한국대기환경학회지 제 26 권 제 6 호
J. KOSAE Vol. 26, No. 6 (2010) pp. 666～682
Journal of Korean Society for Atmospheric Environment

미세먼지 예보시스템 개발

A Development of PM10 Forecasting System

구윤서* · 윤희영 · 권희용[1] · 유숙현[1]
안양대학교 환경공학과, [1]안양대학교 컴퓨터공학과

(2010년 8월 11일 접수, 2010년 10월 5일 수정, 2010년 11월 4일 채택)

Youn-Seo Koo*, Hui-Young Yun, Hee-Yong Kwon[1] and Suk-Hyun Yu[1]
Department of Environmental Engineering, Anyang University
[1]*Department of Computer Engineering, Anyang University*

(Received 11 August 2010, revised 5 October 2010, accepted 4 November 2010)

Abstract

The forecasting system for Today's and Tomorrow's PM10 was developed based on the statistical model and the forecasting was performed at 9 AM to predict Today's 24 hour average PM10 concentration and at 5 PM to predict Tomorrow's 24 hour average PM10. The Today's forecasting model was operated based on measured air quality and meteorological data while Tomorrow's model was run by monitored data as well as the meteorological data calculated from the weather forecasting model such as MM5 (Mesoscale Meteorological Model version 5).

The observed air quality data at ambient air quality monitoring stations as well as measured and forecasted meteorological data were reviewed to find the relationship with target PM10 concentrations by the regression analysis. The PM concentration, wind speed, precipitation rate, mixing height and dew-point deficit temperature were major variables to determine the level of PM10 and the wind direction at 500 hpa height was also a good indicator to identify the influence of long-range transport from other countries. The neural network, regression model, and decision tree method were used as the forecasting models to predict the class of a comprehensive air quality index and the final forecasting index was determined by the most frequent index among the three model's predicted indexes. The accuracy, false alarm rate, and probability of detection in Tomorrow's model were 72.4%, 0.0%, and 42.9% while those in Today's model were 80.8%, 12.5%, and 77.8%, respectively. The statistical model had the limitation to predict the rapid changing PM10 concentration by long-range transport from the outside of Korea and in this case the chemical transport model would be an alternative method.

Key words : PM10, Forecast, Statistical model, Neural network, Regression model

1. 서 론

대기오염물질은 인체, 자연 생태계 및 재산상 등 다방면에 직 · 간접으로 영향을 미치는 것으로 알려

*Corresponding author.
Tel : +82-(0)31-467-0893, E-mail : koo@anyang.ac.kr

져 있으며, 그 중에서도 호흡성 먼지, 또는 미세먼지라고 하는 PM10 (Particulate Matter with a diameter less than 10 μm)은 비교적 대기 중 체류 기간이 길며, 호흡기나 심장 질환이 있는 어린이나 노인 등에게는 직접적으로 건강에 유해한 영향을 끼치며, 빛의 시정장애 유발과 산성비 등의 2차적인 영향을 가져올 수 있다 (Kim, 2010, 2006).

표 1은 최근 전국 주요 대도시의 대기자동측정망에서 미세먼지의 24시간 대기환경기준 초과회수를 나타낸 것이다. 측정소당 미세먼지의 24시간 대기환경기준 초과빈도수는 경기도가 31.7회, 인천이 26.2회, 서울이 27.7회로 다른 광역시에 비해서 수도권지역을 중심으로 높게 나타나고 있다. 따라서 수도권역을 중심으로 미세먼지 저감대책의 수립과 함께 수도권지역을 대상으로 미세먼지에 대한 사전 예방적 예보시스템 구축이 필요한 실정이다.

이와 같이 심각한 수준인 미세먼지는 공사장 및 차량에 의한 비산먼지와 보일러 연소 및 차량 등에 의해서 배출되는 1차 오염물질과 대기 중 오염물질과 상호 반응하여 대기 중에서 생성되는 2차 오염물질로 구성되어 있고, 또한 중국에서 월경하는 미세먼지도 많은 부분을 차지하고 있어 그 발생원에 대한 조사가 어려운 실정이다. 이와 같이 미세먼지의 배출원에 대한 기초 자료가 미흡하고, 2차 생성 기구에 대한 이해 부족으로, 물리적 수치모델 사용에 한계가 있기 때문에 화학수송모델을 적용한 예보가 현실적으로 용이하지 않다. 따라서 대기질 측정망자료와 기상 관측자료와의 상관관계를 분석하여 통계 모델을 이용한 예측기법이 일반적으로 많이 사용하고 있다.

이미 여러 선진국에서 통계모델을 기본으로 한 미세먼지 예보제를 시행하여 국민에게 미세먼지에 대한 정보를 기상예보와 동등한 수준으로 제공함으로써 국민들이 일상생활에서 활용하도록 하고 있다 (Colbourn, 2010; Brian *et al.*, 2009; Kim, 2009; Murphey, 2007; Patricio and Jorge, 2006; Hooyberghs *et al.*, 2005). 특히 미세먼지에 취약한 그룹인 노인, 어린이, 또는 기관지 환자들이 미세먼지 오염도 예보정보를 적극적으로 그들의 생활에 활용하도록 하고 있다.

영국에서는 AQA (Air Quality Archive)에서 대기질 실시간 공개 및 예·경보를 하고 있다. 16개 지역으로 나누어 대기질 지수(band index)를 이용하여 O_3, NO_2, SO_2, CO, PM10을 종합적으로 판단하여 실시간 공개 및 예보를 하고 있다 (http://www.airquality.co.uk). 경보기준 농도(threshold concentration)가 초과시에 즉시 경보를 발령하여 배출원 제어 등의 조치를 수행하고 있고, 이와 더불어 주요 도시 및 주정부에서 24시간 대기질을 예보하고 있다.

Table 1. Number of exceedance of 24 hr average PM10 standard in major cities and region in 2008.

City or Region	Number of exceedance
Seoul (27)[a]	748 (27.7)[b]
Busan (17)	258 (15.2)
Daegu (11)	251 (22.8)
Incheon (15)	393 (26.2)
Gwangju (7)	139 (19.9)
Daejeon (7)	95 (13.6)
Ulsan (13)	297 (22.8)
Gyeonggi (64)	2030 (31.7)

[a]: Total no. of ambient air quality monitoring station in the city or region.
[b]: No. of exceedance of the PM10 24 hour standard per the monitoring station.

미국의 환경보호청은 인구 350,000명 이상인 주요 도시와 주정부에서는 실시간으로 측정된 오존 및 미세먼지를 포함한 기준성 오염물질의 대기질 현황을 AQI (Air Quality Index)로 환산하여 일반 대중에게 공개하고, 기준치 초과시는 경보 (Alarm)를 의무적으로 발령하도록 하고 있다 (http://www.airnow. gov). 또한 대부분의 주요 도시에서 대기질을 AQI로 예보하여 사전경보 (Alert)제를 운영하는 것도 의무화하여 각각의 지자체 특성에 맞는 예보시스템을 구축하고, 사전경보에 따른 배출량 저감을 위한 국민들의 자발적인 참여를 유도하는 여러 형태의 프로그램을 운영하고 있다 (U.S. EPA, 2009).

미국에서는 PM2.5에 대하여 예보를 실시하고 있고, 미국환경청의 예보지침서 (U.S. EPA, 2003)에 의하면 PM2.5와 상관관계가 있는 예보인자들은 표 2와 같다. 지역 특성에 따라서 차이는 있으나, 주로 500 mb 기압에서 고도, 전날의 PM2.5 최대농도, 850 mb 기압에서 온도, 습도와 상관관계가 큰 것으로 알려져 있다. 즉, 대기질농도, 고층 및 지표 기상요소가 미세먼지와 상관성이 큰 것으로 나타나고 있다.

국내 연구에서는 주로 기상인자와 미세먼지의 상

Table 2. Common predictor variables used to forecast PM2.5 (U.S. EPA, 2003).

Variable	Usefulness	Condition for high PM2.5[a]
500 mb height	Indicator of the synoptic-scale weather pattern	High
Surface wind speed	Associated with dispersion and dilution of pollutants	Low
Surface wind direction	Associated with transport of pollutants	–
Pressure gradient	Causes wind/ventilation	Low
Previous day's peak PM2.5 concentration	Persistence, carry-over	High
850 mb temperature	Surrogate for vertical mixing	High
Precipitation	Associated with clean-out	None or light
Relative humidity	Affects secondary reactions	High
Holiday	Additional emissions	–
Day of week	Emissions differences	–

[a]Relative condition is location and season dependent.

