한국대기환경학회지 제 22 권 제 5 호
J. KOSAE Vol. 22, No. 5 (2006) pp. 614~626
Journal of Korean Society for Atmospheric Environment

변수변환을 통한 포항지역 미세먼지의 통계적 예보모형에 관한 연구

A Study on Statistical Forecasting Models of PM10 in Pohang Region by the Variable Transformation

이영섭* · 김현구[1] · 박종석 · 김희경
동국대학교 통계학과, [1]한국에너지기술연구원

(2006년 4월 26일 접수, 2006년 7월 11일 채택)

Yung-Seop Lee*, Hyun-Goo Kim[1], Jong-Seok Park and Hee-Kyung Kim
Department of Statistics, Dongguk University
[1]*Korea Institute of Energy Research*

(Received 26 April 2005, accepted 11 July 2006)

Abstract

Using the data of three environmental monitoring sites in Pohang area (KME112, KME113, and KME114), statistical forecasting models of the daily maximum and mean values of PM10 have been developed. Since the distributions of the daily maximum and mean PM10 values are skewed, which are similar to the Weibull distribution, these values were log-transformed to increase prediction accuracy by approximating the normal distribution. Three statistical forecasting models, which are regression, neural networks (NN) and support vector regression (SVR), were built using the log-transformed response variables, i.e., log (max (PM10)) or log (mean (PM10)). Also, the forecasting models were validated by the measure of RMSE, CORR, and IOA for the model comparison and accuracy. The improvement rate of IOA before and after the log-transformation in the daily maximum PM10 prediction was 12.7% for the regression and 22.5% for NN. In particular, 42.7% was improved for SVR method. In the case of the daily mean PM10 prediction, IOA value was improved by 5.1% for regression, 6.5% for NN, and 6.3% for SVR method. As a conclusion, SVR method was found to be performed better than the other methods in the point of the model accuracy and fitness views.

Key words : PM10 forecast, Regression, Neural network, Support vector regression, Log-transformation

1. 서 론

최근 들어 미세먼지(PM10)가 일반시민들의 건강을 위협하고 각종 질병을 유발하는 등 환경오염물질의 대명사로 부각되고 있다. 갈수록 악화되는 황사나 자동차의 증가로 인한 미세먼지 오염농도의 전반적인 상승 추세는 일상 생활에 직접적인 불편을 끼치고 있으며, 이에 따라 선진국에서는 미세먼지 농도를 사전에 예보함으로써 그 심각성을 주지시킴과 아울

*Corresponding author.
Tel : +82-(0)2-2260-3218, E-mail : yung@dongguk.edu

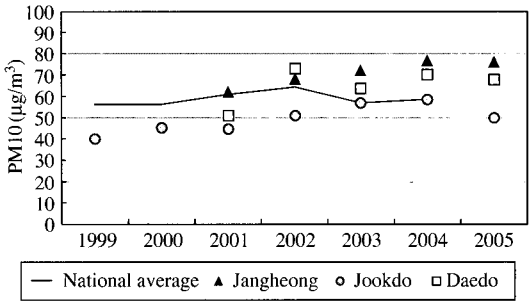


Fig. 1. Trend of annual mean concentrations of PM10 in Pohang region.

러 일기예보와 같이 일상생활의 지침으로 활용할 수 있도록 정보를 제공하고 있다. 즉, 시민들에게 이해가 쉬운 TV, 신문 등 주요 대중매체를 이용하여 예보정보를 제공하고 있으며 특히 일기예보 시간에 오존, 먼지예보 등의 대기질 예보를 병행하고 있다. 또한 예보수치가 건강을 위협하는 수준일 경우 그에 해당하는 시민행동지침을 제공하고 있다. 한편 서울시에서는 2005년 2월부터 당일 최고 먼지농도를 시민들에게 인터넷 (http://dust.seoul.go.kr/) 전광판을 통해 예보하고 예보수준에 따른 시민행동지침을 권고하고 있으며, 환경부의 보도자료 (환경부, 2005b)에 제시된 먼지예보기준과 시민행동요령에 의하면 미세먼지 농도를 5등급으로 구분하여 각 등급별로 시민행동 권고사항을 제시하고 있다.

본 연구의 대상지역인 포항시의 경우, 그림 1 (환경부, 2005a)에 제시된 바와 같이 지난 7년간 미세먼지 연평균 농도가 꾸준한 상승세를 보이고 있으며, 최근에는 장흥동과 대도동의 미세먼지 농도가 전국평균을 상회하고 있다. 포항시의 2004년도 연평균 미세먼지 농도의 전국 측정소별 순위를 살펴보면 전체 177개 측정소 중 포항시 장흥동이 26위, 죽도동이 86위, 대도동이 56위로 중상위권 수준에 해당된다. 포항시는 장기발전계획에 따라 첨단과학도시 및 환경친화적 산업도시를 표방하고 있어, 향후 먼지예보체계와 같은 환경감시 기능을 구비함으로써 장기발전계획에 일조함과 동시에 산업도시가 갖는 오염에 대한 막연한 우려를 과학적인 정보공개에 의한 이성적인 판단 및 적절한 대응으로 전환함으로써 궁극적으로 시민의 삶의 질을 향상시키는 효과가 기대된다.

우리나라도 수도권에서도 이미 2005년 1월부터 먼지경보제를 시행 중에 있으며, 그 이전부터 인터넷, 전광판을 통하여 각종 대기질 측정 정보를 제공해 왔으나 홍보 및 교육의 미흡으로 시민들의 인지도는 매우 낮을 뿐 아니라 지역에 따라서는 심지어 왜곡된 이해를 하는 경우도 있는 것으로 보고되고 있다 (김현구 등, 2006). 따라서 오존, 미세먼지 등의 대기질 정보 및 예 · 경보 상황을 체계적이면서 이해하기 쉽게 전달하는 방법의 개발도 예보모형의 개발 못지않게 중요하며 또 필요한 것이 사실이다.

본 연구에서 개발하고자 하는 예보모형은 다양한 통계적 기법을 적용하여 다음날의 최대농도(또는 평균농도)를 예측함으로써 실생활의 지침으로 제시하고자 한다. 미세먼지 예측모형 개발 및 적용에 대해서는 구윤서 등(2003, 2005)이 서울, 수도권 지역을 대상으로 회귀분석, 신경망 분석, 고농도회귀, 고농도 신경망 방법을 적용한 전일 예보모형의 개발 및 정량적 평가의 선행연구가 국내에서는 유일한 사례이다. 본 연구에서는 포항시 세 곳의 환경측정소에서의 관측자료를 이용하여 회귀분석, 신경망분석 뿐만 아니라 최근 많은 연구가 진행되고 있는 새로운 예측모형 기법인 SVR (Support Vector Regression)을 도입하여 먼지예보 모형을 개발하고, 모형에 따른 예측결과들을 상호비교 함으로써 최종적으로 예측정확도가 가장 우수한 예측모형을 선별하는 것이 목적이다.

