

경기도 파주시 오존농도의 통계모형 연구[†]

이훈자¹

평택대학교 디지털응용정보학과

접수 2009년 8월 20일, 수정 2009년 11월 13일, 게재확정 2009년 11월 21일

요약

지표오존 농도는 국가의 중요한 환경 척도 중의 하나이다. 본 연구에서는 경기도 파주시 오존농도를 자기회귀오차모형과 신경망모형으로 분석하였다. 오존 분석을 위한 설명변수로는 이산화황, 이산화질소, 일산화탄소, 프로메튬10 등의 대기자료와 일 최고온도, 풍속, 상대습도, 강수량, 이슬점온도, 운량, 수증기압 등의 기상자료를 사용하였다. 분석 결과 전반적으로 신경망모형이 좋은 모형으로 나타났다. 자기회귀오차모형도 오존에 영향을 주는 설명변수를 첨가하면 좋은 모형이 될 것으로 생각된다.

주요용어: 기상자료, 대기자료, 신경망모형, 오존농도, 자기회귀오차모형.

1. 서론

지표오존의 고농도 현상은 농작물이나 식물들의 피해뿐 아니라 인간의 건강에도 피해를 주는 것으로 알려져 있다. 지금까지 오존농도의 모형 적합에 관해 연구되어 왔다. Thompson 등 (2000) 에 따르면 오존농도 모형설정은 크게 다음의 3가지 범주로 나뉘질 수 있다. 첫째, regression-based modeling (Feister와 Balzer, 1991), 둘째, extreme value approach (Smith와 Huang, 1993), 셋째, spatio-temporal (Guttorp, 1994) 이다.

오존농도의 모형에 관한 국내연구의 경향을 살펴보면, 하계의 일 최고 오존농도 예측을 위한 신경망모델의 개발 (김용국 등, 1994), 중회귀 예보모델 개발에 관한 연구 (김용준, 1997), wavelet transform을 이용한 단기오존농도 예측에 관한 연구 (김신도, 1998), 전이함수모델 개발에 관한 연구 (김유근 등, 1999), 의정부시 오존농도의 시계열모형 연구 (Lee, 2008), 최적시계열모형에 기초한 오존주의보 날짜 예측 (Park 등, 2009) 등 다양한 방법으로 지표오존 농도모형에 관해 연구하였다.

국외의 경우를 보면, 미국 동부지방 오존 모형을 위해 generalized linear model을 사용했고 (Camalier 등, 2007), 칠레 Santiago 지역의 오존 일 최고농도 예측을 위해 세 가지 모형인 시계열모형, 신경망모형, fuzzy 모형을 사용해 비교 분석하였고 (Jorquera 등, 1998), 오스트리아 Vienna 지역의 오존 농도의 단기예측을 위해 시계열모형인 ARX (Autoregressive Exogenous) 모형을 사용하였다 (Bauer 등, 2001).

본 연구에서는, 경기 북부지역의 파주시 오존농도를 2004년부터 2007년까지 자료 중 오존 농도가 비교적 높은 5월 1일에서 9월 30일까지의 자료를 시계열 모형에 적합하고자 한다. 정확한 오존농도의 분석과 예측을 위해서는 오존과 연관된 대기 및 기상자료를 포함시켜야 한다. 본 논문에서는 오존과 연관이 있는 7종류의 기상자료와 4종류의 대기 자료를 설명변수로 포함한 다변량 분석인 자기회귀오차 (ARE, Autoregressive Error) 모형으로 분석하고 이 모델을 모형 예측도가 높은 신경망모형과 비교하였다.

[†] 이 논문은 2008학년도 평택대학교 학술연구비의 지원에 의하여 연구되었음.

¹ (450-701) 경기도 평택시 용이동 111, 평택대학교 디지털응용정보학과, 부교수. E-mail: esther@ptu.ac.kr

2. 오존자료 및 설명변수 자료

오존자료는 국립환경과학원의 파주시 금천동 관측소의 시간별 자료 (하루 24번)를 이용했고 분석을 위하여 일 최고오존농도를 사용했다. 기간은 2004년부터 2007까지의 5월 1일부터 9월 30일 까지 자료를 사용했다. 본 연구에서 5월에서 9월 까지 자료를 사용한 이유는 1년에 걸쳐 이 기간 동안이 오존 농도가 높기 때문이고 그래서 이 기간 동안의 분석도 중요할 것으로 생각되었다.