관관계에 대한 연구가 이루어졌으며, Hwang *et al.* (2009) 및 Shin *et al.* (2007)의 연구에 따르면 풍향에 따른 차이는 크게 없고 풍속이 0～6 m/s 구간에서는 풍속이 증가할수록 미세먼지 농도가 낮아지는 것으로 나타났고, 안개, 연무, 박무 등의 대기정체일 경우 미세먼지 농도가 높은 것으로 나타났다. PM10의 예보는 Lee *et al.* (2006)이 포항지역을 대상으로 미세먼지 예보모형을 개발하였고, Koo *et al.* (2005, 2003) 및 Yun *et al.* (2008, 2007)은 수도권지역을 대상으로 개발된 미세먼지 예보시스템에 대해서 발표한 바가 있다.

국내에서도 미세먼지를 대상으로 예보시스템을 구축하기 위해서 기상 및 대기질 자동측정망 자료를 활용한 미세먼지 통계예보모형을 개발하였고, 실제 예보시스템 운영을 통해서 그 결과를 분석하였다. 본 논문에서는 주로 미세먼지 예보시스템의 구성과 예보시스템을 구축하기 위한 입력변수와 미세먼지와의 상관관계, 그리고 예보모형 및 예보결과를 중심으로 서술하고자 한다. 한편 실제 예보시스템은 서울 4개 지역, 인천 9개지역, 경기 4개지역으로 나누어 구축되었고, 현재 지자체에서 운영중이나 예보변수 및 예보방법에 차이가 없기 때문에 본 논문에서는 서울 남동지역을 대상으로 한 예보변수 검토결과 및 예보시스템을 중심으로 기술하고자 한다.

2. 예보시스템 구성

미세먼지 예보시스템을 운영하기 위한 자료는 크게 대기오염도 측정자료, 지표 및 고층 기상대 관측자료, 기상예보자료로 나눌 수 있다. 대기오염도 측정 및 기상관측자료는 매시간 미세먼지 예보시스템으로 자료가 입력되어지며, 고층기상관측자료는 03시의 자료를 수신 받아 예보시스템의 입력변수로 활용한다. 예보자료는 12시에 수행되는 MM5 예보자료를 활용하였다.

예보는 전일 17시에 익일 미세먼지 평균농도를 예보하는 내일예보모형과, 당일 오전 9시에 확정 예보하는 당일예보모형으로 2단계로 진행된다.

내일예보모형은 그림 1에 나타낸 바와 같이 16시까지 측정한 대기질 자료와 기상관측자료, 익일 기상예보자료를 활용하여 예보하며, 신경망 모델, 회귀모델, 의사결정모델의 통계기법을 활용하여 종합적으로 예보가 진행된다. 내일예보모형의 실행은 17시에 수행하고 예보결과 표출은 18시에 한다. 예보변수로는 대기자동측정망에서 측정한 대기질 측정자료, 서울 및 오산고층 기상대에서 관측한 기상자료, 그리고 MM5로 계산한 예보자료를 각각 사용하였다. 구체적으로 기술하면, 대기질 측정자료는 SO_2, NO_2, PM10, CO, O_3가 인자로 하여 01～16시의 평균값이 사용되었고, 지표기상 측정자료는 온도, 풍속, 습도의 01～16시의 평균, 01～16시의 누적 강수량합이 사용되었다. 기상예보자료는 예보일(익일)의 01～24시 평균 혼합고, 풍속, 온도, 습도, 500 hpa의 풍향, 500 hpa의 풍속 자료가 사용되었고, 06시의 풍속, 습도, 노점온도를 사용하였으며, 추가로 계절적 특성을 반영하기 위해 월을 변수로 입력하였다.

그림 2에 당일예보모형의 구조를 나타내었다. 당일예보모델은 대기질 및 기상 측정자료를 바탕으로 예

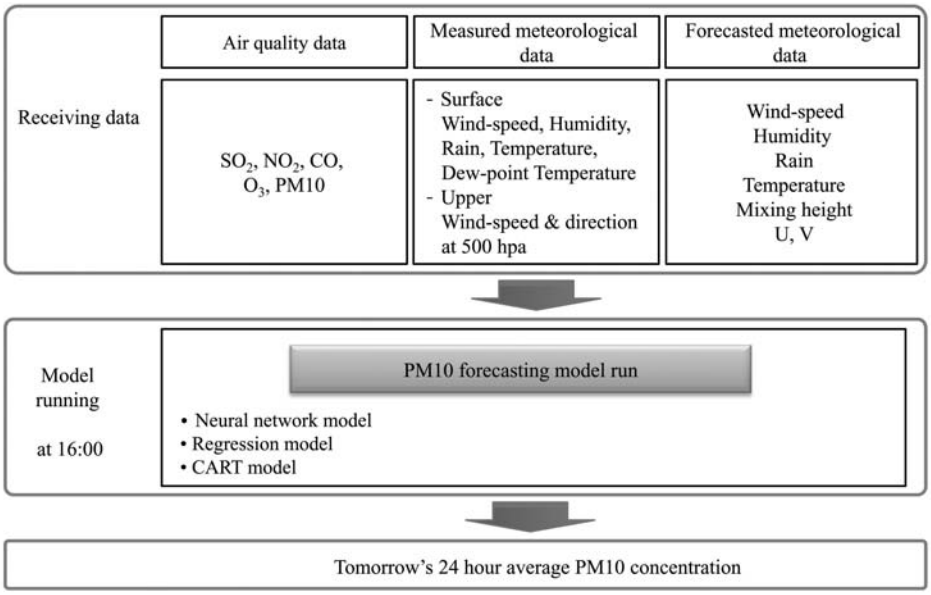


Fig. 1. The outline of tomorrow's PM10 forecasting system.

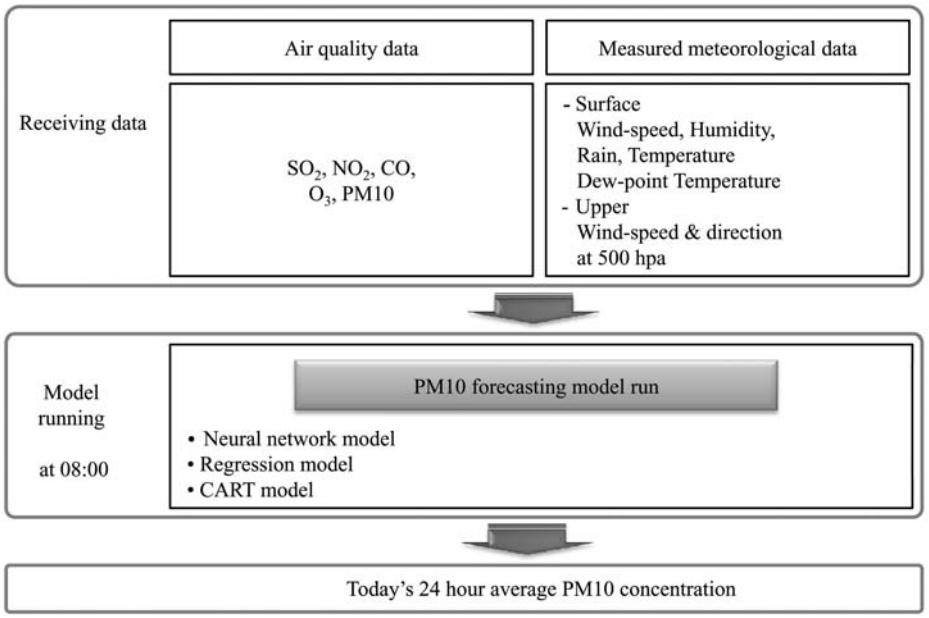


Fig. 2. The outline of today's PM10 forecasting system.

보되며 내일모형과 동일하게 신경망 모델, 회귀모델, 의사결정모델 결과를 종합하여 예보한다. 당일예보모형의 실행은 08시에 수행하고 예보결과 표출은 09시에 한다. 당일예보모형에서 입력변수는 대기질 및 기

상 측정자료만 사용하였다. 대기질 측정자료는 SO_2, NO_2, PM10, CO, O_3를 인자로 하여 01～08시의 평균이 사용되었고, 지표기상 측정자료는 온도, 풍속, 습도의 01～08시의 평균, 전날의 01～24시의 누적 강수량합, 06시의 풍속, 습도, 노점온도가 사용되었다. 고층기상자료는 03시의 500 hpa의 풍향, 풍속자료를 사용하였다. 추가로 계절적 특성을 반영하기 위해 월을 변수로 입력하였다.

3. 예보 입력변수와 미세먼지 농도와의 상관관계 분석

본 연구에서 고려한 예보 입력변수는 대기질 측정자료, 지상 및 고층 기상관측자료, 그리고 기상예보자료이며, 이를 각각에 대해서 예보 대상인 목표 미세먼지 농도(target PM10)와의 관계를 기술하면 아래와 같다. 당일예보에서 목표 미세먼지 농도는 당일 24시간 평균농도이고, 내일예보에서는 익일의 24시간 미세먼지 평균농도이다.

