Journal of the Korean Data & Information Science Society
2009, **20**(2), 293–299

한국데이터정보과학회지

최적 시계열 모형에 기초한 오존주의보 날짜 예측

박철용[1] · 김현일[2]

[12]계명대학교 통계학과

접수 2009년 1월 6일, 수정 2009년 2월 6일, 게재확정 2009년 2월 25일

요 약

이 논문에서는 대구 두 개 동의 시간별 오존농도를 예측하는 모형으로 회귀, 자기회귀누적이동평균, 자기회귀누적이동평균 오차를 가지는 회귀 같은 선형모형들을 고려하였다. 평균제곱오차제곱근에 근거하여 보았을 때 한 개 동에서는 자기회귀누적이동평균 모형이 최적의 모형으로 선택되었고, 다른 동에서는 자기회귀누적이동평균 오차를 가지는 회귀 모형이 최적 모형으로 선택되었다. 이 최적의 모형으로부터 나온 잔차들의 변동석 분석을 수행하였는데 이를 통해 120 ppb를 넘는 오존 주의보 날짜를 예측하였다. 2000년에서 2003년까지의 훈련용 자료에 근거하여 보았을 때 잔차값의 경계값으로 35 ppb를 잡았을 때 오존주의보 날짜를 예측하는데 좋은 결과를 보였다. 하나의 동에서는 2004년의 오존주의보가 발령된 이틀 중 하루와 나머지 주의보가 발령되지 않은 364일을 모두 정확히 예측하였다. 다른 동에서는 2004년의 오존주의보가 발령된 하루와 주의보가 발령되지 않은 365일을 모두 정확히 예측하였다.

주요용어: 시계열 모형, 오존주의보, 자기회귀누적이동평균, 회귀.

1. 서론

오존은 2차 오염물질로서 일사량이 강한 계절이나 낮에 고농도를 나타낸다. 고농도 오존에 대한 오존경보는 환경부에서 대기환경보전법에 의한 '오존오염경보 및 예보제'를 서울지역을 시작으로 실시하였다. 매년 대상지역을 확대하여 현재는 인천지역을 포함한 전국 6대 도시 및 경기, 충북 등에서 오존경보제를 시행 중에 있다. 오존경보는 오염경보제의 일종으로 주의보, 경보 및 중대경보로 나뉜다. 오존농도가 120 ppb (혹은 0.12 ppm) 이상일 때는 주의보를 발령하는데 눈과 코를 자극, 불안감과 두통을 유발하며 호흡수를 증가시킨다. 오존농도가 300 ppb 이상일 때는 경보를 발령하는데 호흡기의 자극, 가슴압박 및 시력감소를 일으킨다. 오존농도가 500 ppb 이상일 때는 중대경보를 발령하는데 폐기능 저하, 기관지 자극 및 폐혈증 등의 인체영향을 미치게 된다.

오존농도 예측과 관련하여 통계 기법을 적용한 국내외 연구로는 다변량 통계분석 기법을 사용한 오존농도 예측 (허정숙과 김동술, 1993), 회귀모형 기법을 사용한 오존 농도 예측 (김용준, 1997; 최성우, 2002; Hubbard 등, 1998), 데이터 마이닝의 신경망 모형을 사용한 오존농도 예측 (김용국과 이정범, 1996; Acuna 등, 1996; Yi 등, 1996), 판별분석 기법을 사용한 오존 고농도일 예측 (박옥현, 1984), 다중회귀모형 기법을 사용한 오존농도 예측 (이기원 등, 1993), 전이함수모델을 사용한 오존농도 예측 (김유근 등, 1999), 일반화가법모형 (generalized additive model)을 이용한 오존 농도 예측 (Niu, 1996),

[1] 교신저자: (704-701) 대구 달서구 신당동 1000번지, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr
[2] (704-701) 대구 달서구 신당동 1000번지, 계명대학교 통계학과, 석사졸업생.