2. 포항시 미세먼지 현황

포항지역은 강우량이 전국 최저 수준인 반면 일조량이 많은 건조한 기후특성을 가지며, 철강산업공단과 시가지가 형산강을 경계로 인접하고 있는 지정학적 배치구조 때문에 원활한 대기유동에 의한 환기구조를 가졌음에도 불구하고 미세먼지 농도가 전국적으로 높은 수준을 보이고 있다. 특히 인공위성 영상분석에 의한 토지이용도 분석결과에 의하면 포항 시가지와 공단지역은 도시건조화가 심각한 상태이며, 철강공단의 특성상 연 · 원료의 대규모 야적에 따른 비산먼지의 발생도 빈번하기 때문에 먼지의 측면에서 여러 가지 불리한 여건이 많다 (김현구, 2005).

포항지역의 시가지 및 공단지역에는 표 1 및 그림 2와 같이 네 지점에 환경측정소가 위치하고 있다. 지난 4년간의 (2001 ~ 2004년) 시간별 측정자료를 이용

Table 1. Specification of environmental monitoring stations in Pohang area.

Area	Monitoring station	Measurement date	Measurement item			Note
			Air quality	Acid rain	Heavy metal	
Manufacturing area	Jangheong (ex-Dongil Steel rooftop)	1990.6	○	○	○	KME112
Commercial area	Jookdo (Jookdo 2dong office rooftop)	1990.7	○			KME113
Residential area	Daedo (Sangdae 2dong office rooftop)	1997.4	○		○	KME114
Residential area	Daesong (Daesong myun office rooftop)	2005.2	○			-

※ Air quality: PM10, SO_2, NO_2, CO, O_3 Heavy metal: Pb, Cd, Cr, Cu, Mn, Fe, Ni

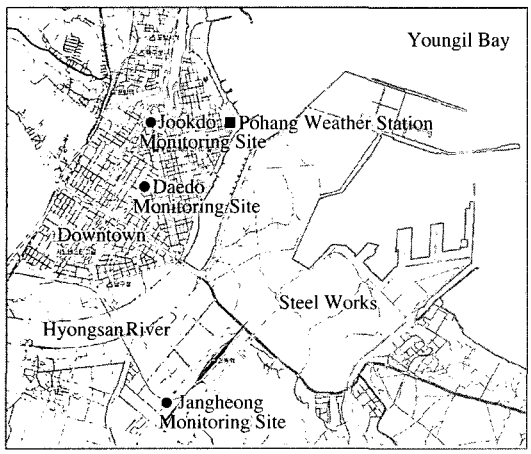


Fig. 2. Location of environmental monitoring stations (black circles) in Pohang area.

하여 측정소 간의 미세먼지 농도에 대한 상관분석을 실시한 결과(대송면 측정소는 2005년 신설된 관계로 제외하였음), 장흥동과 대도동의 미세먼지 농도 사이에 상관계수는 0.26이었으며, 장흥동과 죽도동 간에는 0.33, 대도동과 죽도동 간에는 0.63으로 나타났다. 여기서 상관분석은 미세먼지 농도를 인자로 하였으며, 상관계수는 피어슨(Pearson) 상관계수를 말한다. 그림 2에서 확인되듯이 시내에 위치하고 있는 죽도동과 대도동 환경측정소는 시가지 내에 인접하여 있어 상관관계가 크게 나타난 반면, 장흥동 측정소의 미세먼지 농도 거동은 이들 시내지역과는 독립된 양상으로 나타나고 있다. 이러한 이유는 그림 3에 제시된 대기확산모델링에 의한 미세먼지 농도의 연평균 지면농도 분포도(포항산업과학연구원, 2002) 및 연평균 바람장미로부터 물리적인 해답을 찾을 수 있다. 즉, 대기배출목록(정은희 등, 2005)과 수치기상예측 결과인 기상장(이화운 등, 2004)을 입력자료로 하여 2001년부터 2004년까지 4년간의 대기확산모델링을 수행한 결과에 의하면, 철강공단에서 배출된 먼지는 주풍향인 남서풍에 의해 영일만을 통하여 바다로 원활하게 환기됨으로써 확산영향권이 철강공단 자체 부지 내로 한정되며, 두 지역의 경계인 형산강을 넘어서 공단지역의 먼지가 시내로 확산되는 빈도는 낮기 때문에 두 지역은 서로 별도의 대기질을 형성하고 있는 것이다.

3. 미세먼지 예보모형의 개발

우리나라는 해마다 3~4월이면 황사가 불어와 이때의 미세먼지 농도는 평상시의 수십 배를 초과한다. 따라서 황사일에 측정된 관측치들은 특이값으로 볼 수 있어, 좀 더 정확한 모형구축을 위해 각 연도별로 황사일에 측정된 관측치들은 제외시키고 모형을 구축하였다. 본 연구 자료기간인 2001년부터 2004년 중에서 실제로 황사일로 모형 구축에서 제외된 관측치는 전체 38일이며, 각 연도별로는 2001년에 23일, 2002년에 9일, 2003년에 2일, 2004년에 4일이다.

본 연구에서는 2001년 1월 1일부터 2004년 12월 31일까지 4년간 포항지역 환경 측정소 세 곳, 즉, KME112(장흥동-공단지역), KME113(죽도동-상업지역), KME114(대도동-주거지역)의 시간별 대기오염 측정자료(SO_2, NO_2, CO, O_3, PM10)와 포항기상대의 풍속, 온도, 습도 등 시간별 지면기상자료를 이용하였다(김운수, 2004). 다음날 미세먼지 예측모형을 위한 독립변수 중 대기오염항목은 SO_2, NO_2, CO, O_3의 당일 00:00부터 23:00까지의 관측값 중 최대값을, 당일 00:00부터 11:00까지의 PM10 중 최대값(PM10MA1) 또는 평균값(PM10AV1)과 당일 12:00부터 23:00까지의 PM10 중 최대값(PM10MA2) 또

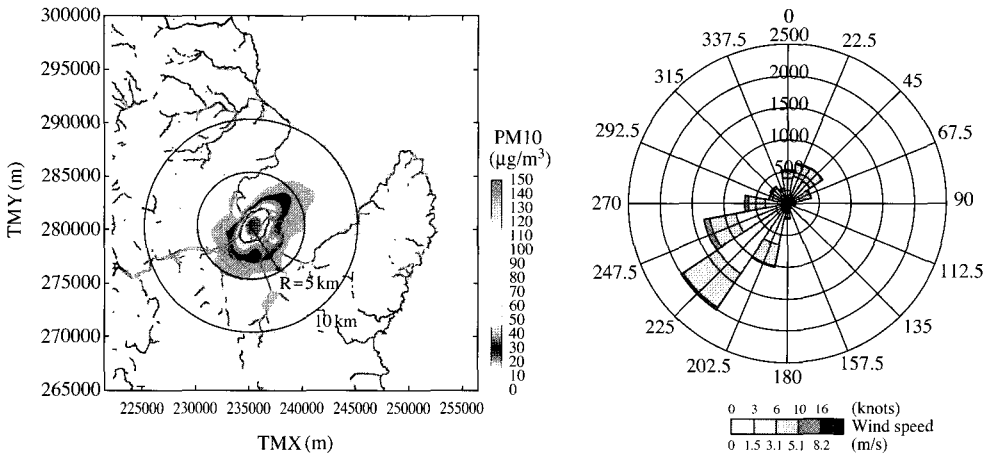


Fig. 3. Ground-level concentration map of PM10 in Pohang region and annual wind rose.