설명변수 중 기상 자료는 시간별 자료로 파주에서 가장 가까운 곳에 위치한 기상청의 파주시 문산 지점 자료를 사용하였다. 기상 자료도 오존과 같이 2003년-2007년 5월부터 9월까지의 자료이다. 8종류의 기상 자료는 김유근 등 (1999)을 참고로 하여 일 최고온도 (maximum temperature), 풍속 (wind speed), 상대습도 (relative humidity), 강수량 (rainfall), 이슬점온도 (dew point temperature), 운량 (amount of cloud), 수증기압 (water vapor pressure)을 사용했으며 표 3.1에 나타나 있다.

설명변수인 4종류의 대기 자료로는 이산화황 (SO_2), 이산화질소 (NO_2), 일산화탄소 (CO), 프롬테움10 (PM10)를 사용했고, 오존자료와 같이 국립환경과학원의 파주시 monitoring site를 이용하여, 2004년-2007년의 5월1일-9월 30일의 시간별 자료를 이용하였다. 사용된 대기 자료도 표 3.1에 나타나 있다.

7종류의 기상 자료와 4종류의 대기 자료의 시간별 선택과 차수 선택은 Hubbard (1998)를 참고로 하여 각 자료를 시간별 혹은 일별로 오존과의 상관관계가 가장 높은 시간대를 사용하였다. 예를 들면, 운량 (cloud) 변수는 오존과 상관관계가 높은 일별 9시-15시의 평균을 사용하였다. 각 설명변수의 구체적 인 시간대는 표 3.1에 나타나 있다. 표 3.1의 Notation은 5장의 모형식 변수표시이다.

3. 오존농도의 ARE 모형과 적합성

반응변수에 영향을 주는 설명변수가 있을 때는 설명변수를 분석에 이용하는 것이 효율적이다. ARE 모형은 설명변수를 분석에 사용 할 수 있는 모형으로 시계열자료를 회귀모형에 적합 시킬 때 적합한 모형이다. 반응변수에 영향을 주는 PM10농도나, 일별 최고기온 같은 설명변수를 사용하여 자료를 분석한다. ARE 모형은 전이함수모형의 특수한 경우로, 오차항 ϵ_t 가 독립이 아니라 서로 상관관계를 갖게 되며, 특히 AR (Autoregressive) 형태를 갖는다. 일반적인 k 차 ARE 모형은 다음과 같다.

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_p X_{tp} + \epsilon_t, \quad t = 1, 2, \dots, n,$$

이며

$$\epsilon_t = e_t - \phi_1 \epsilon_{t-1} - \phi_2 \epsilon_{t-2} - \dots - \phi_k \epsilon_{t-k}, \quad \text{이고 } e_t \sim i.i.dN(0, \sigma^2) \quad (3.1)$$

식 (3.1)에서 오차항 ϵ_t 는 서로 상관관계가 있고 AR(k) 형태를 갖는다. 또한 식 (3.1)에서 Y_t 는 반응변수인 오존 농도를 나타내며, $X_{t,j}$, $j = 1, 2, \dots, p$ 는 Y_t 에 영향을 주는 PM10, 일 최고온도 등의 설명변수들을 나타낸다. 위의 ARE 모형과 회귀모형의 차이점은 ARE 모형은 식 (3.1)에서 오차항 ϵ_t 가 서로 독립이 아니라 시차에 따른 자기 상관을 갖고 있으며 특히 AR(k) 모형의 형태를 갖는다는 점이다. 따라서 식 (3.1)을 ARE (자기회귀오차)모형이라 부르며 시계열자료에 회귀모형을 적합 시킬 경우 많이 사용된다.