데이터마이닝과 시계열 모형을 결합한 합성모형을 사용한 오존농도 예측 (Kim과 Park, 2008), 자기상관 오차를 가지는 회귀모형을 이용한 일 최고 오존농도 예측 (Lee, 2008) 등이 있었다. 이와 더불어 오존 농도 예측을 위한 여러 통계적 모형을 비교 연구하는 시도 (Robeson 등, 1990; Jorquera 등, 1998)도 있었다.

이 연구에서는 먼저 대구지역 시간별 오존농도를 예측할 수 있는 모형을 찾고자 한다. 앞에서 소개된 여러 가지 통계적인 방법이 사용될 수 있겠지만 이 연구에서는 상대적으로 쉽게 접근할 수 있는 선형모형들인 회귀, ARIMA 및 회귀+ARIMA 모형을 고려하였다. 여기서 회귀+ARIMA 모형은 ARIMA 오차를 가지는 회귀모형을 나타내는 표기법이다. 최적 모형의 선택 방법으로는 여러 가지 기준이 존재하지만 (Hyndman과 Koehler, 2006) 이 연구에서는 RASE(평균제곱오차제곱근; root average squared error)를 사용하였다. 그런데 이렇게 구한 최적의 모형이라도 오존농도의 급격한 상승기에는 예측값이 충분히 실제값에 미치지 못하는 경우가 허다하였다. 따라서 이 연구에서는 추가적으로 최적 예측모형으로부터 얻은 잔차의 변동성분석을 통해 오존농도가 향후 몇 시간 후에 오존주의보 기준 이상이 되는지를 예측할 수 있는 방법을 찾아보고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 이 연구에서 사용하는 자료에 대한 설명과 함께 자료분석의 결과를 제시한다. 3절에서는 이 연구의 요약, 결론 및 토의를 제시한다.

2. 자료분석

2.1. 자료 설명

본 연구에 사용된 자료는 대구광역시의 수창동, 만촌동의 2000년부터 2004년까지 시간별로 관측된 총 87,696개의 대기오염 자료와 기상 자료이다. 반응변수는 대기오염 자료 중의 하나인 오존이며, 설명변수는 반응변수에 영향을 미치는 오존을 제외한 나머지 대기오염 자료 및 기상 자료이다. 표 2.1에 원자료에 대한 간단한 설명이 주어져 있다.

표 2.1 대기오염 자료와 기상자료의 변수설명

구분	변수명	설명
대기오염 자료	O3	오존(단위: ppb)
	SO2	이산화황(단위: ppb)
	CO	일산화탄소(단위: ppb)
	PM10	미세먼지(단위: ㎍/㎥)
기상 자료	TEMP	온도(단위: 0.1℃)
	WINDDIR	바람 방향(1:동, 2:남, 3:서, 4:북)
	WINDSPD	바람 속도(단위: 0.1m/s)
	HUMID	습도(단위: %)
	CLOUD	전운량(단위: 1/10)

최적의 오존농도 예측 모형을 찾기 위해 각 동마다 훈련용 자료(training data)로 2000년부터 2003년까지의 35,064개의 자료(80%)를, 평가용 자료(validation data)로 2004년의 8784개의 자료(20%)를 잡았다.

2.2. 최종 예측모형 선택

서론에서도 잠시 언급하였듯이 이 연구에서는 비교적 쉽게 접근할 수 있는 대표적 선형모형인 회귀, ARIMA(자기회귀누적이동평균; autoregressive integrated moving average) 및 회귀+ARIMA 모형을

고려하였다. 기존 연구에서 개발된 여러 가지 방법대신 선형모형을 사용한 이유는 모형 적합이 쉬우며 모형 적합 후에 수행되는 변동성 분석을 용이하게 할 수 있는 장점이 있었기 때문이다.

회귀모형은 표 2.1에 주어진 설명변수를 이용하여 반응변수 O3를 설명하는 선형모형을 설정하는데 오차는 서로 상관이 없다고 가정하는 모형이며, ARIMA 모형은 이 설명변수에 의한 효과를 고려하지 않고 O3 시계열의 상관성만 고려한 모형이다. 그리고 회귀+ARIMA 모형은 설명변수를 이용하여 반응변수 O3를 설명하는 회귀모형을 설정하되 오차가 ARIMA 모형을 따르는 ARIMA 오차를 가지는 회귀모형을 나타내는 표기법이다.

