韓國大氣保全學會誌 第 7 卷 第 2 號
J. KAPRA Vol. 7, No. 2 (1991) pp. 127~136

大氣汚染濃度에 관한 確率모델

A Stochastic Model for Air Pollutant Concentration

김 해 경

연세대학교 이과대학 수학과

(원고접수 : 1991. 7. 24)

Hae Kyung Kim

Department of Mathematics, College of Sciences, Yonsei University

(Received 24, July 1991)

Abstract

This paper is concerned with the development and application of a stochastic model for daily sulphur dioxide (SO_2) concentrations in urban area (Seoul). For this, the characteristics of the regression trend, periodicity and dependence of the daily SO_2 concentration are investigated by a statistisical analysis of the daily average SO_2 values measured in Seoul area during 1989~1990. Based on these, nonlinear regression time series model for the prediction of daily SO_2 concentrations is derived. A statistical procedure for using the model to predict the concentration level is also proposed.

1. 서 론

大氣汚染은 도시의 산업화, 인구의 밀집화 등 경제발전의 가속화과정에서 수반되는 환경오염의 하나이다. 이 오염의 증가는 생태계는 물론 인간생활의 모든 분야에 많은 영향을 주게 되며, 이로 인한 재해를 최소화하기 위한 정확한 예측방법이 요구된다.

大氣汚染濃度나 오염상태의 정확한 설명은 관련된 화학 및 물리적작용 그리고 여기에 영향을 주는 모든 지배인자 사이에서 일어나는 力學現象 등의 종합적인 記述로 이루어질 수 있다. 따라서 복잡한 화학 및 물리현상과 또 다양한 기상현상들의 영향을 받는 이 대기오염현상은 몇가지 간단한 決定모델로는 만족스럽게 설명되지 않는다.

대기오염의 상태는 氣象狀態와 같이 늘 변하면서도 어떤 규칙성을 지니고 있다. 이런 규칙성의 존재는 현재의 상태는 과거에, 미래는 과거와 현재에 의하여 從屬되어 있고, 이 從屬性이 시간에 따라 일정한 변화를 하고 있음을 의미한다. 또, 이러한 종속성에서 기인한 규칙은 같은 조건하에서 같은 결과가 주어지지 않는 현상으로 決定函數보다는 오히려 確率函數로 기술하는 것이 더 합리적이다.

본 연구의 목적은 대기오염현상의 確率 및 統計的 특성을 파악하고 대기오염농도의 예측을 위한 確率모델을 확립하는데 있다. 이를 위해, 먼저 대기오염농도의 回歸趨勢, 週期性 및 從屬性 등 確率 및 統計的 특성을 파악하고, 다음으로 이 결과들을 기초로 하여 아황산가스 농도수준의 예측을 위한 確率모델을 완성하고 또 그 適合度를 확인하였다. 마지막으로 유도된 確率모델을 이용한 오염농도예측의 統計的절차를 제안하고 그 實例를 들었다.

2. 資料와 分析모델

본 분석은 1989, 1990년 환경청이 관측한 서울지역 20개지점(광화문, 면목동, 신설동, 길음동, 불광동, 마포동, 문래동, 대치동, 신림동, 잠실동, 한남동, 구의동, 성수동, 쌍문동, 남가좌동, 구로동, 오

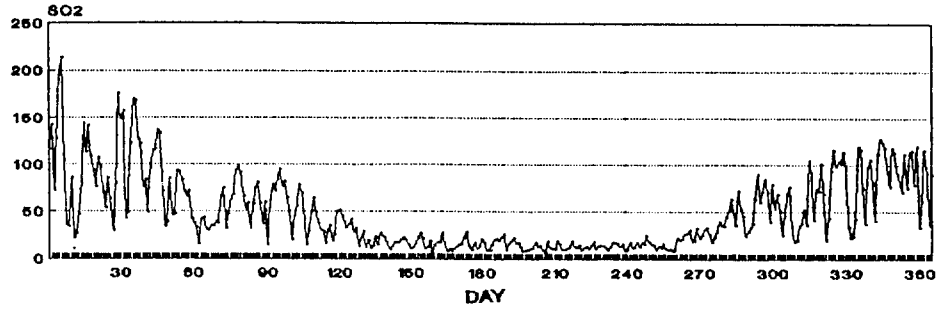

Fig. 1. Original daily SO_2 series (unit : ppb): Jamsil, 1989. 1. 1~1989. 12. 31.

류동, 반포동, 잠실1동, 방이동)에서의 시간별 아황산가스(SO_2) 농도 및 풍속, 풍향 등 기상지배인자의 측정치와, 같은 기간동안 기상청에서 관측한 서울지역 일별(최저, 최고, 평균) 기온과 강우량, 습도의 측정치를 포함하여 과거 30년동안의 기온자료를 기초로 하였다. 특히 1989. 1. 1~1989. 12. 31(잠실)의 자료가 수치계산의 사례로 사용되었다. Fig. 1은 자료중 SO_2의 농도시계열 일부를 나타낸 것이다.

대기중 SO_2의 농도는 다양한 요인으로 방출된 오염물질이 시간의 변화에 따라 여러가지 기상지배인자들의 영향을 받고 확산이동되는 상태에서 결정된다. 따라서 농도의 분석에는 기본적으로 방출요인 또는 방출량에 영향을 주는 지역의 특성, 계절과 기온, 지형은 물론이고 확산과 이동에 영향을 주는 풍속, 풍향, 습도, 강우량 등과 같은 기상 지배인자들도 포함되어야 한다. 더욱이, 이와같은 요인에 의해 결정된 濃度水準이 시간의 변화에 따라 量的으로 스스로 영향을 주며 부단히 변화하고 있음이 고려되어야 한다.