ARE 모형들의 적합성을 측정하는데 있어 결정계수 R^2 와 잔차 분석의 RMSE를 사용하였다. 결정계수 R^2 은 오존농도 Y_t 가 설명변수에 의한 설명력으로 값이 1에 가까울수록 설명력이 높은 좋은 모형이

표 3.1 설명변수로 사용된 대기 자료와 기상 자료

Variable	Contents and Timing	Unit	Notation
O_3 -1	Previous day maximum 1 hour O_3	ppb	Y_{t-1}
SO_2	Previous day maximum 1 hour SO_2	ppb	S_{t-1}
NO_2	Previous day maximum 1 hour NO_2	ppb	N_{t-1}
CO	Previous day maximum 1 hour CO	0.1ppm	I_{t-1}
PM10	Daily maximum 1 hour PM10	$\mu g/m^3$	P_t
Temperature	Daily maximum 1 hour surface temperature	$^{\circ}C$	T_t
Wind	Average wind speed (from 9:00 to 15:00)	m/s	W_t
Cloud	Average cloud amount (from 9:00 to 15:00)	-	C_t
Humidity	Average relative humidity (from 9:00 to 13:00)	%	H_t
Rainfall	Sum of precipitation (from 1:00 to 24:00)	mm	A_t
Dew	Average dew point temperature (from 9:00 to 15:00)	$^{\circ}C$	D_t
Waterp	Average water vapor pressure (from 9:00 to 15:00)	hPa	V_t

다. 예측값과 실제값의 분산정도를 나타내는 RMSE (root mean square error)의 공식은 다음과 같다.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2}, \quad (3.2)$$

o_i 와 p_i 는 각각 i 일의 관측 오존량과 추정 오존량을 의미하며, n 은 통계량에 사용된 일수 (day)이다. RMSE는 값이 작을수록 좋은 모형이다.

ARE 모형과 신경망 모형 결과를 평가하기 위한 척도로 RMSE 외에도, NMSE (normalized mean square error) 와 MAE (mean absolute error)를 사용하여 비교하였다. NMSE 와 MRE는 식 (3.3)과 (3.4)로부터 계산되며, NMS는 RMSE에 비해 고농도 오존 부근에서의 오차를 더 많이 반영하는 특성이 있다. MAE는 실제값과 추정치간의 상대적인 예측정도를 평균한 값이다.

$$NMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}}{\frac{1}{n} (\sum_{i=1}^n O_i) (\sum_{i=1}^n P_i)}, \quad (3.3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |O_i - P_i| \quad (3.4)$$

4. 오존농도의 신경망모형

인간의 뇌신경으로부터 영감을 얻어 시작된 신경망분석은 자료를 비선형모형 (non-linear model)으로 분석하며 정확한 예측을 목적으로 하고 있다. 신경망모형은 많은 설명변수를 가지고 있으며 설명변수와 반응변수 간의 관계가 복잡한 비선형 형태를 가질 때 유용하다. 회귀모형이나 ARE 모형은 반응변수를 수식으로 표현하여 어떤 설명변수들이 오존에 얼마나 영향을 주는지 등에 관한 해석을 할 수 있는 장점이 있다. 신경망모형은 반응변수인 오존에 대한 예측은 다른 모형보다 정확한 장점이 있는 반면에 반응변수를 수식으로 나타내기가 어렵고 따라서 설명변수들에 관한 해석도 어렵다. 본 연구에서는 신경망모형 중 자료 분석을 위해 가장 널리 사용되는 모형인 MLP (Multi-layer Perceptron, 다층인식차) 신경망으로 분석하였다. MLP는 입력층, 은닉층, 출력층으로 구성되어 있다. 은닉층은 여러 개의 은닉마디로 구성되어 있고, 각 은닉마디는 입력층으로 부터 받은 설명변수들의 선형 결합을 비선형함수로 분석

하여 출력층 또는 다른 은닉층으로 전달된다. 예를 들면, 입력층, 하나의 은닉층에 두 개의 은닉마디를 가지는 MLP의 구조는 다음과 같다.

$$H_1 = f_1(b_1 + w_{11}X_1 + w_{21}X_2 + \dots + w_{k1}X_k)$$

$$H_2 = f_2(b_2 + w_{12}X_1 + w_{22}X_2 + \dots + w_{k2}X_k)$$

$$Y = g(b_0 + w_{30}H_1 + w_{40}H_2)$$

여기서 H_1 과 H_2 는 은닉층의 은닉마디를 나타내며 X_i ($i = 1, 2, \dots, k$)는 각 설명변수 Y 는 반응변수인 오존을 나타낸다.