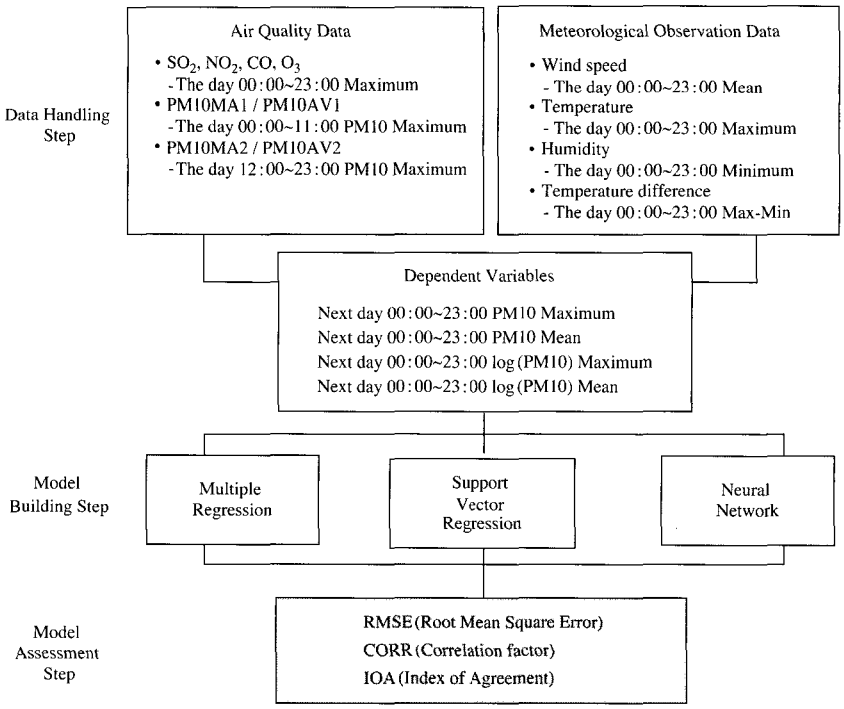


Fig. 4. Schematic diagram for the statistical prediction models of PM10.

는 평균값(PM10AV2)으로 나누어 입력자료로 사용하였다. 지면기상항목은 당일 00:00부터 23:00까지의 평균풍속, 최대온도, 최대습도, 최대온도와 최소온도의 차이를 사용하였다. 또한 종속변수는 다음날 00:00부터 23:00까지의 PM10 관측값 중 최대값(또는 평균값)을 사용하였다. 예측모형에 대한 개괄적 개요

를 도식화하면 그림 4와 같다.

통계예측모형은 전통적 방법인 선형회귀분석방법(Regression)과 인공지능기법을 사용한 비선형 방법이면서 정확도가 뛰어난 신경망분석(Neural Networks, NN)방법, 그리고 신경망분석의 과도 적합과 신경망 구조의 설계에 많은 시간과 노력이 필요한 단점을 해결하는 방법으로써 많이 쓰이는 SVR모형을 적용하였다. 본 연구에 적용된 통계모형에 대해 간략히 소개하면 다음과 같다.

3. 1 선형회귀분석 (Linear Regression, LR)

자연이나 사회현상의 규명에 있어서, 관련된 변수들간의 상호 관련성을 수학적인 함수의 형태로서 찾으려고 시도할 경우가 많다. 회귀분석은 이러한 변수들간의 함수적 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계학적 분석방법을 말한다. 일반적으로 이렇게 추정된 모형을 사용하여 필요한 예측을 하거나 관심 있는 통계학적 추정과 검정을 하게 된다(Draper and Smith, 1998). 즉, 선후관계가 명백한 독립변수(또는 입력변수)와 종속변수(또는 출력변수) 간의 의존도를 평가하는 방법이다. 일반적으로 연속성 변수의 종속변수(Y)와 독립변수들(X)과의 관계를 수학적 공식으로 함수화(예 : $Y=a \cdot X+b$)한 회귀식을 추정하는 것이 목적이다. 선형회귀분석이란 두 변수의 관계가 직선적이라는 가정하에 회귀분석을 시도하는 방법으로 대부분의 회귀분석이 선형을 가정하기 때문에 일반적으로 선형(linear)이라는 단어를 앞에 생략하고 사용한다. 회귀분석은 분석을 시작하기 전에 분석자료가 직선적 관계에 있는지 확인해야 하는데 산점도나 오차분석에서 두 변수의 관계가 비선형이거나, 각 변수가 아주 극심한 비정규분포일 경우에는 독립변수나 종속변수를 적절히 변환(transformation)시킨 후 회귀분석을 시행해야 한다. 일반적으로 회귀분석에서는 거리의 제곱의 합이 가장 작은 직선식을 구하게 되는데, 이를 최소제곱법(least square method)에 의한 회귀식이라고 한다.

3. 2 신경망분석 (Neural Network, NN)

신경망분석은 회귀분석과는 달리 어떠한 통계적인 분포도 가정하고 있지 않다. 인간의 뇌 기능에 착안하여 개발된 이 분석법은 패턴인식의 한 분야로 과거의 경험이나 지식을 습득함으로써 오류를 최소화하는 과정들을 포함하고 있다. 신경망분석은 각 층(layer)에 있는 마디(node)로 이루어져 있으며 신경망의 형성 과정에 따라 그 종류가 다양하고 복잡하다. 신경망에는 여러 가지 다양한 모형이 있으나, 그 중에서도 자료분석을 위해 가장 널리 사용되는 모형은 MLP(Multi-Layer Perceptron, 다층인식자) 신경망이다. MLP의 구조는 다층인식자 신경망으로 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있다. 입력층은 각 입력변수에 대응되는 마디들로 구성되어 있으며 명목형(nominal) 변수는 가변수를 사용한다. 은닉층은 여러 개의 은닉마디로 구성되어 있으며 각 은닉마디는 입력층으로부터 전달되는 변수값들의 선형결합(linear combination)을 비선형함수로 처리하여 출력층 또는 다른 은닉층으로 전달한다. MLP신경망은 이 비선형함수를 시그모이드(sigmoid) 형태의 함수로 사용하는 것을 말한다. 출력층은 출력변수에 대응하는 마디들을 갖는다. 이를 도식화하면 그림 5와 같다. 신경망분석의 장점은 자료들간의 비선형적인 관계를 찾아 낼 수 있다는 것이다. 그러나 자료를 과대적합하는 경향이 있기 때문에 새로운 자료가 주어졌을 때 적당하지 않을 수 있다는 단점이 있다. 또 다른 단점은 회귀 분석기법과는 달리 모형의 결과를 해석하기가 어렵다는 것이다(Ripley, 1996).