이러한 관점에서 볼 때, 오염농도의 예측에는 두 가지 형태의 모델개발이 가능하다. 그 중 하나는 오염농도에 관련된 주요인자들 그리고 그 관련된 決定的 또는 確率的 형태를 밝히고 그 관계를 모델개발에 이용하는 이른바 傳達函數確率모델의 개발이다. 이에 반하여 시간에 따라 내존하는 오염농도의 확률적 종속관계를 파악하고 그 관계를 예측모델에 이용하여 미래농도의 예측에는(관련인자의 정보 또는 지식보다는) 과거와 현재의 오염농도만을 기초로 하는 單變數確率모델이 생각될 수 있다.

본 논문에서는 다음의 두 가지 이유에서 단변수확률모델의 개발에 역점을 둔다. 첫째, 傳達函數 모델의 경우는 관련된 주요 독립인자가 많을 뿐 아니라, 대부분의 이 인자들은 기온이나 풍속 등과 같이 각각 고유의 回歸趨勢, 週期性 그리고 從屬性 등 복잡한 확률구조를 가지며 어떤 시점에서 오염농도의 예측에는 적어도 같은 시점에서 이 변수들의 정확한 예측값을 필요로 한다. 이것은 이 방법이 예측의 효율면에서 단변수모델보다 불리할 수도 있음을 의미한다. 둘째, 단변수확률모델의 개발은 전달함수모델의 개발과정에서 필수적으로 요구된다. 왜냐하면, 단변수모델에서 얻어지는 殘差의 分散은 많은 독립인자중에서 독립변수를 선별하는 전달함수모델의 효율평가에 기준이 되며, 또 전달함수의 수정과 보완을 필요로 하는 독립인자의 異常상태나 자료의 왜곡에 대한 파악은 단변수확률모델의 殘差를 통해서만이 확인되기 때문이다.

3. SO_2 濃度의 單變數確率모델

單變數確率모델을 이용한 대기오염농도의 분석을 위해서 시간에 따라 나타나는 자료의 외관상 특징을 먼저 관찰한다. Fig. 1에서 보는 바와 같이, 먼저 SO_2의 일별농도는 전반부(1월~3월)와 후반부(9월~12월)에서는 중반부(4월~8월)에 비해 매우 큰 分散을 가지고 진동한다. 그리고 농도수준의 전체적인 경향을 전반부부터 점차 하강하여 7월~8월경에는 최저상태를 보이고, 그후 다시 점차 상승하는 어떤 趨勢를 내포하며, 동시에 趨勢주위를 천천히 상하진동하며 어떤 기간을 두고 반복되는 週期性이 나타난다. 또한 오염농도의 수준은 연속변화로서 순간적 도약을 갖지 않기 때문에 농도수준 상호간의 從屬性의 존재가 가정된다.

그러나, 시간에 따라 일정치 않은 오염농도의 분산은 분석결과의 효율을 감소시킬 뿐 아니라 분석방법의 기본가정인 분산의 定常을 위한 변수변환이 필요하다. 적당한 변수변환은 Fig. 2에서 보는 바와 같이 SO_2농도 X_t의 對數變換 $\ln X_t$이다. 따라서, 對數 SO_2 농도의 관찰치 $Y_t = \ln X_t$(t는 시간)는

$$Y_t = D_t + C_t + S_t + E_t$$

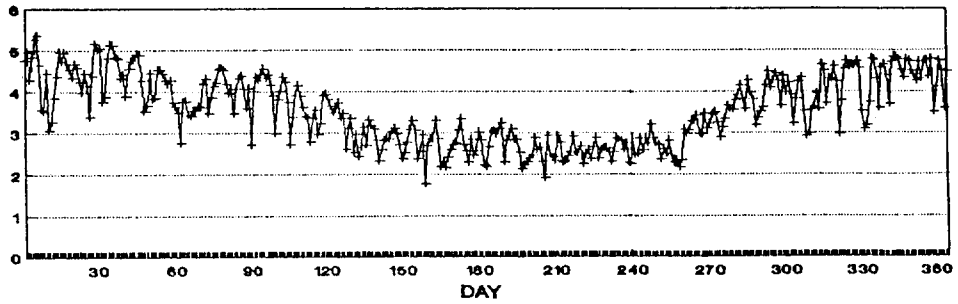

Fig. 2. Transformed series of SO_2 by natural logarithm.

의 기본假定式으로 표현된다. 여기서에, D_t는 回歸趨勢성분, C_t는 週期성분, S_t는 確率성분 그리고 E_t는 측정오차를 포함하는 確率誤差성분이다. 또, 확률성분 S_t와 측정오차 E_t의 합, $Z_t=S_t+E_t$는 보통定常 또는 弱定常 ARMA(p,q) 모델

$$\sum_{j=0}^{p} a_j Z_{t-j}=\sum_{k=0}^{q} b_k e_t \quad (p, q는 양의 정수) \quad (3.1)$$

로 설명된다. 여기에서 a_j와 b_k는 상수(단 $a_1=b_1=1$)이고 e_t는 (正規)白色誤差이다. 따라서, 위 기본가정식은 Y_t와 식(3.1)을 만족하는 Z_t를 사용하여

$$Y_t=D_t+C_t+Z_t \quad (3.2)$$

로 표시된다.

