AIC (Akaike information criterion), BIC (Bayesian information criterion) 등 다양한 통계량을 이용하여 수행된다. 그러나 모형에는 다수의 모수가 포함되어 있어 추정에 어려움이 있을 수 있다. 따라서 가장 흔하게 사용되는 모형은 MAR(1) 모형이다.

(2) 오차수정 모형(vector error correction model)

오차수정 모형은 비정상 시계열을 정상으로 만든 후 분석하지 않고 직접 시계열 간의 관계를 찾아 분석한다. 전체적인 분석 방법은 MAR 모형 분석과 유사하지만 오차수정 모형 분석에서는 공적분 검정(co-integration test)을 실시한 후 그 결과를 이용하여 분석한다. 만약 비정상 시계열 간의 선형결합이 정상 시계열을 만족한다면 이 결과를 이용하여 시계열의 장기평형(long run relation)관계를 찾아낼 수 있다. 다음이 흔히 분석에 사용되는 오차수정 모형이다.

$$\Delta Z_t=\Pi Z_{t-1}+\sum_{i=1}^{p-1}\Phi_i^*\Delta Z_{t-1}+a_t \qquad (1.8)$$

여기서 $Z_t=(Z_{1t},Z_{2t},\cdots,Z_{mt})'$이고, $\Delta Z_t=(Z_{1t}-Z_{1t-1}, Z_{2t}-Z_{2t-1},\cdots,Z_{mt}-Z_{mt-1})'$이다. 즉 $\Delta=(1-B)$이고 $\Delta Z_t=(1-B)Z_t$는 차분을 의미한다. $\Pi=\Phi_p+\Phi_{p-1}+\cdots+\Phi_1-I$로 ΠZ_{t-1}를 흔히 오차 수정항이라고 부른다. 또한, $\Phi_j^*=-\sum_{i=j+1}^{p}\Phi_i$, $j=1,\cdots,p-1$이다. 위의 모형을 이용하게 되면 오차 수정항에 의해 장기평형관계를 파악할 수 있게 된다. 이 분석도 SAS 등 통계 패키지를 이용하면 분석이 가능하다.

2. 환경생태 자료의 시계열 분석

2절에 시계열 분석에서 사용할 수 있는 다양한 모형을 설명하였다. 그러나 국내외를 통틀어 환경생태 분야에서 시계열 분석 기법을 이용한 분석은 많지 않다. 이 절에서는 간단히 환경생태 분야에서 시계열 분석 기법이 사용된 논문을 살펴보았다.

1) Box-Jenkins 분석 방법을 이용한 단변량 시계열 분석

국외에서 연구된 논문으로 Ives *et al.* (2010)은 생태 자료 분석을 위해 ARMA(p,q) 모형을 적합하고 분석하는 방법을 자세히 설명하였다. 국내에서 연구된 단변량 시계열 분석 연구는 Lee *et al.* (2007), Kim and Ahn (2010), Choi *et al.* (2013)

