서울特別市 保健環境硏究院報 第49號 (2013)
Report of S.I.H.E., 49:214~230 (2013)

시계열모형을 이용한 한강 하류(행주)의 단기 수질변동 예측

수환경생태팀

최낙경 · 이호찬 · 임귀철 · 박진아 · 최인석 · 이광식
김갑수 · 김경희 · 박창호 · 이태호 · 배경석 · 전재식

Forecasting Short-Term Water Quality Variations of Lower Han River(Haeng-ju) by Using Time Series Models

Aquatic Ecology Team

Nak-gyong Choi, Ho-chan Lee, Gyu-chul Lim, Jin-a Park, In-seok Choi, Kwang-sig Lee, Gab-soo Kim, Kyung-hee Kim, Chang-ho Park, Tae-ho Lee, Kyung-seok Bae and Jae-sik Jeon

Abstract

Using a number of variables, including water temperature, pH, DO, TN, TP, TOC, and EC, collected between 2002~2013 at Haeng-ju station on the lower Han river, a time series model with the best fit was obtained utilizing an autoregressive integrated moving average(ARIMA) model and an exponential smoothing method. As a result, the models with the best fit derived for each variable measured were ARIMA(1, 0, 0)(2, 1, 0)$_{12}$ for DO, ARIMA(0, 0, 2)(0, 1, 1)$_{12}$ for pH, ARIMA(1, 0, 0)(0, 1, 1)$_{12}$ for TN, ARIMA(1, 0, 0)(0, 1, 1)$_{12}$ for TP after applying a logarithmic transformation, Winters' Multiplicative for TOC, and ARIMA(1, 0, 0)(0, 1, 1)$_{12}$ for EC. In order to examine the model's forecasting ability, the researchers eliminated the actual 2013 data, and computed the predicted value for 2013, based on each of the established models. The resulting MAPE values were: 12.36 for DO, 2.39 for pH, 22.58 for TN, 12.86 for TP, 18.21 for TOC, and 6.73 for EC. As can be seen from the data, the model was highly accurate at predicting pH and EC, it accurately predicted DO, TP, and TOC values, and, although the forecasting ability was relatively low for the TN value, the estimated value was acceptable. The monthly values for 2014 were forecasted using the model derived including the data in 2013.

Key words : ARIMA model, Exponential smoothing method, Time series analysis

서 론

서울시는 1975년 노량진에 수질측정소를 처음 설치한 이래 여러 지점에 측정소를 설치하여 운영하고 있다. 1997년부터 상수원수의 수질감시를 위해 잠실수중보 상류는 상수도사업본부에서 운영하고 있으며 생태계 보호 등 수질관리를 위해 잠실수중보 하류는 보건환경연구원에서 관리 및 운영을 하고 있다. 잠실수중보 하류의 지점은 한강 본류 3곳(행주, 노량진, 선유)과 지천 3곳(안양천, 중랑천, 탄천)으로 총 6개 측정소로 운영되고 있다. 잠실수중보 하류의 측정자료는 서울시 리스크 관리 시스템에 실시간 측정자료로 제공되며, 2014년부터 서울시 열린데이터 광장(data.seoul.go.kr)을 통해 공공 데이터를 시민에게 제공할 계획이다. 보건환경연구원에서는 수질자동측정망 상시 상황관제를 운영하고 있으며, '수질오염감시경보를 위한 측정소별 측정항목과 항목별 경보기준'(1)에 따라 수질오염 감시기준을 설정하고 있다. 수질오염 감시경보는 총 4단계(관심, 주의, 경계, 심각)로 구분하며 수소이온농도(pH), 용존산소(DO), 총질소(TN), 총인(TP), 전기전도도(EC), 총유기탄소(TOC), 페놀, 시안, 수은 항목과 생물감시 측정값을 바탕으로 각각의 기준을 설정하고 있다. 측정항목별 감시기준은 측정 수계의 수질환경기준, 배출허용등급, 각 측정소에서 6개월~1년간 실측한 항목별 측정값 중 유효측정시간의 평균, 최고, 최저값 등을 고려하여 설정한다(2).

각 측정소의 측정값은 일정한 시간 간격에서 추출한 일종의 시계열 자료이며, 측정값을 생성하는 프로세스를 파악하여 미래의 측정값을 정확하게 예측할 수 있다. 하지만 시간의 흐름에 따라 발생하는 다양한 변수의 모든 값 뿐 아니라 그 정확한 수학적 표현은 알 수 없기 때문에 측정값을 생성하는 과정을 적절히 모사하는 수학적 표현인 시계열 모형을 찾아 기존 데이터를 적합 및 예측할 수 있다. 동일한 시간 간격으로 측정된 과거 값들이 존재할 경우, 시간의 흐름에 나타난 시계열의 패턴을 파악하여 그 패턴이 미래에도 계속 적용되는 가정에서 예측하는 방법을 시계열분석이라 한다. 시계열분석은 여러 가지 형태의 방법이 존재한다. 예측할 변수 하나를 선정한 후 해당변수의 과거

Fig. 1. The location of Haeng-ju station.

자료를 근거로 해당변수의 미래값을 예측하는 방법을 일변량 시계열분석 방법이라 한다(3). 본 연구에서는 일변량 시계열분석에 자주 사용하는 두 개의 모형인 ARIMA모형과 지수평활법을 사용하여 수소이온농도(pH), 용존산소(DO), 총질소(TN), 총인(TP), 전기전도도(EC), 총유기탄소(TOC) 등의 항목별 최적모형을 구하고, 앞으로 전개될 2014년 월별 측정값을 예측하여 수질측정망 상시 상황관제 운영 및 감시기준 설정의 참고자료로 활용하고자 한다.

재료 및 방법

1. 분석 자료의 범위와 자료수집

분석 및 예측에는 월별 결측치가 가장 적은 행주측정소 자료가 사용되었으며 사용된 자료는 2002년 1월부터 2013년 12월까지의 월별 평균 측정치이다. 다음 표 1은 수소이온농도(pH), 용존산소(DO), 총질소(TN), 총인(TP), 전기전도도(EC), 총유기탄소(TOC)의 기술통계량이며, 대부분 불검출로 자료해석이 어려운 페놀, 시안, 수은 항목과 생물감시항목은 제외하였다.

2. 자료처리방법

2002년부터 2013년까지의 월별 자료들로 ARIMA 모형을 만드는데 사용하였고, 2013년 자료를 제외한 2002년부터 2012년 자료로 다시 모형을 만들어 추정된 계수의 안정성을 평가하였다. 또한 대체모형인 지수평활법과의 모형 적합도 및 예측력 비교를 위해 2002년부터 2012년 자료들로 만들어진 각각의 모형으로 기존의 2013년 자료를 예측해 보았다. 모형의 적합도는 BIC, MAE, RMSE값 등 모형적합통계량을 사용하여 비교하였고, 예측력 비교를 위한 통계량으로 MAPE값을 이용하였다. 대체모형과의 비교를 통해 최종적으로 모형을 선택하고 2002년부터 2013년까지의 월별자료를 이용하여 2014년 월별 측정값을 예측하였다. 통계분석은 SPSS 21 통계프로그램을 사용하였다.

결과 및 고찰

1. ARIMA 모형

일변량 ARIMA(autoregressive integrated moving average) 모형은 특히 단기예측에 적합하다. 왜냐하면 ARIMA 모형은 먼 과거보다는 최근 시점에 가까운 과거 관측값에 더 많은 비중을 주기 때문이다. 또한 ARIMA 모형을 설정하기 위해서는 적절한 표본크기가 필요하다. Box와 Jenkins(4)는 최소 50개 이상의 관측값이 필요하다고 제안했다. 물론 계절적 변동이 존재하는 자료는 특히 표본의 수가 더 많아야 한다. ARIMA 모형 구축을 위해 Box와 Jenkins는 정상성-식별-추정-모형진단-예측 절차를 제안하였다. 여기서 모형진단 후 모형이 만족스럽지 않으면 식별단계로 다시 수행하게 된다(3).