최종 예측모형을 선택하는 과정은 다음과 같다. 먼저, 각 동의 2000년부터 2003년까지의 시간별 훈련용 자료를 이용하여 회귀, ARIMA 및 회귀+ARIMA 적합 모형을 찾는다. 그 다음 훈련용 자료에 적합된 각 모형을 2004년의 시간별 평가용 자료에 적용하여 나온 잔차의 RASE(평균제곱오차제곱근; root average squared error)를 계산하여 이 값이 적은 모형을 최종 예측모형으로 선택한다.

각 선형모형의 적합 모형을 찾는 과정은 다음과 같다. 회귀모형은 최소제곱법에 의해 적합 모형이 간단히 계산되며, ARIMA, 회귀+ARIMA 모형은 시차 50까지의 자기상관과 부분자기상관의 절대값이 모두 0.02보다 작게 되는 모형을 적합 모형으로 선택하였다. 이는 훈련용 자료의 크기가 35,000 정도로 너무 커서 자기상관이 0.01 정도면 신뢰구간을 벗어나는 어려움이 있어 표본크기 10,000 정도에 해당되는 기준값을 사용한 것이다.

수창동, 만촌동에서 각각 ARIMA, 회귀, 회귀+ARIMA 모형을 적합시키고 RASE에 의해 비교한 결과가 표 2.2에 주어져 있다.

표 2.2 세 가지 모형의 RASE 비교

동	자료	회귀	ARIMA	회귀+ARIMA
수창동	훈련용	10.94	4.31	4.15
	평가용	12.52	4.54	4.80
만촌동	훈련용	11.67	4.52	4.33
	평가용	12.28	5.13	5.04

표 2.2에 의하면 훈련용에서는 수창동, 만촌동 모두 회귀+ARIMA 모형의 RASE 값이 각각 4.15, 4.33으로 최소값을 보인다. 그러나 평가용에서는 수창동의 ARIMA모형의 RASE 값이 4.54로 최소값을 보여 최종 예측모형으로 선택되었으며, 만촌동의 회귀+ARIMA 모형의 RASE 값이 5.04로 최소값을 보여 최종 예측모형으로 선정되었다. 이 최종 예측모형의 RASE값에서 전반적으로 예측 오차가 그리 크지 않다는 것을 알 수 있다.

수창동의 최종 예측모형으로 선택된 ARIMA 모형의 모형식은 다음과 같다.

$$O_t = [\phi(B)\Phi(B)]^{-1}\Theta(B)e_t$$

단, 여기서

$$\begin{aligned}\phi(B) &= (1 - 1.096B + 0.236B^2 - 0.038B^3 - 0.018B^5 - 0.014B^{19} - 0.012B^{20} \\ &\quad + 0.026B^{26} + 0.027B^{23} - 0.014B^{24} + 0.004B^{25} - 0.018B^{26}), \\ \Phi(B) &= (1 - 0.993B^{24}), \quad \Theta(B) = (1 - 0.950B^{24})\end{aligned}$$

이다.

마찬가지로 만촌동의 최종 예측모형으로 선택된 회귀+ARIMA 모형의 모형식은 다음과 같다.

$$O_t = \hat{X}_t + [\phi(B)\Phi(B)]^{-1}\Theta(B)e_t.$$

단, 여기서

$$\begin{aligned}\hat{X}_t &= 26.088 - 0.374 \times \text{이산화황} - 0.607 \times \text{미세먼지} \\ &\quad + 0.052 \times \text{온도} - 0.073 \times \text{풍향3} + 0.040 \times \text{풍속} - 0.127 \times \text{습도}, \\ \phi(B) &= (1 - 1.066B + 0.200B^2 - 0.010B^3 - 0.030B^{20} - 0.043B^{22} - 0.001B^{24} \\ &\quad - 0.008B^{25} + 0.023B^{26} + 0.012B^{27} - 0.003B^{34} - 0.028B^{46} + 0.002B^{48}), \\ \Phi(B) &= (1 - 0.992B^{24}), \quad \Theta(B) = (1 - 0.957B^{24})\end{aligned}$$