신경망분석과 같은 비선형모형은 일반적으로 수치해석적인 반복적 최적화방법을 사용하여 계수값을 추정한다. SAS E-miner 에서는 Levenberg-Marquardt, quasi-Newton, conjugate-gradient 등의 다양한 알고리즘이 있으나 본 연구에서는 오존과 같은 대기오염물질 농도를 예측하기에 적절한 방법으로 (Boznar 등, 1993) 알려진 역전파 (back-propagation) 알고리즘을 사용했다. 역전파 알고리즘은 다층의 구조를 갖으며 계수 추정을 위해 개발된 하강 알고리즘의 하나이다. 본 연구에서는 1개의 은닉층, 2개의 은닉층에 각각 3개의 은닉마디를 사용해 계수를 추정한 후, 결과를 비교해 RMSE 가 적은 은닉층을 선택하여 분석하였다. 모수 추정에는 MLE 방법을 사용하였다. 구체적인 내용은 김대수 (2005)를 참고하기 바란다. 신경망모형은 원래 대용량의 자료에 적합한 것으로 알려져 있다. 본 연구의 오존자료와 같이 자료가 크지 않은 자료를 신경망 모형으로 분석하는 이유는 예측이 뛰어난 신경망분석의 결과를 응용하기가 용이한 ARE 모형의 결과와 비교하고자 하는데 있다.

5. 파주 오존농도의 ARE 모형과 신경망모형

파주지역의 오존농도를 연도별, 월별로 box plot 그림 5.1 통해 살펴보았다. 먼저, 연도별로 살펴보면 오존농도 중앙값이 2006년도가 좀 높게 나타났고 사분위수 범위 (IQR)도 넓게 퍼져있다. 2006년을 제외한 각 년도마다 오존 농도 100ppb를 넘는 이상치가 3일~4일 정도 씩 나타났다. 오존 농도를 월별로 살펴보면, 5월, 6월이 7, 8, 9월에 비해 오존농도가 높게 나타났다. 최고온도가 5월, 6월이 7월, 8월, 9월 보다 낮음에도 불구하고, 오존 농도가 높게 나타났다.

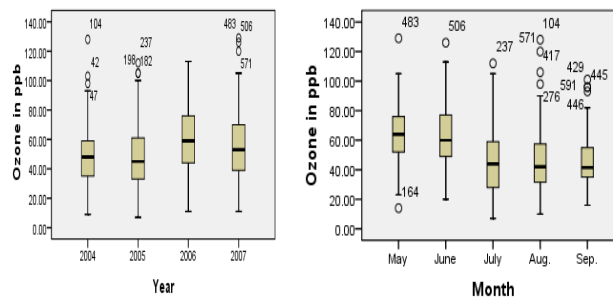


그림 5.1 파주 지역의 연도별, 월별 오존 농도의 box-plot

5.1. 전체 오존농도 자료에 관한 ARE 모형

파주지역의 오존농도 자료 분석을 위하여, 오존 전체 자료인 2003년부터 2007년의 5월-9월 자

료 609일을 사용하였다. ARE 모형의 변수 선택은 11종류의 설명변수들 간의 다중공선성 (multi-collinearity)를 고려하면서 설명변수들 간의 다양한 조합 모형을 분석한 후 RMSE 값이 작은 모형을 선택하였다. ARE 모형의 차수 선택은 표 3.1에 나타나 있듯이 대기변수인 SO_2 , NO_2 , CO, PM10은 전일 최대 농도를 선택하기 때문에 전일자료 ($t-1$)를 사용하였고, 기상변수인 최고온도, 풍속, 상대습도 등은 오존과 상관관계가 높은 시간대의 당일 자료(t)를 사용하였다.

ARE 모형의 모수 추정에는 YW (Yule-Walker), ULS (Unconditional Least Square), MLE (Maximum Likelihood Estimate) 등의 방법이 있지만, 본 연구에서는 오존 모형과 같이 과거시점의 종속변수 (lagged dependent variable)가 설명변수로 사용되는 경우에 적합한 MLE 방법을 사용하였다.