다음 절에서는 모델(3.2)에 대한 각 성분의 최적함수형태나 확률구조를 파악하여 오염농도의 回歸趨勢, 週期性 및 從屬性 등 確率 및 統計的 특성을 기술하고, 이것을 기초로 하여 농도수준의 예측을 위한 確率모델을 완성한다. 특히, 趨勢성분 D_t의 분석은 時系列 線型回歸분석(Shumway, 1988)을 통하여, 성분 C_t에 대한 週期함수식은 週期圖분석(Priestley, 1981)과 非線型回歸분석(Gallant, 1987)에 의하여 결정된다. 그리고 확률성분 Z_t의 결정에는 Box & Jenkins 방법(Abraham & Ledolter, 1983 또는 Box & Jenkins, 1976)을 이용한다.

3.1 回歸趨勢의 분석

回歸趨勢성분 D_t의 분석에서 자료 $\{Y_t\}$, t=0, 1, ……, T-1, 의 전체경향을 설명하는 다항식을 결정하고 그 誤差의 확률구조를 파악한다. 가능한 함수식은 2 또는 4차식으로, 보편화된 최소제곱방법으로 함수식의 계수를 推定하고 그 有意性을 檢定할 수 있다. 그러나 線型回歸 모델의 確率誤差項으로 볼 수 있는 $C_t+Z_t(=U_t)$가 從屬이기 때문에 추정의 一致性을 위하여, U_t의 分散共分散行列 Σ가 加重行列인 一般化된 加重최소제곱방법이 적절하다. 즉 σ^{ts}가 Σ^{-1}의 (t, s)-원소일 때

$$\sum_{t=0}^{T-1}\sum_{s=0}^{T-1}\sigma^{ts}[Y_t-\beta_0-\beta_1 t-\beta_2 t^2-\beta_3 t^3-\beta_4 t^4][Y_s-\beta_0-\beta_1 s-\beta_2 s^2-\beta_3 s^3-\beta_4 s^4] \quad (3.3)$$

를 최소화하는 β_i(i=1, 2, 3, 4)를 결정한다. 여기에서 σ^{ts}은 未知이지만 다음과 같이 그 값을 추정한다. 먼저 식(3.3)에 $\sigma^{ts}=1$(t,s=0, 1, ……T-1)을 사용한 최소제곱추정값 $\hat{\beta}_i$(i=0, 1, 2, 3, 4)를 구한다. 그리고, 이때 계산된 殘差 $\{R_t^{(1)}\}$, t=0, 1, ……, T-1(단, $R_t^{(1)}=Y_t-\hat{\beta}_0-\hat{\beta}_1 t-\hat{\beta}_2 t^2-\hat{\beta}_3 t^3-\hat{\beta}_4 t^4$)의 標本 分散共分散行列 Σ에서 σ^{ts}의 推定값 $\hat{\sigma}^{ts}$를 결정한다. 다음에, 추정된 $\hat{\sigma}^{ts}$를 식(3.3)에 사용하여 加重최소제곱추정값 $\hat{\hat{\beta}}_i$(i=0, 1, 2, 3, 4)를 얻을 수 있다. 그러나, 본 절의 목적이 최종모델(3.2)에 대한 계수의 初期값과 그 殘差시계열의 주기성분석에 필요한 확률적 특징을 파악하는데 있으므로 $\sigma^{ts}=1$(t,s=0, 1, ……, T-1)의 (常)최소제곱추정값으로 충분하다. 통계팩키지(예 ; SAS)를 이용한 상최소제곱 추정값은

$\hat{\beta}_0=4.1937(0.1184)$ $\hat{\beta}_1=0.0209(0.0045)$
$\hat{\beta}_2=-0.0004(0.00005)$ $\hat{\beta}_3=0.0000019(0.0000002)$
$\hat{\beta}_4=-0.000000002(0.000000000)$

이다(괄호안은 推定의 標準誤差(SE)). 이때의 잔차제곱합은 71.354이다.

추정값중 $\hat{\beta}_3$, $\hat{\beta}_4$는 비교적 작은 값을 보이고 있어 실제로 이 값이 零인지 아닌지에 관한 有意性 검정이 요구된다. 그러나, 對立假說 "H_1 : $\beta_i\neq 0$"에 대한 歸無假說 "H_0 : $\beta_i=0$"(i=3, 4)는 t-검정(또는 t^2되는 F-검정)의 값

$$t^*=\frac{\hat{\beta}_{(i)}-0}{SE(\hat{\beta}_{(i)})}$$

으로 t-분포(d.f.=360)에서 계산되는 P-값=0.0001으로 모두 쉽게 棄却된다.

Fig. 3에서 보는 바와 같이, 일평균 SO_2의 對數농도가 4차곡선의 回歸趨勢를 가진다. 그러나, 실제의 일일 평균농도수준은 이 回歸趨勢성분 뿐 아니라 다음절들에서 결정되는 週期성분 그리고 確率성분의 총화로 결정된다.

3.2 週期性 분석

週期성분 C_t의 분석은, 시계열$\{Y_t\}$에서 回歸趨勢성분이 제거된 殘差$\{R_t^{(1)}\}$의 주기성분석으로 이루어진다. Fig. 4에서 보는 바와 같이 시계열$\{R_t^{(1)}\}$는 어떤 週期성분을 함유하고 있다. 이 사실은 스펙트럼의 추정치인 週期圖 또는 標本스펙트럼(Fig. 5)에서도 관찰된다.

