Journal of the Korean Data & Information Science Society
2015, **26**(2), 289–299

http://dx.doi.org/10.7465/jkdi.2015.26.2.289
한국데이터정보과학회지

서울지역 PM10 농도 예측모형 개발[†]

손건태[1] · 김다홍[2]

[12]부산대학교 통계학과

접수 2015년 2월 21일, 수정 2015년 3월 18일, 게재확정 2015년 3월 19일

요 약

본 연구는 PM10 농도에 대한 계량치 예측모형 개발을 목적으로 한다. 세 종류의 자료 (기상관측자료, 세계기상통신망 중국 관측자료, 대기질 화학수치모델자료)를 예측인자로 사용하였으며, 일일 단기예보 시스템에 쉽게 적용할 수 있도록 시간자료를 일자료로 변환하였고 시차변환을 수행하였다. 상관분석과 다중공선성 진단을 통하여 예측인자를 선택하고 두 종류의 모형 (중회귀모형, 문턱치 회귀모형)을 각각 적합하였다. 모형 안정성 검사를 위하여 모형검증을 수행하였으며, 전체자료를 사용하여 모형을 재추정한 후 예측치와 관측치 사이의 산점도와 시계열그림, RMSE, 예측성 평가측도를 작성 및 산출하여 두 모형을 비교하였다. 문턱치 회귀모형의 예측력이 고농도 PM10예측에서 다소 우수한 결과를 보였다.

주요용어: 계량치 예측, 문턱치 회귀모형, 미세먼지 농도, 예측성 평가측도.

1. 서론

대기 중 미세먼지 (particulate matter, PM)는 대기오염의 주요원인으로 천식과 같은 호흡기계 질병의 악화, 가축의 피해, 반도체와 같은 정밀기기의 불량률 증가, 항공기 및 선박 운항제한 등 사회 · 경제 전반에서 악영향을 끼치고 있다. 미세먼지는 크기에 따라 PM10, PM2.5, PM1 등으로 구분하며, PM10의 경우 직경이 10μm 이하인 미세먼지를 통칭한다. PM10의 발생원인은 자연적인 원인과 인위적인 원인으로 구분되며, 대부분 자동차 배기가스, 발전시설 및 공장의 배출물질과 황산염 (SO_4^{2-}), 질산염 (NO_3^-), 암모늄 (NH_4^+), 탄소화합물, 금속 화합물 등의 화학반응 과정에 의한 인위적인 원인에 의해 발생한다.

현재 기상청에서 β-ray 장비로 국내 28지점에서 PM10 농도를 관측하고 있으며, 환경부에서는 PM10에 대한 네 범주 예보를 CMAQ (community multiscale air quality) 모형에 의해 생성되는 예측치를 이용하여 실시하고 있다 (Moon 등, 2006; Cho 등, 2007; Itahashi 등, 2012). CMAQ 모형은 미국 환경보호청에서 개발한 수치예보모형으로 국내 PM10 범주예보모형으로 사용시 한반도와 동아시아 지역의 국지 기상정보를 수정해 사용해야 한다. 하지만 그 세부조건을 정확히 설정하는데 어려움이 있어 현재 PM10 범주예보 정확도가 높지 못한 한계점이 있으므로 한반도 실정에 맞는 PM10 예보모형 개발의 필요성을 시사한다.

[†] 이 논문은 2014년도 한국기상산업진흥원 연구지원사업인 '기상산업 지원 및 활용기술개발 사업' 중 '통계적 기법을 이용한 연무정보 예측 서비스 개발 (KMIPA 2014-11030)'의 위탁과제인 '연무예측 통계모델 개발'의 지원으로 수행된 연구 결과입니다.

[1] 교신저자: (609-735) 부산 금정구 부산대학로63번길 2, 부산대학교 통계학과, 교수.
E-mail: ktsohn@pusan.ac.kr
[2] (609-735) 부산 금정구 부산대학로63번길 2, 부산대학교 통계학과, 석사과정.

본 연구는 서울지역 PM10 농도에 대한 계량치 예측치를 생산하는 통계모형 개발을 목적으로 하고 있으며, PM10에 초점을 맞추어 연구설계와 분석이 수행되었다. PM10 생성에는 복잡한 화학반응 과정이 포함되어 있어 통계모형 개발에서 이를 모형에 포함시키기 어려우므로, 화학반응과정에 기초하여 모의실험을 수행하는 화학수치모델에서 생산된 예측치를 예측인자로 사용하고자 하였다. 이는 복잡한 화학반응 과정을 간접적으로 통계모형에 포함시킬 수 있어 통계모형의 예측성 (predictability)을 향상시킬 것으로 기대하기 때문이다. 한국기상청 자료에 따르면 PM10 발생 원인물질 중 30-50%는 서풍을 통해 중국에서 유입되는 것으로 알려져 있어 세계기상통신망 (global telecommunication system, GTS)에 의해 수집된 중국 기상관측자료를 사용하여 모형의 타당성을 확보하도록 하였다. 추가하여 국내에서 관측된 다양한 기상정보도 예측인자에 포함시켰다.

