SARIMA모형을 이용한 대기 중 오존농도 예측 모델 구축

강정구*, 박석천**, 김종현***
*가천대학교 일반대학원 모바일소프트웨어학과
**가천대학교 컴퓨터공학과 정교수(교신저자)
***위세아이텍 대표이사
e-mail:webclub09@gmail.com

Implementation of Ozone Concentration Prediction Model Using SARIMA Model in Atmospheric

Jung-Ku Kang*, Seok-Cheon Park**, Jong-Hyun Kim***
*Dept. of Mobile Software, Gachon University
**Dept. of Computer Engineering, Gachon University(Corresponding Author)
***Representative Director, WISEITECH co., ltd

요 약

우리나라는 지난 40년간 급속한 경제 성장의 과정에서 에너지 소비가 급증하고 있으며, 이로 인해 온실가스 배출량은 1990년~2005년 사이 두 배 이상 증가하였고, 이는 OECD 국가 중 가장 높은 증가율이다. 2차 오염물질인 오존은 1990년부터 2012년까지 연평균 3% 상승하고 있으며, 반복 노출 시 폐에 피해를 줄 수 있는 오염 물질로 예방 대책이 필요하다. 이를 위해 본 논문에서는 계절성 특성을 지닌 오존농도 시계열 데이터를 바탕으로 SARIMA 모형을 활용하여 예측 모형을 구축 하였다.

1. 서론

우리나라는 지난 40년간 급속한 경제 성장의 과정에서 에너지 소비가 급증하였으며, 총 에너지 소비는 2000년~2012년 기간 중 연평균 3.1% 증가하고 있는 추세이다. 이로 인해 우리나라 온실가스 배출량은 1990년~2005년 사이 두 배 이상 증가하고 있다. 이는 OECD 국가 중 가장 높은 증가율이다[1].

현재 우리나라에서 기준성 오염물질은 미세먼지(PM_{10}), 오존(O_3), 이산화황(SO_2), 이산화질소(NO_2), 일산화탄소(CO), 납, 벤젠 등 7개 항목을 꼽고 있으며[2], 이 중 오존은 반복 노출 시 인간의 건강과 식물에 피해를 주며, 인간의 삶에 악영향을 미치는 것으로 잘 알려져 있다.

오존은 주로 대기 중 배기가스와 공장에서 배출되는 오염물질이 강한 태양광선으로 인해 광화학반응을 일으켜 발생하는 2차 오염물질로, 1990년부터 2012년까지 연평균 3% 상승하고 있으며, 오존주의보 발령 가능성이 꾸준히 증가하고 있다. 따라서 오존 농도 예측을 통해 오존의 피해에 대비가 필요하다.

또한 오존은 기온이 높고, 일사량이 많으며, 풍속이 약한 5월~6월, 오후 2시~5시에 높은 수준으로 형성 되어 뚜렷한 계절성의 특성을 가지고 있다고 판단 할 수 있다.

따라서 본 논문에서는 시계열 데이터 분석 기법 중 하나인 ARIMA모형에 계절성을 부여한 SARIMA모형을 통해 대기 중 오존농도 예측 모형을 구축하고자 한다.

2. 관련연구

2.1 오존

오존은 대기 중에 배출된 NOx와 휘발성 유기화합물 등이 자외선과 광화학 반응을 일으켜 생성된 PAN, 알데하이드, Acrolein 등의 광화학 산화물의 일종으로 2차 오염물질에 속한다. 전구물질인 휘발성 유기화합물은 자동차, 화학공장, 정유공장과 같은 산업시설과 자연적 생성 등 다양한 배출원에서 발생한다.

오존에 반복 노출 시에는 폐에 피해를 줄 수 있는데, 가슴의 통증, 기침, 메스꺼움, 목 자극, 소화 등에 영향을 미치며, 기관지염, 심장질환, 폐기종 및 천식을 악화시키고, 폐활량을 감소시킬 수 있다. 특히 기관지 천식 환자나 호흡기 질환자, 어린이, 노약자 등에게는 많은 영향을 미치므로 주의해야 할 필요가 있다. 또한 농작물과 식물에 직접적으로 영향을 미쳐 수확량이 감소되기도 하며 잎이 말라죽기도 한다.