이 있다. 다른 분석 방법에 비해 단변량 시계열 분석을 이용한 다수의 연구가 수행되었다. 이 논문 중에서 Choi *et al.* (2013)의 논문은 시계열 모형을 이용하여 한강 하류(행주)의 단기 수질 변동을 예측하였다. 본 논문에서 사용한 자료는 2002년 1월부터 2013년 12월까지의 월별 자료이며 수소이온농도(pH), 용존산소(DO), 총질소(TN), 총인(TP), 전기전도도(EC), 총유기탄소(TOC)가 분석에 사용되었다. 이 논문은 ARIMA(p,d,q) 모형을 이용하여 모형을 식별하고, 모수를 추정한 후, 모형 검진을 실시하여 최종 모형을 만들었다. 다른 변수에 비해 총인(Tp) 분석 결과가 자세히 설명되어 있다. 구체적으로 이 논문에서는 TP 자료를 이용하여 얻어진 모수 추정값과 표준오차, p-값 등이 수록되었으며 최종 모형은 $ARIMA(1,0,0)(0,1,1)_{12}$로 나타났다. 또한 그림을 이용하여 예측값과 실제값을 비교하였다. 그러나 그림을 살펴보면 분산이 시간의 흐름에 따라 분산이 점점 줄어드는 것을 파악할 수 있으며 이는 변환이 필요하다는 판단을 내릴 수 있는 근거가 된다. 따라서 향후 분석에서는 변환한 자료를 이용하여 분석한 후 이를 재변환하는 방법을 사용하는 것이 타당하다고 판단된다. 변환은 이미 설명한 것처럼 Box-Cox 변환을 사용하면 되고, 분석이 완성되면 다시 원 scale로 바꾸기 위해 재변환을 실시해야 한다. 최근 Kang and Shin (2015)은 재변환을 수행할 때 발생할 수 있는 문제를 연구하였다. 특히 로그 변환을 실시할 경우 재변환을 실시할 때 로그 함수의 역함수인 지수함수(exponential function)를 사용하여 최종적인 예측을 수행할 수 있다. 즉 변환된 자료를 이용하여 예측을 한 후 얻어진 예측값 z_{t+k}에 지수를 취하여, 즉 $\exp(z_{t+k})$를 적용하여 최종 예측값을 얻을 수 있다. 반면 지수함수 대신에 $\exp\left(z_{t+k}+\frac{1}{2}\sigma^2\right)$을 사용할 수도 있다. 자료에 따라 다른 결과가 나올 수도 있지만 일반적인 경우에는 $\exp\left(z_{t+k}+\frac{1}{2}\sigma^2\right)$을 사용할 경우 MSE (mean squared error)가 더 우수해진다는 것을 확인할 수 있다. Jang *et al.* (2007)은 지하수 수질 관리를 위한 연구에서 AR(1) 모형을 이용하여 분석을 실시하였으나 이 자료 분석에서도 ARMA(p,q) 모형 분석 전에 변환을 먼저 실시하고 얻어진 변환된 자료를 이용하여 분석을 하였다면 더 정확한 결과를 얻을 수 있을 것으로 판단된다.

2) 자기회귀오차 모형 분석

자기회귀오차 모형은 독립변수에 시간 변수를 사용하여 추세를 분석하기에 적합한 모형이다. Lee (2011)는 미세먼지인 PM10 (Particulate Matter with a diameter less than 10 μm)을 종속변수로, 독립변수로는 일 최고온도(maximum temperature), 풍속(wind speed), 상대습도(relative humidity), 강수량(rainfall), 일사량(radiation), 운량(amount of cloud)을 사용했다. 또한 Kwon *et al.* (2013)은 한산거제만 해역의 수질 장기변동 특성을 다양한 통계 분석을 실시하면서 시계열 분석도 실시하였다. Kwon *et al.* (2013)의 논문에서는 장기자료에 대한 시간적 변동 추세를 파악하기 위해 4-이동 평균 자료를 구하고 이 자료를 이용하여 회귀분석을 실시하였다. 물론 이 방법에서 얻어진 추세 유무 및 추세의 크기는 자체로 의미가 있을 수 있으나 m-이동 평균 방법을 사용하지 않고, 원 자료에 직접 자기회귀오차 모형을 이용하여 분석을 실시하였다면 더욱 정확한 분석 결과가 얻어질 수 있을 것으로 판단된다. Kim *et al.* (2012)은 호소 수질 자료의 장기변동 추세 분석을 위하여 시계열 분석법을 이용한다고 하였으나 논문에서는 구체적인 모형을 제시하지 않았고, 단지 그림과 단순 통계량을 이용한 분석을 실시하였다. 따라서 이 경우에도 자기회귀오차 모형을 이용하여 분석한다면 통계적으로도 타당한 결론을 얻을 수 있을 것으로 판단된다. 또한 독립변수에 다양한 형태의 시간 함수를 적용할 수 있다. 즉 직선 추세, 2차 추세 등과 같은 추세함수, 계절적 요인을 분석할 수 있는 변수 등을 추가할 수 있다. 이러한 방법을 분해시계열 분석이라 하며 이를 위해 자기회귀오차 모형을 이용할 수 있다. 분해시계열 분석과 관련된 내용은 Worrall *et al.* (2003)을 살펴보기 바란다.

3) 개입 모형 분석

개입 모형은 시간이 흐르면서 얻어진 시계열 자료에 특정 시점에서 알려진 사건이 일어남으로써 시계열이 달리지게 된 경우에 분석하는 방법이다. 단변량 시계열 분석, 특히 ARIMA 모형을 이용한 시계열 분석에서는 다수의 논문이 출간 되었음에도 국내외에서 개입 모형을 이용하여 시계열 분석을 실시한 논문은 찾기가 어렵다. Kim (2006)은 안양천의 COD 자료를 ARIMA 모형을 이용하여 분석하였다. 분석 과정에서 두 개의 개입이 있다고 판단되어 개입 모형 분석도 실시하였다. 그러나 본 분석은 개입에 관련된 특정 사건이 어떤 것인지 사건 내용 및 이유를 밝히지 않아 통계학적으로 개입 모형을 이용한 분석 결과이지 환경생태학적인 이유를 파악하기 위해 사용했다고 판단하기는 어렵다.