시계열모형에 의한 예측은 과거로부터 현재까지의 변화패턴이 미래에도 유지된다는 가정아래 이루어진다. 이를 뒷받침하기 위하여 모형검증이 수행된다. 모형훈련자료로 모형을 추정한 후 추정된 모형을 검증자료에 적용하여 예측모형의 안정성을 검토한다. 안정성이 있다고 판단되면 전체자료로 재추정하여 예측모형으로 사용한다.

단기예보 시스템에서는 예측의 정확도 (accuracy)가 중요하며, PM10 농도 예측의 경우에 전체적 정확도보다 고농도에 대한 예측 정확도에 초점을 맞추어야 한다. 이를 위해 문턱치 회귀모형 (threshold regression model)을 고려하여 고농도에서 예측의 정확도를 향상시키도록 하였다. 문턱치 회귀모형은 다양한 종류가 있으며, 본 연구에서는 문턱치 (threshold)에 의하여 두 가지 모형으로 구성되는 문턱치 회귀모형을 고려하였다. 문턱치를 고려하지 않았을 때의 중회귀모형을 적합시켜 문턱치 회귀모형의 결과와 비교하였다. Tong (1983)에서 다양한 형태의 문턱치 회귀모형들을 소개하고 있으니 참고하기 바란다. 모형비교를 위한 예측성 평가측도 (skill score)에 대하여는 Murphy (1993), Storch와 Zwiers (1999)를 참고하기 바란다.

2. 자료와 연구방법

2.1. 모형 개발에 사용된 자료

PM10 농도예측은 통계모형에 의해 생성된 예측치를 통해 이루어진다. 통계모형 개발을 위하여 세 종류의 자료 (기상관측치, 중국 GTS 자료, 대기질 화학수치모델에 의하여 생산된 대기질 자료 예측치)를 예측인자로 사용하였다. 중국 GTS 자료는 고층 (850 hPa) 기상관측이 가능하고, 한반도 수도권지역 PM10 농도에 영향을 준다고 생각되는 베이징과 칭다오 지역의 관측치를 이용하였다. 모형개발 대상지점은 서울지역이며, 사용된 예측인자의 종류, 변수명, 단위는 Table 2.1에 요약하였다.

Table 2.1 Meteorological variables with their notations and units in parentheses

KMA observations	wind direction (WD, 22.5°), wind speed (WS, m/s), pressure (PA, hPa), sea-level pressure (PS, hPa), air temperature (TA,°C), dew point temperature (TD,°C), surface temperature (TS,°C), daily range of temperature (Tdif,°C), relative humidity (HM,%), water vapor pressure (PV, hPa), amount of precipitation (RN, mm), total amount of cloud (Ctot, 1/10), visibility (VS, 10m), sunshine time (SS, hr), solar radiation (SI, MJ/m^2), haze (HZ)
China GTS observations	wind direction (WD), wind speed (WS, m/s), haze (HZ)
Air pollution materials	nitrogen oxide (NOX, ppmV), sulfate (SULF, ppmV), PM10 concentration (PM10, μg/m^3)

한국기상청에서 시정이 10 km미만이고 상대습도가 75%미만인 경우에 연무발생으로 정의하므로 연무발생을 산출하여 예측인자로 사용하였다. 중국에서 발생된 PM10 원인물질이 서풍에 의하여 한반도

PM10 농도에 영향을 미치므로 고층자료인 850 hPa지점의 풍향 중 서풍에 해당하는 자료만 사용하였다. 즉, 서풍계열은 WD=1로, 아니면 WD=0으로 이진자료를 생성하여 사용하였다.

모든 자료는 시간자료 (hourly data)이며, 자료기간은 세 종류의 자료가 모두 존재하는 기간으로 2012년 5월 1일부터 2013년 8월 12일, 2014년 1월 1일부터 2014년 7월 31일까지의 자료를 사용하였다. 세 종류의 자료에서 기준시간을 한국표준시 (korean standard time, KST)로 통일시킨 후, 결측치 보정작업을 수행하여 자료의 등간격화를 실시하였다. 17시에 기상청예보가 발표된다는 가정아래 국내 기상관측치의 경우 전날 0시부터 14시까지, 중국 GTS 자료는 전날 0시부터 12시까지, 대기질 수치모델 자료는 해당일 전체의 자료를 각각 사용하였다.

본 연구는 PM10 농도 일일 단기예보를 목표로 하므로 시간자료를 일자료로 변환하여 사용하였다. 가능한 PM10 농도증가에 영향을 주는 값을 예측인자로 사용하기 위하여 기상청 관련부서와 협의하여 다음과 같이 변환작업을 수행하였다. 국내 기상관측치의 경우 기온을 이용하여 일교차를, 시정과 상대습도를 이용하여 연무발생 여부를 결정하였다. 풍향, 해면기압은 일중 평균치를, 강우량, 시정, 온도의 경우 일중 최소치를 사용하였고 나머지 예측인자들은 일중 최대치를 사용하였다. 중국 GTS 자료는 풍향, 연무발생은 최대치를, 풍속의 경우 평균치를 사용하였고, 수치모델 예측치는 모두 평균치를 사용하였다. 예측 대상인 PM10 농도는 기상관측소 관측치이며, 자료 변환시 일중 최대치를 사용하였고, 단위는 $\mu g/m^3$이다.