이다. 여기서 풍향1, 풍향2, 풍향3은 각각 풍향이 1(동), 2(남), 3(서)이면 1, 풍향이 4(북)이면 -1, 그 외는 0으로서 풍향 동, 남, 서와 북쪽 방향을 비교하는 편차 방식의 가변수이다.

2.3. 변동성 분석

실제 오존농도가 주의보수준인 120 ppb를 넘고 나서 오존주의보를 발령하는 것은 오존의 위험성에 대처할 수 있는 시간이 부족하여 실효성이 떨어진다. 오존주의보가 실효성을 가지기 위해서는 최소한 한 시간 전에 오존주의보를 예측할 수 있어야 할 것이다.

앞에서 구한 최종 예측모형을 이용하여 한 시간 후의 오존 예측값으로 오존주의보를 발령할 수도 있을 것이다. 그러나 이 최적의 모형들도 오존의 급격한 상승기에는 오존 예측값이 실제 오존값에 못 미치는 경향이 나타나서 오존주의보를 예측하기에는 다소 미흡한 결과가 나타났다.

따라서 최종 오존농도 예측 모형의 잔차를 분석하여 이 잔차의 변동성이 커지는 점에서 오존주의보를 발령하는 방법을 추가로 도입하기로 하였다. 실제로 여러 개의 잔차 경계값을 시도해 본 결과 경계값으로 35 ppb를 사용할 때 분류정확도가 좋은 결과를 나타내어 이것을 최종 경계값으로 결정하였다. 두 개 동에서 이 경계값을 사용하였을 때 오존주의보 발령에 따른 오분류표를 정리한 것이 표 2.3과 표 2.4에 주어져 있다.

표 2.3 수창동의 오존주의보 오분류표

구분	실제	예측		계
		주의보	미주의보	
훈련용	주의보	1	0	1
	미주의보	1	1459	1460
	계	2	1459	1461
평가용	주의보	1	1	2
	미주의보	0	364	364
	계	1	365	366

수창동의 훈련용 결과를 먼저 살펴보기로 한다. 오존주의보 발령 기준인 오존이 120 ppb를 넘는 날이 하루(146 ppb) 존재했는데 우리의 모형으로 예측할 수 있었으며, 따라서 민감도가 1이 되었다. 오존주의보가 발령되지 않은 나머지 1460일 중에는 110 ppb인 하루에 우리 모형으로 오존주의보를 예측했기 때문에 특이도가 1459/1460이 되었다. 수창동의 평가용 경우에는 이틀(127 ppb, 121 ppb) 오존주의보가 발령되었으나, 우리의 모형으로 127 ppb는 제대로 예측하였으나, 121 ppb는 잔차값이 30 ppb 정도 나와 오존주의보를 제대로 예측하지 못해 민감도가 0.5가 되었다. 또한 오존주의보가 발령되지 않은 나머지 364일 중에서 우리 모형으로 오존주의보가 예측된 날이 없기 때문에 특이도는 1이 되었다.

만촌동의 경우도 훈련용 결과를 먼저 살펴보기로 한다. 오존주의보가 발령된 날이 하루(152 ppb) 있었는데, 우리 모형으로 이 날 오존주의보를 예측하여 민감성이 1이 되었다. 오존주의보가 발령되지 않

표 2.4 만촌동의 오존주의보 오분류표

구분	실제	예측		계
		주의보	미주의보	
훈련용	주의보	1	0	1
	미주의보	1	1459	1460
	계	2	1459	1461
평가용	주의보	1	0	1
	미주의보	0	365	365
	계	1	365	366

은 1460일 중 103 ppb인 하루에 우리 모형으로 오존주의보를 예측하여, 특이도가 1459/1460이 되었다. 만촌동 평가용의 경우 실제 오존주의보가 발령된 날이 하루(125 ppb) 있었는데, 우리 모형으로 이 날에만 오존주의보를 정확히 예측하여 민감도와 특이도가 모두 1로 나타났다.