오존은 대기 중 2차 오염물질인 2차 유기탄소 에어로졸 입자, 황산염, 질산염 입자 등을 생성하는 데 직접적으로 기여하는 물질이므로 미세입자 제어 전략 수립을 위해서는 반드시 제어해야 하는 물질이다. 그러나 오존 생성의 전구물질 제어도 중요하지만 오존은 장거리 이동이 가능한 물질이므로 외부로부터 해당 지역에 유입되는 부분에 대해서는 해당 지방자치단체 또는 정부에서 해결이 어려운 부분이 존재한다[3].

2.2 시계열 모형

두 변수 간에 성립되는 인과관계 또는 함수관계를 규명하는데 주된 목적이 있는 회귀분석 방안은 어떤 경제현상을 특수 함수관계로 파악하고 또 성공적으로 그 관계를 추정할 경우 그 추정 결과에 근거하여 장래에 대한 예측을 시행한다. 반면, 시계열 분석방법론은 인과관계에 관한 어떠한 이론적 배경이 없이도 예측을 수행할 수 있는 방법론으로서, 한 변수의 미래 값을 단지 자신의 과거 관측치 값에 근거하여 파악할 수 있는 방법이다. 이는 미래 예측을 주 관심사로 볼 때 오히려 회귀분석방법보다 쉽게 활용할 수 있다.

시계열 분석은 기본적으로 한 변수의 변동 내용을 장기적 추세변동(Secular Trend)과 주기적 변동(Cyclical Variation) 및 계절적 변동(Seasonal Variation) 그리고 불규칙 변동 등으로 구성되어 있는 것으로 보고 이를 분해하여 추정하는 방법을 사용한다.

대표적인 시계열 분석방법론에는 자기회귀(AR : Auto Regressive)모형과 이동평균(MA : Moving Average)모형 그리고 자기회귀 이동평균(ARMA : Auto Regressive Moving Average)모형과 ARIMA(Auto Regressive Integrated Moving Average)모형 그리고 SARIMA(Seasonal Auto Regressive Integrated Moving Average)모형 등이 있다[4].

2.3 SARIMA모형

일반적인 적분된 자기회귀 이동평균(Autoregressive Integrated Moving Average, ARIMA) 모형은 (식 1)과 같이 정의된다.

$$\phi_p(L)(1-L)^d_{y_t} = \theta_q(L)u_t \qquad (1)$$

(식 1)은 ARIMA(p, d, q)로 정의되며 p는 자귀 회귀(Autoregressive, AR)항의 차수, q는 이동평균(Moving Average, MA)항의 차수, d는 단위근의 개수를 의미한다. ARIMA모형은 일반적으로 불안정적인 시계열을 대상으로 하는 시계열 모형으로 단위근을 감안하는 모형이다. $\phi_p(L)$과 $\theta_q(L)$은 자기회귀항(AR tern) 및 이동평균항(MA tern)의 래그 다항식을 의미한다. 대표적인 사례로서 ARIMA(1, 1, 1) 모형은 p, d, q의 차수가 모두 1인 경우이다. ARIMA(1, 1, 1)의 수식은 (식 2)와 같다.

$$(1-\phi_1 L)(1-L)_{y_t} = (1+\theta_1 L)u_t \qquad (2)$$

ARIMA 모형은 특정 연도의 연속적인 월별 자료 간의 특성을 고려하는 모형이며 SARIMA 모형은 이외에도 연속 연도의 동일 시점(월) 자료의 특성을 모두 고려하는 모형이다. SARIMA 모형은 크게 두 부분으로 정의되며 첫 번째 부분은 ARIMA 부분으로 (식 1)과 같다. 예를 들어, 첫 번째 부분은 2013년 12월 오존농도는 2013년 11월 자료와 관계가 있음을 의미한다. 계절성을 고려한 부분은 예를 불면 2013년 5월 오존 농도는 2013년 5월 오존농도와 관련 있음을 의미하며 (식 3)과 같이 정의된다.

$$\Phi_p(L)(1-L^s)^D_{y_t} = \Theta_Q(L)u_t \qquad (3)$$

이때 s는 만약 분기별 자료라면 4가 될 것이고, 월별 자료라면 12가 될 것이다. 또한 D는 계절 단위근의 차수, P는 계절 자기회귀항의 차수, Q는 계절 이동평균항의 차수이다. 또한 $\Phi_P(L)$와 $\Theta_Q(L)$은 각각 계절 자기회귀항(SAR tern)과 계절 이동평균항(SMA tern)의 다항식을 의미한다.