전체 오존자료에 관한 최종 ARE모형은 표 5.1과 같은 설명변수를 포함한 모형이 선택되었다. 표 5.1에서 Y_t 는 t 일 최고오존농도, Y_{t-1} = $t-1$ 일의 최고오존농도, N_{t-1} = $t-1$ 일의 최고 NO_2 농도, T_t = t 일의 최고온도, C_t = t 일의 9시-15시의 평균 운량 등이며 기타 변수에 관한 해석은 표 3.1을 참고하기 바란다. 모형의 결정계수 R^2 는 0.80으로 오존 농도 Y_t 는 Y_{t-1} , NO_2 , 최고온도, 운량에 의해 80% 설명되고 RMSE 는 16.54ppb로 나타났다. 설명변수 중 NO_2 , 최고온도의 계수는 양수로 나타나 NO_2 나 온도가 증가하면 오존도 증가하는 것으로 나타난 반면, 운량의 계수는 음수로 나타나 운량이 증가하면 오존은 감소하는 것으로 나타났다.

ARE모형의 적합성 검정으로는 오존모형과 같이 과거시점의 종속변수가 설명변수로 사용되는 경우에 적합한 더빈-h 검정을 사용해 검정하였고 또한 잔차의 ACF (Auto-Correlation Function)와 PACF (Partial Auto-Correlation Function)의 결과를 검정하였다 (조신섭 등, 1997 참고).

표 5.1 오존 전체농도의 ARE 모형

Period	ARE Model	R^2	RMSE
Full data (n=609 days)	$Y_t = 0.237Y_{t-1} + 0.277N_{t-1} + 1.779T_t - 2.418C_t + \epsilon_t$, $\epsilon_t = 0.168\epsilon_{t-2} + 0.118\epsilon_{t-6} + 0.156\epsilon_{t-8} + e_t$	0.8	16.54

5.2. 월별 오존 농도 자료에 관한 ARE 모형

그림 5.1의 상자그림에서 월별로 오존농도가 조금씩 다르게 나타나 월별로 오존농도 모형을 분석할 필요를 느껴 분석한 결과를 표 5.2에 나타냈다. 월별 오존분석은 2004년-2007년의 4개월 자료를 모두 모형구축기간으로 사용했다. 원래 검증기간을 고려하려 했으나 자료수가 적은 관계로 실시하지 못했다. 표 5.2를 보면, 오존을 설명하는 변수는 5월-9월 모두 전일 오존농도 (Y_{t-1}), 일 최고온도 (T_t)와 운량 (C_t)이 포함되었고, 최고온도의 계수는 양수로, 운량의 계수는 음수로 나타났다. 5월에는 전일 오존농도, 일 최고온도, 운량, 외에 전일 PM10농도 (P_{t-1})가 9월에는 전일 농도 (N_{t-1})가 오존 Y_t 를 설명하는 변수로 나타났다. 결정계수 R^2 는 모두 82%이상으로 설명력이 높게 나타났다.

위의 ARE 모형 결과를 예측 정확도가 높은 모형으로 알려진 신경망모형과 비교하였다. 신경망모형은 표 5.2의 ARE 모형과 같이 오존을 설명변수가 포함된 수식으로 나타내기가 어려워, 적합성 검정인 RMSE, NMSE, MAE 값들을 이용하여 두 모형을 비교한 결과를 표 5.3에 나타냈다. 5월 오존농도 ARE 모형의 RMSE가 13.22ppb로 NMSE는 0.384ppb 로 신경망모형의 RMSE 13.78ppb, NMSE 0.398ppb 보다 조금 작게 나타났으나 MAE 값은 ARE 모형이 10.46pp, 신경망모형이 8.75ppb로 나타났다.

반면에, 6월-9월에는 신경망모형의 RMSE 값, NMSE 값, MAE 값들이 모두 ARE모형의 값들보다 더 적게 나타나 신경망모형이 더 적절한 것으로 나타났다. 6월과 9월은 두 모형의 값들이 큰 차이가 없어 ARE모형도 적절한 것으로 나타났다. 7월과 8월에는 신경망모형의 적합성 값 들이 ARE 모형의 값