상관관계 분석 기간은 2006. 01. 01～2009. 05. 31이고, 서울남동지역을 대상으로 분석하였다. 지표기상자료는 서울기상대 자료를 활용하였고, 고층기상자료는 오산기상대 자료를 사용하였다. 한편 대기질 측정자료는 서울남동지역에 위치한 반포동, 도곡동, 방이동 및 천호동 대기자동측정소 자료를 이용하였다. 예보기상자료는 MM5 모델링을 수행하여 3 km 해상도의 모델링 결과로부터 서울남동지역에 해당되는 자료를 추출하여 사용하였다. 주요 변수들의 상관성에 대해서 간략히 기술하면 다음과 같다.

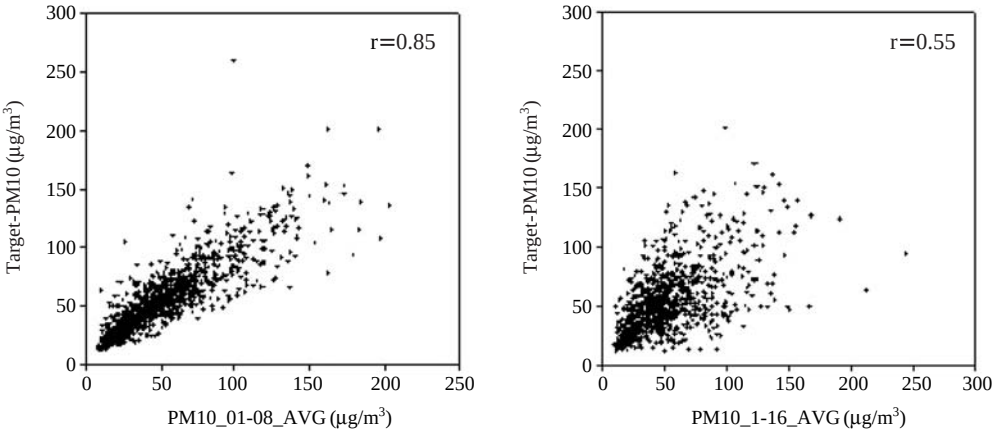
대기질 측정자료와 목표 미세먼지 농도와 상관성은 매우 크게 나타났고, 그중에서 대표적으로 PM10과의 상관성을 그림 3의 (a)와 (b)에 각각 나타났다. 그림 변수명은 “오염물질_통계시간_통계방법”으로 나타내었다. 예로써, 그림 3(a)에서 PM10_01-08_AVG는 01시부터 08시까지 측정된 시간별 PM10의 평균농도이다. 그림 3(a)는 당일예보에서 예보시점이 9시이므로 08시까지의 미세먼지 평균농도와 당일목표농도와의 상관관계를 나타낸 것이고, 그림 3(b)는 내일예보인 경우에는 예보의 시점이 17시이므로 16시까지의 미세먼지 측정자료 평균과 익일의 24시간 평균 목표 미세먼지 농도와의 상관관계를 나타낸 것이다. 특히 당일 01시부터 08시까지 측정된 미세먼지와 그날의 24시간 평균 미세먼지와의 상관관계는 0.85로 08시 이후에 기상상황이 급격하게 변하지 않으면, 측정된 PM10의 농도가 그날의 PM10 농도를 결정할 것으로 예상된다. 한편 내일예보인 경우에는 당일예보와 유사한 상관성은 보이고 있으나, 상관도가 상대적으로 낮다. 이는 미세먼지 농도가 급격한 기상패턴의 변화와 중국에서 월경성 미세먼지의 영향으로 급격하게 변하는 경우에는 그 상관성이 떨어지는 것으로 판단된다.

미세먼지 농도에 가장 영향이 큰 것으로 예상되는 기상자료에 대해서 검토한 결과, 주요 변수는 풍속, 강수량, 노점편차, 그리고 혼합고이다. 풍속이 약하고, 강수량이 적고, 혼합고가 낮을수록 미세먼지의 농도는 높아지는 경향을 보이고 있다. 미세먼지 농도가 증가할수록 박무, 안개, 연무 등의 기상빈도가 증가(Shin *et al.*, 2007)하므로 이를 반영할 수 있는 인자로 노점편차를 추가하였다. 06시의 온도와 이슬점온도와의 차이인 노점편차(dew-point deficit)는 음의 상관성을 나타내고 있는데, 이는 노점편차 값이 작을수록 안개가 생성될 가능성이 높다는 것을 의미하는 것이다.

그림 4에 있는 500 hpa (약 고도 5 km)에서 고층기상요소와 미세먼지와의 상관성을 분석한 것으로 풍향이 270도 근방인 서풍계열의 바람이 형성될 경우에 미세먼지 농도가 높은 것으로 나타났다. 이는 서풍일 경우에 중국에서 월경하는 미세먼지의 영향으로 서울이 미세먼지 농도가 높아지는 것을 의미하는 것으로 정성적으로 판단된다. 그러나 북한을 포함한 경기도 북부지역 또는 인천지역에서 배출된 오염원의 영향 등에 대한 보다 자세한 해석을 위해서는 기상 및 화학수송모델을 결합한 수치모델의 적용이 필요한 것으로 판단되며, 이는 향후 수치모델을 이용한 예보모형을 적용하면 보다 명확한 결과를 얻을 것으로 생각된다.

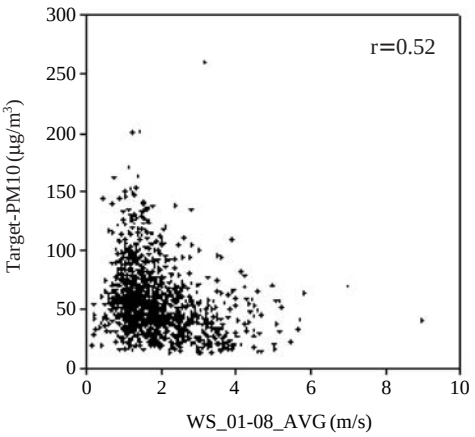
4. 미세먼지 통계예보모형의 이론

미세먼지 통계모형은 앞에서 설명된 예보변수를 입력자료로 하여, 신경망 모형, 회귀모형, 의사결정모형을 각각 수행하여 예보농도와 예보지수를 각각 계산하고, 세 개의 모형의 예보지수중에서 가장 빈도가

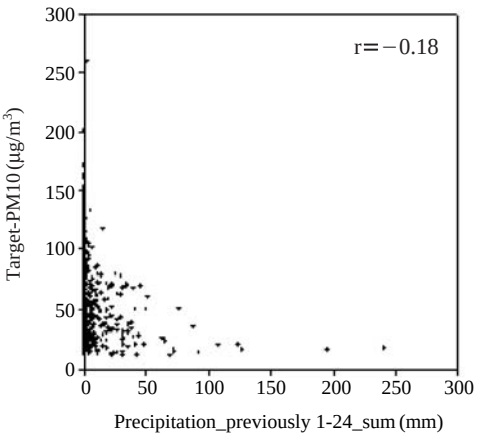


(a) Today's PM10

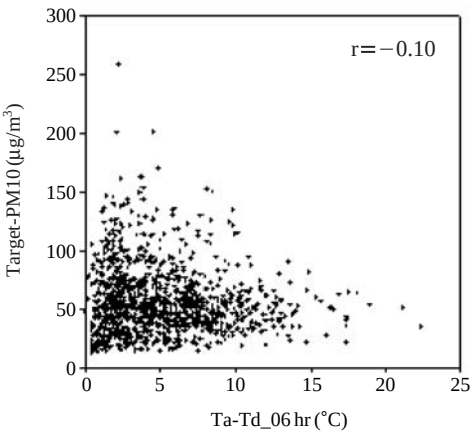
(b) Tomorrow's PM10



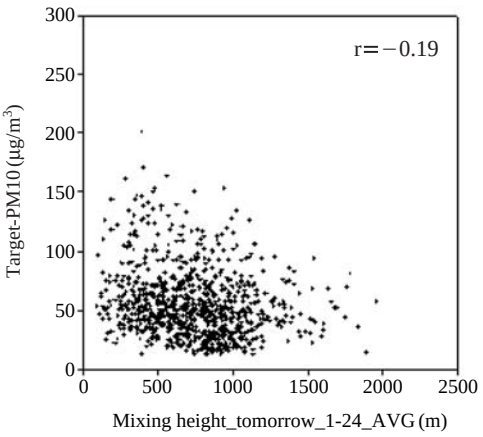
(c) Wind speed



(d) Precipitation

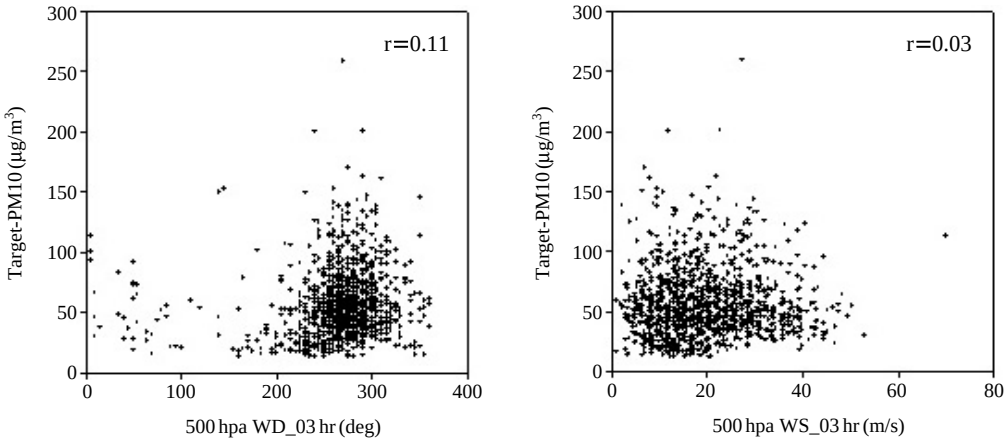


(e) Dew-point deficit



(f) Mixing height

Fig. 3. Correlation between Target PM10 and measured input variables.