잠실지점의 1989. 1. 1~1989. 12. 31의 자료에서는 91.3일이 主週期요소로, 8.3일, 13.0일, 24.5일 그리고 182.5일이 副週期요소로 나타나고 있다. 91.3일의 週期는 서울지역 20개 전역에서 그리고 다른년도의 모든 자료에서도 主週期로 나타나고, 副週期는 지역과 시기에 따라 비교적 다양한 결과가 나타난다. 그러나, 대부분의 지역과 시기에서 대체로 8.3일과 24.3일 그리고 182.5일이 副週期로 나타나고 있다. 이러한 관점에서 보면, 서울지역에 있어서 SO_2 농도의 대표주기는 91.3일, 8.2일 그리고 24.3일로 요약될 수 있다. 한편, 같은 년도에서 일일(최저) 기온의 주기성이 겨울철에는 40.0일과 12.0일이 그리고 일년전체에는 121.6일, 182.5일이 각각 강한

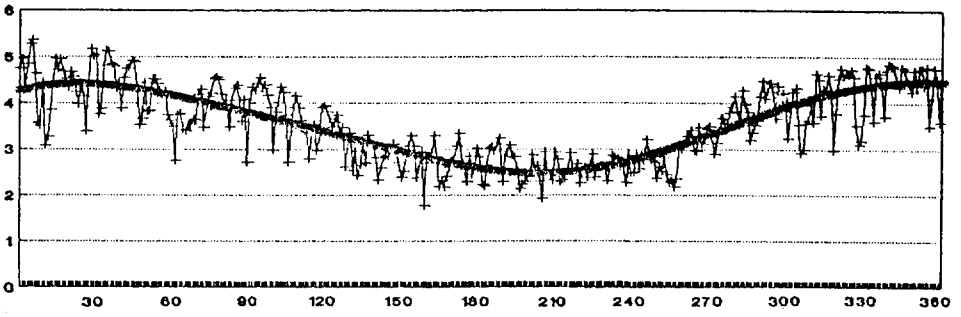

Fig. 3. In SO_2 series and polynomial trend: 1989. 1. 1.~1989. 12. 30.

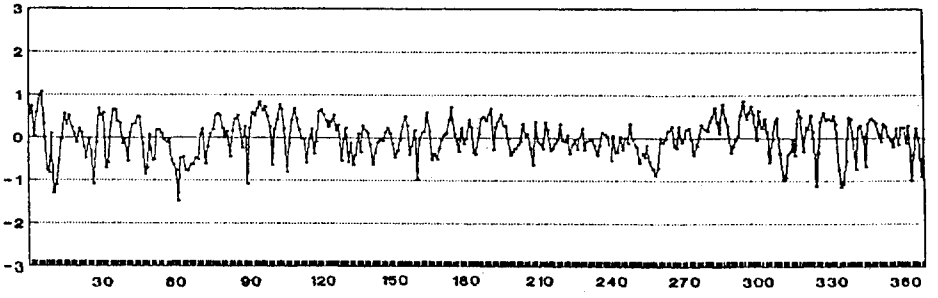

Fig. 4. Plot of residual series $\{R_t^{(1)}\}$: 1989. 1. 1.~1989. 12. 30.

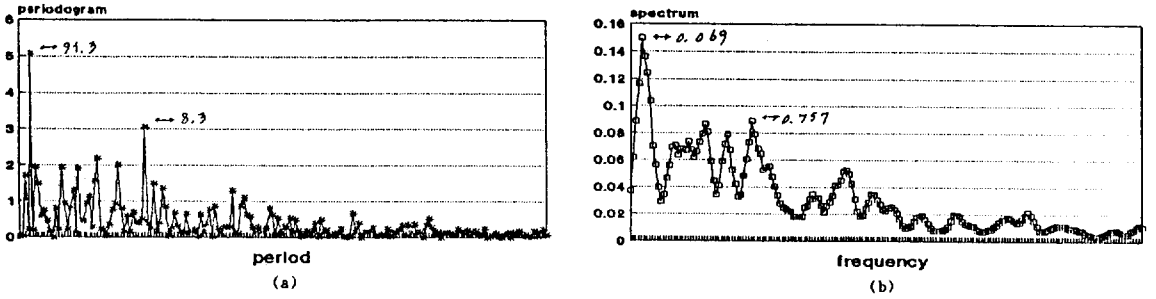

Fig. 5. (a) : Periodogram; (b) : Sample spectrum, of the residual series $\{R_t^{(1)}\}$.

主, 副週期임이 관찰된다. 오염농도에 있어서 副週期 182.5일은 主週期 91.3일의 倍數週期인 調和 현상의 결과로 볼 수 있기 때문에 기온과 오염농도의 週期는 일반적으로 완전히 일치하지 않음을 알 수 있다. 최저기온이 대기오염농도에 영향을 주는 대표적인 기상인자로 알려져 있지만(Kim, 1991), 이 두 변수사이의 종속관계는 주기성성분에서 설명되지 않음을 말한다. 오염농도의 주기중 91.3일과 8.2일은 각각 季節과 週단위로 형성되는 생활패턴과 관련된다고 볼 수 있다.