최종적으로 첫 번째 부분과 두 번째 부분을 모두 함께 고려하면 (식 4)와 같이 정의된다.

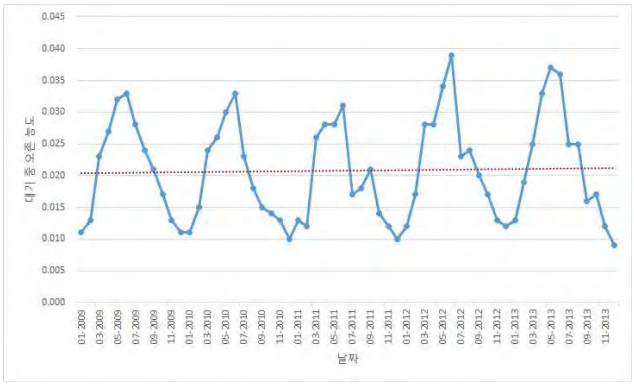
$$\phi_p(L)\Phi_P(L)(1-L)^d(1-L^s)^D y_t = \theta_q(L)\Theta_Q(L)u_t \qquad (4)$$

(식 4)은 SARIMA 모형으로 차수는 (p, d, q)×(P, D, Q)로 정의된다[5].

3. SARIMA모형 구축 및 검증

3.1 자료수집

본 연구에서 사용한 자료는 서울시의 2009년부터 2013년까지 총 5년간의 월평균 대기 중 오존농도 자료이다. 2009년 1월부터 2013년 12월까지의 오존 농도 시계열 그래프는 (그림 1)과 같다.



(그림 1) 오존농도 시계열 그래프

3.2 정상성 및 계절성 존재여부 점검

대기 중 오존농도 시계열 그래프인 (그림 1)을 살펴보면 평균 참조선을 기준으로 일정한 패턴을 보이고 있기 때문에 정상성임을 알 수 있다. 따라서 비계절적인 차분을 실시하지 않고 계절성 존재 여부 판단 한다.

관찰되는 이 시계열은 뚜렷한 계절성 패턴을 가지고 있다. 따라서 계절성 차분(D=1)이 필요함을 알 수 있다.

3.3 모형의 식별 및 추정

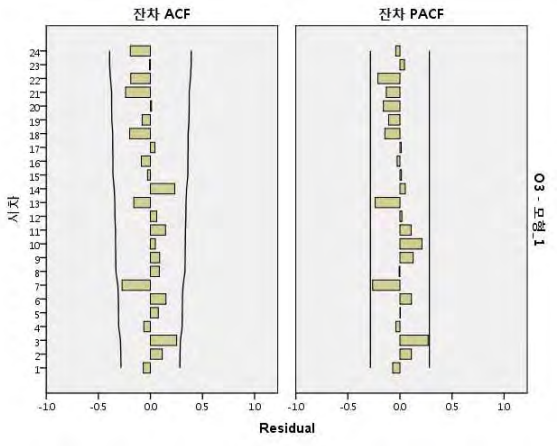
$SARIMA(p,d,q)(P,D,Q)_s$ 모형을 추정하는데 있어 가장 중요한 부분은 p,d,q,P,D,Q의 값을 찾는 것이다. d,D 의 값은 앞에서 정상성과 계절성 존재 여부에서 d=0, D=1 이라는 것을 알 수 있었다. 다음으로 추정해야 하는 값은 p,q,P,D 값을 추정해야 한다. p,q,P,D 값을 추정하기 위해 베이지안 방법을 이용한 정규화된 베이지안 정보 판단 기준(Normalized Bayesian Information Criterion, Normalized BIC)를 고려하는데 <표 1>과 같이 $SARIMA(1,0,0)(0,1,1)_{12}$ 모형이 가장 작으므로 우선적으로 최적 모형으로 고려하되 다른 모형들의 경우도 함께 살펴보기로 한다.

<표 1> ARIMA 모형의 정규화된 BIC

모형	정규화된 BIC
SARIMA(1,0,0)(0,1,1)$_{12}$	-11.516574
SARIMA(1,0,0)(1,1,1)$_{12}$	-11.356426
SARIMA(1,0,1)(0,1,1)$_{12}$	-11.390542
SARIMA(1,0,1)(1,1,1)$_{12}$	-11.296022

3.4 모형의 진단

모형 식별에서 선택된 모형을 기준으로 다음과 같이 모수를 추정 하였다. 대기 중 오존 농도 예측 모형으로 $SARIMA(1,0,0)(0,1,1)_{12}$모형을 생성한 결과 (그림 2)와 같이 나타났으며, 모두 신뢰한계선 범위 안에 존재하므로 $SARIMA(1,0,0)(0,1,1)_{12}$모형을 대기 중 오염농도 예측 모형으로 선정하였다.