2.2. PM10 농도 예측모형 개발방법과 기준

일자료로 변환된 자료를 사용하는 예측모형개발에서 다음과 같은 원칙과 기준들을 고려하였다.

첫째, 예측모형 개발이므로 미래시점에 대한 예측치를 생산하기 위하여 현실적으로 사용가능한 자료들을 예측인자로 사용하여야 한다. 본 연구에서는 일일 예보시스템에 적용이 가능하도록 모형 개발을 하였다. 즉, t 시점의 예측치 생산을 위하여 $t-1$ 시점의 관측치들을 사용한다. 관측자료의 경우 시차변환된 기상인자들을 예측인자로 활용하였다. 반면 수치모델 예측치는 $t-1$ 시점에 t 시점의 예측치가 생산되므로 시차변환 없이 예측인자로 활용하였다.

둘째, PM10 농도와 유의한 상관을 보이는 기상인자들을 선택하기위해 예측인자들에 대하여 상관분석을 실시하여 유의한 (p-value<0.15) 변수들만 예측모형 개발에 사용하였다.

셋째, 모형개발에 사용되는 예측인자들 사이에 다중공선성의 해결을 위하여 문제가 되는 변수를 제거하였다. 분산팽창인수가 10보다 큰 경우 다중공선성이 존재한다고 판단하였다.

넷째, PM10 농도예측을 위하여 두 가지 모형 (중회귀모형, 문턱치 회귀모형)을 적용하고 비교한다. 문턱치 회귀모형을 고려한 것은 PM10 농도가 지속성이 있을 것이라는 가정 하에서 이용하였다. 본 연구에서는 다음의 모형식으로 표현되는 문턱치 회귀모형을 적합하였다.

$$\mathrm{PM}(t) = \begin{cases} f_1(X_1(t)) & \text{if } \mathrm{PM}(t-1) \le \mathrm{TH} \\ f_2(X_2(t)) & \text{if } \mathrm{PM}(t-1) > \mathrm{TH} \end{cases}$$

여기서 $X_1(t)$과 $X_2(t)$는 예측인자들, $\mathrm{PM}(t)$는 t시점의 PM10 농도, TH는 문턱치이다. 전날 PM10 관측치가 문턱치보다 작다면 를 예측인자로 하는 중회귀모형을 적합시키고, 전날 PM10 관측치가 문턱치보다 크거나 같다면 를 예측인자로 하는 중회귀모형을 적합한다. 이를 통해 고농도 PM10에 대한 예측의 정확도를 향상시키는데 초점을 맞추었다.

다섯째, 자료를 모형 훈련자료와 검증자료로 나누어 모형검증을 실시하여 모형의 안정성을 검사한다. 먼저 훈련자료로 모형을 적합시킨 뒤 예측치를 생산한다. 연속형자료인 예측치를 중앙값을 기준으로 하여 이진자료로 변환한다. 변환된 자료를 이용하여 작성된 두 이원분할표의 동일성검정을 실시하여 두 분할표가 동질적이라면 모형의 안정성이 확보되었다고 판단한다. 두 이원분할표의 동일성을 검정하기

위해 반응변수를 다항분포로 가정하는 로그선형모형 (log-linear model)을 고려하였다. 로그선형모형에 대한 자세한 설명은 Jeong과 Choi (2009)를 참고하길 바란다.

여섯째, 전체자료를 이용해 중회귀모형과 문턱치 회귀모형 (문턱치=120 $\mu g/m^3$)을 각각 독립적으로 적합시킨다. 두 모형 모두 단계적 회귀방법으로 유의한 (p-value<0.1) 변수들만 선택하였다. 문턱치 선정기준은 문턱치를 기준으로 데이터를 분할했을 때 데이터개수가 적절히 선택되도록, PM10이 고농도일 때 모형의 예측성을 높일 수 있도록 선택하였다.

일곱째, 두 모형의 예측성을 비교하기 위하여, 두 모형에서 생산된 PM10 농도 예측치를 사용하였으며, 다음 절에 설명되는 예측성 평가 측도를 산출하여 비교하였다.

2.3. 모형에 대한 예측성 평가

적합된 두 모형의 비교는 다음과 같은 예측성 평가 기준을 고려한다.

첫째, 예측치와 관측치 사이의 산점도 및 시계열그림을 작성하여 예측의 정확도를 가시적으로 비교한다. 시계열그림을 통해 예측의 정확도 및 추세, 주기성을 살펴보고, 두 모형에서 관측치가 고농도일 때 예측의 예측정도를 비교한다.