Table 1. Descriptive statistics of each water quality parameter at Haeng-ju of lower Han river

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
DO(℃)	144	3.0	14.9	8.30	2.73	0.51	-0.59
pH	144	6.3	8.6	7.22	0.407	0.74	0.76
TN(mg/L)	144	3.027	17.551	6.721	2.494	1.16	2.28
TP(mg/L)	144	0.129	0.699	0.354	0.116	0.68	-0.07
TOC(mg/L)	108	2.00	17.58	4.89	3.06	2.08	4.07
EC(μS/cm)	144	95	622	286	88.3	0.71	0.72

1) 평균과 분산의 정상성

ARIMA 모형을 구축하기 위해서는 우선 주어진 시계열 자료에 대한 정상성을 만족시켜 주어야 한다. 여기서 정상성이란 시계열을 일정한 주기로 나누었을 때, 각 주기에 해당하는 평균과 분산이 일정하다는 의미이다. 만일 시계열의 분산과 평균이 비정상적일 경우, 정상성을 만족시키기 위해서 각각 변수변환 및 차분을 취해 준다(3). 평균과 분산을 시각적으로 확인하기 위해 그림 2에 각 항목들의 평균선을 참조선으로 하여 시도표를 나타내 보았다.

여러 항목 중 TP항목을 예로 살펴보면 분산이

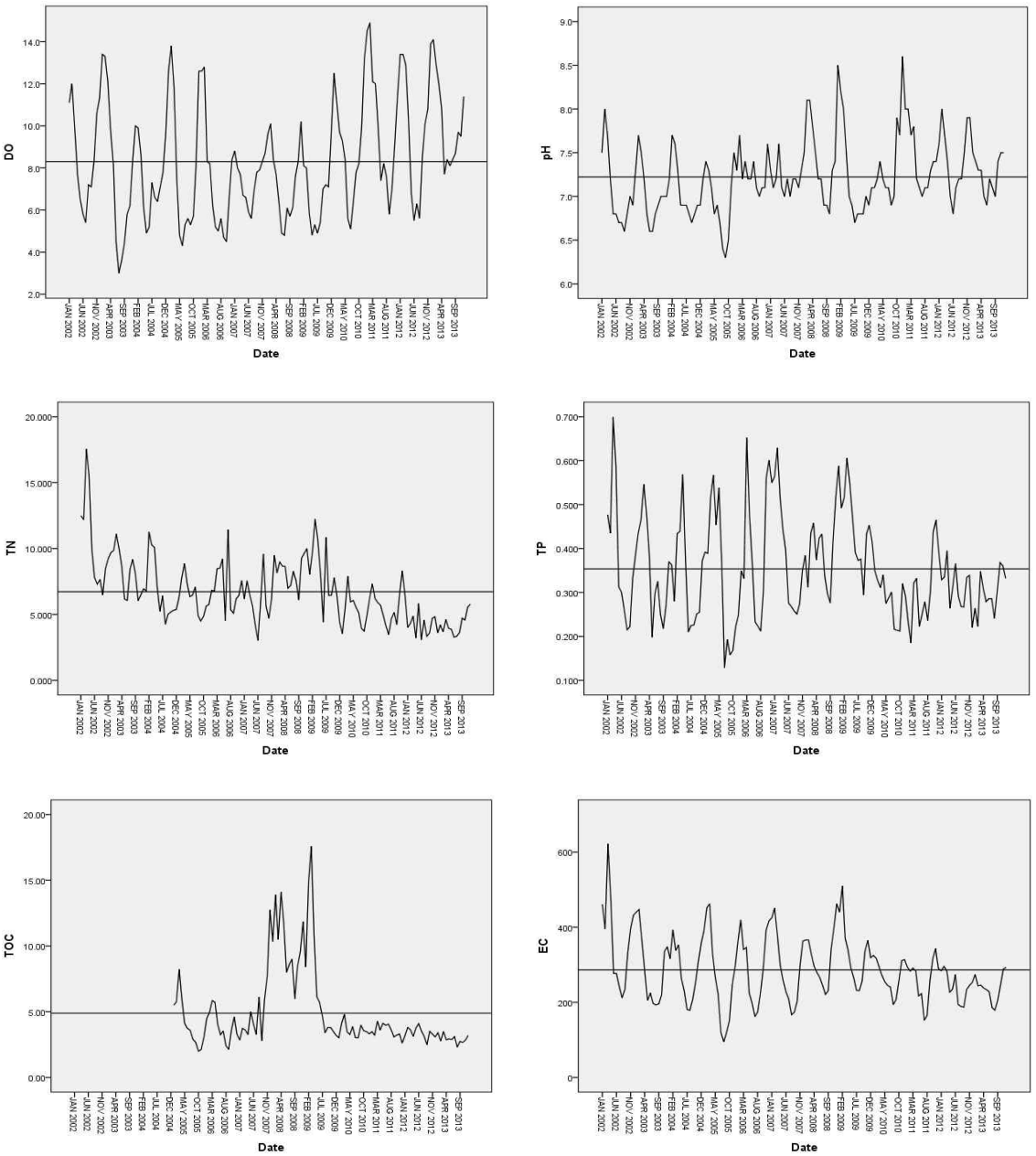

Fig. 2. Sequence plots of each water quality parameter at Haeng-ju of lower Han river.

시간에 따라 변하고 있다. 또한 시도표에 참조선으로 표현한 평균선에서 등락을 반복하는 것으로 보아 평균은 비교적 일정한 것을 알 수 있었다. 정상성을 만족시키기 위한 분산 조정을 위해 로그변환을 수행하였다. 때때로 자연로그 변환만으로 분산을 정상적으로 만들 수 없는 경우가 있지만 여기서는 SPSS 21의 자동 변환을 이용하기 위해 자연로그 변환을 수행하여 분산을 조정하였다. 이후 변환 된 TP의 추정된 자기상관함수와 편자기상관함수를 36시차까지 그림 3과 표 2에 나타내보았다. Box와 Jenkins는 추정된 자기상관계수를 사용하기 위해 필요한 최대개수는 관측값 수의 약 25%를 제시하였다(4). TOC(n=108)는 두 번째 계절시차인 24시차, 그 외의 항목(n=144)은 3번째 계절시차인 36시차를 설정하여 계산하였다.

평균의 정상성을 위해 그림 3과 표 2를 살펴보면 TP의 계절시차 중 시차 12와 24, 36의 자기상관계수가 서서히 감소하므로 계절적 차분이 필요하다. 비계절적 시차의 경우 추정된 첫 5개 내지 7개 시차까지의 추정된 자기상관계수들이 0을 향해 서서히 감소하는 패턴이면(즉, t-검증통계량(자기상관/표준오차)의 절대값이 1.6보다 크면)시계열의 평균이 비정상적이며 차분을 통해서 평균을 정상적으로 만든다(3). TP의 경우 자기상관함수가 시차 2 이후로 t-통계량 값이 1.6 이하로 비계절 차분 불필요하였다. 다음 표 3과 그림 3은 계절적 1차 차분을 행한 후 추정한 자기상관함수와 편자기상관함수이다.