殘差 $\{R_t^{(1)}\}$의 主週期(91.3일)로부터 週期성분 C_t의 함수형태는 삼각함수식

$$C_t=\beta_{10}\sin(\beta_{11}+\beta_{12}\,t) \qquad (3.4)$$

으로 가정되고, 일반화된 가중최소제곱 삼각함수식은

$$\sum_{t=0}^{T-1}\sum_{s=0}^{T-1}\sigma^{ts}(R_t^{(1)}-C_t)\,(R_s^{(1)}-C_s)$$

또는

$$\sum_{t=0}^{T-1}\sum_{s=0}^{T-1}\sigma^{ts}[Y_t-(D_t+C_t)][Y_s-(D_s+C_s)]$$

을 최소화하는 β_{10}, β_{11}, β_{12}에 의하여 결정된다. 이때 σ^{ts}에는 확률성분 Z_t의 逆分散共分散行列의 (t,s)-원소이지만, 回歸趨勢분석에서와 같이 $\sigma^{ts}=1$인 최소제곱추정값을 이용한다.

1989. 1. 1~1989. 12. 31의 잠실자료의 경우, 추정된 계수는

$$\hat{\beta}_{10}=-0.1688(0.0315)\quad \hat{\beta}_{11}=-1.0263(0.3874)$$
$$\hat{\beta}_{12}=-0.0667(0.0018)$$

이고(이때의 잔차제곱합(SSE)은 66.124), 추정된 週期성분 C_t의 함수식은

$$\hat{C}_t=-0.1688\sin(-1.0263-0.0667\,t)$$

이다. Fig. 6은 殘差 $\{R_t^{(1)}\}$와 推定式의 값 $\{\hat{C}_t\}$를 동시에 표시한 것이다. 두번째 殘差 $\{R_t^{(2)}\}$, 단 $R_t^{(2)}=R_t^{(1)}-\hat{C}_t$는 Fig. 6에 나타내었다. 殘差 $\{R_t^{(2)}\}$의 週期圖 또는 標本스펙트럼에서는 週期 91.3의 피리오도그램과 標本스펙트럼이 5.0486과 0.1497에서 0.0065와 0.0497로 각각 감소된다. 이 사실은 $\{R_t^{(1)}\}$에 내존하던 91.3의 主週期성분의 제거됨을 의미하고, 따라서 推定된 삼각함수식의 적합함이 확인된다.

3.3 從屬性 분석

모델(3. 2)의 確率성분 Z_t는 시계열 Y_t에 내존하는 종속성을 설명한다. 確率성분 Z_t의 確率모델은 Y_t에서 回歸성분 D_t와 週期성분 C_t가 모두 제거된 殘差

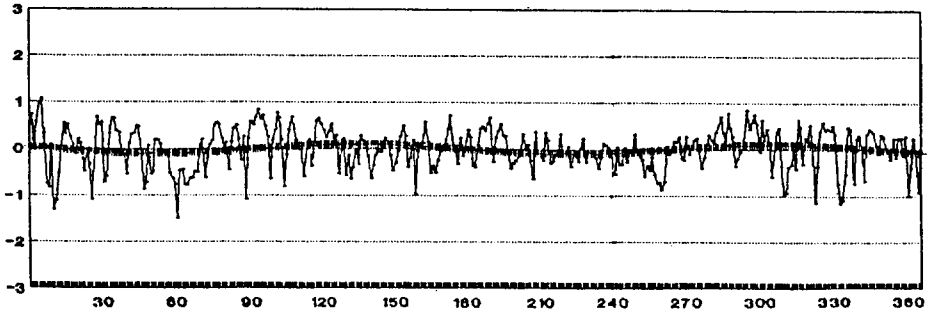

Fig. 6. Comparison of residual series $\{R_t^{(1)}\}$ (···) and fitted cyclic components $\{C_t\}$ (***).

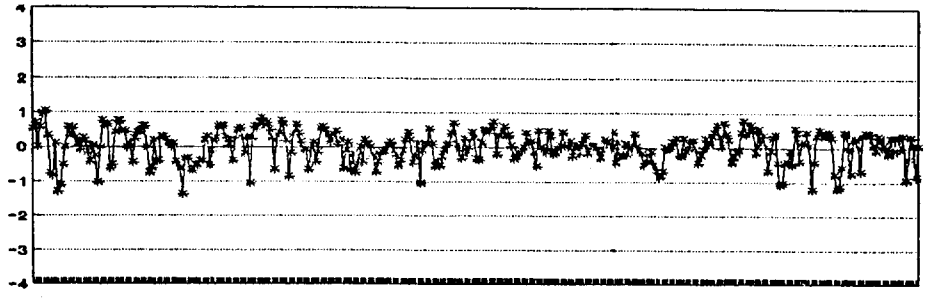

Fig. 7. Residual series $\{R_t^{(2)}\}$.

Name of variable=$R_t^{(2)}$.
Mean of the series=0.003464, Standard deviation=0.436128
Number of observations=365

Autocorrelations

Lag	Covariance	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.190208	1.00000	**********
1	0.091509	0.48110	**********
2	0.030017	0.15781	***
3	0.010469	0.05504	* .
4	0.0007326	0.00385	.
5	0.0053980	0.02838	* .
6	−0.0052323	−0.02751	.
7	−0.0040135	−0.02110	.
8	0.011269	0.05924	* .
9	0.013582	0.07141	* .
10	0.010076	0.05297	* .

"·" marks two standard errors

(a)

Partial Autocorrelations

Lag	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.48110	**********
2	−0.09582	.
3	0.02356	.
4	−0.02568	.
5	0.05120	*.
6	−0.08016	.
7	0.03177	*.
8	0.08082	**
9	0.01002	.
10	0.00294	.