둘째, RMSE (root mean square error)를 비교한다. RMSE는 예측오차의 변동성에 대한 수치적인 값으로 전체 변동성에서 모형에 의해 설명되는 변동성을 제외한 값을 제공해준다. 즉 RMSE가 작을수록 전체적인 예측의 정확도가 높다고 판단한다.

셋째, 계량치 예측치를 문턱치 120을 기준으로 이범주자료로 변환하여 Table 2.2와 같이 이원분할표를 작성한 후, 예측성 평가측도를 산출하여 비교한다. Murphy (1993)는 좋은 예측모형의 조건을 제시하면서 예측성 평가측도의 사용을 강조하였다. 본 연구에서는 예측품질과 가치에 중점을 두고 예측정확도 (Accuracy), 탐지확률 (probability of detection, POD), 허위경고율 (false alarm ratio, FAR)을 고려하였다. Accuracy는 모형의 전체적인 정확도를 나타내고, POD는 실제 고농도 PM10이 관측되었을 때 모형에서 고농도 PM10이라고 예측할 확률을 나타낸다. FAR은 모형에서 고농도 PM10이라고 예측한 경우 대비 실제 관측치는 저농도 PM10일 때의 비율이다. Accuracy와 POD는 높을수록, FAR은 낮을수록 예측의 품질이 좋다. 각각의 계산식은 다음과 같다.

$$\text{Accuracy} = \frac{\text{hits+correct negatives}}{\text{total}},\ \text{POD} = \frac{\text{hits}}{\text{observed high}},\ \text{FAR} = \frac{\text{false alarms}}{\text{forecast high}}$$

Table 2.2 Contingency table for forecasts vs. observations

		Observed		
		high (>120)	low (≤120)	total
Forecast	high (>120)	hits	false alarms	forecast high
	low (≤120)	misses	correct negatives	forecast low
	total	observed high	observed low	total

3. PM10 농도 예측모형 개발 결과

3.1. 중회귀모형 적합 결과

중회귀모형에 대한 모형 적합과정을 소개하도록 한다. 2.1절에서 제시한 기준에 따라 시간자료를 일자료로 변환하였으며, 2.2절의 기준에 따라 시차변환된 예측인자와 PM10 농도의 상관분석을 통해 예측인자를 선정한 결과는 Table 3.1과 같다. 예측인자 중 LPM과 LHZ는 PM10 농도와 연무발생 관측치의 시차변환된 변수고, TSmin은 하루 중 지면온도의 최소값이다. 중국 기상관측치 중 베이징과 칭다오

지역에서의 연무발생, 풍속이 PM10 관측치와 유의한 상관을 가짐으로서, 중국 기상관측치 사용의 타당성을 확보하였다. 수치모델 예측치는 세 변수 (PM10, SULF, NOX) 모두 유의한 상관을 나타냈으며, 이는 수치모델 예측치 사용이 모형개발에 타당함을 나타낸다. 선택된 예측인자 (Table 3.1의 변수들) 사이에 다중공선성 문제가 발생하므로 분산팽창인수가 10보다 큰 변수를 제거하였다.

2.2절에서 언급한 PM10 농도가 지속성이 있을 것이라는 가정은 전날과 그 다음날의 PM10 농도 관측치의 상관계수가 0.484로 가장 높은 상관관계를 나타나 가정에 대한 타당성을 확보하였다.

Table 3.1 Significant variables based on correlation analysis for Seoul

variable	LPM	PM10	LHZ	VS	SULF	Tdif	NOX	HZ_BJ	WS_QD
corr. coef.	0.484	0.481	0.343	-0.288	-0.204	0.200	0.177	0.165	-0.142
p-value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
variable	TSmin	HZ_QD	PV	Ctot	WS_BJ	SS	PS	HM	
corr. coef.	-0.128	0.111	-0.109	-0.085	-0.085	0.066	0.064	0.062	
p-value	<.001	0.004	0.004	0.026	0.027	0.086	0.097	0.107	

예측모형의 안정성을 평가하기 위해 자료를 훈련자료 (2012 ∼ 2013년)와 검증자료 (2014년)로 나누어 모형검증을 실시하였다. 2.2절의 기준에 따라 중회귀모형을 적합시키고 예측치를 생산하였으며 예측치와 관측치를 중앙값 (67 $\mu g/m^3$)을 기준으로 나누어 이원분할표로 만든 결과는 각각 Table 3.2와 Table 3.3과 같다.

Table 3.2 Contingency table of multiple regression model for training data

		Observed		
		high (>67)	low (≤67)	total
Forecast	high(>67)	154	71	225
	low (≤67)	61	183	244
	total	225	254	469

Table 3.3 Contingency table of multiple regression model for validation data

		Observed		
		high (>67)	low (≤67)	total
Forecast	high(>67)	105	30	135
	low (≤67)	16	61	77
	total	121	91	212

카이제곱 검정통계량에 의한 검정결과 (p-value=0.072) 두 이원분할표는 유의수준 0.05에서 다르다고 할 수 없다. 훈련자료로 적합시킨 모형이 검증자료에서 일관성을 보이므로 모형이 안정적이라고 판단할 수 있다. 모형의 안정성이 확인되었으므로 전체자료를 이용하여 예측모형을 적합하였다.