2) 식별

식별 단계에서는 시계열 내 관측값들 사이에 존재하는 상관관계를 측정하여 ARIMA(p,d,q)모형을 구성하는 자기회귀요소(autoregressive, AR) 요소인 p와 이동평균(Moving Average, MA)요소인 q를 임시적으로 결정한다. 여기에서 자기상관함수(Autocorrelation function, ACF)와 편자기상관함수(Partial Autocorrelation function, PACF)를 이용하여 상관관계를 측정한다. 추정된 자기상관함수에 대한 t-통계량(자기상관/표준오차)의 절대값의 실전적 경고수준은 다음과 같다. 첫째 식별단계에서 t-통계량의 절대값이 '1.6'을 초과하는 비계절적 자기상관계수에 주목한다. 이들 시차의 계수들은 추정단계에 거의 통계적으로 유의하다. 둘째 식별단계와 모형검진단계에서 t-통계량의 절대값이 '1.25'를 초과하는 계절적 자기상관계수에 주목한다. 만일 계절시차(s, 2s, …)의 잔차 자기상관함수가 통계적으로 '0'이면 계절적 주기의 반절인 시차(0.5s, 1.5s, …)와 계절시차들에 근접한 시차(s+1, s-1, 2s+1, 2s-1, …)의 잔차 자기상관함수에 대한 t-통계량의 절대값이 '1.25'를 초과하는지를 주목하여 적용한다. 셋째 모형검진단계에서 단기시차(1, 2, 3 등)의 잔차 자기상관계수에 대한 t-통계량의 절대값이 '1.25'를 초과하면, 이 단기시차에 대응하는 계수를 추정한다. 넷째

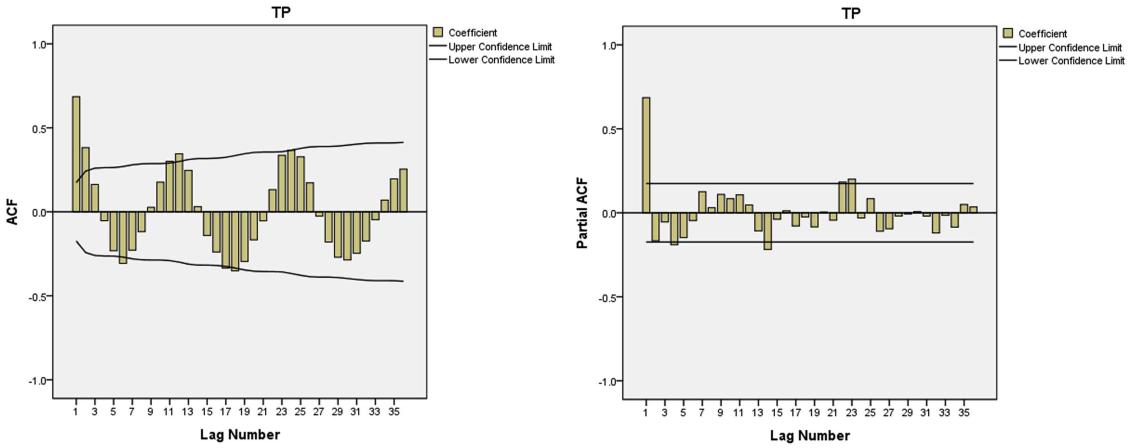

Fig. 3. ACF and PACF plots of TP.

Table 2. Estimated ACF and PACF of TP

Lag	Autocorrelation	Std. Error[a]	Partial Autocorrelation	Std. Error
1	0.686	0.087	0.686	0.087
2	0.382	0.121	-0.166	0.087
3	0.163	0.130	-0.054	0.087
4	-0.053	0.132	-0.190	0.087
5	-0.231	0.132	-0.147	0.087
6	-0.307	0.135	-0.046	0.087
7	-0.228	0.140	0.126	0.087
8	-0.118	0.143	0.031	0.087
9	0.027	0.143	0.109	0.087
10	0.177	0.144	0.084	0.087
11	0.301	0.145	0.107	0.087
12	0.345	0.150	0.047	0.087
13	0.246	0.156	-0.108	0.087
14	0.031	0.159	-0.218	0.087
15	-0.140	0.159	-0.038	0.087
16	-0.240	0.160	0.012	0.087
17	-0.335	0.162	-0.079	0.087
18	-0.351	0.167	-0.024	0.087
19	-0.295	0.173	-0.084	0.087
20	-0.167	0.177	0.005	0.087
21	-0.054	0.178	-0.043	0.087
22	0.131	0.178	0.183	0.087
23	0.336	0.179	0.200	0.087
24	0.366	0.184	-0.030	0.087
25	0.327	0.189	0.084	0.087
26	0.172	0.193	-0.109	0.087
27	-0.026	0.194	-0.095	0.087
28	-0.180	0.194	-0.019	0.087
29	-0.271	0.196	-0.007	0.087
30	-0.286	0.198	0.007	0.087
31	-0.247	0.202	-0.020	0.087
32	-0.174	0.204	-0.120	0.087
33	-0.046	0.205	-0.015	0.087
34	0.069	0.205	-0.085	0.087
35	0.196	0.205	0.049	0.087
36	0.254	0.207	0.035	0.087

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

편자기상관함수에 대한 t-검증통계량의 절대값의 실전적 경고수준은 '2.0'이다(3).

TP를 계절차분 후 편자기상관함수(그림 4)를 살펴보면 시차1에서 뚜렷한 스파이크를 발견할 수 있었으며 그 이후 단기시차에서 0으로 절단하는 형태이므로, 비계절적 부분을 AR(1)모형으로 고려할 수 있었다. 또한 자기상관함수의 시차12에서 뚜렷한 스파이크를 발견할 수 있었고(표 3에서 자기상관함수 12시차의 t-통계량의 절대값이 3.32(|-0.525/0.158|)로 계절시차의 경고수준인 1.25보다 훨씬 크다), 시차 24에서는 0으로 절단하는 형태이므로 $MA(1)_{12}$모형을 고려할 수 있었다. 따라서 $ARIMA(1,0,0)(0,1,1)_{12}$ 모형을 추정하였다.

3) 추정

추정 단계에서는 식별단계에서 선택한 모형의 계수를 정확히 추정하며, 이 추정된 계수가 통계적으로 유의한지 여부를 판가름 한다. 추정된 계수들의 절대값 크기를 통해서 정상성과 가역성 만족 여부를 판가름 한다. 추정된 계수가 유의하지 않을 경우, 다시 식별단계로 돌아가서 다른 모형을 임시로 선별한다(3).

표 4에 $ARIMA(1,0,0)(0,1,1)_{12}$ 모형의 계수를 추정하였다. 여기서 상수항이 모형에 포함되었을 때 상수에 대한 p값(0.457)이 유의수준 5%보다 훨씬 크기 때문에 모형에서 상수항을 제거한 후 표기하였다.

표 4를 살펴보면 두 개의 추정된 계수가 모두 통계적으로 매우 유의하며(각 p값<0.05), 각 계수의 정상성, 가역성을 살펴보면 |0.642|<1, |0.800|<1로 정상성, 가역성 조건을 모두 만족하였다(3).

4) 모형검진

모형검진 단계에서는 추정된 모형이 통계적으로 적절한지 여부를 결정한다. 여기서 백색잡음의 독립성 가정 여부를 잔차 자기상관함수, t-검증통계량, Box와 Ljung의 카이제곱검증, 잔차도표 등으로 점검한다.