4) 전이함수 모형 분석

전이함수 모형은 개입 모형을 확장한 모형이다. 국내에서는 전이함수 논문과 관련하여 발표된 논문은 거의 없는 것으로 판단된다. Bakker *et al.* (2014)은 날씨 자료를 입력 계열로 water demand를 출력 계열로 하여 전이함수 모형을 이용하여 분석하였다. 이 연구에서는 구체적인 모형 식별과정과 추정값 등은 발표되지 않았으나 입력 계열을 사용하는 전이함수 모형 분석이 매우 우수한 결과를 주고 있음을 밝

혔다. Choi and Moon (2009)는 계절 식물의 개화 시기 변화를 연구하면서 지수 평활법(exponential smoothing)을 이용한 시계열 분석을 실시하였다. 이 논문에서는 기온과 개화시기와의 관련성을 분석하기 위해 상관분석을 이용하여 두 변수의 관계를 분석하였다. 만약 기온을 입력 계열로 개화시기를 출력 계열로 하여 전이함수 모형을 이용한 분석을 실시하였다면 통계적으로 매우 타당한 결과가 얻어질 수 있을 것으로 판단된다.

5) 다변량 시계열 모형 분석

다변량 시계열 분석은 다수의 모수가 모형에 포함되기 때문에 매우 긴 기간 동안 자료가 모아져야 하며 모형 식별 및 모수 추정이 복잡하다는 단점이 있다. 따라서 매우 많은 수의 모수가 포함된 MAR(p) 모형보다는 간단하지만 효과적인 MAR(1) 모형이 주로 사용된다. Hall *et al.* (2009)에서는 여러 통계적 기법을 이용하여 분석하면서 시계열 변수가 3개인 3차원 MAR(p) 모형을 이용한 분석도 실시하였다. 사용된 시계열 변수는 uninfected host class의 밀도, infected host class 그리고 diluting Daphnia competitor 밀도이다. 그러나 조사 시점의 길이에 관한 정보가 없고, 모형 식별, 즉 차수 p를 결정하는 과정이나 추정량 및 표준오차를 정확히 수록하지 않아 분석의 신뢰성이 떨어진다.

한편 Knape *et al.* (2013)은 야생에서 휘파람새의 모니터링 기록을 토대로 시계열에 따른 개체군 풍부도 추정을 위해 상태-공간 모델(state-space model)을 이용하여 분석하였다. 상태-공간 모델은 최근에는 잘 사용되지 않는데, 이는 상태-공간 분석이 다변량 시계열 분석의 하나이기 때문이다. 물론 과거에는 다변량 시계열 분석이 거의 불가능했기 때문에 비교적 분석이 용이한 상태-공간 모델을 사용하였으나, 지금은 다변량 시계열 분석을 사용한다. 특히 SAS에서 PROC VARMAX를 이용하여 분석이 가능하다.

반면 Keightley *et al.* (2011)은 LAMBDA 소프트웨어를 사용하여 4차원의 MAR(1) 모형을 분석하였다. 분석에 사용된 자료는 pacific ocean perch (*Sebastes alutus*), Canary rockfish (*S. pinniger*), Sablefish (*Anoplopoma fimbria*), Pacific hake (*Merluccius productus*)이다. 또한 4차원 시계열에 영향을 줄 수 있는 5개의 공변량도 분석에 사용하였다. 흔히 경제 경영 시계열 분석에서는 다변량 분석에서 가장 일반적인 모형을 VARMAX 모형이라 부른다. 먼저 다변량을 뜻하는 multivariate 대신에 vector를 사용하고, AR 부분과 MA 부분이 있으며 전이함수 모형처럼 입력 계열이 있을 때 기호 X를 사용하게 된다. 이 논문에서는 각 행렬의 추정 계수와 추정량의 신뢰구간, 모형 식별 방법이 설명되어 있으며 공변량의 추정 계수와 신뢰구간도 구해져 있어 통계적으로는 매우 타당한 분석 방법을 보여주고 있다. 여기서 LAMBDA는 open source Matlab이고 2007년 MAR workshop ESA에서도 LAMBDA가 사용되었다. 물론 통계 패키지인 SAS에서는 PROC VARMAX를 이용하여 이 모형을 분석할 수 있다.