(b)

Fig. 8. (a) : Sample autocorrelation function; (b) : Sample partial autocorrelation function, of residual series $\{R_t^{(2)}\}$.

$R_t^{(2)}$의 분석에서 결정되며 $R_t^{(2)}$는 앞절에서와 같이 $R_t^{(1)}-\hat{C}_t$ 또는 $Y_t-(D_t+C_t)$으로 정의된다. Fig. 7에서와 같이 시간에 따른 일정한 평균과 분산 등은 시계열 $\{R_t^{(2)}\}$의 定常 또는 弱定常 상태를 의미하며, 이 사실은 時差에 따른 標本 自己相關함수와 標本 部分自己相關함수로도 확인된다.

Fig. 8(a)와 (b)에서 보는 바와 같이, 標本 自己相關함수는 時差가 증가할수록 급속히 零에 가까워지고, 標本 部分自己相關함수는 시차 1 이후에서 切捨하고 있다. 이것은 Z_t가 1계의 自己回歸過程(AR(1))

$$Z_t=a_1Z_{t-1}+e_t \qquad (3.5)$$

로 설명됨을 의미한다. (造件)最小自乘法을 이용한 a_1의 추정값은

$$\hat{a}_1=0.48112(SE=0.04595)$$

이고(잔차제곱합은 51.27), 얻어지는 確率성분 Z_t의 2계 自己回歸모델은

$$Z_t=0.48112\,Z_{t-1}+e_t \qquad (3.6)$$

로 주어진다. 이 값들은 1계 自己回歸過程의 定常條件인 부등식 $|a_1|<1$을 만족한다. 더욱이, 모델(3.6)의 확률성분$\{Z_t\}$는 殘差$\{R_t^{(2)}\}$를 기초로 하여 다음과 같이 추정된다. 즉, 추정값 $\hat{Z}_t$는

$$\hat{Z}_t=\begin{cases}R_0^{(2)} & t=0\\ 0.48112\,\hat{Z}_0 & t=1\\ 0.48112\,\hat{Z}_{t-1} & t\geq 2\end{cases}$$

로 주어진다. 새 殘差$\{R_t^{(3)}\}$, $R_t^{(3)}=R_t^{(2)}-\hat{Z}_t$의 正規白色誤差의 성립은 식(3.5)가 해당년(1989년도) SO_2 농도수준의 確率성분모델로써 적합함을 의미하며, 이것은 다음 節에서 확인된다.

從屬성분이 AR(1)로 설명됨은 SO_2농도수준의 "1日從屬性"을 의미한다. 즉 SO_2에 있어서 오늘의 오염농도는 그 수준에 있어서 내일의 오염농도에만 직접적인 영향을 주고 있음을 말한다. 더욱이 확률성분의 계수가 식(3.6)에서와 같이 주어질때 현재의 농도수준이 1일을 넘어선 미래의 수준에 미치는 영향은 급속히 감소됨은 쉽게 확인된다. SO_2 농도수준의 이와같은 현상은 과거의 일일 오염자료를 기초로 한 1일을 넘어선 미래수준의 예측에는 적어도 이 확률성분의 크기정도의 예측오차가 필연적으로 동반됨을 말해준다. 이와같은 결과는 3절의 實例분석을 통해 확인된다.

그러나, 오염상태의 종속기간을 설명하는 自己回歸過程의 階數는 관찰지역이나 관찰년도에 따라 다소의 차이를 보인다. 예컨데, 앞에서 확인된 바와 같이 1989년도 잠실의 경우는 AR(1)이지만 광화문, 마포, 문래지역 등은 같은 년도에서 AR(3)로 설명된다. 그러나, 서울의 대부분지역에서 自己回歸過程의 階數는 3을 넘지 않음이 확인된다. 이러한 결과는 서울지역에 있어서 SO_2농도의 영향 지속기간은(지역에 따라 다소의 차이가 있으나) 일반적으로 3일(72시간)을 넘지 않음을 의미한다. 이와같은 현상은 오염농도의 지속기간이 해당지역의 기상상태나 지리적여건에 의하여 결정되기 때문이다. 이러한 관점에서 보면, 단변수확률모델의 경우에도 간접적이긴 하지만 오염농도에 영향을 미치는 인자들의 정보가 이용되고 있는 것이다.

3.4 單變數모델과 適合度

위 분석에서 얻어진 성분 D_t, C_t, S_t 등의 確率모델들은 (3.2)에 대입하여 시계열 Y_t의 確率모델이 완성된다. 즉 농도시계열 X_t의 確率모델은

$$\begin{aligned}\ln X_t=&\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3+\beta_4 t^4\\&+\beta_{10}\sin(\beta_{11}+\beta_{12}t)\\&+a_1 Z_{t-1}+a_2 Z_{t-2}+a_3 Z_{t-3}+e_t\end{aligned}\qquad(3.7)$$

이다. 여기서 Z_t는 殘差 $R_t^{(2)}$로 近似化될 수 있는 確率變數이고, e_t는 正規$(0,\sigma^2)$ 確率變數이다.

특히 Z_t가 AR(1)이면 $a_2=a_3=0$이고, AR(2)이면 $a_3=0$이다.

測定誤差의 分散 σ^2은 殘差 $\{R_t^{(3)}\}$의 分散으로 推定될 수 있다.