전체자료를 이용하여 중회귀모형을 적합시킨 결과는 Table 3.4와 같다. 모형의 전체적인 설명력 R^2는 41% 정도이며, 최종 예측인자는 중국 기상관측치의 경우 칭다오 지역의 변수만 유의하였고, 상대적으로 거리가 먼 베이징의 경우 유의하지 않음을 확인하였다. 변수의 단위를 표준화하여 상대적인 영향력을 나타내는 표준화계수를 살펴본 결과 PM10-LPM-NOX-Tdif-VS-WS-HZ-SS-SULF 순으로 모형에서의 상대적인 영향력이 크다고 해석할 수 있다.

PM10 생성 원인물질인 황산화물과 질소산화물의 수치모델 예측치인 SULF와 NOX의 회귀계수 부호가 음수로써 원인물질이 적을수록 PM10 농도가 높다는 해석을 할 수 있다. 이는 다른 예측인자의 영향력이 상대적으로 커서 SULF와 NOX가 억제변수 역할을 하여 부호가 음수이거나, SULF와 NOX가 PM10 생성 원인물질이기 때문에 PM10으로 화학반응을 많이 일으키면 그만큼 원인물질은 합성되어 사라지기 때문에 회귀계수 부호가 음수라고 해석할 수 있다.

Table 3.4 Estimates of parameters in multiple regression model

Parameter	Estimate	SE	P-value	Standardized Estimate	R^2 (adj R^2)	RMSE
Intercept	48.129	5.700	<.001	-		
PM10	0.626	0.061	<.001	0.454		
LPM	0.291	0.036	<.001	0.291		
NOX	-136.058	28.921	<.001	-0.192		
Tdif	2.095	0.632	0.001	0.141	0.4122	28.864
VS	-0.006	0.002	0.004	-0.105	(0.4043)	
WS_QD	-0.215	0.076	0.005	-0.086		
HZ_QD	6.949	2.739	0.011	0.078		
SS	-7.135	3.859	0.065	-0.075		
SULF	-8013674	4115817	0.052	-0.064		

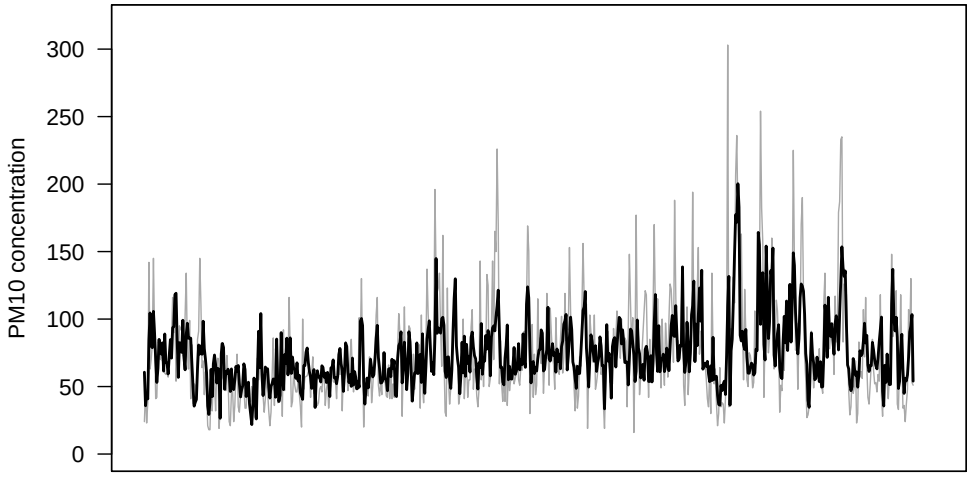

Figure 3.1 Time series plot of multiple regression model: observations (gray) and forecasts (black)

관측치와 예측치의 시계열그림은 Figure 3.1과 같다. 특별한 추세나 주기성은 나타나지 않으며, 서울지점 최근 자료에서 고농도 PM10의 빈도가 높아짐을 확인할 수 있다. PM10 농도가 저농도에서는 예측치가 잘 맞는 것을 알 수 있으나, 고농도에서는 예측치가 과소예측되는 한계점이 있었다.

3.2. 문턱치 회귀모형 적합 결과

문턱치 회귀모형에 대한 모형 적합과정을 소개하도록 한다. 기준시간 통일, 자료 등간격화 및 상관분석을 통한 예측인자의 사용은 3.1절과 동일하다. 예측모형의 안정성을 평가하기 위해 모형검증 3.1과 같은 방법으로 실시한 결과는 각각 Table 3.5와 Table 3.6과 같다.