6) 장기간 식생변화 탐지를 위한 위성영상 활용 사례

생태계에서 식생은 육상생태계 내 다른 생물들의 서식처로서 중요한 기반이 되며, 우점하는 식생은 그 지역의 비생물적 환경과 상호작용하면서 생태계 내 다른 개체군의 자원과 토대에 큰 영향을 미친다(Turner *et al.* 2001). 따라서 기후변화에 대비한 생태계 취약성 평가를 위해 육상생태계의 식생 상태와 변화에 대한 연구가 필요한데, 이와 관련하여 원격탐사 기반의 생태계 모니터링 기법의 활용이 증가하고 있는 추세이다. 국립생태원(NIE 2015)에서는 '기후변화에 따른 취약생태계 적응전략 수립' 연구를 위해 MODIS (Moderate Resolution Imaging Sepectroradiometer) 위성 센서가 취득한 15년간(2000년~2014년)의 영상을 분석하여 NDVI (Normalized Difference Vegetation Index) 추세를 분석하였다. 그러나 본격적인 시계열 분석을 이용하지는 않았으며, 단순히 15년 동안의 NDVI의 선형회귀 결과만을 도출함으로써 미래 상황을 예측하지는 않았다.

결론 및 제언

본 연구에서는 조사 시점의 길이가 충분히 긴 시계열 자료를 분석하는 방법과 환경생태 자료에서 어떻게 시계열 자료 분석이 수행되었는지 살펴보았다. 연구 결과 시계열 분석을 이용한 논문이 많지 않은 것을 알 수 있었다. 또한 정확히 시계열 분석 방법을 파악하고 분석하지 않았던 연구 논문도 있었다. 이는 경제 경영분야처럼 긴 시계열 자료를 얻는 것이 어려워 분석할 자료가 많지 않았기 때문으로 판단된다. 또한 단변량 시계열 분석 즉 ARIMA 모형과 자기회귀오차 모형을 이용한 분석이 많은 것에 비해 전이함수 모형, 다변량자기회귀 모형을 이용한 분석은 상대적으로 미미하였다. 그러나 향후 지구 온난화 등 기후 변화로 많은 환경 및 생태가 변하기 때문에 전이함수 모형을 이용한 시계열 분석은 반드시 필요할 것으로 판단된다. 즉 기후 자료를 입력 계열로 그리고 환경, 생태 자료가 출력 계열이 되는 자료의 분석이 반드시 필요하다고 판단된다. 또한 단변량 시계열 분석의 경우 SAS/PROC ARIMA, 자기회귀오차 모형은 SAS/PROC AUTOREG, 전이함수 모형, MAR 모형은 SAS/PROC VARMAX 등으로 어렵지 않게 분석을 수행할 수 있다. 다만 다변량 분석의 경우 모형에 포함된 모수의 수가 많기 때문

에 차원의 크기가 크지 않으면서 MAR(1) 또는 MAR(2) 모형 등 차수가 크지 않는 모형을 사용한다면 우수한 분석 결과를 얻을 수 있을 것으로 판단된다. 또한 향후 중요하게 사용될 수 있는 오차수정 모형도 PROC VARMAX를 이용하여 분석할 수 있다.

적 요

환경생태 자료 분석에 사용된 많은 자료가 시간에 따라 얻어지고 있다. 조사된 시점의 수가 적은 경우에는 자료가 충분한 정보를 주지 않기 때문에 반복 측정하거나 여러 지점을 조사하여 종합적인 분석을 수행하게 된다. 이때 사용하는 방법이 경시적 자료 분석(longitudinal data analysis) 또는 혼합모형(mixed model) 분석이다. 그러나 시점의 수가 많아 정보의 양이 충분하다면 반복적인 자료가 필요하지 않으며 이러한 자료는 시계열 분석 기법을 이용하여 분석하게 된다. 특히 현재와 같이 다수의 시점에서 얻어진 자료의 수가 많아지고 있는 상황에서 각 변수 간에 서로 어떤 영향을 주는지 또는 향후 어떤 경향을 띠게 되는지 예측을 원한다면 시계열 분석 기법을 사용하여 자료를 분석해야 한다. 본 연구에서는 단변량 시계열 분석(univariate time series analysis), 개입 분석(intervention time series model), 전이함수 모형 분석(transfer function model), 다변량 시계열 분석(multivariate time series model) 기법을 소개하고 현재까지 진행된 국내외 연구 논문을 살펴보았다. 또한 향후 환경생태 자료 분석에서 중요하게 사용될 수 있는 오차수정 모형(error correction model)을 소개하였다.