3. 3 Support Vector Regression (SVR)

SVR은 1998년 Vapnik에 의해 개발된 학습기법으로, 입력공간의 비선형문제를 고차원 특징공간의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분

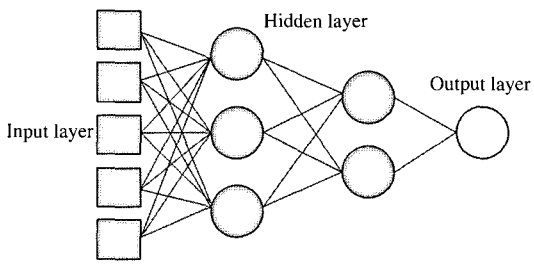


Fig. 5. Structure of MLP (Multi-Layer Perceptron) neural network system.

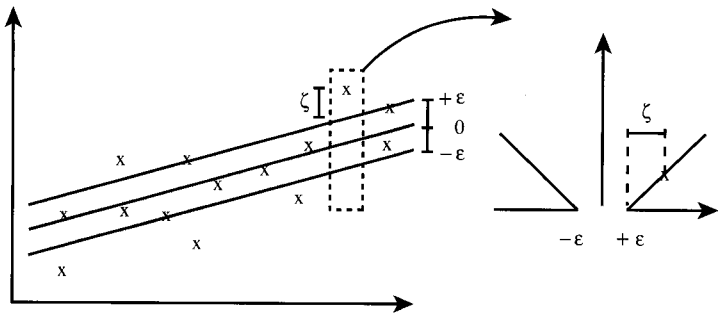


Fig. 6. ε-Insensitive loss function.

석하는 것이 수월하며 조정해야 할 모수(parameter)의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다(Vapnik, 1998). 그리고 구조적 위험을 최소화함으로써 과대적합(overfitting) 문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 국부해를 구할 수 있다는 점에서 신경망기법보다 성능이 좋은 학습기법으로 주목 받고 있다(Schölkopf and Smola, 2001).

예를 들면, 모형구축을 위한 자료 $\{(x_1, y_1), \cdots(x_N, y_N)\} \subset \aleph \times \Re$가 주어져 있다고 가정한다. 여기서 N은 모형구축 자료의 개수, x는 입력변수벡터, y는 출력변수, $\aleph$는 m차원의 실수공간인 입력 공간 $\Re^m$을 나타낸다. Vapnik이 제안한 ε-SVR은 모든 모형구축 자료에 대해서 실제 출력 변수값 y_i들로 부터 최고 ε보다 작은 모형구축 자료의 에러는 무시하게 된다. 이를 도식화하면 그림 6과 같고 다음과 같은 볼록최적화 문제(convex optimization problem)를 구성할 수 있다.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\zeta_i + \zeta_i^*)$$
$$\text{subject to} \quad \begin{Bmatrix} y_i - < w, x_i > - b \le \varepsilon + \zeta_i \\ < w, x_i > + b - y_i \le \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \ge 0 \end{Bmatrix} \tag{1}$$

여기서 ζ_i와 ζ_i^*는 학습데이터에 대한 어느 정도의 오차를 허용하는 여유변수(slack variable)이며, 상수 C는 추정오차의 패널티(penalty)로 모형의 복잡도를 결정하는 모수이다.

위의 식(1)을 만족하는 선형 함수 $f(x)$를 식(2)와 같이 쓸 수 있다.

$$f(x) = < w, x > + b \text{ with } w \in \aleph, b \in \Re \tag{2}$$

여기서 $< , >$은 입력 공간에서의 내적을 나타내며, w는 가중치, b는 상수항을 나타낸다.

4. 미세먼지 예보모형의 평가

포항지역 내 각 환경측정소에서 측정된 자료로부터 입력자료의 특성을 잘 반영하기 위해 단순임의추출을 하였으며 자료의 70%를 모형 구축을 위한 훈련용 자료(training data)로 사용하였고 나머지 30%를 구축된 모형의 평가를 위한 검증용 자료(test data)로 사용하였다. 또한 세 기법의 결과로 나온 예측값에 대한 정량평가를 위하여 검증용 자료를 이용하여 RMSE(Root Mean Square Error), IOA(Index of Agreement), CORR(Correlation factor) 등의 지수를 구하였으며, 이들 지수에 의한 단순 정확도 평가로 모형을 상호비교하였다.

각 정량분석 항목의 계산식은 식(3)과 같으며 P_i는 예측값, O_i는 관측값을 의미한다.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$$
$$IOA = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - O_{mean}| + |O_i - O_{mean}|)^2} \tag{3}$$

$$CORR=\frac{N\left(\sum_{i=1}^{N}O_iP_i\right)-\left(\sum_{i=1}^{N}O_i\right)\left(\sum_{i=1}^{N}P_i\right)}{\sqrt{\left[N\left(\sum_{i=1}^{N}O_i^2\right)-\left(\sum_{i=1}^{N}O_i\right)^2\right]\left[N\left(\sum_{i=1}^{N}P_i^2\right)-\left(\sum_{i=1}^{N}P_i\right)^2\right]}}$$

참고로 RMSE 값은 0에 가까울수록, IOA와 CORR은 1에 가까울수록 모형이 좋은 예측력을 가진다는 것을 의미한다.

4.1 최대농도 종속변수 모형

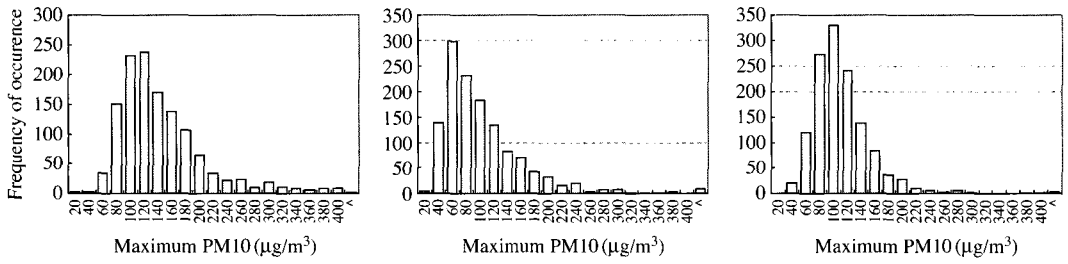
포항지역 내 각 환경측정소별로 세 개의 통계적 예측모형을 구축하였다. 주어진 자료를 모형구축용 자료(training data)와 검증용 자료(test data)로 나눈 후에 모형구축용 자료를 이용하여 구축된 모형을 검증용 자료에 적용하여 다음날의 PM10 최대농도 예측값을 구하였다. 이 예측값과 검증용 자료의 PM10 최대농도 실제 관측값을 비교하여 RMSE, CORR, IOA의 측정지수값을 구하여 그 결과를 표 2에 종합하였다. 표 2에서 보는 바와 같이 KME112, KME113 측정소에서는 회귀모형(LR)의 예측력이 가장 뛰어났으며, KME114에서는 신경망모형(NN)의 예측력이 가장 뛰어났음을 알 수 있었다. 그런데 예상했던 것과는 달리 SVR이 근소한 차이로 예측력이 떨어지는 것으로 나타났다. 또한 전반적으로 모형기법에 상관없이 예측결과가 우수한 편이 아니므로 이러한 원인을 밝혀내기 위하여 관측된 PM10의 일별 최대농도의 분포를 살펴보았다.