TP 모형으로 추정된 $ARIMA(1,0,0)(0,1,1)_{12}$ 모형의 타당성을 검진해 보았다. 그림 5에서 잔차의 자기상관함수와 편자기상관함수를 살펴보면 단기시차에서 신뢰한계선 밖으로 튀어나온 스파이크는 더 이상 발견할 수 없었다. 시차 23에서 다소 큰 스파이크를 발견할 수 있었으나 단기시차에 해당하지 않았으며 불규칙 변동에 기인하는 것으로 간주하였다. 표 5에서 Box-Ljung 검증통계량의 p값이 모두 5% 유의수준보다 월등히 크기 때문에 백색잡음항의 독립성을 만족하였다. SPSS에서는 Ljung-Box Q(18)항의 유의확률로 간단히 백색잡음항의 독립성을 살펴볼 수 있었는데, Box와 Ljung

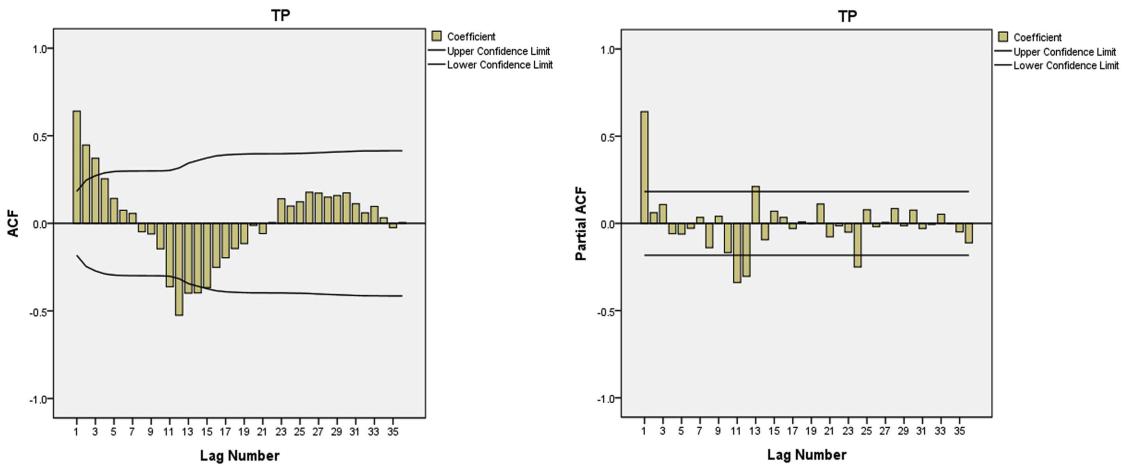

Fig. 4. ACF and PACF plots of TP by seasonal differences.

Table 3. Estimated ACF and PACF of TP by seasonal differences

Lag	Autocorrelation	Std. Error[a]	Partial Autocorrelation	Std. Error
1	0.641	0.091	0.641	0.091
2	0.447	0.123	0.061	0.091
3	0.372	0.136	0.108	0.091
4	0.254	0.144	-0.059	0.091
5	0.142	0.148	-0.061	0.091
6	0.075	0.149	-0.028	0.091
7	0.056	0.149	0.034	0.091
8	-0.048	0.150	-0.139	0.091
9	-0.060	0.150	0.040	0.091
10	-0.145	0.150	-0.168	0.091
11	-0.362	0.151	-0.339	0.091
12	-0.525	0.158	-0.303	0.091
13	-0.398	0.172	0.211	0.091
14	-0.397	0.180	-0.093	0.091
15	-0.366	0.187	0.069	0.091
16	-0.251	0.193	0.034	0.091
17	-0.196	0.195	-0.029	0.091
18	-0.144	0.197	0.008	0.091
19	-0.116	0.198	-0.003	0.091
20	-0.012	0.198	0.111	0.091
21	-0.058	0.198	-0.077	0.091
22	0.005	0.199	-0.014	0.091
23	0.140	0.199	-0.050	0.091
24	0.098	0.199	-0.250	0.091
25	0.123	0.200	0.078	0.091
26	0.178	0.200	-0.018	0.091
27	0.172	0.202	0.005	0.091
28	0.150	0.203	0.085	0.091
29	0.159	0.204	-0.013	0.091
30	0.174	0.205	0.076	0.091
31	0.112	0.206	-0.030	0.091
32	0.060	0.207	-0.006	0.091
33	0.096	0.207	0.052	0.091
34	0.032	0.207	-0.002	0.091
35	-0.024	0.207	-0.049	0.091
36	0.004	0.207	-0.111	0.091

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

(5)는 한 집합으로 여기는 모든 잔차 자기상관 함수를 근거로 한 검증통계량을 제시하였다. Box와 Ljung의 Q통계량은 값이 크면 한 집합으로서 잔차 자기상관계수가 통계적으로 0이 아님을 의미한다. 그래서 추정된 모형의 백색잡음들이 서로 상관되므로 다른 모형을 고려해야한다. Ljung-Box Q(18)항의 유의확률의 값이 5%보다 크면 백색잡음항이 독립이다는 귀무가설을 채택하게 된다(3). 이는 표 4에서 Ljung-Box Q(18)란의 유의확률의 값(0.715)으로 확인할 수 있었다.

또한 ARINMA모형의 안정성을 검진하기 위해 기간을 다르게 설정하여 변화 된 계수값 차이를 살펴보았다. 이는 모형 검진의 또 다른 방법으로 최근에 가까운 관측값의 일부를 제거하고 이에 대해 동일한 모형으로 재추정한다. 만일 축소된 관측값에 근거하여 계산된 계수의 추정값이 모든 자료에 근거하여 계산된 계수의 추정값과 거의 차이가 없다면(±0.1 이내), 최근에 가까운 관측값들이 초기 관측값들과 마찬가지로 똑같은 프로세스에서 생성되었다고 결론지을 수 있다(3). TP의 경우 표 4와 표 6을 비교해보면 각각의 계수는 0.642, 0.800이 0.648, 0.841로 그 차이가 0.006, 0.041이며, 따라서 시계열자료에 대해 안정적인 모형으로 간주 할 수 있었다. 그림 6에서 ARIMA$(1,0,0)(0,1,1)_{12}$ 모형을 실제 관측값과 적합시킨 그림으로 관측값과 잘 적합함을 시각적으로 확인 할 수 있었다. 위와 같은 방법으로 TP이외의 항목에 대하여 정상성, 식별, 추정, 모형진단을 수행하여 ARIMA모형을 구하고 추가로 계수비교를 수행하였다.

Table 4. Parameter estimates table for ARIMA$(1,0,0)(0,1,1)_{12}$ model of TP

ARIMA Model Parameters(used 2002~2013 TP data)											
					Estimate	SE	t	p	Model Statistics Ljung-Box Q(18)		
									Statistics	DF	p
ARIMA $(1,0,0)(0,1,1)_{12}$	TP	Natural Log	AR	Lag 1	0.642	0.068	9.504	0.000	12.414	16	0.715
			Seasonal Difference		1						
			MA, Seasonal	Lag 1	0.800	0.100	7.979	0.000			

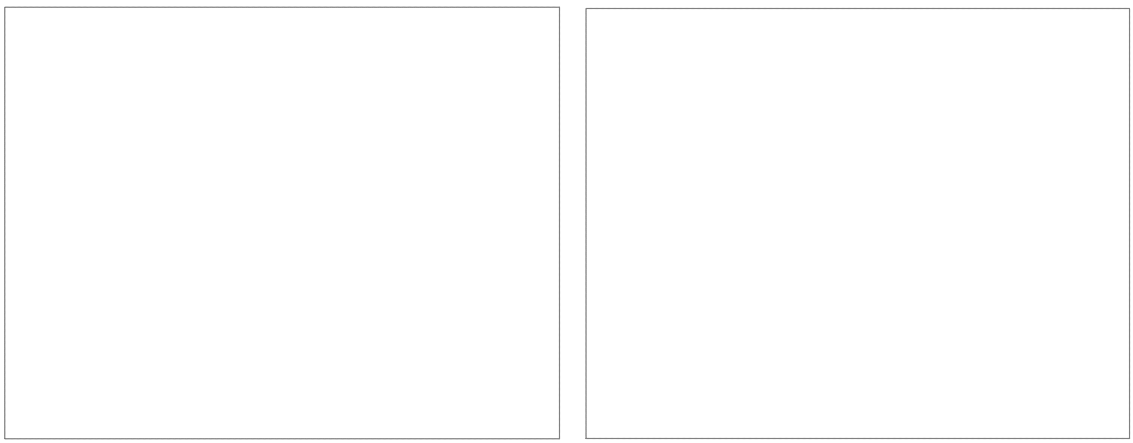

Fig. 5. ACF and PACF plots of residuals by ARIMA$(1,0,0)(0,1,1)_{12}$ model of TP.