(a) Wind direction (b) Wind speed

Fig. 4. Correlation between Target PM10 and observed 500 hpa upper-air meteorological input variables.

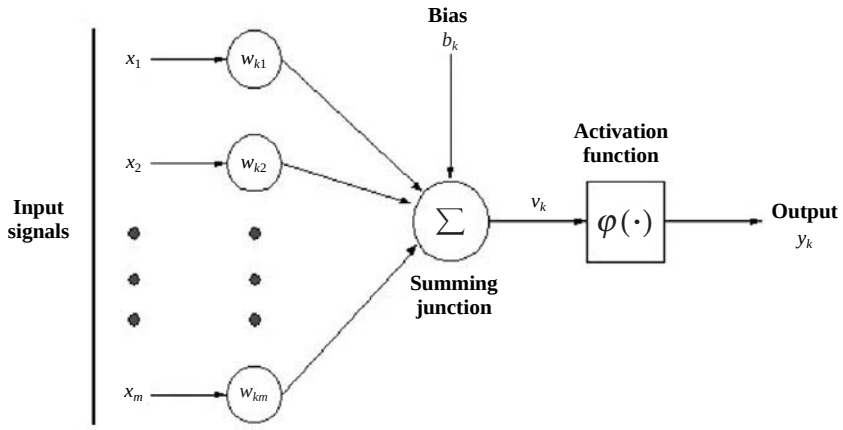


Fig. 5. Structure of MLP (Multi-Layer Perceptron) neural network system.

높은 지수를 이용하여 최종 예보치를 확정하였다. 예보모형은 본 장에서 설명하고, 예보지수는 다음장에 기술하였다. 본 연구에서 사용된 예보모형을 간략히 설명하면 다음과 같다.

4.1 신경망 모형

신경망은 과거의 경험을 학습시켜 미지의 입력에 대한 출력을 생성하는 비모수 모형으로, 입력 인자 간, 또는 입력과 출력간의 상호 인과 관계가 불분명할 경우, 그들 간의 관계를 효율적으로 찾는 특성이 있다. 본 연구에서는 신경망 모형 중 가장 범용성이 높은 오류 역전파 학습(Error Back Propagation) 기능을 갖는 다층 인식자(Multi-Layer Perceptron) 신경망 모형을 기본 모형으로 하였고, 신경망의 구조는 주어진 문제가 전형적인 비선형 관계인 점을 고려하여 입력층과 은닉층(1), 은닉층(2), 출력층의 네 개 층으로 구성하였으며 출력층을 제외한 각층의 노드 수는 실험 환경 즉 입력 인자들에 따라 다양하게 변하도록 하였고, 전체 구조는 그림 5와 같다(Rumelhart *et al.*, 1986; Parker, 1985).

이 때 층간의 연결선은 두 뉴론 간의 신호가 전달될 때 곱하기 형식으로 가중치로 작용한다. 따라서 이 연결선의 가중치를 적절히 변화시켜 특정 입력에 원하는 출력이 나오도록 조정하는 것이 학습 과정이며 또한 문제 해결 과정이 된다. 이 가중치가 원하는 답을 얻도록 학습 시키는 과정은 오류 역전파 학습으로 하였다. 오류 역전파 학습은 과정 ① 입력 자료를 입력 노드에 적용, ② 입력에 따른 출력을 계산, ③ 출력과 원하는 출력간의 오차를 계산, ④ 오차를 줄이도록 가중치의 증감 여부 결정, ⑤ 각각의 가중치 변화량 결정, ⑥ 가중치를 갱신(변경) ⑦ 모든 학습 자료에 대해 오차가 적정 수준으로 감소하기까지 1단계에서 6단계를 반복이다.

4. 2 회귀모형

회귀 분석(Regression Analysis)은 먼저 변수들 간의 관계를 나타내는 타당한 수학적 모형을 설정하고, 변수들의 측정된 값을 이용하여 그 모형을 추정한 다음, 추정한 모형에 의해 변수들 간의 관계를 설명하든지 또는 예측 등의 분석에 응용하는 통계적 방법 중 하나이다(Colbourn, 2010; U.S. EPA, 2003). 예를 들면, 일평균 미세먼지량(Y)과 전일 평균 습도(X)의 관계에 대한 수학적 모형 Y=f(X)를 추정하였다면 먼지량과 습도간의 관계를 설명할 수 있을 뿐만 아니라, 특정일 전날 평균 습도를 입력하면, 그 날의 먼지 량을 예측할 수 있다. 회귀 분석에서 Y=f(X)와 같이 변수들 간의 관계를 나타내는 수학적 모형을 회귀식(regression equation)이라고 하며, 서로 관계를 가지고 있는 변수들 중에서 먼지의 량 Y와 같이 다른 변수에 의해 영향을 받는 변수를 종속 변수라고 하며, 이는 주어진 문제에서 설명하고자 하는 변수로써, 주로 다른 변수들이 주어지는 경우 그에 대한 반응으로 관측되는 변수이므로 반응 변수라고도 한다. 이러한 종속 변수에 영향을 주는 습도 X와 같은 변수는 독립 변수라고 하며, 종속 변수를 설명하는데 이용하는 변수이므로 설명 변수라고도 한다. 선형 회귀에는 회귀식에 사용된 독립 변수의 수에 따라 하나인 경우인 단순선형회귀와 둘 이상인 경우인 중선형회귀가 있다. 본 연구에서는 중선형회귀 모형을 사용하였으며, 중선형 회귀모형은 실제 응용에 있어서 단순 회귀모형 보다 실제적이다. 그 이유는 독립 변수 1개로 특정 현상을 설명할 수 있는 경우는 드물고, 대부분의 경우, 종속 변수는 다수의 독립 변수에 의해 영향을 받기 때문이다. 중선형회귀 모형에서는 종속 변수 Y와 k개의 독립 변수 X_1, …, X_k가 다음과 같은 관계식을 가진다고 가정한다.

$$Y_i=a_0+a_1X_{i1}+\cdots+a_kX_{ik}+e_i \quad (1)$$

즉, 종속 변수는 각 독립 변수의 일차 함수로 나타내어지며, 여기에 오차항을 나타내는 확률 변수 e가 더해진다. 이때 오차항을 최소화하기 위해 최소 제곱법은 잔차의 제곱의 합을 최소화하는 b를 구해 아래와 같은 회귀방정식을 구한다.

$$Y=b_0+b_1X_1+\cdots+b_kX_k \quad (2)$$

본 연구에서는 SYSTAT v. 10.2를 사용하여 회귀방정식을 산정하였다.

4. 3 의사결정모형

의사결정법(Decision Tree)이란 의사결정규칙을 나무구조로 도표화하여 대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석 방법이다.

의사결정법의 장점은 모형을 쉽게 이해할 수 있고 나무 구조로부터 어떤 예측변수가 목표변수를 설명하는데 더 중요한지를 쉽게 파악할 수 있다는 것과 두개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 알 수 있으며, 이상치에 민감하지 않다는 것이다(U.S. EPA, 2003).

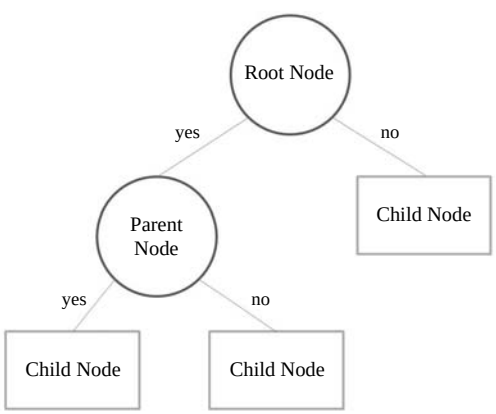


Fig. 6. Structure of Decision Tree.

Table 3. Health category classification of PM10 concentration range to CAI.