각 측정소별 PM10 최대농도의 농도값에 따른 분포를 살펴보면 그림 7에 나타난 바와 같이 KME112의 경우에는 120 μg/m^{3}이 최빈값이며, KME113는 60 μg/m^3, 그리고 KME114는 100 μg/m^{3}이 최빈값이며, 전반적으로 와이블(Weibull) 분포에 가까운 형태를 보이는 것으로 분석되었다. 이에 확률적 분포형상에 착안하여 PM10 최대농도가 정규분포에 근접하도록 로그변환(log-transformation) 시켜보았다. 그 결과, 그림 8에서와 같이 정규분포에 많이 근접하였기에 본 연구에서는 종속변수값으로 로그변환된 PM10 최대농도값을 사용하였다. 로그변환을 통한 자료의 정규화는 분포의 치우침을 완화시킴으로서 회귀모형의 경우 정규성 가정을 만족시켜 예측력을 높일 수 있으며, 신경망분석이나 SVR과 같이 분포를 가정하지 않는 모형의 경우도 치우친 특정 자료에 대해 모형의 적합이 집중되는 것을 완화시켜주므로 모형의 예측력을 높이는 효과를 기대할 수 있다.

표 3은 로그변환 시킨 PM10 최대농도, 즉, log(Max(PM10))를 종속변수로 한 예측모형을 구축한 후 구축된 모형을 검증용 자료에 적용시킨 결과로, 각각의 측정소별 RMSE, CORR, IOA의 측정지수값

Table 2. Quantitative assessment of the maximum of PM10 for each models.

Model	Measure	Monitoring sites		
		KME112	KME113	KME114
LR	RMSE	72.6846	63.2651	64.7749
	CORR	0.5053	0.6538	0.3862
	IOA	0.6074	0.7289	0.4690
NN	RMSE	72.8694	70.2838	63.5025
	CORR	0.5019	0.5347	0.4261
	IOA	0.6132	0.6241	0.4886
SVR	RMSE	73.5792	72.8580	65.9821
	CORR	0.4866	0.5165	0.3645
	IOA	0.4834	0.6092	0.3997



(a) KME112 (b) KME113 (c) KME114

Fig. 7. Distribution of the maximum of PM10 for each environmental monitoring stations.

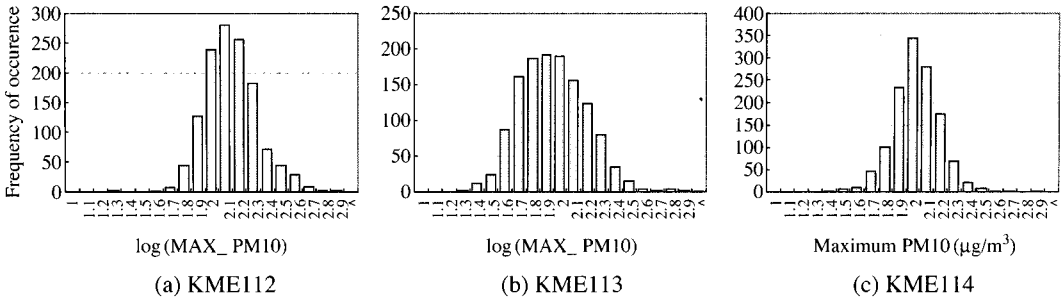


Fig. 8. Distribution of the maximum of PM10 after log transformation.

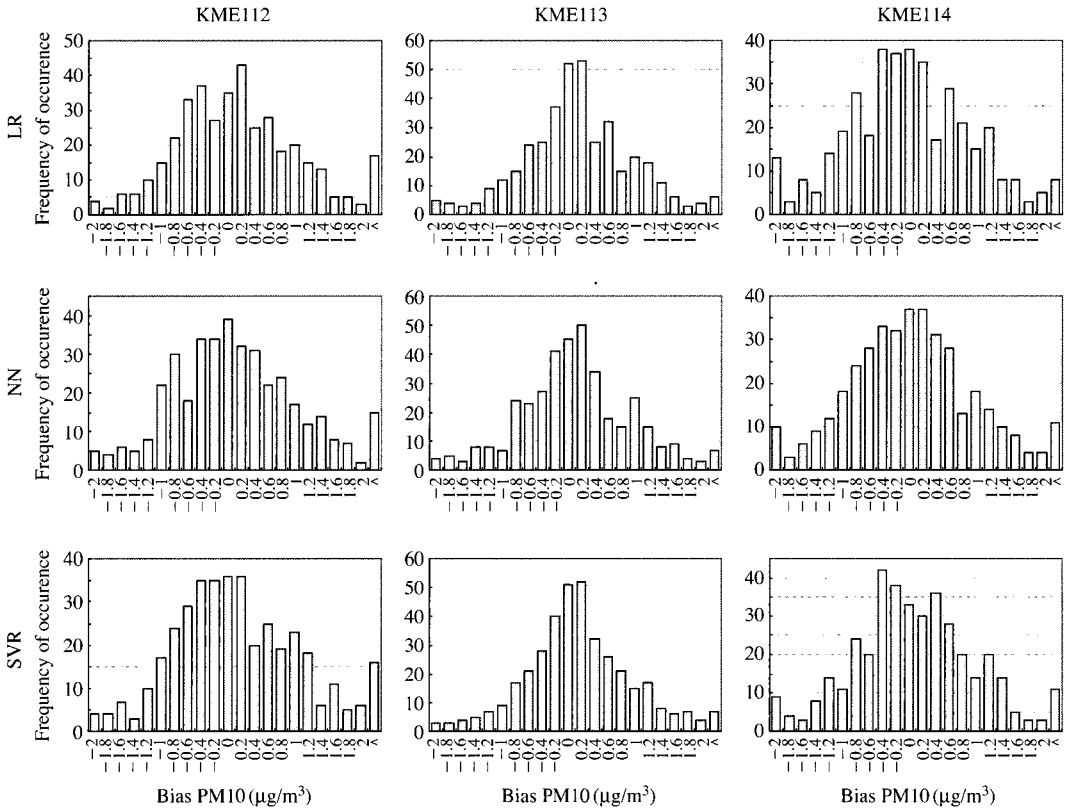


Fig. 9. Distribution of the bias for the maximum of PM10 for each models after log-transformation.