確率모델(3.7)은 오염의 豫測방정식을 구하는데 이용될 수 있다. 그러나 앞 절에서 밝혀진 바와 같이 回歸推勢 D_t나 週期성분 C_t는 해당년 고유의 특성을, 自己回歸確率 성분 S_t는 해당지역의 階數를 갖게 되므로 豫測방정식은 기본적으로 해당지역의 豫測年의 자료만으로 결정되는 것이 바람직하다.

豫測年 전반기 對數농도자료 $\{Y_t\}$, $t=0,1,\cdots\cdots$, T-1을 사용 모델(3.7)를 통한 후반기 농도수준의 豫測방정식은 다음과 같이 몇 단계로 결정된다.

1단계 : 回歸趨勢성분 $D_t=\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3+\beta_4 t^4$에 대한 최소제곱추정값 $\hat{\beta}_i(i=1,2,3,4)$를 구한다. 이때, 이 趨勢성분은 決定函數로 간주하여 과거의 자료 모두를 이용하여 初期값을 결정한다.

2단계 : 殘差 $\{R_t^{(1)}\}$, $R_t^{(1)}=Y_t-\hat{D}_t$의 主, 副週期를 週期圖 또는 스팩트럼을 통하여 구하고, 이것을 β_{i2} $(i=1,\cdots\cdots,P)$의 初期값으로 이용하여 週期성분

$$C_t=\sum_{i=1}^{p}\beta_{i0}\sin(\beta_{i1}+\beta_{i2}t)$$

의 최소제곱추정값 $\hat{\beta}_{i0}$, $\hat{\beta}_{i1}$, $\hat{\beta}_{i2}$, $(i=1,\cdots\cdots,P)$를 구한다. 이때 p는 主, 副週期의 갯수이다.

3단계 : 위 두 단계에서 결정된 D_t와 C_t를 동시에 사용한 함수 D_t+C_t에 대한 미지모수의 최소제곱추정값을 對數농도자료 $\{Y_t\}$, $t=0,1,\cdots\cdots$, T-1로부터 구하고, 여기에서 얻어진 殘差 $\{R_t^{(2)}\}$, $R_t^{(2)}=Y_t-(D_t+C_t)$의 標本 自己相關 및 部分自己相關 함수를 관찰하여 自己回歸過程의 階數 q를 결정한다. 또 식(3.5)에서와 같은 방법으로 이 종속성분의 미지계수에 대한 최소제곱추정값을 구한다.

4단계 : 위에서 결정된 각 성분들로부터 豫測방정식을 완성한다. 즉 $t\geq T+1$에 대하여

$$\begin{aligned}\ln X_t=&\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3+\beta_4 t^4\\&+\sum_{i=1}^{p}\beta_{i0}\sin(\beta_{i1}+\beta_{i2}t)+\sum_{u=1}^{q}a_u Z_{t-u}\end{aligned}\qquad(3.8)$$

모델 (3.8)이 농도시계열의 모델로써 적합한지 아닌지의 확인은 이 모델을 실제 자료에 적용한 후 얻어지는 殘差 $\{R_t^{(3)}\}$, $R_t^{(3)}=Y_t-(D_t+C_t+Z_t)$의 분석으로 이루어진다. 확률모델(3.1)에서 가정된 바와 같이 誤差項 $\{e_t\}$의 正規白色誤差의 성립은 모델의 適合함을 의미한다.

Name of variable=$R_t^{(3)}$
Mean of the series=0.001855, Standard deviation=0.382344
Number of observations=365

Autocorrelations

Lag	Covariance	Correlation	− 1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.1461870	1.00000	**********
1	0.0067624	0.04626	. *.
2	−0.0121460	−0.08308	** ..
3	−0.0019039	−0.01302	.. ..
4	−0.0067044	−0.04586	.* ..
5	0.0088215	0.06034	.. *..
6	−0.0071688	−0.04904	.* ..
7	−0.0078315	−0.05357	.* ..
8	0.0092595	0.06334	.. *..
9	0.0064669	0.04424	.. *..
10	0.0070832	0.04845	.. *..

"." marks two standard errors

(a)

Partial Autocorrelations

Lag	Correlation	− 1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.04626	. *.
2	−0.08541	** ..
3	−0.00495	. ..
4	−0.05256	.* ..
5	0.06447	. *..
6	−0.06496	.* ..
7	−0.03791	.* ..
8	0.05756	. *..
9	0.03640	. *..
10	0.04546	. *..

(b)

Fig. 9. (a) : Sample autocorrelation function; (b) : Sample partial autocorrelation function, of residual series $\{R_t^{(3)}\}$.

Fig. 9(a), (b)는 잠실자료에 대한 殘差시계열의 標本 自己 및 部分自己 相關函數로 모두 白色誤差의 특성을 보여준다. 標本 自己相關係數들로 계산되는 Box-Pierce Chi-square 統計量(Box & Jenkins, 1976) 등 다른 檢定統計量들도 이 시계열이 白色誤差임을 확인하여 준다. 또한, 이 시계열의 正規性은 正規確率紙 또는 適合度檢定을 통해 쉽게 확인된다.