Category	Good	Moderate	Unhealthy for sensitive groups	Unhealthy	Very unhealthy
PM10 (μg/m^3)	0 ~ 30	31 ~ 80	81 ~ 120	121 ~ 200	201 ~ 300

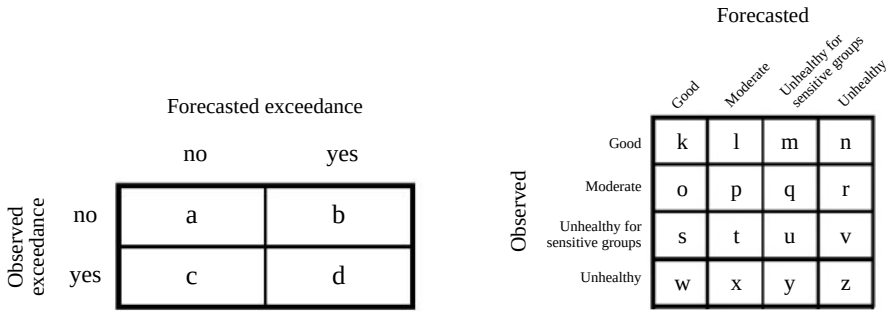


Fig. 7. Contingency table for category forecast.

그림 6과 같은 의사결정나무의 구성요소는 나무처럼 형성되며 나무구조가 시작되는 마디로 전체 자료로 이루어지는 뿌리마디(Root Node), 하나의 마디로부터 분리 되어 나간 2개 이상의 마디들을 의미하는 자식마디(Child Node), 각 자식마디의 상위 마디를(Parent Node)라고 한다. 하나의 마디로부터 끝마디까지 연결된 일련의 마디들을 가지(Branch)라고 하며 가지를 이루고 있는 마디의 개수를 깊이(Depth)라고 한다.

의사결정나무의 분석과정은 하나의 부모마디로부터 자식 마디들이 형성될 때, 예측변수를 선택하는 기준인 분리기준과, 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러 가지 규칙을 적용하여 의사결정나무를 형성하고, 형성된 의사결정나무에서 적절하지 않은 마디를 제거한다. 이렇게 결정된 의사결정나무는 타당성을 평가하고 실제 자료를 적용하여 해석 및 예측을 하게 된다. 본 연구에서는 SYSTAT v. 10.2를 사용하여 의사결정모형을 개발하였다.

5. 미세먼지 예보모형의 평가

앞에서 설명된 예보입력변수를 적용하여 예보모형 학습을 통해서 예보모델을 완성하고, 완성된 예보모형을 적용하여 모델의 예보 정합도를 평가하였다.

Table 4. Statistics used to evaluate category forecasts.

Statistic name	Equation
Accuracy (%)	A=(k+p+u+z)/N∗100
Upward accuracy (%)	A_{UP}=(l+m+n+q+r+v)/N∗100
Downward accuracy(%)	A_{down}=(o+s+t+w+x+y)/N∗100
Bias	B=(b+d)/(c+d)
False alarm rate (%)	FAR=b/(b+d)∗100
Provavility of detection (%)	POD=d/(c+d)∗100

5.1 미세먼지 예보모형 평가 방법

지자체에서 운영되는 미세먼지 예보결과는 농도구간을 인체의 위해성과 연계하여 분류한 표 3의 통합대기환경지수(CAI, Comprehensive Air-quality Index)로 예보된다. 통합대기환경지수는 좋음, 보통, 민감군 영향, 나쁨, 매우 나쁨으로 구분되어 있고, 관련 국립환경과학원 보고서(NIER, 2006) 및 한국환경공단에서 운영하는 AirKorea site (http://www.airkorea.or.kr/)에 자세한 설명이 있다. 한편 신경망 및 회귀 모형은 예보결과로 미세먼지 농도가 계산되므로, 농도값에 대한 시계열 비교 등을 통해서 1차적으로 정합도를 평가하고, 최종적으로는 계산된 미세먼지 농도를 통합대기환경지수로 변환하여 평가를 수행하였다.

그림 7은 예보평가를 위한 지수구간을 나누는 것을 표시한 것이며, 표 4는 지수구간별로 통계분석하는 방법을 정리한 것이다. 지수일치도는 각 구간이 정확히 일치하는 확률을 계산한 것으로 100에 가까

울수록 좋다. 편의는 민감군 이상의 농도(이하 기준값)를 초과한 측정일수 대비 예보일수로 1에 가까울수록 좋다. 거짓경보율은 기준값을 초과한 예보일수 대비 기준값보다 낮게 나타난 비율로 0이면 고농도를 예보한 것은 전부 맞춘 것이다. 감지확률은 기준값을 초과한 측정일수 대비 정확히 예보한 일수로 100이면 측정된 고농도를 모두 예보한 것이다. 한편 감지확률과 거짓경보율은 미세먼지 농도를 그림 7의 좌측에 표시된 바와 같이 2개의 구간으로 구분하여 경계치는 보통과 민감군 영향의 구분경계인 80 μg/m^3을 기준으로 하였다.

5.2 미세먼지 예보모형 학습 결과

미세먼지 예보모형의 학습기간은 당일예보모형 2006. 01. 01～2009. 05. 31, 내일예보모형 2006. 07. 01～2009. 05. 31으로 하였으며, 기간 중 황사발생일은 제외하였다. 내일예보모형의 2006. 01. 01～2006. 06. 30 기상예보자료가 부족하여 학습기간이 단축되었다.

예보모형의 학습결과를 확인하기 위하여 학습기간에 측정한 미세먼지 농도와 예측한 미세먼지 농도를 비교하였다. 그림 8과 9에 있는 자기 학습기간 중에 모형예보농도는 측정농도를 잘 모사하는 것을 알 수 있다. 특히 고농도치도 잘 따라가고 있다. 회귀모형의 결과는 그림 10과 11에 각각 나타내었다. 회귀모형은 신경망모형에 비해서 고농도 영역에서 저평가 되는 경향을 보이고 있다.

한편 신경망 및 회귀모형에서 계산된 목표 미세먼지 농도를 통합대기환경지수로 환산하고, 통합대기환경지수로 예보되는 의사결정모형결과를 종합하여 최종예보(final forecast)를 평가한 것을 표 5에 정리하였다. 지자체 운영시 최종예보는 예보자의 판단에 의해 최종 예보하면 예측도가 높아질것으로 판단되나, 시스템상에서 자동 예보는 위의 세가지 모형중에 가장 빈도가 높은 지수를 최종예보하므로, 본 논문에서는 세 모형중 빈도가 높은 지수를 최종예보지수로 평가하였다. 당일예보모형의 지수일치도는 82.7%, 거짓경보율은 22%, 그리고 감지확률은 69.6%이고, 내일예보모형은 지수일치도가 70.8%, 거짓경보율이 32.1%, 감지확률이 41.7%로 당일예보에 비해서 다소 정확도가 떨어지는 것으로 나타났다. 이는 당일예보모형에 비하여 내일예보모형이 급격하게 변하는 기상인자 및 월경성 미세먼지를 반영하지 못하는 한계에 기인한 것으로 판단된다.

5.3 미세먼지 예보모형 평가 결과

앞에서 개발된 예보모형을 실제 미세먼지 예보시스템을 구축하여 현업에 적용하였다(http://cleanair.seoul.go.kr/main.htm). 실제 모형의 적용 평가 기간은 예보모형개발에 사용하지 않은 2010. 04. 01～2010. 07. 31이며, 예보된 미세먼지 농도 및 통합대기환경지수를 측정값과 비교하였다. 단, 비교기간 중 황사발생

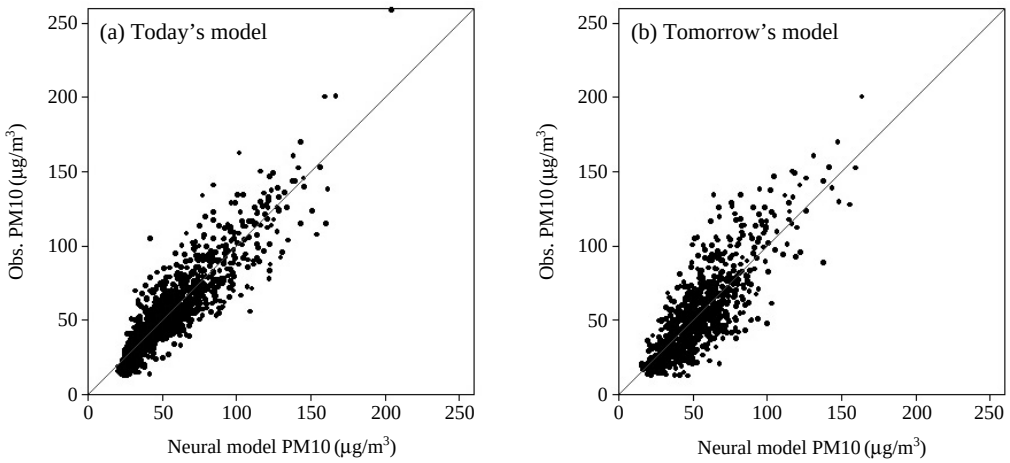


Fig. 8. Comparison of observed PM10 with predicted PM10 by neural network model for a self-studying period.