3. 결론 및 토의

이 연구에서는 먼저 대구지역 시간별 오존농도를 예측할 수 있는 회귀, ARIMA 및 회귀+ARIMA 모형을 고려한 후 그 중 최적의 오존농도 예측모형을 찾았다. 최적 모형의 선택 방법으로는 RASE(root average squared error)를 사용하였는데 수창동에서는 ARIMA, 만촌동에서는 회귀+ARIMA 모형이 최적의 모형으로 선택되었다. 그런데 이 최적의 모형이라도 오존농도의 급격한 상승기에는 예측값이 충분히 실제값에 미치지 못하는 경우가 발생하여, 추가적으로 최적 예측모형으로부터 나온 잔차의 변동성 분석을 통해 오존농도가 향후 몇 시간 후에 오존주의보 기준 이상이 되는지 예측할 수 있는 방법을 제시하였다. 구체적인 방법은 어떤 잔차가 35 ppb (혹은 0.035 ppm) 이상일 때 오존주의보를 예측하는 것이다.

이 오존주의보 예측 방법에 의한 예측결과를 요약하면 다음과 같다. 수창동 훈련용의 경우 오존주의보가 발령된 하루를 우리의 모형으로 정확히 예측하여 민감도가 1이 되었다. 오존주의보가 발령되지 않은 나머지 1460일 중에는 하루에 우리 모형으로 오존주의보를 잘 못 예측했기 때문에 특이도가 1459/1460이 되었다. 수창동 평가용의 경우에는 이틀 오존주의보가 발령되었으나, 우리의 모형으로 하루는 제대로 예측하였으나 다른 하루는 오존주의보를 예측하지 못해 민감도가 0.5가 되었다. 또한 오존주의보가 발령되지 않은 나머지 364일 중에서 우리 모형으로 오존주의보가 예측된 날이 없기 때문에 특이도가 1이 나왔다.

만촌동 훈련용의 경우 오존주의보가 발령된 날이 하루 있었는데, 우리 모형으로 이 날 오존주의보를 정확히 예측하여 민감성이 1이 되었다. 오존주의보가 발령되지 않은 1460일 중 하루에 우리 모형으로 오존주의보를 예측하여 특이도가 1459/1460이 되었다. 만촌동 평가용의 경우 실제 오존주의보가 발령된 날이 하루 있었는데, 우리 모형으로 이 날에만 오존주의보를 정확히 예측하여 민감도와 특이도가 모두 1로 나타났다.

따라서 이 연구에서 제시한 오존주의보 예측 방법은 특이성이 거의 1에 가까워 오존주의보가 발령되지 않은 날들을 거의 100% 예측하였다. 더욱 중요한 민감성의 경우 1인 경우가 세 번, 0.5인 경우가 한 번 나타나 전체적으로 결합한 민감도가 4/5=0.8로 나타나 좋은 예측력을 보였다.

이 연구에서 제시된 결과는 수창동, 만촌동의 2000년부터 2004년까지의 시간별 시계열 자료에 한정된 결과이기 때문에 일반적으로 확대해석하는 것은 경계해야 할 것이라 생각된다. 특히 오존주의보가 발령된 날짜가 그리 많지 않았고 잔차분석을 통해 경계값인 35 ppb를 결정하였기 때문에 일반적 확대해

석에 신중하여야 할 것이다.

지금 잠정적으로 예측력을 높일 수 있는 방안으로 생각할 수 있는 것은 다음과 같다. 먼저, 이 연구에서 사용된 대기 자료가 수창동, 만촌동의 자료가 아닌 대구광역시 전체의 자료여서 각 동의 특성을 정확하게 반영할 수 없는 단점이 있다. 이런 이유로 ARIMA에 비해 회귀모형에서의 예측력이 떨어진 이유가 되었으리라 짐작된다. 또한 오존농도에 영향을 많이 주는 일사량과 NO, NO2변수를 사용하지 못하였는데 향후 분석에서 이 변수들을 포함시킨다면 예측력을 높일 수 있으리라 생각한다.