사 사

본 연구는 환경부 “기후변화대응환경기술개발사업(2014001310008)”의 지원으로 수행되었습니다.

REFERENCES

Bakker M, H van Duist, K van Schagen. J Vreeburg and L Rietveld. 2014. Improving the performance of water demand forecasting models by using weather input. Procedia Eng. 70:93-102.

Boero F, AC Kraberg, G Krause and KH Wiltshire. 2015. Time is an affliction: Why ecology cannot be as predictive as physics and why it needs time series. J. Sea Res. 101:12-18.

Box GEP and DR Cox. 1964. An analysis of transformation. J. R. Stat. Soc. Series B 26:211-252.

Choi CM and SG Moon. 2009. Changes of Flowering Time in the Weather Flora in Busan Using the Time Series Analysis. J. Environ. Sci. 18:369-374.

Choi NG, HC Lee, GC Lim, JA Park, IS Choi, KS Lee, GS Kim, KH Kim, CH Park, TH Lee, KS Bae and JS Jeon. 2013. Forecasting Short-Term Water Quality Variations of Lower Han River (Haeng-ju) by Using Time Series Models. Report of S.I.H.E. 49:214-230.

Hall SR, CR Becker, JL Simonis, MA Duffy, AJ Tessier and CE Caceres. 2009. Friendly competition: evidence for a dilution effect among competitors in a planktonic host-parasite system. Ecology 90:791-801.

Ives AR, KC Abbott and NL Ziebarth. 2010. Analysis of ecological time series with ARMA(p,q) models. Ecology 91: 858-878.

Jang EA, JS Woo, YK Sung, SK Kim, IS Seo, JS Kim and GH Kim. 2007. A Study on Groundwater Quality Management Using Time Series Analysis. Report of G.I.H.E. 20:163-173.

Kang JH and KI Shin. 2015. Re-transformation of power transformation for ARMA(p,q) model-simulation study. Korean J. Appl. Statist. 28:511-527.

Keighley SJ, AM Edwards and CA Holt. 2011. Potential for using multivariate autoregressive models to investigate dynamics of british columbia groundfish communities, including appraisal of the LAMBDA software package. Canadian Technical Report of Fisheries and Aquatic Sciences Fs97-6/2986E.

Kim BH. 2006. A study on water pollution by time series analysis. Dissertation of Masteral Degree, Keimyung University.

Kim ES, JH Yoon, JW Lee and HI Choi. 2012. Analysis of Long-Term Trends in Lake Water Quality Observations. J. Korean Soc. Hazard Mitig. 12:231-238.

Kim KS and JH Ahn. 2010. A Study on the Real Time Forecasting for Monthly Inflow of Daecheong Dam using Seasonal ARIMA Model. Proceedings of the Korea Water Resources Association 1395-1399.

Knape J, P Besbeas and P de Valpine. 2013. Using uncertainty estimates in analyses of population time series. Ecology 94:2097-2107.

Kwon JN, YC Park and KH Eom. 2013. The Characteristic of Long Term Variation of the Water Quality from Hansan-Geoje bay, Korea. J. Korean Soc. Mar. Environ. Energy. 16:189-201.

Lee H. 2011. Analysis of PM10 Concentration using Auto-Regressive Error Model at Pyeongtaek City in Korea. Asian J.

Atmos. Environ. 27:358-366.

Lee J, J Lee, Y Lee, R Kim and J Han. 2007. A Time Series Analysis for O_3 Concentration in Seoul Using ARIMA Model. Proceedings of the Asian Journal of Atmospheric Environment 1363-1364.

NIE. 2015. Vulnerable ecosystems strategy for adapting to climate change. National Institute of Ecology.

Turner MG, RH Gardner and RV O'neill. 2001. Landscape ecology in theory and practice (Vol. 401). Springer, New York.

Worral F, WT Swank and TP Burt. 2003. Changes in stream nitrate concentrations due to land management practices, ecological succession, and climate: Developing a systems approach to integrated catchment response. Water Resour. Res. 39:1177-1190.

Received: 21 December 2016
Revised: 26 December 2016
Revision accepted: 26 December 2016