을 정리하였다. 표 3을 살펴보면 KME112는 RMSE, CORR, IOA의 세가지 기준 모두에서 신경망기법을 이용한 모형이 가장 뛰어난 예측력 및 적합도를 가지는 것으로 나타난 반면, KME113과 KME114에서는 SVR모형의 예측력 및 적합도가 가장 뛰어난 것으로 나타났다.

또한 대부분의 평가측정값이 로그변환 이전인 표 2의 결과보다 대폭 향상되었기에 로그변환한 PM10 종속변수를 사용한 모형이 더 적합하다는 결론을 얻을 수 있다.

모형의 예측경향성을 파악하기 위해서 실제 관측값과 모형에 의한 예측값의 차이인 편의(bias=O_i-P_i) 분포를 살펴보았다. 그림 9는 로그변환 시킨 PM10 최대농도와 각각의 모형에 의한 예측값의 차이에 대한 히스토그램을 작성한 것이다. SVR모형 적합 결과를 살펴보면 KME114를 제외하고는 편의분포가 0에서 매우 높은 빈도를 나타내는 것을 알 수 있다.

표 4는 그림 9에 나타낸 편의의 각 분포들의 평균과 표준편차를 나타낸 것으로, 모든 분포들이 평균 0과 표준편차 1에 가까운 것을 알 수 있다. 편의에 대한 평균과 표준편차값을 보면 모형간에 차이가 작은 것으로 나타났으나, 이는 실제로 로그 변환에 따른 단위의 변화로 인해 자료의 값들이 전체적으로 작아졌기 때문이다. SVR에 의한 편의의 표준편차가 KME113과 KME114에서 모두 가장 작은 것으로 나타났다. 따라서 모형의 적합결과 대체로 과소나 과대추정의 경향 없이 타당하다는 것을 확인할 수 있다.

Table 3. Quantitative assessment of the maximum of PM10 for each models after log-transformation.

Model	Measure	Monitoring sites		
		KME112	KME113	KME114
LR	RMSE	0.364220	0.508314	0.329076
	CORR	0.576301	0.480560	0.501586
	IOA	0.714179	0.666054	0.653574
NN	RMSE	0.360829	0.458125	0.314520
	CORR	0.581710	0.568367	0.553694
	IOA	0.706885	0.711421	0.696096
SVR	RMSE	0.361543	0.457236	0.305469
	CORR	0.580891	0.574518	0.583573
	IOA	0.696129	0.721888	0.711607

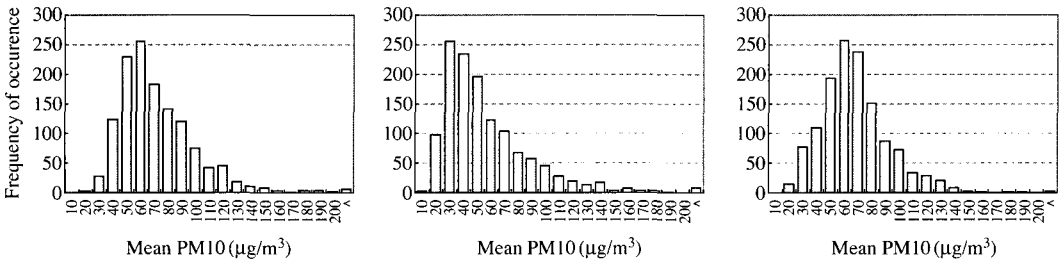
4.2 평균농도 종속변수 모형

먼지예보에 있어서 다음날의 최대농도 예보와 함께 평균농도 역시 중요한 예보항목이기 때문에 다음날의 PM10 최대농도를 종속변수로 하는 대신 다음날의 PM10 평균농도를 종속변수로 사용하는 모형을 구축하였다. 또한 독립변수 중 당일 00:00부터 11:00까지의 PM10 평균값(PM10AV1)과 당일 12:00부터 23:00까지의 PM10 평균값(PM10AV2)으로 나누어 입력자료로 사용하였다. 그 외의 독립변수들과 모형평가방법 및 절차는 이전의 최대농도 종속변수 모형과 동일하게 하였다. 표 5는 PM10 평균농도에 대한 예측모형 구축 후, 구축된 모형을 검증용 자료에 적용시킨 결과로, KME112에서는 회귀모형의 예측력이 가장 뛰어났으며, KME113과 KME114에서는 SVR의 예측력이 가장 뛰어난 것으로 나타났다.

한편 각 측정소별 PM10 평균농도의 농도값에 따른 분포를 살펴보면 그림 10에 나타난 바와 같이 KME112의 경우 평균농도 60이 최빈값이며, KME113의 경우는 평균농도 30이 최빈값, KME114의 경우 평균농도 60 μg/m^{3}이 최빈값으로 Weibull

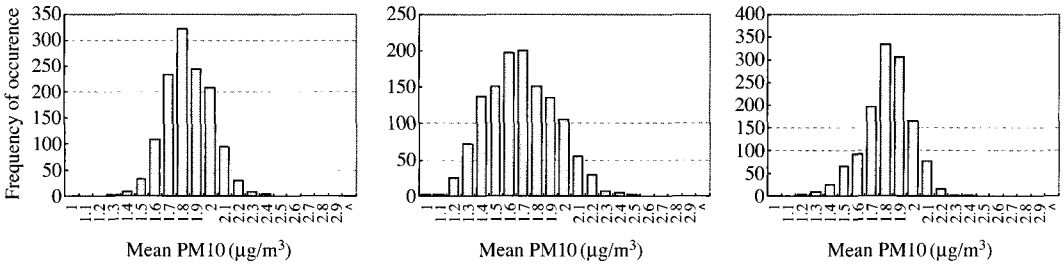
Table 4. Mean and standard deviation of the bias for the maximum of PM10 for each models after log-transformation.

Models		KME112	KME113	KME114
LR	Mean	0.050049	−0.00677	−0.11449
	STD	0.973286	1.039564	1.055966
NN	Mean	0.033535	−0.01055	−0.07657
	STD	0.988756	0.986741	1.009484
SVR	Mean	0.062201	0.043209	−0.02782
	STD	0.999025	0.971679	0.984417



(a) KME112 (b) KME113 (c) KME114

Fig. 10. Distribution of the mean of PM10 for each environmental monitoring stations.



(a) KME112 (b) KME113 (c) KME114

Fig. 11. Distribution of the mean of PM10 after log transformation for each environmental monitoring stations.

Table 5. Quantitative assessment of the mean of PM10 for each models.

Model	Measure	Monitoring sites		
		KME112	KME113	KME114
LR	RMSE	22.17775	27.01100	20.10637
	CORR	0.64137	0.59956	0.65746
	IOA	0.75401	0.75480	0.76454
NN	RMSE	22.32317	25.12076	19.85227
	CORR	0.63538	0.64112	0.67282
	IOA	0.74843	0.75659	0.75784
SVR	RMSE	22.22731	24.73738	19.57617
	CORR	0.63971	0.64831	0.68365
	IOA	0.74950	0.75424	0.77592

Table 6. Quantitative assessment of the mean of PM10 for each models after log-transformation.