Table 5. ACF and PACF of residuals by ARIMA(1,0,0)(0,1,1)$_{12}$ model of TP

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	-0.047	0.087	0.302	1	0.582
2	-0.020	0.087	0.355	2	0.837
3	0.111	0.087	2.034	3	0.565
4	0.001	0.088	2.034	4	0.729
5	-0.036	0.088	2.218	5	0.818
6	-0.067	0.088	2.855	6	0.827
7	0.012	0.089	2.877	7	0.896
8	-0.068	0.089	3.540	8	0.896
9	0.109	0.089	5.241	9	0.813
10	0.102	0.090	6.735	10	0.750
11	-0.071	0.091	7.470	11	0.760
12	-0.010	0.092	7.484	12	0.824
13	0.060	0.092	8.022	13	0.842
14	-0.080	0.092	8.977	14	0.833
15	-0.104	0.092	10.623	15	0.779
16	-0.012	0.093	10.644	16	0.831
17	-0.105	0.093	12.349	17	0.779
18	0.021	0.094	12.414	18	0.825
19	-0.093	0.094	13.774	19	0.797
20	0.063	0.095	14.392	20	0.810
21	-0.002	0.095	14.392	21	0.852
22	-0.036	0.095	14.598	22	0.879
23	0.172	0.095	19.373	23	0.679
24	-0.016	0.098	19.417	24	0.729
25	0.037	0.098	19.637	25	0.766
26	0.053	0.098	20.099	26	0.787
27	0.027	0.098	20.223	27	0.821
28	-0.010	0.098	20.240	28	0.856
29	-0.052	0.098	20.706	29	0.870
30	0.086	0.098	21.984	30	0.855
31	-0.054	0.099	22.496	31	0.867
32	-0.085	0.099	23.777	32	0.852
33	0.116	0.100	26.167	33	0.795
34	-0.030	0.101	26.328	34	0.823
35	-0.035	0.101	26.550	35	0.847
36	-0.004	0.101	26.554	36	0.875

a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.

b. Based on the asymptotic chi-square approximation.

각 항목의 ARIMA모형은 표 7과 같으며, 2013년을 제외하고 추정한 ARIMA모형과 계수의 차이는 0.004~0.041로 똑같은 프로세스에서 생성되었다고 볼 수 있었으며, 안정적인 ARIMA모형으로 간주할 수 있었다.

2. 대체모형

예측에 앞서 단변량 시계열분석의 또 하나의 기법인 지수평활법을 ARIMA 모형의 대체모형으로 고려하여 보았다. ARIMA 모형과 더불어 대표적으로 일변량 시계열을 예측하는 데 사용하는 지수평활법(exponential smoothing method, ESM)은 각 실현값들에 다른 가중값을 부여하는 예측방법이다. 지수평활법은 시계열을 표현하는 모수들이 시간의 흐름에 따라 서서히 변화할 때 가장 효율적으로 사용할 수 있기 때문에 중기 및 단기 예측에 사용된다. 그러나, 어느 특정적 통계모형 또는 통계이론에 근거한 것은 아니며, 적절한 예측값을 얻는 데 직관적이며 경험적인 방법으로 사용된다. 지수평활법의 주된 장점은 단순하고, 직관적이며, 쉽게 이해될 수 있다는 점이다. 일반적으로 과거 관측값의 수가 적을 때도 사용할 수 있으며, 다양한 응용 사례에서 좋은 예측값을 힘들이지 않고 산출할 수 있는 방법이기도 하다(3).

Table 6. Parameter estimates table for ARIMA(1,0,0)(0,1,1)$_{12}$ model of TP

ARIMA Model Parameters(used 2002~2012 TP data)								
					Estimate	SE	t	p
ARIMA (1,0,0)(0,1,1)$_{12}$	TP	Natural Log	AR	Lag 1	0.648	0.072	8.962	0.000
			Seasonal Difference		1			
			MA, Seasonal	Lag 1	0.841	0.139	6.046	0.000

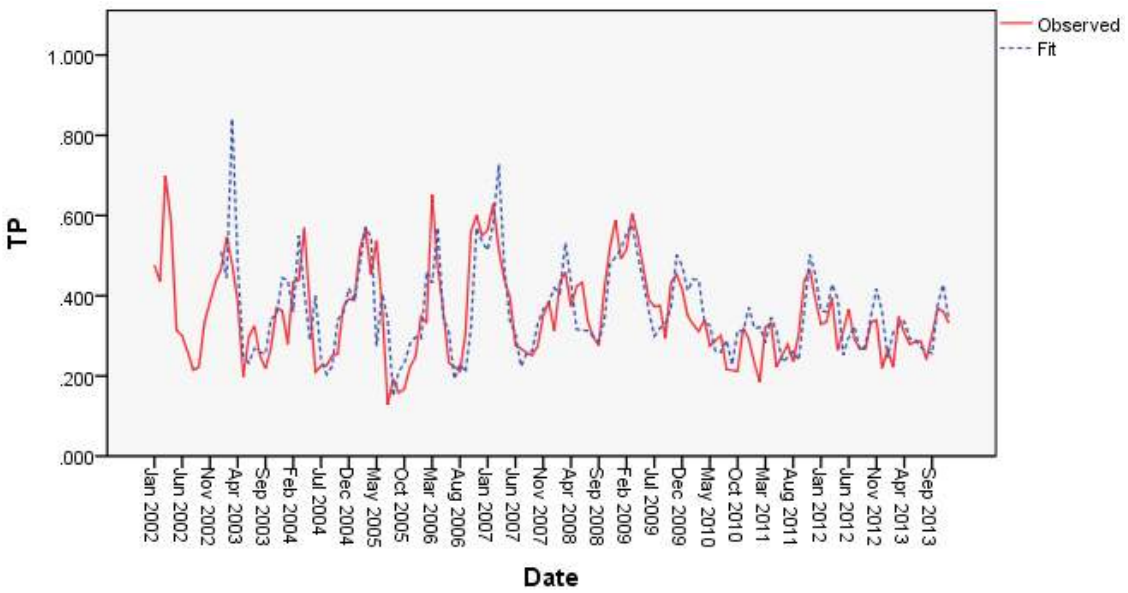

Fig. 6. Comparison of observed and fit values by ARIMA(1,0,0)(0,1,1)$_{12}$ model of TP.