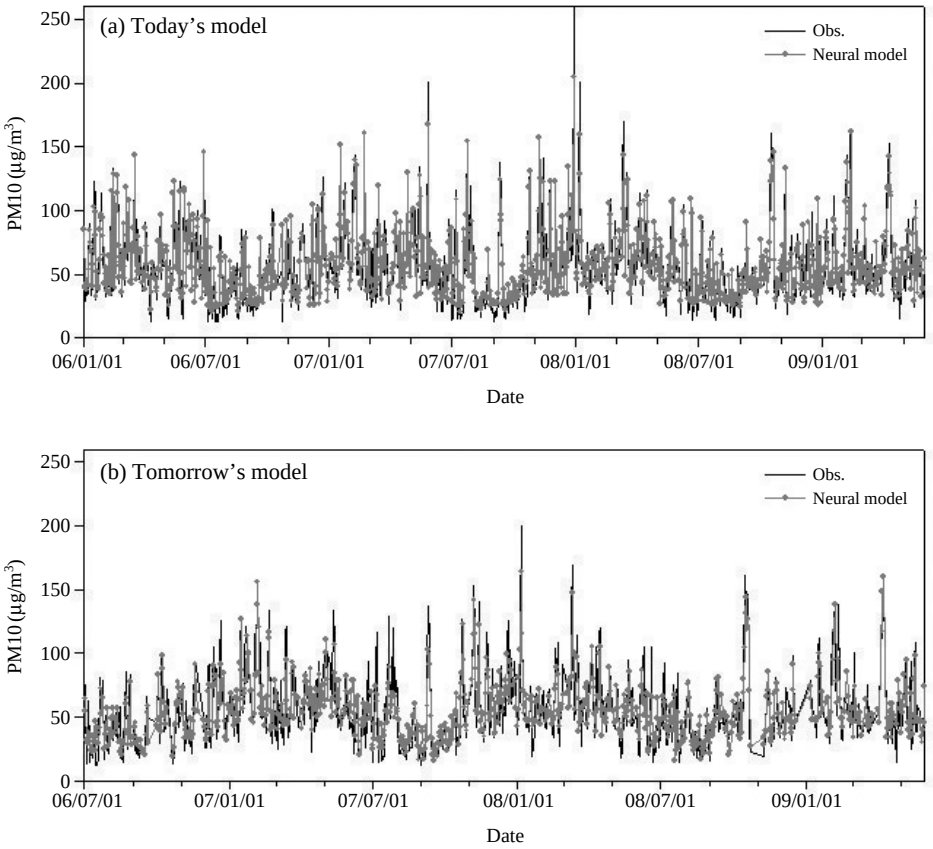


Fig. 9. Time-series comparison of observed PM10 with predicted PM10 by neural network model for a self-studying period.

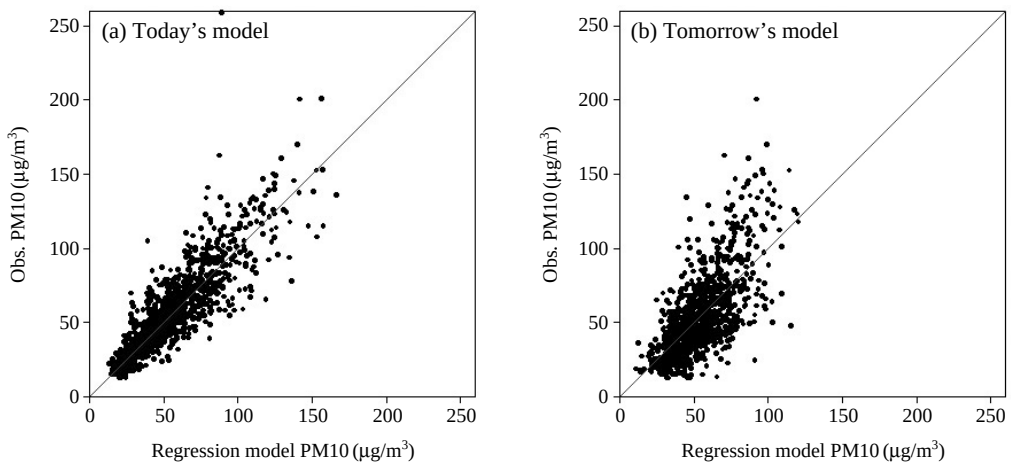


Fig. 10. Comparison of observed PM10 with predicted PM10 by regression model for a self-studying period.

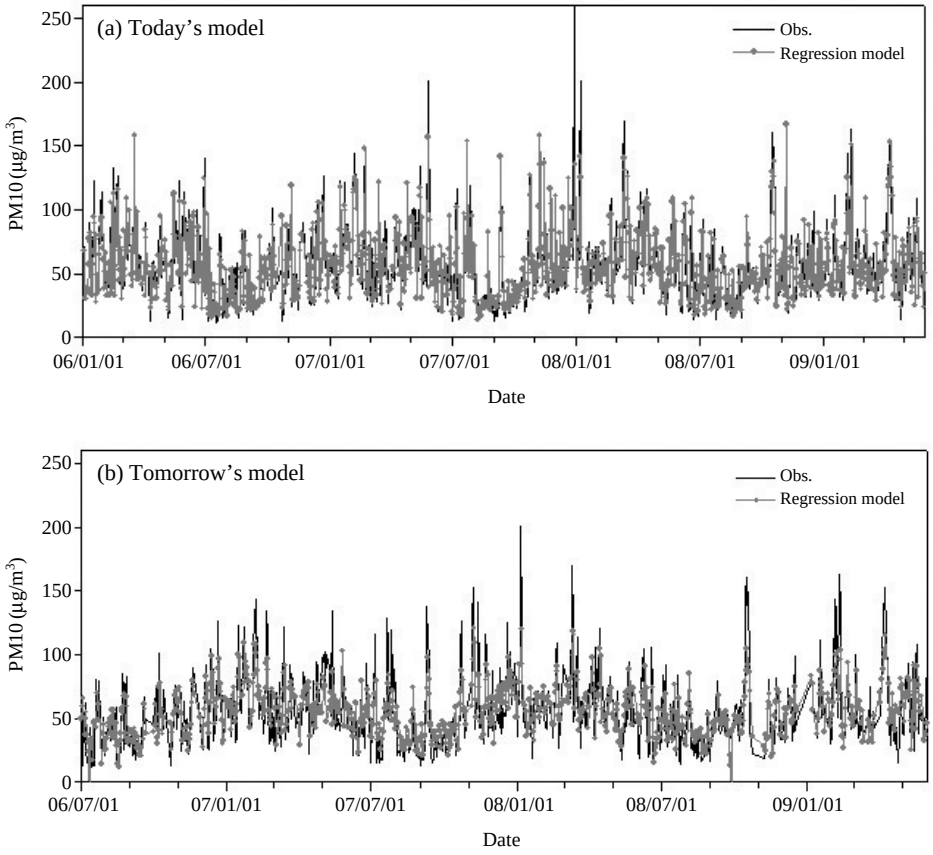


Fig. 11. Time-series comparison of observed PM10 with predicted PM10 by regression model for a self-studying period.

Table 5. Statistics of model performance for self-studying period.

Statistic name	Neural model		Regression model		Decision tree model		Final forecast	
	Today's	Tomorrow's	Today's	Tomorrow's	Today's	Tomorrow's	Today's	Tomorrow's
Accuracy (%)	81.5	75.4	82.6	69.7	79.9	61.9	82.7	70.8
Upward accuracy (%)	11.2	13.5	8.3	16.3	4.5	21.1	8.1	16.3
Downward accuracy (%)	7.2	11.1	9.1	14.0	15.5	17.0	9.2	12.8
Bias	0.99	0.74	0.92	0.61	0.69	0.69	0.89	0.61
False alarm rate (%)	26.4	26.6	22.9	35.9	15.7	51.7	22.0	32.1
Probability of detection (%)	72.8	54.3	71.2	39.4	58.2	33.1	69.6	41.7

일은 제외 하였다.