Table 3.5 Contingency table of threshold regression model for training data

		Observed		
		high (>67)	low (≤67)	total
Forecast	high (>67)	153	75	228
	low (≤67)	62	179	241
	total	215	254	469

Table 3.6 Contingency table of threshold regression model for validation data

		Observed		
		high (>67)	low (≤67)	total
Forecast	high (>67)	104	28	132
	low (≤67)	17	63	80
	total	121	91	212

3.1절과 같이 로그선형모형을 고려하여 두 분할표의 동일성 검정결과 (p-value=0.072) 두 이원분할표는 유의수준 0.05에서 다르다고 할 수 없다. 문턱치 회귀모형에서도 모형의 안정성이 확보되었으므로 전체자료를 이용하여 모형을 적합시켰다. 전날 PM10농도 (LPM)에서 문턱치를 기준으로 자료를 나누어 회귀모형을 각각 적합한 결과는 각각 Table 3.7, Table 3.8과 같다.

Table 3.7 Estimates of parameters in threshold regression model (LPM≤120)

Parameter	Estimate	SE	P-value	Standardized Estimate	R^2 (adj R^2)	RMSE
Intercept	51.365	6.129	<.001	-		
PM10	0.651	0.064	<.001	0.494		
NOX	-154.553	28.761	<.001	-0.242		
LPM	0.247	0.056	<.001	0.175	0.3658	27.028
VS	-0.008	0.002	<.001	-0.151	(0.3574)	
Tdif	1.670	0.523	0.002	0.122		
WS_QD	-0.215	0.076	0.005	-0.095		
SULF	-9138422	3955155	0.021	-0.082		
HZ_QD	6.455	2.728	0.018	0.079		

Table 3.8 Estimates of parameters in threshold regression model (LPM>120)

Parameter	Estimate	SE	P-value	Standardized Estimate	R^2 (adj R^2)	RMSE
Intercept	2628.311	966.412	0.009	-		
PM10	0.649	0.150	<.001	0.492		
PS	-2.544	0.942	0.009	-0.473	0.3553	41.615
PV	-2.109	1.044	0.048	-0.348	(0.3109)	
LPM	0.363	0.141	0.013	0.277		

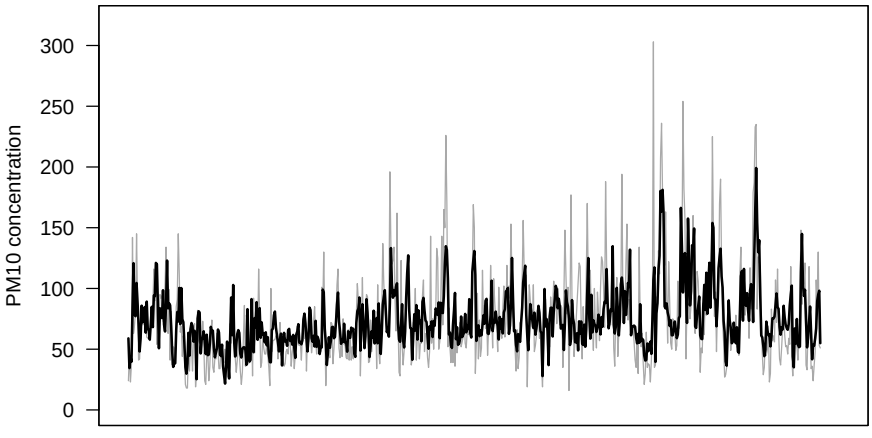

Figure 3.2 Time series plot of threshold regression model: observations(gray) and forecasts(black)

LPM이 문턱치보다 작거나 같을 때 모형적합 결과는 Table 3.7과 같다. 선택된 최종 예측인자의 표준화계수는 PM10-NOX-LPM-VS-Tdif-WS-SULF-HZ_QD 순으로 모형에서의 상대적인 영향력이 크다고 해석할 수 있다. LPM이 문턱치보다 클 때 모형적합 결과는 Table 3.8과 같으며, 모형의 RMSE가 41.615로 상대적으로 큰 값을 나타낸다. 예측인자의 표준화계수의 절대값은 PM10-PS-PV-LPM 순으로 컸으며, 이를 통해 고농도 PM10 자료만 사용하여 선택된 예측인자와 중회귀모형의 예측인자 사이에 많은 차이가 있음을 확인하였다. Figure 3.2의 시계열그림을 살펴보면 문턱치 회귀모형에서도 여전히 고농도 PM10 예측이 과소예측되는 한계점을 살펴볼 수 있다.

3.3. 모형 비교

모형 비교를 위하여 수치모형, 중회귀모형, 문턱치 회귀모형 예측치들의 관측치와의 상관계수와 RMSE를 산출하여 Table 3.9에 정리하였다. 문턱치 회귀모형의 RMSE 산출을 위한 계산식은 식 3.1과 같다. 대기질 화학수치모형인 CMAQ의 PM10 농도예측치와 관측치의 상관계수는 0.481에 불과하고, RMSE는 47.724로 크게 나타나 예측모형으로 사용하는데 문제가 있다. 중회귀 모형은 상관계수가 0.642로 증가하였으며 RMSE는 28.864로 감소하였다. 문턱치 회귀모형의 RMSE는 중회귀모형보다 0.292 감소하였다.