참고문헌

김용국, 이종범 (1996). 하계의 일최고 오존농도 예측을 위한 신경망 모델의 개발. <한국대기보전학회지>, **10**, 224-232.

김용준 (1997). 현업 운영 가능한 서울지역의 일 최고 대기오염도 예보모델 개발 연구. <한국대기보전학회지>, **13**, 79-89.

김유근, 손건태, 문윤섭, 오인보 (1999). 서울지역의 지표오존농도 예보를 위한 전이함수모델 개발. <한국대기환경학회지>, **15**, 779-789.

박옥현 (1984). 간단한 대기확산 모델과 통계학적 방법을 병용한 도시 대기오염의 예측. <대한환경공학회지>, **6**, 2001-2012.

이기원, 권숙표, 정용 (1993). 서울시 대기 중 오존오염도의 연도별 변화와 그 영향인자 분석. <한국대기보전학회지>, **9**, 107-115.

최성우 (2002). 다중회귀분석을 통한 대구지역 오존농도 예측. <한국환경과학회지>, **11**, 687-696.

허정숙, 김동술 (1993). 다변량 통계분석을 이용한 서울시 고농도 오존의 예측에 관한 연구. <한국대기보전학회지>, **9**, 207-215.

Acuna, G., Jorquera, H. and Perez, R. (1996). Neural network model for maximum ozone concentration prediction. *Lecture Notes in Computer Science*, **1112**, 263-268, Springer, New York.

Hubbard, M. C. and Cobourn, W. G. (1998). Deveolopment of a regression model to forecast ground-level ozone concentration in Louiville, KY, U.S.A. *Atmospheric Environment*, **32**, 2637-2647.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22**, 679-688.

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.

Kim, H. and Park, C. (2008). An optimal hybrid model for predicting hourly ozone concentration level. *Journal of the Korean Data & Information Science Society*, **19**, 209-217.

Lee, H. (2008) Analysis of time series models for ozone concentrations at the Uijeongbu city in Korea. *Journal of the Korean Data & Information Science*, **19**, 1153-1164.

Niu, X.-F. (1996). Nonlinear additive models for environment time series, with application to ground-level ozone data analysis. *Journal of the American Statistical Association*, **91**, 1310-1321.

Robeson, S. M. and Steyn, D. G. (1990). Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment*, **24B**, 303-312.

Yi, J. and Prybutok, V. R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialised urban area. *Environmental Pollution*, **92**, 349-357.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 293–299

한국데이터정보과학회지

Predicting ozone warning days based on an optimal time series model

Cheolyong Park[1] · Hyunil Kim[2]

[12]Department of Statistics, Keimyung University

Received 6 January 2009, revised 6 February 2009, accepted 25 February 2009

Abstract

In this article, we consider linear models such as regression, ARIMA (autoregressive integrated moving average), and regression+ARIMA (regression with ARIMA errors) for predicting hourly ozone concentration level in two areas of Daegu. Based on RASE(root average squared error), it is shown that the ARIMA is the best model in one area and that the regression+ARIMA model is the best in the other area. We further analyze the residuals from the optimal models, so that we might predict the ozone warning days where at least one of the hourly ozone concentration levels is over 120 ppb. Based on the training data in the years from 2000 to 2003, it is found that 35 ppb is a good cutoff value of residulas for predicting the ozone warning days. In one area of Daegu, our method predicts correctly one of two ozone warning days of 2004 as well as all of the remaining 364 non-warning days. In the other area, our methods predicts correctly all of one ozone warning days and 365 non-warning days of 2004.

Keywords: ARIMA, ozone warning, regression, time series model.

[1] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr
[2] Master of Science, Department of Statistics, Keimyung University, Daegu 704-701, Korea.