(그림 2) 잔차 ACF, 잔차 PACF

3.5 모형의 검증

최종적으로 선정된 $SARIMA(1,0,0)(0,1,1)_{12}$ 모형의 접합성을 분석하기 위해 <표 2>와 같이 모수 추정치를 비교하였다.

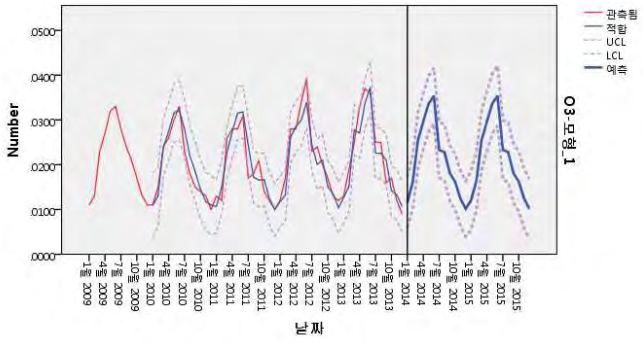
<표 2> SARIMA 모형 통계량

모형	정상 R제곱	R^2	RMSE	MAPE
$SARIMA(1,0,0)(0,1,1)_{12}$	0.250	0.878	0.00291	11.581
	통계량	자유도	유의확률	정규화된 BIC
	22.659	16	0.123	-11.517

$SARIMA(1,0,0)(0,1,1)_{12}$ 모형은 87.8%의 설명력을 보이고 있으며 비교적 낮은 BIC로 추측되어 지고, 유의확률에서 0.05보다 큰 것으로 나왔기 때문에 백색잡음이 독립적으로 존재 하므로 예측 모형으로서 적합하다고 할 수 있다.

3.6 예측

최종적으로 선정된 $SARIMA(1,0,0)(0,1,1)_{12}$ 모형을 이용하여 2014년 1월부터 2014년 12월까지의 대기 중 오존 농도를 예측 하였다. 그 결과는 (그림 3)과 같다.



(그림 3) $SARIMA(1,0,0)(0,1,1)_{12}$ 모형을 통한 예측

4. 결론

대기 오염은 인간의 삶의 질 향상에 직접적으로 연관되어 있으며, 그 중 오존은 반복 노출 시 폐에 피해를 줄 수 있다. 특히 어린이와 노약자에게 많은 악영향을 줄 수 있는 오염 물질로, 예방 대책이 필요하다.

따라서 본 논문에서는 대기 중 오존농도 예측 모형을 제시 하였다. 계절성의 특성을 갖고 있는 오존농도 예측 방법에 계절성 시계열 모형인 SARIMA모형으로 예측을 진행하였고, 그 중 $SARIMA(1,0,0)(0,1,1)_{12}$모형을 채택하였다. 그 결과 87.8%의 설명력을 보였으며, 유의확률에서 0.05보다 큰 것으로 확인 되었다. 따라서 백색 잡음이 독립적으로 존재하므로, 예측모형에 적합하다는 결과를 얻을 수 있었다.

사사의 글

본 연구는 미래창조과학부의 2015년 고용계약형 SW석사과정 지원 사업(과제번호:H0116-15-1003)으로부터 지원받아 수행한 결과입니다.

참고문헌

[1]김희재, 전명진, “도시 특성과 대기오염 수준과의 관계 분석 연구”, 대한국토도시계획학회지, 제 49권, 제 7호, pp. 151-167, 2014
[2] 한국대기환경학회, “대기오염에 대한 올바른 이해”, 2011
[3] 정원삼, “광주지역의 대기질 특성에 관한 연구”, 2011
[4] 김윤식, “SARIMA모형과 VAR모형을 이용한 철도 여객의 수송수요예측”, 2014
[5] 이재민, 권용재, “계절성을 감안한 ARIMA모형을 이용한 교통수요 동태적 변화 연구”, 대한교통학회지, 제 29권 제 5호, pp. 139-155, 2011