확률모델의 적합성이 곧 관련 예측방정식의 정확성을 의미하지는 않는다. 일반적으로, 예측방정식에서 발생되는 예측오차의 크기는 모델의 함수적형태와 그리고 적용된 알고리즘에 의하여 결정된다. 이런 관점에서 볼 때, 모델(3.8)의 週期성분과 從屬성분에서 階數 p, q를 알맞게 결정하는 것은 매우 중요하다. 그러나, 앞에서 언급한 바와 같이 自己回歸過程을 따르는 從屬성분에서 발생되는 예측오차는 모델의 적합도나 알고리즘의 성공적인 개발로도 해결될 수 없다. 階數가 q인 自己回歸過程에 있어서, 예측간격이 q를 넘을 때는 평균(여기에서는 零)이 가장 바람직한 예측값이기 때문이다. 따라서, 이 경우에 적어도 從屬성분크기의 예측오차가 반드시 동반된다. 이러한 현상은 自己回歸過程의 從屬성분을 갖는 모든 확률현상의 특징이기도 하다. 이런 의미에서, 서울지방 SO_2 오염농도의 경우, 종속성분의 오차까지도 고려된 적절한 예측간격은 보통 q=1인 경우는 2, 3일, q=2인 경우는 3, 4일, 그리고 q=3 경우는 4, 5일로 관찰된다. 이와같이, 예측간격은

기본적으로 自己回歸성분의 階數에 종속되는 예측년 또는 예측지역의 고유치로 볼 수 있다. 이러한 의미에서, 해당년 그리고 해당지역 적정 예측간격을 초과한 농도예측은 回歸推勢와 週期성분만을 나타내는 예측으로 한정될 수 밖에 없다. 이와같은 현상은 Fig. 10에서도 쉽게 확인된다.

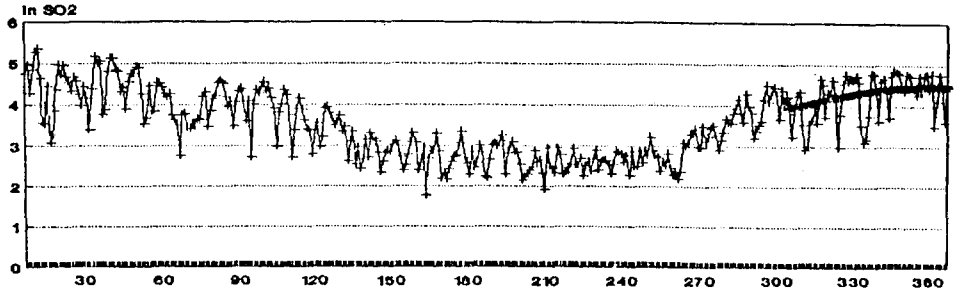

Fig. 10. Actual ln SO_2 series (⁺⁺) and predicted series (***) using data: 1989. 1. 1~1989. 10. 31.

Fig. 10는 1989년도 잠실자료에 대한 예측값과 관찰값을 동시에 나타낸 것이다. 일년중 전반부 300일 관찰치만을 기초로 하여 얻은 후반부 65일까지의 전방예측을 실시하여 그 예측값과 실제의 관찰값을 동시에 나타낸 것이다. 여기에서 확인되는 바와 같이 q=1 경우로 예측간격 2, 3일까지는 비교적 정확한 예측이 이루어지지만 이 간격을 넘어선 예측은 回歸推勢와 週期성분만을 기초로 한 예측으로 한정되고 있다.

위 분석에서는 主週期성분(91. 3일)만이 고려되었지만, 主週期와 副週期를 모두 이용할 경우 얻어지는 p=2(또는 3)의 모델은 예측오차를 더 감소시킬 수 있다. 이러한 경우에도 종속성의 확률구조에는 같은 결과로 주어진다.

4. 結　論

서울지방 일평균 SO_2 농도의 統計 및 確率的 특징과 그 예측모델은 다음과 같다.

첫째, SO_2 농도수준은 년중 4차곡선의 回歸推勢를 가지며, 일년중 回歸趨勢상의 일평균 최고농도는 매년 1월 30일 전후 15일 그리고 최저농도는 매년 7월 16일 전후 15일 내에서 나타난다.

둘째, SO_2 농도는 관찰년도에 따른 週期性을 가지며, 그 週期는 대체로 91.3일, 8.2일 그리고 24.3일로 요약될 수 있다.

셋째, SO_2 농도수준은 관찰년도와 관찰지역에 따른 固有階數 p의 自己回歸過程으로 설명되는 從屬性을 갖는다. 서울지역의 경우 p는 3을 넘지 않는 整數로 SO_2 농도의 영향 지속기간이 많아야 72시간 임을 의미하며, 적당한 예측간격은 q에 따라 q=1인 경우는 2, 3일, q=2인 경우는 3, 4일, 그리고 q=3 경우는 4, 5일이다.

넷째, 서울지방 SO_2 농도수준의 確率모델은

$$\ln X_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \sum_{i=1}^{p} \beta_{i0} \sin(\beta_{i1} + \beta_{i2} t) + \sum_{u=1}^{q} a_u Z_{t-u}$$

이다.

〈본 연구는 1991년도 연세대학교 학술연구조성비의 지원으로 이루어졌음〉

참 고 문 헌

Abraham, B. & Ledolter, J. (1983) *Statistical Methods for Forecasting*. John Wiley & Sons, 257 — 258.

Box, G.E.P. & Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, 2nd ed. San Francisco: Holdon-Day, 171 — 207.

Gallant, A.R. (1973) *Nonlinear Statistical Models*. John Wiley & Sons, 267 — 397.

Kim, Hae Kyung (1991) A Dynamic-stochastic Model for Air Pollutant Concentration. Preprint.

Priestley, M.B. (1981) *Spectral Analysis and Time Series*. Academic Press, 394 — 405.

Shumway, R.H. (1988) *Applied Statistical Time Series Analysis*. Prentice Hall, 117 — 193.