Model	Measure	Monitoring sites		
		KME112	KME113	KME114
LR	RMSE	0.26876	0.41716	0.28688
	CORR	0.68890	0.66266	0.69348
	IOA	0.80990	0.77317	0.80560
NN	RMSE	0.26846	0.40816	0.27477
	CORR	0.68971	0.68092	0.72859
	IOA	0.81151	0.78959	0.80948
SVR	RMSE	0.26883	0.41726	0.275244
	CORR	0.68697	0.66652	0.72235
	IOA	0.80239	0.79574	0.82623

분포에 가까운 형태를 보이는 것으로 나타났다. 따라서 PM10 평균농도가 정규분포에 근접하도록 로그변환을 시켰으며, 그 결과로 그림 11에서와 같이 정규분포에 많이 근접한 분포를 얻었기 때문에 본 연구에서는 종속변수값으로 로 그 변환된 PM10 평균농도값을 사용하였다.

표 6은 로그변환 시킨 PM10 평균농도에 대한 예측모형 구축 후, 구축된 모형을 검증용 자료에 적용시킨 결과로 각각의 측정소별 RMSE, CORR, IOA의 측정지수값을 정리하여 나타내었다. KME112의 경우 RMSE, CORR, IOA의 세가지 기준 모두에서 NN 기법을 이용한 모형이 가장 뛰어난 예측력을 가지는 것으로 나타났다. KME113과 KME114의 경우 평가기준에 따라 최적의 모형에 약간씩 차이가 있는 것으로 나타났으나 RMSE와 CORR을 기준으로 한 경우 NN 모형의 예측력이 가장 뛰어난 것으로 나타났다. 반면 IOA를 기준으로 한 경우는 SVR 모형의 예측력이 뛰어난 것으로 나타났다. 전체적으로 보면 NN 모형의 예측력이 가장 뛰어난 것으로 나타났다.

그림 12는 앞에서와 마찬가지로 모형의 예측경향성을 알아보기 위해 실제 관측값을 로그변환 시킨 PM10 평균농도와 각각의 모형에 의한 예측값의 차이인 편의를 범위별로 히스토그램을 작성한 것이다. 회귀모형의 경우 KME113에 대해서는 다소 과대예측하는 경향이 있는 반면 KME114에 대해서는 과대예측하는 경향이 나타났다. 신경망모형은 KME112에 대해서는 다소 음수쪽으로 치우쳐 과대예측하는 경향이 있는 반면 KME113과 KME114에 대해서는 양수쪽으로 치우쳐 다소 과소예측하는 경향이 있는 것으로 나타났다. SVR모형의 경우는 모든 측정소 자료에 대해서는 다소 과소예측하는 경향이 있는 것으로 나타났다. 그러나 전반적으로 정규분포형태를 이루고 있으며 치우침이나 퍼짐의 정도도 그렇게 크지 않음을 알 수 있다. 보다 자세히 알아보기 위하여 분포를 검정하여 보았다.

표 7은 그림 12에 나타나있는 편의 분포들의 평균

과 표준편차를 나타낸 것이다. 모든 분포들이 평균이 0에 가까우며 표준편차가 1보다 작은 값을 가지는 것을 알 수 있다. 편의의 평균은 모든 측정소 자료에서 회귀모형에 의한 결과가 가장 0에 가까운 것으로 나타났으며, 편의의 표준편차는 KME112에 대해서는 신경망모형에 의한 결과가, KME113과 KME114에 대해서는 SVR에 의한 결과가 가장 작은 것으로 나타났다. 이는 앞의 표 6의 결과에서 NN모형이 가장 예측력이 뛰어난 것으로 나타난 것과 달리 SVR모형이 로그변환한 자료에 대해 가장 고른 예측력을 나타낸다고 할 수 있다.

Table 7. Mean and standard deviation of the bias for the mean of PM10 for each models after log-transformation.

Model		KME112	KME113	KME114
LR	Mean	0.02205	−0.02543	−0.00856
	STD	0.72886	0.74908	0.72093
NN	Mean	0.05454	−0.05588	−0.01351
	STD	0.72733	0.74760	0.69161
SVR	Mean	0.03205	−0.03514	−0.04696
	STD	0.72767	0.73251	0.68895

구윤서 등(2005)의 전일 먼지예보에 대한 연구에서는 기상청의 예보 기상자료(강수량 최대값, 역정층 유무, 지표풍속, 환기지수, 혼합고, 대기오염지수 등)를 도입함으로써 예측성능을 향상시킨 것으로 보고하고 있다. 구윤서 등(2005)의 예보결과에 대한 정량평가 결과를 보면 회귀모형 예측결과의 상관도가 0.23~0.68, IOA가 0.51~0.78이며 신경망회로 예측결과의 상관도가 0.50~0.62, IOA가 0.69~0.77로 본 연구결과에 비하여 상당히 낮은 수준으로 나타나고 있다. 이는 구윤서 등(2005)의 모형이 다수개의 환경측정소를 평균하여 지역에 대하여 예보모형을 적용하였기 때문일 것으로 예상되지만 또한 특이값을 제외하고 로그변환을 통한 확률적인 정규분포 화