Table 7. Parameter estimates table for ARIMA models of each variable

ARIMA Model Parameters												
model					2002.1 ~ 2012.12				2002.1 ~ 2013.12			
					Estimate	SE	t	p	Estimate	SE	t	p
DO	ARIMA $(1,0,0)(2,1,0)_{12}$	No Transformation	AR	Lag 1	0.759	0.062	12.240	0.000	0.745	0.060	12.490	0.000
			AR, Seasonal	Lag 1	-0.610	0.094	-6.455	0.000	-0.597	0.088	-6.805	0.000
				Lag 2	-0.335	0.098	-3.420	0.001	-0.316	0.093	-3.395	0.001
			Seasonal Difference		1				1			
pH	ARIMA $(0,0,2)(0,1,1)_{12}$	No Transformation	MA	Lag 1	-0.577	0.081	-7.173	0.000	-0.599	0.078	-7.701	0.000
				Lag 2	-0.509	0.082	-6.194	0.000	-0.490	0.078	-6.262	0.000
			Seasonal Difference		1				1			
			MA, Seasonal	Lag 1	0.758	0.096	7.934	0.000	0.769	0.085	9.012	0.000
			Constant		-0.415	0.117	-3.538	0.001	-0.411	0.102	-4.017	0.000
TN	ARIMA $(1,0,0)(0,1,1)_{12}$	No Transformation	AR	Lag 1	0.614	0.073	8.371	0.000	0.609	0.070	8.695	0.000
			Seasonal Difference		1				1			
			MA, Seasonal	Lag 1	0.886	0.178	4.97	0.000	0.865	0.127	6.806	0.000
TP	ARIMA $(1,0,0)(0,1,1)_{12}$	Natural Log	AR	Lag 1	0.648	0.072	8.962	0.000	0.642	0.068	9.504	0.000
			Seasonal Difference		1				1			
			MA, Seasonal	Lag 1	0.841	0.139	6.046	0.000	0.800	0.100	7.979	0.000
TOC	ARIMA $(5,1,0)(0,0,1)_{12}$	Natural Log	AR	Lag 5	-0.274	0.101	-2.707	0.008	-0.263	0.096	-2.747	0.007
			Difference		1				1			
			MA, Seasonal	Lag 1	-0.259	0.105	-2.472	0.015	-0.236	0.100	-2.364	0.020
EC	ARIMA $(1,0,0)(0,1,1)_{12}$	No Transformation	AR	Lag 1	0.687	0.069	10.02	0.000	0.691	0.065	10.705	0.000
			Seasonal Difference		1				1			
			MA, Seasonal	Lag 1	0.638	0.097	6.557	0.000	0.633	0.088	7.172	0.000

본 연구에서는 SPSS 21의 자동 모형 생성기를 사용하여 지수평활법을 적용하였다. 자동 모형 생성기에서는 7개의 지수평활법(4개의 비계절모형과 3개의 계절모형)을 적합시킨 후 각 모형에 대한 (모형적합 통계량 중 하나인) BIC를 계산하여 가장 작은 값을 갖는 모형을 선택한다(3). 표 8에서 대체모형으로 선택한 지수평활법과 이전에 구해진 ARIMA모형을 여러 가지 모형적합통계량으로 비교해보고 Box-Ljung Q통계량으로 백색잡음항의 독립성 강도를 살펴보았다. 여기서 정상 R제곱, R제곱, Ljung-Box Q(18)통계량의 p값은 클수록, RMSE, MAPE, MAE, MaxAPE, MaxAE, 정규화된 BIC값은 작을수록 더 우세한 모형이 된다(3).

3. 최종 모형 선택 및 예측

표 8을 살펴보면 모형적합통계량의 관점에서는 지수평활법에 의한 대체모형이 대부분 우위에 있었고, Box-Ljung Q통계량의 관점에서는 ARIMA 모형이 우위에 있었다. TP의 경우 모형적합통계량을 비교해보면 8개 항목 중 2개 항목을 빼놓고 모두 지수평활법이 모형적합력에서 우수하였고 Box-Ljung Q통계량이 나타내는 백색잡음의 독립성은 ARIMA모형이 우수하였다. 두 모형 중 최종모형을 선택하기 위해 마지막으로 예측력을 고려하였다.

좋은 모형은 과거 자료를 잘 적합시키지만 더 중요한 것은 미래의 값을 잘 예측해야 한다. 즉, 예측력을 말한다. 현실적으로 이를 점검하는 것은

Table 8. Model statistics of each variable for ARIMA model and Exponential smoothing method

	Model Type	Transfor-mation	Model statistics								Ljung-Box Q(18)
			Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	p
DO	ARIMA $(1,0,0)(2,1,0)_{12}$	None	0.669	0.802	1.242	13.06	0.972	56.03	3.144	0.553	0.693
	Simple Seasonal	None	0.524	0.858	1.033	11.24	0.820	62.07	2.851	0.138	0.361
pH	ARIMA $(0,0,2)(0,1,1)_{12}$	None	0.659	0.625	0.257	2.407	0.177	12.15	0.969	-2.597	0.635
	Simple Seasonal	None	0.549	0.695	0.233	2.382	0.175	11.76	1.000	-2.837	0.122
TN	ARIMA $(1,0,0)(0,1,1)_{12}$	None	0.473	0.278	1.790	20.43	1.327	73.83	6.755	1.284	0.766
	Simple Seasonal	None	0.621	0.610	1.553	17.93	1.156	78.15	5.139	0.954	0.474
TP	ARIMA $(1,0,0)(0,1,1)_{12}$	Natural log	0.606	0.436	0.088	20.05	0.065	164.4	0.367	-4.782	0.773
	Simple Seasonal	None	0.529	0.605	0.075	17.34	0.056	168.6	0.249	-5.096	0.177
TOC	ARIMA $(5,1,0)(0,0,1)_{12}$	Natural log	0.133	0.683	1.800	20.42	1.101	127.1	8.561	1.271	0.769
	Winters' Multiplicative	None	0.648	0.804	1.413	20.70	0.998	78.70	4.490	0.834	0.406
EC	ARIMA $(1,0,0)(0,1,1)_{12}$	None	0.539	0.677	47.80	12.67	33.75	80.23	292.8	7.814	0.675
	Simple Seasonal	None	0.613	0.783	42.27	10.87	30.81	55.36	233.3	7.562	0.158

불가능하다. 왜냐하면 예측값을 구하기 위해서는 시간이 경과하여 미래 시점에 해당하는 관측값을 얻어야 하기 때문이다. 이 경우 최근에 가까운 실현값의 일부(일반적으로 최근에 가까운 10% 관측값)를 제거하고 제거된 시점에 대한 관측값을 추정된 모형으로 예측하여 실제 관측값과 어느 정도 차이가 있는지를 비교한다. 제거 된 최근 10%의 실현값을 예측하여 모형별로 실현값에 더 가까운 예측값을 갖는 모형이 우선시된다(3).

각 항목의 자료에서 2013년 자료를 제외한 후, 2002년 1월부터 2012년 12월 데이터를 바탕으로 ARIMA모형과 지수평활법을 사용한 2013년 예측치와 실제 2013년 자료를 표 9에 나타내었다. 예측의 정확도를 검증하기 위하여 MAPE(Mean Absolute Percentage Error)방법을 다음과 같이 사용하였다.