표 5.2 월별 (5월-9월) ARE 모형

Month	ARE Model	R ²
May	$Y_t = 0.189Y_{t-1} + 0.034P_{t-1} + 2.428T_t - 1.417C_t + \epsilon_t,$ $\epsilon_t = 0.210\epsilon_{t-2} + 0.271\epsilon_{t-20} + e_t$	0.90
June	$Y_t = 0.150Y_{t-1} + 2.471T_t - 2.131C_t + \epsilon_t,$ $\epsilon_t = 0.242\epsilon_{t-2} + e_t$	0.92
July	$Y_t = 0.240Y_{t-1} + 2.113T_t - 2.849C_t + \epsilon_t,$ $\epsilon_t = 0.239\epsilon_{t-15} + e_t$	0.82
August	$Y_t = 0.232Y_{t-1} + 1.806T_t - 2.592C_t + \epsilon_t,$ $\epsilon_t = 0.344\epsilon_{t-18} + e_t$	0.92
September	$Y_t = 0.151Y_{t-1} + 0.246N_{t-1} + 1.601T_t - 1.606C_t + \epsilon_t,$ $\epsilon_t = 0.321\epsilon_{t-22} + e_t$	0.89

표 5.3 월별 ARE 모형과 신경망모형의 적합성 검증

		May	June	July	August	September
RMSE	ARE	13.22	14.93	19.02	17.66	12.46
	NN	13.78	14.09	13.13	13.70	11.63
NMSE	ARE	0.384	0.451	1.071	1.016	0.707
	NN	0.398	0.429	0.739	0.790	0.656
MAE	ARE	10.46	11.55	15.32	13.22	9.37
	NN	8.75	8.10	8.06	8.06	7.14

들보다 비교적 많이 낮게 나타났다. 8월 오존농도 ARE 모형과 신경망모형의 실제값과 추정치의 산점도를 그림 5.2에 나타내었다. 신경망모형은 추정이 잘되었으나 ARE 모형은 과소추정되는 경향이 있었다. 그러나 ARE 모형은 오존에 영향을 주는 설명변수를 수식으로 표현하는 장점이 있으므로, 모형에 일사량 등의 오존에 영향을 주는 설명변수를 포함시켜, 더 개발하면 좋은 결과를 가질 수 있을 것으로 기대한다.

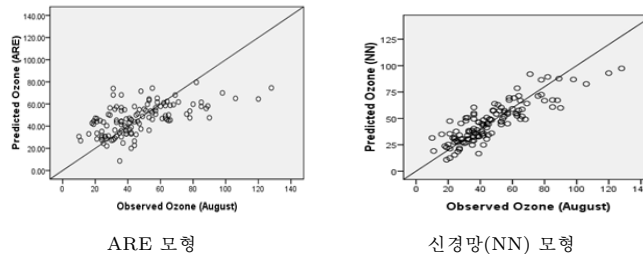


그림 5.2 과주 지역 8월 ARE 모형과 신경망모형의 실제값과 추정치의 산점도

6. 결론

경기 북부지역인 과주의 오존농도를 시계열 모형인 ARE 방법과 정확한 예측모형의 장점을 갖고 있는 신경망모형으로 분석, 비교하였다. ARE 모형과 신경망모형은 설명변수를 분석에 사용할 수 있는 모형으로 시계열자료를 적합 시킬 때 적절한 모형이다. 본 연구에서는, 오존농도에 영향을 주는 7종류의 기상자료와 4종류의 대기자료를 설명변수로 사용하였다. 7종류의 기상자료는 일 최고온도, 풍속, 상대습도, 강수량, 이슬점온도, 운량, 수증기압이며 4종류의 대기자료는 SO₂, NO₂, CO, PM10이다. 기상

자료 중 일사량은 오존에 영향을 많이 주는 변수로 알려져 있으나 자료의 부재로 사용하지 못한 아쉬움이 있다.

연구기간인 2004년부터 2007년의 5월1일에서 9월 30일까지의 자료를 사용하였다. 전체자료 ARE모형에서는 NO_2 농도, 일 최고온도, 운량에 의해 영향을 받는 것으로 분석되었다. 월별로 오존농도를 분석한 결과 5월-9월 모두 일 최고온도와 운량이 오존농도에 영향을 주는 것으로 나타났고, 일 최고온도와 운량이외에 5월에는 PM10 농도가 9월에는 NO_2 농도가 영향을 주는 것으로 분석되었다. ARE모형과 신경망모형의 적합도를 비교하면, 5월, 6월, 9월에는 두 모형의 적합성 값들이 거의 비슷해, 두 모형 다 적합하고, 7월과 8월은 신경망모형의 적합성 값이 많이 작게 나타나 신경망모형이 더 적합한 것으로 분석되었다, 그러나 ARE 모형이 정확도에 있어서는 신경망모형보다 떨어지지만, 표 5.2와 같이 오존에 어떤 설명변수들이 영향을 주는지 등의 수식으로 간단히 표현할 수 있는 장점이 있으므로, 본 연구에서 사용하지 못한 일사량 등의 설명변수 들을 첨가하여 분석하면 더 좋은 모형이 되리라 기대한다.