신경망예보모형의 미세먼지 농도예보값과 측정값을 비교하여 그림 12와 13에 정리하였고, 회귀모형의 미세먼지 농도예보값과 측정값을 비교하여 그림 14와 15에 각각 나타내었다. 평가기간 중 80 μg/m^3 이상의 농도 발생일수가 9회로 고농도일에 대한 예보능력을 평가하기에는 다소 미흡하지만, 당일예보에서는 고농도가 잘 모사되고 있는 반면에, 내일예보모형에

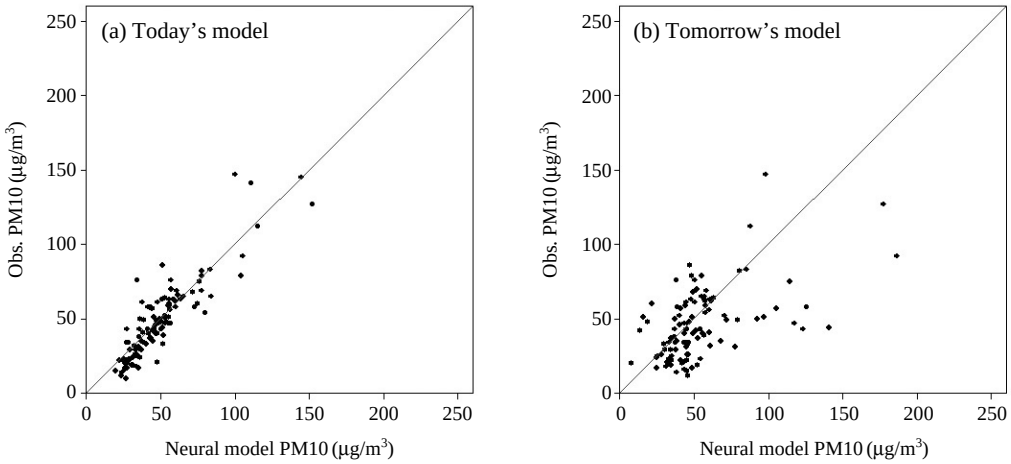


Fig. 12. Comparison of observed PM10 with predicted PM10 by neural network model for an operation period.

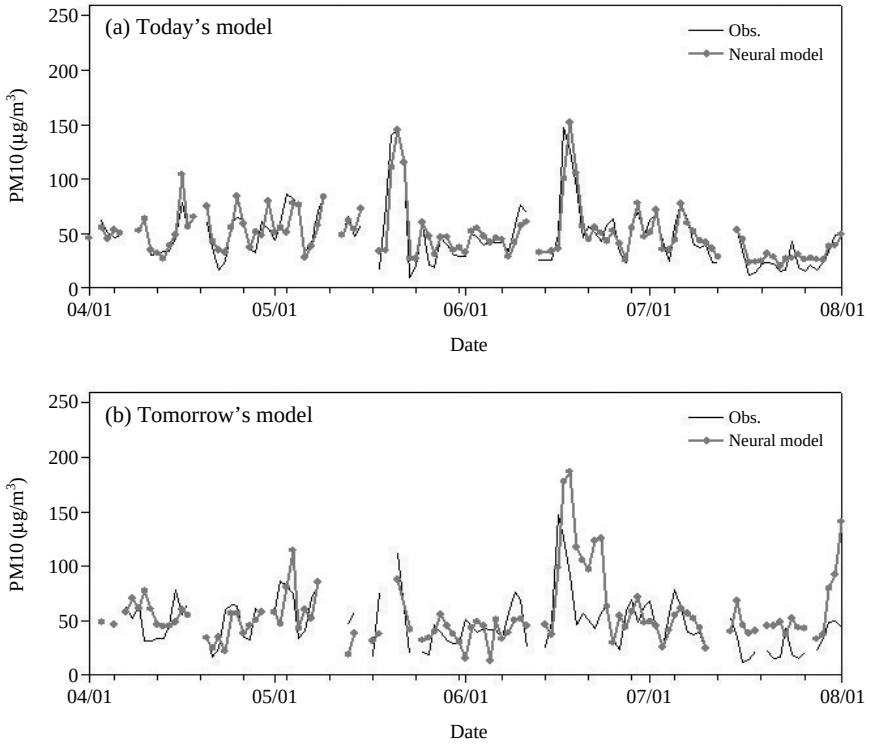


Fig. 13. Time-series comparison of observed PM10 with predicted PM10 by neural network model for an operation period.

서 고농도일에 대해서 신경망모형은 과대평가, 회귀모형은 과소평가를 하는 경향이 있다.

한편 표 6에 있는 신경망 및 회귀모형, 의사결정모형으로 예보된 통합대기환경지수중에 가장 빈도수가

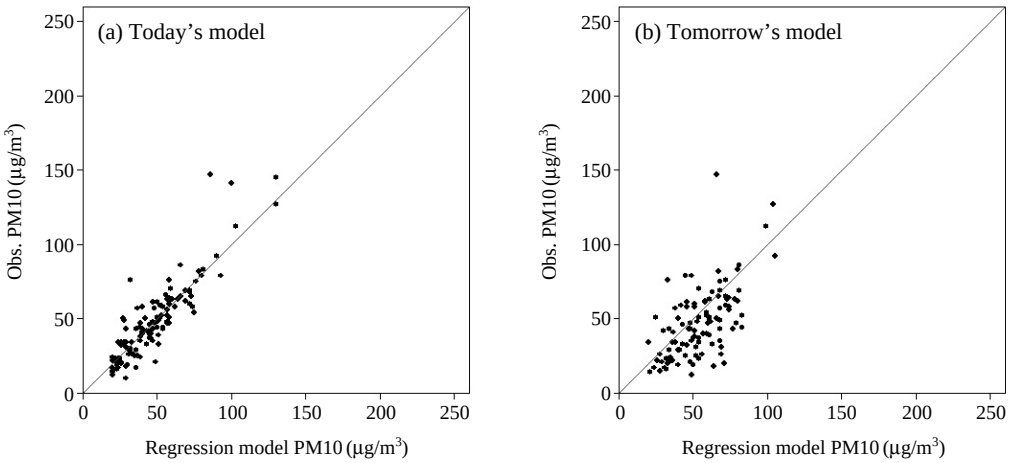


Fig. 14. Comparison of observed PM10 with predicted PM10 by regression model for an operation period.

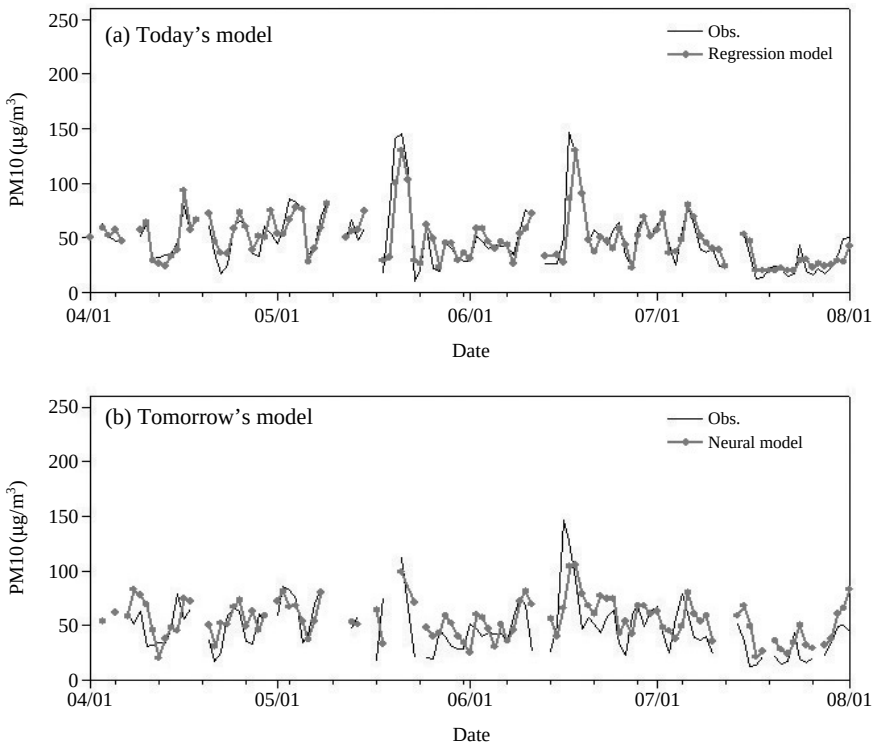


Fig. 15. Time-series comparison of observed PM10 with predicted PM10 by regression model for an operation period.

높은 지수로 예보를 결정하는 최종예보치를 기준으로 판단하면, 당일예보모형의 지수일치도는 80.8%, 거짓경보율은 12.5%, 그리고 감지확률은 77.8%이고, 내일예보모형은 지수일치도가 72.4%, 거짓경보율이

Table 6. Statistics for operating performances of the forecasting system.

Statistic name	Neural model		Regression model		Decision tree		Final forecast	
	Today's	Tomorrow's	Today's	Tomorrow's	Today's	Tomorrow's	Today's	Tomorrow's
Accuracy (%)	77.5	62.2	79.2	70.4	82.5	66.3	80.8	72.4
Upward accuracy (%)	15.8	30.6	10.8	22.4	5.8	17.3	10.8	20.4
Downward accuracy (%)	6.7	7.1	10.0	7.1	11.7	16.3	8.4	7.2
Bias	1.00	1.86	0.89	0.86	0.78	0.29	0.89	0.43
False alarm rate (%)	22.2	53.8	12.5	33.3	14.3	100	12.5	0.0
Probability of detection (%)	77.8	85.7	77.8	57.1	66.7	0.0	77.8	42.9

0.0%, 감지확률이 42.9%로 학습기간에 대해서 평가한 결과와 지수일치도는 유사하나, 거짓경보율 및 감지확률은 오히려 학습기간에 비해서 양호한 수치를 보이고 있다. 이는 모델링 평가기간에 월경성에 의한 고농도 발생횟수가 적은 것에 기인한다고 판단된다.