Table 9. Future forecasting(2013) of each variable by ARIMA model and Exponential smoothing method

	2013	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	MAPE
DO	Observed	14.1	12.9	12	10.8	7.7	8.4	8.1	8.4	8.7	9.7	9.5	11.4	
	ARIMA(1,0,0)(2,1,0)$_{12}$	15.4	14.5	12.7	11.3	8.8	6.5	6.8	6.7	7.7	8.8	10.2	13.1	12.36
	Simple Seasonal	15	14.9	13.3	11.6	10	8.7	8.9	9.4	9.7	10.4	11.7	13.8	13.18
pH	Observed	7.5	7.4	7.3	7.3	7	6.9	7.2	7.1	7	7.4	7.5	7.5	
	ARIMA(0,0,2)(0,1,1)$_{12}$	7.9	7.8	7.6	7.5	7.1	7	7	7	7	7.3	7.4	7.6	2.389
	Simple Seasonal	8	8.3	8.1	8	7.6	7.5	7.5	7.4	7.4	7.6	7.8	7.9	6.898
TN	Observed	4.204	3.675	4.62	3.952	3.876	3.278	3.346	3.614	4.715	4.574	5.552	5.784	
	ARIMA(1,0,0)(0,1,1)$_{12}$	3.502	4.976	6.11	5.156	4.19	2.665	4.794	3.667	2.975	4.019	4.625	4.69	22.58
	Simple Seasonal	4.353	5.504	6.413	5.337	3.957	2.398	4.516	3.353	2.586	3.474	4.027	4.195	26.86
TP	Observed	0.264	0.223	0.348	0.308	0.279	0.286	0.286	0.241	0.303	0.368	0.361	0.332	
	ARIMA(1,0,0)(0,1,1)$_{12}$ Natural log	0.25	0.304	0.398	0.368	0.323	0.31	0.281	0.258	0.263	0.341	0.407	0.374	12.86
	Simple Seasonal	0.258	0.284	0.338	0.301	0.245	0.222	0.196	0.184	0.187	0.243	0.3	0.293	18.77
TOC	Observed	3.4	2.77	3.5	2.87	2.94	2.89	3.11	2.31	2.74	2.67	2.85	3.2	
	ARIMA(5,1,0)(0,0,1)$_{12}$	3.43	3.92	3.59	3.67	3.94	4.09	3.94	4.01	3.88	4.29	4.31	4.41	36.66
	Winters' Multiplicative	3.2	3.95	4.35	3.74	2.99	2.38	2.24	2.2	1.67	2.24	2.72	3.32	18.21
EC	Observed	274	244	246	238	234	228	186	179	205	246	289	293	
	ARIMA(1,0,0)(0,1,1)$_{12}$	272	299	284	255	232	243	191	184	213	270	305	301	6.725
	Simple Seasonal	281	301	278	226	160	149	105	91	117	174	239	257	25.39

$$MAPE = \frac{1}{N}\sum\left|\frac{X_t - F_t}{X_t}\right| \times 100$$

여기서 X는 실제값, F는 예측값, 그리고 N은 실제값의 기간 수를 각각 나타낸다. 또한, 계산된 MAPE 값P은 수치에 따라 $0\% \leqq MAPE < 10\%$: 매우 정확한 예측, $10\% \leqq MAPE < 20\%$: 비교적 정확한 예측, $20\% \leqq MAPE < 50\%$: 비교적 합리적 예측, $MAPE \geqq 50\%$: 부정확한 예측으로 해석된다(6). TP의 경우를 살펴보면 MAPE값을 비교해 볼 때 예측력에서 ARIMA모형이 더 우수한 것을 알 수 있었다. 이는 그림 7과 그림 8을 비교해 보면 시각적으로도 확인 할 수 있었다.

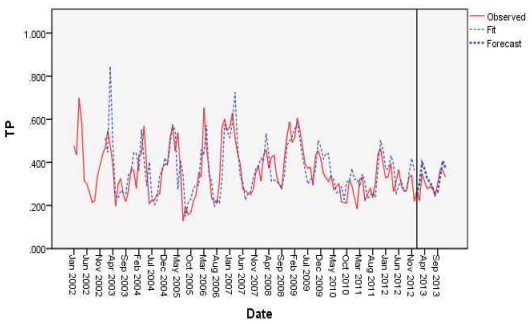

Fig. 7. The future forecasting(2013) plot of TP by ARIMA model.

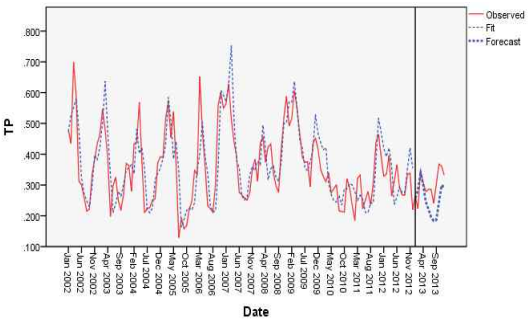

Fig. 8. The future forecasting(2013) plot of TP by Exponential smoothing method.

예측력을 교려하여 최종 선택 된 ARIMA모형으로 TP항목을 2002년 1월부터 2013년 12월 데이터를 사용하여 2014년 측정값을 예측하여 그림 9에 나타내었다.

다른 항목들도 ARIMA모형과 지수평활모형을 비교하여 선택한 최종모형으로 2014년 예측값을 그림 10과 표 10으로 표시하였다. 또한 SPSS 21에서 제공하는 적합값 및 예측값의 근사된 95% 신뢰구간을 신뢰구간 상한(UCL)과 신뢰구간 하한(LCL)으로 표시하였다.

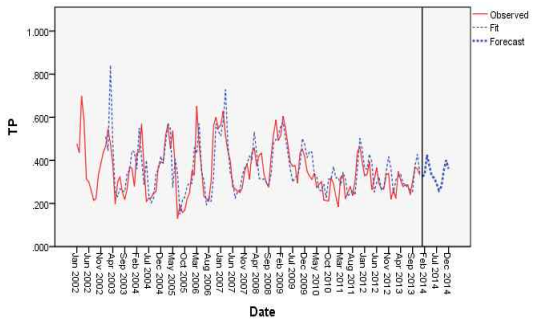

Fig. 9. The future forecasting(2014) plot of TP by ARIMA model.

결 론

본 연구에서는 2002년부터 2013년까지의 행주측정소 월별 수질측정자료를 단변량 시계열 모형을 사용하여 2014년 월별 수질예측 모형을 구축하고, 2014년 항목별 측정값을 예측하였으며, 다음과 같은 결과를 얻을 수 있었다.

1. 행주측정소의 2014년 항목별 예측을 위한 단변량 시계열 최종모형으로 DO는 ARIMA(1,0,0)$(2,1,0)_{12}$모형, pH는 ARIMA(0,0,2)$(0,1,1)_{12}$모형, TN은 ARIMA(1,0,0)$(0,1,1)_{12}$모형 TP는 자연로그 변환 후 ARIMA(1,0,0)$(0,1,1)_{12}$모형, TOC는 윈터스승법, 전기전도도는 ARIMA(1,0,0)$(0,1,1)_{12}$모형을 선택하였고, 선택한 모형을 적용하여 2014년 항목별 월별 측정값을 예측하였다.

2. 2002년부터 2012년 자료에 최종모형을 적용하여 계산한 2013년 예측치의 MAPE값은 DO가 12.36, pH가 2.389, TN이 22.58, TP가 12.86, TOC가 18.21, 전기전도도가 6.725로 pH와 전기전도도가 매우 정확한 예측력을 보

여 주었고, DO, TP, TOC는 비교적 정확한 예측, TN은 가장 예측력이 떨어지지만 비교적 합리적 예측력을 보여 주었다.

3. 본 연구에서는 예측할 변수 하나를 선정한 후 해당변수의 과거 자료를 근거로 해당 변수의 미래 관측값을 예측하는 일변량 시계열 분석방법으로 ARIMA모형과 지수평활법을 사용하였다. 향후 추가적으로 일변량 시계열에 영향을 줄 수 있는 사건이 발생하였을 경우, 이러한 외적인 요인들의 효과를 일변량 시계열에 포함시켜 모형을 구축하는 개입모형으로 모형의 정확도를 높이려는 노력과, 폭넓고 다양한 상황을 다룰 수 있는 독립변수가 포함된 다변량 시계열 모형을 제시함으로써 예측의 정밀성을 높이려는 노력은 계속되어야 할 것이다.