앞으로, 오존농도에 사용되는 설명변수에 관한 다양한 연구와 함께 다양한 모형에 관한 연구가 지역별로 활발히 이루어져야 된다고 생각된다.

참고문헌

- 김신도 (1998). <오존예보모델 및 예보시스템의 개선>, 오존예보시스템에 관한 전문가 토론회, 16-23.
- 김용국 (1994). 하계의 일 최고 오존농도 예측을 위한 신경망모델의 개발. <한국대기보존학회지>, **10**, 224-232.
- 김용준 (1997). 현업운영 가능한 서울지역의 일 최고 대기오염 예보모델 개발연구. <한국대기보존학회지>, **13**, 79-89.
- 김유근, 손건태, 문윤섭, 오인보 (1999). 서울지역의 지표오존농도 예보를 위한 전이함수모델 개발. <한국대기환경학회지>, **15**, 779-789.
- 조신섭, 이정형 (2007). <SAS/ETS를 이용한 경제시계열 분석>, 자유아카데미, 경기.
- Bauer, G., Deistler, M. and Scherrer, W. (2001). Time series models for short term forecasting of ozone in the eastern part of Austria. *Environmetrics*, **12**, 117-130.
- Camalier, L., Cox, W. and Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, **41**, 1-11.
- Feister, U. and Balzer, K. (1991). Surface ozone and meteorological predictors on a subregional scale. *Atmospheric Environment*, **25**, 1781-1790.
- Guttorp, P., Meiring, W. and Samson, D. P. (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, **5**, 241-254.
- Hubbard, M. and Cobourn, W. (1998). Development of a regression model to forecast ground-level ozone concentration in Louisville. KY, *Atmospheric Environment*, **32**, 2637-2647.
- Jorquera, H., Perez, R., Cipriano, A., Espejo, A., Letelier, M. V. and Acuna, G. (1998). Forecasting ozone daily maximum levels at Santiago, Chile. *Atmospheric Environment*, **32**, 3415-3424.
- Lee, H. (2008). Analysis of time series models for ozone concentrations at the Uijeongbu city in Korea. *Journal of the Korean Data & Information Science Society*, **19**, 1153-1164.
- Park, C. and Kim, H. (2009). Prediction ozone warning days based on optimal time series model. *Journal of the Korean Data & Information Science Society*, **20**, 293-299.
- Smith, L. S. and Huang, L. S. (1993). *Modeling the threshold exceedence of urban ozone*, Technical Report No.6, National Institute for Statistical Science, Research Triangle Park, NC.

Analysis of statistical models for ozone concentrations at the Paju city in Korea[†]

Hoonja Lee¹

Department of Information Statistics, Pyeongtaek University

Received 20 August 2009, revised 13 November 2009, accepted 21 November 2009

Abstract

The ozone data is one of the important environmental data for measurement of the atmospheric condition of the country. In this article, the Autoregressive Error (ARE) model and Neural Networks (NN) model have been considered for analyzing the ozone data at the northern part of the Gyeonggi-Do, Paju monitoring site in Korea. In the both ARE model and NN model, seven meteorological variables and four pollution variables are used as the explanatory variables for the ozone data set. The seven meteorological variables are daily maximum temperature, wind speed, relative humidity, rainfall, dew point temperature, steam pressure, and amount of cloud. The four air pollution explanatory variables are Sulfur dioxide (SO_2), Nitrogen dioxide (NO_2), Cobalt (CO), and Promethium 10 (PM10). The result showed that the NN model is generally better suited for describing the ozone concentration than the ARE model. However, the ARE model will be expected also good when we add the explanatory variables in the model.

Keywords: Autoregressive error model, explanatory variable, meteorological variables, neural networks model, ozone concentration, pollution variable.

[†] This paper was supported by the Research Fund, 2008, Pyeongtaek University in Korea.

¹ Associate Professor, Department of Information Statistics, Pyeongtaek University, Pyeongtaek 450-701, Korea. E-mail: esther@ptu.ac.kr