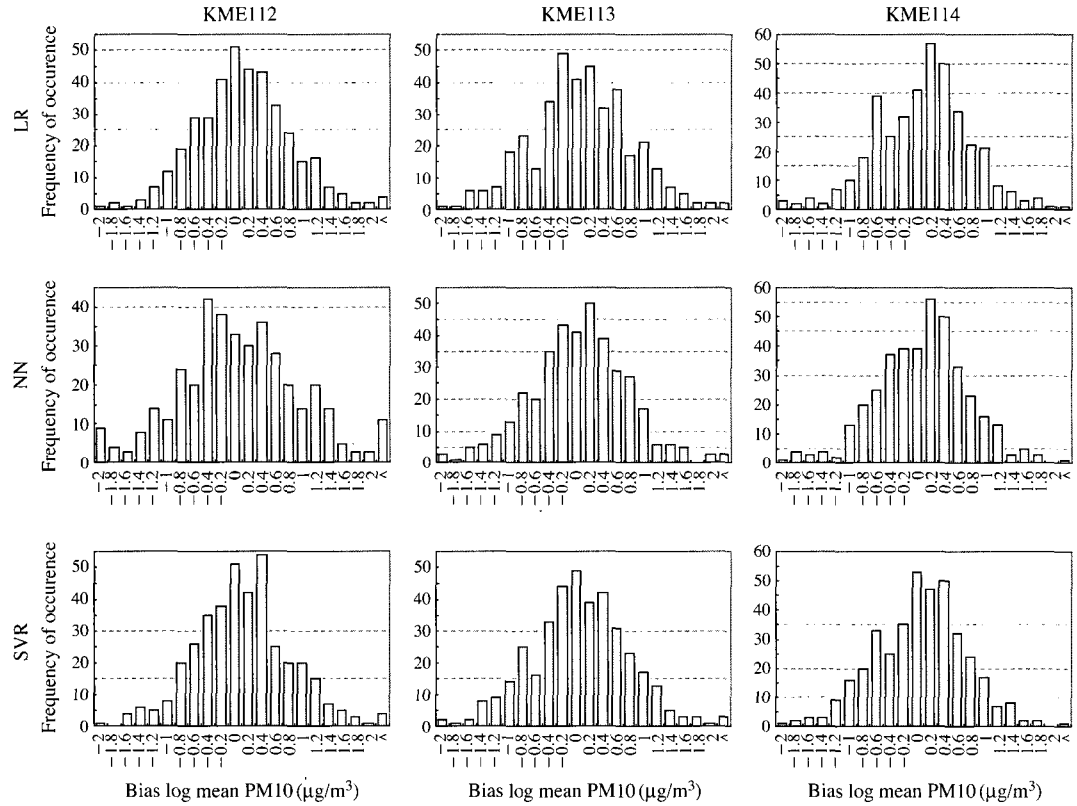


Fig. 12. Distribution of the bias for the mean of PM10 for each models after log-transformation.

작업을 하지 않은 영향도 포함되었기 때문으로 판단된다. 실제로 본 연구에서 로그변환을 적용한 경우, 최대농도 예측에서는 IOA의 상승률이 평균적으로 회귀모형은 16.2%, NN이 23.8%, SVR이 46.7%로 특히 SVR의 정확도가 대폭 상승할 수 있었다. 그러나 평균농도의 경우에는 로그변환을 적용함에 따른 IOA의 상승률이 평균적으로 회귀모형은 5.1%, NN이 6.5%, SVR이 6.3%로 최대농도의 정확도 상승효과에 비하여 전반적으로 효과가 크지 않은 것으로 나타났다.

5. 결론 및 토의

본 연구에서는 새로운 통계모형인 SVR을 적용하여 그 예측성능을 회귀모형 및 신경망모형과 비교해 본 결과 로그변환한 미세먼지 최대농도에 대한 결과에서 SVR모형이 예측력 및 적합성이 가장 뛰어났으며, 로그변환한 미세먼지 평균농도에 대한 결과에서는 NN이 뛰어난 예측력을 보인 반면 SVR모형은 자료에 대해 가장 고른 예측력을 보이는 것을 확인하였다. 특히 미세먼지 최대농도 및 평균농도가 치우친 분포를 띄고 있어, 이를 로그변환 시킨 후 모형을 적합함으로써 예측정확도를 대폭 향상하였다. 특히 SVR의 경우 자료의 로그변환을 통해 모형의 예측력 및 적합도가 매우 향상된 것으로 나타났는데, SVR 모형이 로그변환전 치우친 분포를 나타내는 원자료에 대해서는 고른 예측력을 가지지 못하다가 로그변환 후 정규분포에 가까워지면서 자료 전반에 걸쳐 고른 예측력을 가지게 된 것으로 풀이된다. 이러한 향상은 PM10의 최대농도와 평균농도 예측에서 다른 모형들보다 예측결과로 나타난 편의의 표준편차가 가장 작았던 결과로도 확인할 수 있었다.

본 연구를 통하여 구축된 먼지예보 모형은 전일의 대기오염도 측정자료와 기상관측자료를 이용하여 당일의 미세먼지 최대농도 및 평균농도를 예측할 수 있으므로 조간신문 또는 일기예보 아침방송 시간에 기상정보와 같이 제공될 수 있을 것이다. 1차적으로는 대기환경측정소 세곳(표 1)에 대한 예보를 실시하고, 차후 2차적으로 대기확산모델링에 의해 대기환경측정소와 권역별 미세먼지 농도의 상관성을 정량적으로 파악한 후 권역별 예보로 발전시켜나가고자 한다. 한편 최근에는 한국천식알레르기협회(http://www.kaaf.org/)에서 꽃가루, 황사, 오존, 기온/습도 등의 항목에 가중치를 부여하는 방식으로 천식지수를 예보하고 있는데, 여기에 중요한 인자인 먼지의 영향도가 포함될 수 있을 것으로 기대된다. 또한 본 연구에서 사용한 통계적 예측모형 뿐만 아니라 시계열 분석기법 또는 고농도 사례에 가장 큰 영향을 미치는 인자를 분류해 내기 위한 분류회귀나무(CART) 기법 등을 사용하여 PM10 예측모형을 구축하고 그 정확도를 본 연구에서 사용한 통계적 기법들과 비교하여 보는 것을 향후과제로 남겨두고자 한다.

감사의 글

본 연구는 '통신해양기상위성 기상자료처리시스템 개발' 사업의 지원으로 수행되었습니다.

참 고 문 헌

구윤서, 권희용, 윤희영(2003) 통계모델을 이용한 실시간 오염도 예보 시스템 개발(PM-10), 한국대기환경학회 추계학술대회논문집, 445-446.

구윤서, 윤원정, 권희용, 양재문, 최종혁, 윤희영(2005) 전일 미세먼지(PM10) 예보시스템 개발, 한국대기환경학회 춘계학술대회논문집, 403-404.

김운수(2004) 서울시 미세먼지 배출량 조사·분석 및 관리방안 연구, 시정연 2004-R-22, 서울시정개발연구원.

김현구(2005) 기상조건별 비산먼지 관리체계 최적화 연구, 한국대기환경학회지, 21(5), 573-583.

김현구, 이영섭, 구자문, 고유나(2006) 사회통계조사에 의한 대기환경 체감지수 개발에 관한 연구, 한국대기환경학회 춘계학술대회논문집, 142-143.

이화운, 정우식, 김현구, 이순환(2004) 대기오염 확산해석을 위한 포항지역 기상장 연구-바람장 수치해석, 한국대기환경학회지, 20(1), 1-15.

정은희, 김현구, 전희동(2005) 포항지역 대기오염물질 배출량 산정(2003년도), 한국대기환경학회 춘계학술대회논문집, 356-357.

포항산업과학연구원(2002) 포항시 환경보전 종합계획 2002-2011, 포항시청.

환경부(2005a) 연도별 대기오염도 변화추이(측정소별), 환경부 대기환경연보.

환경부 (2005b) 미세먼지 농도별 행동요령, 환경부 2005. 1. 26 보도자료, http://dust.seoul.go.kr.

Draper, N. and H. Smith (1998) Applied Regression Analysis (3rd ed), John Wiley & Sons, New York.

Ripley, B.D. (1996) Pattern Recognition and Neural Networks, Cambridge University Press.

Schölkopf, B. and A.J. Smola (2001) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press.

Vapnik, V. (1998) Statistical Learning Theory, John Wiley & Sons, New York.