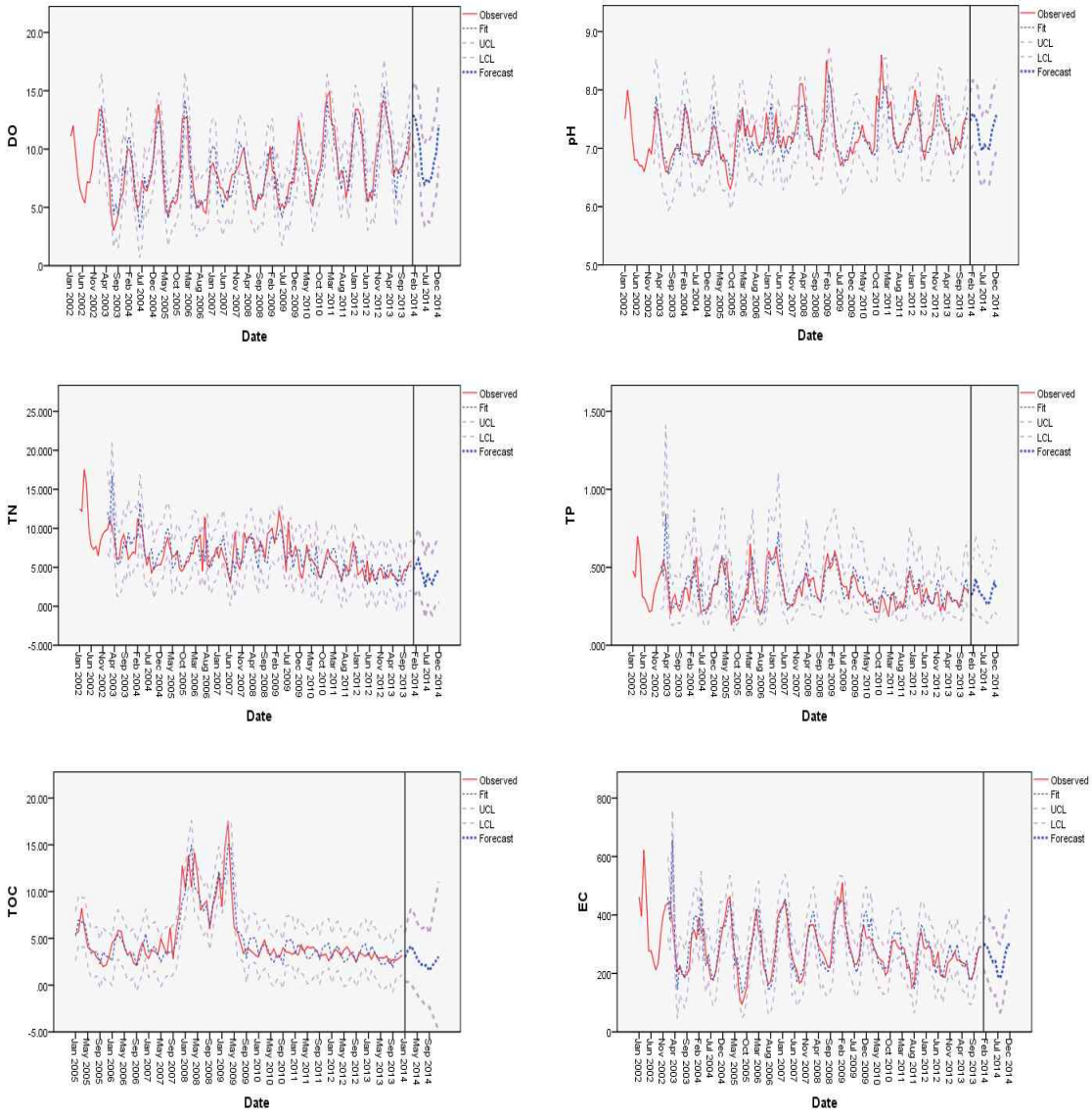

Fig. 10. Future forecasting(2014) plots of each water quality parameter at Haeng-ju of lower Han river by time series models.

Table 10. Future forecasting(2014) of each water quality parameter at Haeng-ju of lower Han river by time series models

2014		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
DO	Forecast	12.9	12.8	11.6	10.6	7.8	7.0	7.4	7.2	7.6	8.9	9.7	12.1
ARIMA $(1,0,0)(2,1,0)_{12}$	UCL	15.2	15.7	14.9	13.9	11.2	10.5	10.9	10.7	11.1	12.4	13.2	15.6
	LCL	10.5	9.9	8.4	7.2	4.3	3.5	3.9	3.7	4.1	5.3	6.1	8.5
pH	Forecast	7.5	7.6	7.5	7.5	7.1	7.0	7.1	7.0	7.0	7.3	7.4	7.6
ARIMA $(0,0,2)(0,1,1)_{12}$	UCL	8.0	8.2	8.2	8.1	7.7	7.6	7.7	7.6	7.6	7.9	8.0	8.2
	LCL	7.0	7.0	6.9	6.8	6.5	6.3	6.4	6.4	6.4	6.7	6.8	7.0
TN	Forecast	4.741	5.300	6.031	4.882	3.998	2.513	4.282	3.326	2.886	3.782	4.435	4.467
ARIMA $(1,0,0)(0,1,1)_{12}$	UCL	7.904	9.003	9.914	8.830	7.969	6.493	8.266	7.310	6.872	7.767	8.420	8.451
	LCL	1.577	1.598	2.148	0.934	0.026	-1.468	0.299	-0.659	-1.099	-0.204	0.450	0.482
TP (Natural log)	Forecast	0.327	0.332	0.426	0.374	0.327	0.318	0.293	0.261	0.277	0.353	0.401	0.361
ARIMA $(1,0,0)(0,1,1)_{12}$	UCL	0.494	0.539	0.708	0.629	0.552	0.538	0.495	0.442	0.470	0.598	0.679	0.611
	LCL	0.207	0.191	0.237	0.205	0.178	0.173	0.159	0.142	0.151	0.192	0.218	0.196
TOC	Forecast	3.01	3.77	4.16	3.57	2.84	2.27	2.12	2.10	1.57	2.10	2.52	3.07
Winters' Multiplicative	UCL	5.64	7.19	8.21	7.61	6.70	6.01	6.11	6.43	5.48	7.36	8.96	11.03
	LCL	0.37	0.35	0.10	-0.46	-1.01	-1.47	-1.86	-2.23	-2.34	-3.16	-3.92	-4.90
EC	Forecast	300	297	283	257	239	242	192	184	211	262	299	298
ARIMA $(1,0,0)(0,1,1)_{12}$	UCL	384	400	393	371	354	358	309	301	328	379	416	415
	LCL	215	194	172	143	123	126	75	67	94	145	182	181

참고문헌

1. 환경부 : 수질오염감시경보를 위한 측정소별 측정항목과 항목별 경보기준, 2013.
2. 환경부 : 수질 및 수생태계 보전에 관한 법률, 2013.
3. 정동빈 : SPSS(PASW) 시계열 수요예측 I, 1판. 한나래출판사, 서울, 2009.
4. Box, GEP, Jenkins, GM and Reinsel, GC : Time Series Analysis: Forecasting and Control, 3rd ed. Prentice Hall, New Jersey, 1994.
5. Ljung, GM and Box, GEP : On a Measure of Lack of Fit in Time Series Models, Biometrika, 64:297, 1978.
6. 이충기 : 관광조사연구론, 일신사, 2003.