Table 3.9 Model comparison

Model	N	Number of parameters	SSE	RMSE	corr. coef.
CMAQ	681	-	1551067	47.724	0.481
Multiple regression	681	9	559031	28.864	0.642
Threshold regression (LPM≤120)	618	8	444892	28.572	0.653
Threshold regression (LPM>120)	63	4	100443		

$$\sqrt{\frac{\mathrm{SSE}_1 + \mathrm{SSE}_2}{n_1 + n_2 - p_1 - p_2 - 1}} = \sqrt{\frac{444892 + 100443}{618 + 63 - 8 - 4 - 1}} = 28.572 \tag{3.1}$$

여기서 n_1과 n_2는 자료수이고, p_1과 p_2는 모수의 갯수이다. 두 모형에 대한 예측치와 관측치 사이의 산점도는 각각 Figure 3.3, Figure 3.4과 같다. PM10 농도가 저농도일때는 큰 차이를 보이지 않는 것처럼 보이나, 고농도일때는 문턱치를 고려했을 때 예측의 정확도가 조금 향상되었음을 알 수 있다. 종합적으로 판단할 때 문턱치 회귀모형이 중회귀모형보다 향상된 결과를 보이고 있다.

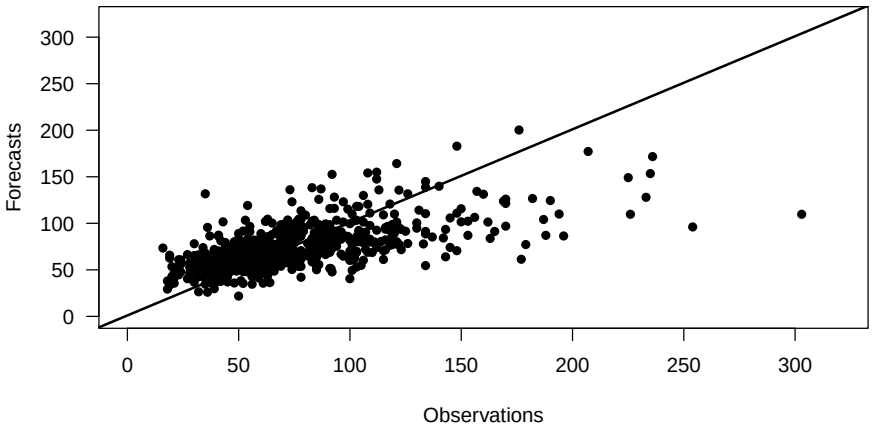

Figure 3.3 Scatter plot of multiple regression model: forecasts vs. observations

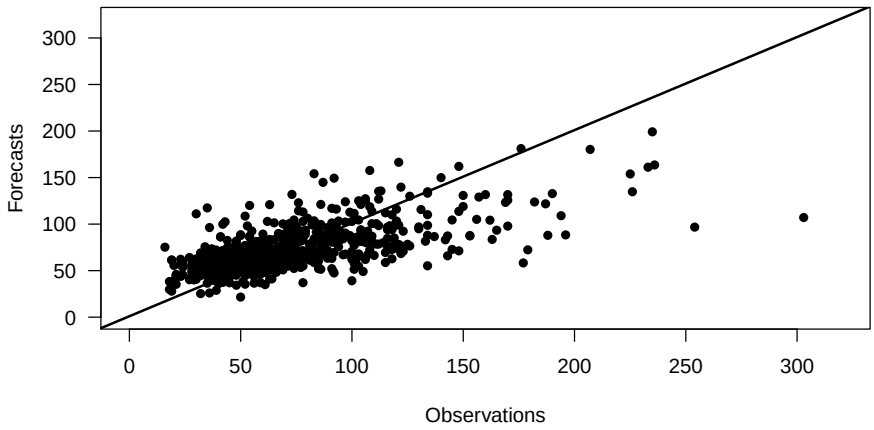

Figure 3.4 Scatter plot of threshold regression model: forecasts vs. observations

Table 3.10 Contingency table of obsetvations and forecasts using multiple regression model

		Observed		
		high (>120)	low (≤120)	total
Forecast	high (>120)	20	16	36
	low (≤120)	43	602	645
	total	63	618	681

Table 3.11 Contingency table of obsetvations and forecasts using threshold regression model

		Observed		
		high (>120)	low (≤120)	total
Forecast	high (>120)	23	16	39
	low (≤120)	40	602	642
	total	63	618	681

Table 3.12 Skill scores for multiple regression model and threshold regression model

	Accuracy	POD	FAR
Multiple regression model	91.34%	31.75%	44.44%
Threshold regression model	91.78%	36.51%	41.03%