6. 결 론

수도권지역에 농도가 높고 위해성이 큰 미세먼지를 대상으로 사전예방적 차원에서 미세먼지 예보시스템을 개발하였다. 미세먼지 예보시스템은 오후 5시에 익일의 일평균 미세먼지를 예보하는 내일예보와 내일 예보치를 예보 당일 오전 9시에 당일예보를 수행하여 수정 예보할 수 있도록 구축하였다.

대기질 측정자료, 지표 및 고층 기상자료 및 예보기상자료를 예보변수로 사용하였다. 예보변수 중에서 대기질 측정자료 중에서는 예보시점 이전까지 측정된 미세먼지 농도와 목표 예보농도와 상관성이 가장 크게 나타났고, 기상자료는 풍속, 강수량, 06시의 노점편차, 그리고 혼합고가 주요 변수이다. 풍속이 약하고, 강수량이 적고, 혼합고가 낮을수록 미세먼지의 농도는 높아지는 경향을 보이고 있다. 노점편차 값이 작을수록 안개가 생성될 가능성이 높고, 아침에 형성된 안개가 연무가 되어 고농도를 유발하는 것으로 판단된다. 또한 500 hpa 고도에서 서풍계열의 바람이 형성될 경우에 미세먼지 농도가 높은 것으로 나타났다. 500 hpa 고도에서의 풍향은 대류권의 평균 흐름에 대한 값으로 서풍일 경우는 중국에서 월경한 미세먼지의 영향으로 서울이 미세먼지 농도가 높아지는 것으로 해석할 수 있다.

예보시스템에 이용된 예보모형은 신경망모형, 회귀모형 및 의사결정모형이고, 각각의 모형에서 예보된 통합대기환경지수 중에서 가장 높은 빈도를 나타내는 지수값을 최종 예보치로 확정하였다.

각각의 모형을 입력변수로 학습을 시키는 기간은 2006. 07. 01부터 2009. 05. 31까지의 측정된 입력변수들을 이용하여 각각의 모형을 학습시키고, 2010. 04. 01부터 2010. 07. 31까지 실제 예보시스템을 운영하여 예보정합도를 평가하였다.

평가기간 중에 80 μg/m^3 이상의 농도 발생일수가 9회로 고농도일에 대한 예보능력을 평가하기에는 다소 미흡하지만, 당일예보에서는 고농도가 잘 모사되고 있는 반면에 내일예보모형에서 고농도일에 대해서는 신경망모형을 과대평가, 회귀모형은 과소평가를 하는 경향이 있다. 한편 대기환경지수 중에 가장 빈도수가 높은 지수로 예보를 결정하는 최종예보치를 기준으로 운영결과를 판단하면 당일예보모형의 지수일치도는 80.8%, 거짓경보율은 12.5%, 그리고 감지확률은 77.8%이고, 내일예보모형은 지수일치도가 72.4%, 거짓경보율이 0.0%, 감지확률이 42.9%로 학습기간에 대해서 평가한 결과와 지수일치도는 유사하나 거짓경보율 및 감지확률은 오히려 학습기간에 비해서 양호한 수치를 보이고 있다. 이는 평가 기간 중에 월경성 오염물질의 영향이 상대적으로 적었기 때문이다.

일반적으로 통계예보치가 낮게 예측하는 경우에는 주로 중국 등에 의해서 외부오염물질이 국내에 영향을 미치는 월경성에 의한 고농도가 발생한 경우로 판단된다. 그러나 본 연구에서는 구체적으로 서풍계열의 바람이 경기도 북부지역 또는 인천지역을 통과하면서 그 지역의 영향이 반영된 것에 대한 구체적인 구분을 할 수 없는 한계를 갖고 있다. 이와 같이 급격하게 외부 요인에 의해서 고농도가 발생하는 경우나

또는 경기도 북부지역 또는 인천지역의 배출원의 영향을 구별하여 통계모형을 구축하는데 한계가 있기 때문에 향후 수치예보모형을 도입한 예보모형의 개발이 필요한 것으로 판단된다.

감사의 글

본 연구의 일부는 국립환경과학원의 연구비에 의해서 지원되었으며, 또한 주저자의 연구년 기간 중에 수행되었습니다.

참 고 문 헌

Brian, E., K. Daiwen, S.T. Rao, R. Mathur, Y. Shaocai, O. Tanya, S. Ken, W. Richard, J. Scott, D. Paula, M. Jeff, and B. George (2009) A demonstration of the national air quality forecast capability, U.S. EPA's 2009 National Air Quality Conference.

Colbourn, W.G. (2010) An enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory concentrations, Atmospheric Environment, 44(25), 3015-3023.

Hooyberghs, J., C. Mensink, G. Dumont, F. Rierens, and O. Brasseur (2005) A neural network forecast for daily average PM10 concentrations in elgium, Atmospheric Environment, 39(18), 3279-3289.

Hwang, Y.J., S.J. Lee, H.S. Do, Y.K. Lee, T.J. Son, T.G. Kwon, J.W. Han, D.H. Kang, and J.W. Kim (2009) The analysis of PM10 concentration and the evaluation of influences by meteorological factors in ambient air of Daegu area, J. Korean Soc. Atmos. Environ., 25(5), 459-471. (in Korean with English abstract)

Kim, M.W., L. Hyanlin, and C. Kevin (2009) Development of PM2.5 forecasting model and a web-based air quality mapping and forecasting system for Ohio, U.S. EPA's 2009 National Air Quality Conference.

Kim, Y.P. (2006) Air pollution in Seoul caused by aerosols, J. Korean Soc. Atmos. Environ., 22(5), 535-553. (in Korean with English abstract)

Kim, Y.P. (2010) Analysis of the trend of atmospheric PM10 concentration over the Seoul Metropolitan Area between 1999 and 2008, Korean J. of Environ. Impact Assessment, 19(1), 5-74. (in Korean with English abstract)

Koo, Y.S., H.Y. Kwon, and H.Y. Yun (2003) Development of real-time air quality forecasting system using the statistical Model (PM10), Proceeding of the 36th Meeting of KOSAE, 445-446.

Koo, Y.S., W.J. Yoon, H.Y. Kwon, J.M. Yang, J.H. Choi, and H.Y. Yun (2005) Development of PM10 forecasting system of the day before, Proceeding of the 39th Meeting of KOSAE, 403-404.

Lee, J.B., T.H. Cheon, and Y.S. Koo (2008) study on the prediction of atmospheric ozone concentration using neural network, Proceeding of the 46th Meeting of KOSAE, 231-232.

Lee, Y.S., H.G. Kim, J.S. Park, and H.K. Kim (2006) A study on statistical forecasting models of PM10 in Pohang region by the variable transformation, Korean J. of Environ. Impact Assessment, 22(5), 614-626. (in Korean with English abstract)

Murphey, B. (2007) A comparison between statistical models and three dimensional air quality models: is the more the better, U.S. EPA's 2008 National Air Quality Conference.

NIER (National Institute of Environmental Research) (2006) A study of development of comprehensive evaluation technique of air quality, 123-133.

Parker, D.B. (1985) Learning Logic, Technical Report TR-47, Center for Computational Res. in Economics and management Science, MIT.

Patricio, P. and R. Jorge (2006) An integrated neural network model for PM10 forecasting, Atmospheric Environment, 40, 2845-2851.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams (1986) Learning Internal Representation by Error Propagation: in Parallel Distributed Processing (Eds. David E. Rumelhart and James L. McClelland), The MIT Press, vol.1, 318-362.

Shin, M.K., C.D. Lee, H.S. Ha, C.S. Choe, and Y.H. Kim (2007) The influence of meteorological factors on PM10 concentration in Incheon, J. Korean Soc. Atmos. Environ., 23(3), 322-331. (in Korean with English abstract)

U.S. EPA (2003) Guidelines for developing an air quality (Ozone and PM2.5) forecasting program, EPA-456/R-03-002.

U.S. EPA (2009) Technical assistance document for the reporting of daily air quality - the Air Quality Index (AQI), EPA-454/B-09-001.

Yun, H.Y., Y.S. Koo, H.Y. Kwon, and S.H. Yu (2007) The study on the PM10 statistical forecasting model using the MM5 data, Proceeding of the 45th Meeting of KOSAE, 438-439.

Yun, H.Y., Y.S. Koo, H.Y. Kwon, and S.H. Yu (2008) A study on the PM10 statistical forecasting model using the MM5 output, Proceeding of the 47th Meeting of KOSAE, 402-403.