중회귀모형과 문턱치 회귀모형에서 생산된 예측치와 관측치를 문턱치 (120 $\mu g/m^3$)로 나눈 결과는 각각 Table 3.12와 Table 3.13과 같고, 분할표를 이용하여 예측의 정도를 비교한 결과는 Table 3.14와 같다. 문턱치 회귀모형이 중회귀모형 대비 Accuracy가 0.44%P 높았고, POD가 4.76%P 높아 모형의 품질이 좋아졌음을 알 수 있다. FAR 또한 문턱치 회귀모형이 중회귀모형 대비 3.41%P 낮아 허위경고율이 더 낮음을 확인하였다. 두 모형에서의 예측치와 관측치의 이원분할표의 도수가 큰 차이는 없었지만, 고농도 PM10에서 문턱치를 고려한 모형이 중회귀모형보다 향상되었음을 확인하였다.

4. 결론

본 논문은 서울지역 PM10 농도 예측모형 개발을 위하여 중회귀모형과 문턱치 회귀모형을 각각 적합하고 비교한 결과이다. 모형의 타당성을 확보하기 위하여 국내 기상관측치, 중국 GTS 관측치, 대기질 수치모델 예측치를 예측인자로 고려하였다. 문턱치 회귀모형에서 PM10 농도가 고농도일때 예측성이

중회귀모형보다 향상되었으나, 고농도에서 과소예측되는 점은 여전히 해결해야할 문제로 남아있다. 예측모형 개발은 고농도 PM10으로 인한 피해로부터 국민건강 보호 및 경제적 피해 감소에 초점을 맞추고 있으므로 전체적인 정확성보다 고농도일 때의 정확성을 향상시키는데 중점을 두는 것이 바람직하다. 이번 연구에서 문턱치 회귀모형을 고려했을 때 고농도 PM10에 대한 예측성 향상에 대한 가능성을 보았다. 이를 바탕으로 PM10 농도예측을 전국으로 확대하고자 한다. 지역에 따라 미세먼지의 원인이 다양하게 구성되므로 이에 대한 고려가 필요하다. 예를 들어, 부산지역은 선박에 의한 대기오염이 유의한 영향을 미칠 것으로 예상되어 관련된 자료의 추가가 요구된다. 현재 미세먼지 농도 예측을 위한 통계모형에서 예측의 정확도가 미흡한 상태이므로 더 향상된 예측품질을 위하여 향후 연구로 다양한 모형을 고려하고자 한다. 비선형 회귀모형, 다중 문턱치 모형 등 예측모형에 대한 지속적인 연구가 필요하며, 교차상관분석을 통하여 예측인자를 추가하고자 한다.

References

Cho, C., Chun, Y., Ku, B., Park, S., Lee, S. and Chung Y. (2007). Comparison of ADAM's(Asian Dust Aerosol Model) Results with Observed PM10 Data. *Atmosphere*, **17**, 87-99.

Itahashi, S., Uno, I. and Kim, S. (2012). Source Contributions of Sulfate Aerosol over East Asia Estimated by CMAQ-DDM. *Environmental Science & Technology*, **46**, 6733-6741.

Jeong, K. M. and Choi, Y. S. (2009). *Categorical Data Analysis Using SAS*, Free Academy, Seoul.

Moon, N., Kim, S., Byun, D. W. and Joe, Y. (2006). *Air Quality Modeling System Ⅰ-Development of Emissions Preparation System with the CAPSS*, Korea Environment Institute.

Murphy, A. H. (1993). What is Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, **8**, 281-293.

Storch, H. V. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*, Cambridge, Cambridge University Press, Cambridge.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*, Springer, New York.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 289–299

http://dx.doi.org/10.7465/jkdi.2015.26.2.289
한국데이터정보과학회지

Development of statistical forecast model for PM10 concentration over Seoul[†]

Keon Tae Sohn[1] · Dahong Kim[2]

[12]Department of Statistics, Pusan national University

Received 21 February 2015, revised 18 March 2015, accepted 19 March 2015

Abstract

The objective of the present study is to develop statistical quantitative forecast model for PM10 concentration over Seoul. We used three types of data (weather observation data in Korea, the China's weather observation data collected by GTS, and air quality numerical model forecasts). To apply the daily forecast system, hourly data are converted to daily data and then lagging was performed. The potential predictors were selected based on correlation analysis and multicollinearity check. Model validation has been performed for checking model stability. We applied two models (multiple regression model and threshold regression model) separately. The two models were compared based on the scatter plot of forecasts and observations, time series plots, RMSE, skill scores. As a result, a threshold regression model performs better than multiple regression model in high PM10 concentration cases

Keywords: PM10, quantitative forecast, skill score, threshold regression model.

[†] This work was supported by the project 'Development of regional haze forecast based on statistical approach' (KMIPA 2014-11030).
[1] Corresponding author: Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea. E-mail: ktsohn@pusan.ac.kr
[2] Master's course, Department of Statistics, Pusan National University, Busan 609-735, Korea.