مقایسه کاربرد روش‌های شبکه عصبی مصنوعی و رگرسیون خطی چندمتغیره براساس تحلیل مؤلفه‌های اصلی برای پیش‌بینی غلظت میانگین روزانه کربن‌مونوکسید: بررسی موردی شهر تهران

روح‌اله نوری[۱*]، خسرو اشرفی[۲] و ابوالفضل اژدرپور[۳]

[۱] دانشجوی دکترای مهندسی محیط‌زیست، دانشگاه تهران، ایران
[۲] استادیار گروه مهندسی محیط‌زیست، دانشکده محیط زیست، دانشگاه تهران، ایران
[۳] دانشجوی کارشناسی ارشد مهندسی بهداشت محیط، دانشگاه تربیت مدرس، تهران، ایران

(دریافت: ۸۵/۱۲/۲۰ ، پذیرش نهایی: ۸۶/۱۰/۲۵)

چکیده

هدف از این مقاله، پیش‌بینی میانگین غلظت روزانه کربن‌مونوکسید در هوای شهر تهران با استفاده از دو مدل شبکه عصبی مصنوعی و رگرسیون خطی چندمتغیره برحسب تحلیل مؤلفه اصلی (PCA) است. از روش PCA برای از بین بردن هم‌راستایی چندگانه (multicolinearity) بین متغیرهای ورودی و تفسیر بهتر نتایج مدل رگرسیونی استفاده شده است. همچنین با استفاده از شبکه عصبی Feed-Forward با یک لایه پنهان نیز مدل مناسب برای این امر ایجاد شده است. به‌منظور پیش‌بینی غلظت کربن‌مونوکسید آمار سال‌های ۱۳۸۳ و ۱۳۸۴ ایستگاه قلهک واقع در شمال تهران مورد استفاده قرار گرفته است. پس از اجرای مدل‌های پیش‌گفته، ضریب همبستگی (R)، شاخص میانگین نسبی خطای مطلق (MARE) و خطای میانگین مجموع مربعات (RMSE) در شبکه عصبی برای مرحله آزمون، به ترتیب برابر با ۰/۷۱۶، ۰/۱۵۸ و ۰/۹۶۹ به‌دست آمده که در مقایسه با مدل ترکیبی رگرسیونی (R =۰/۵۸۱ ، MARE = ۰/۱۸۹ و RMSE = ۱/۱۳۸) حاکی از برتری مطلق نتایج شبکهٔ عصبی نسبت به مدل ترکیبی رگرسیونی است.

واژه‌های کلیدی: تحلیل مؤلفه اصلی، شبکه عصبی مصنوعی، کربن‌مونوکسید، رگرسیون خطی چندمتغیره، تهران

Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: a case study of Tehran

Noori, R[1]., Ashrafi, Kh[2]. and Ajdarpour, A[3].

[1]*Ph.D. student, Faculty of Environment, University of Tehran*, Iran
[2]*Assistant professor, Faculty of Environment, University of Tehran, Iran*
[3]*M.Sc. student, Faculty of Environmental Health Engineering, Tarbiat Modares University, Tehran, Iran*

(Received: 11 Mar 2007 , Accepted: 15 Jan 2008)

Abstract

CO is the important air pollutant in Tehran. Two forecasting techniques are presented in this paper for prediction of average daily CO concentration. One of them, Multivariate Linear Regression (MLR) is based on Principal Component Analysis (PCA). The other technique is Artificial Neural Network (ANN) model. With this regard to six pollutants,

*نگارنده رابط: تلفن: ۰۹۱۷-۹۳۳۱۲۰۸ دورنگار: ۰۲۱-۶۶۴۰۷۷۱۹ E-mail: noori_2200@yahoo.com

i.e., PM_{10}, NO_x, SO_2, THC, CH_4 and O_3, and six meteorological variables, i.e., wind speed, wind direction, temperature, air pressure, humidity, solar radiation are used. These variables were measured daily throughout 2004 and 2005 at Gholhak Monitory Station, one of the eleven monitory stations in the Tehran area.

Among all the ANNs available paradigms, a Feed-Forward Multi-Layer Perceptron (FFMLP) was considered to be the best choice for this study because it is the most popular architecture for an ANNs. Therefore, in our research, one hidden layer FFMLP was used for the average daily CO concentration prediction. The activation functions chosen were the sigmoid hyperbolic tangent function in the hidden and output layers. The error correction learning with the Levenberg–Marquardt (L–M) algorithm was chosen for training the networks.

Regression model in matrix form can be shown as:

$$Y = X\beta + e \qquad (1)$$

where β is regression coefficient matrix, e is fitting error matrix and Y is response matrix. By solving equation for β we will have:

$$\beta = (X'X)^{-1}(XY') \qquad (2)$$

where X' is transpose of X.

For calculating the inverse of $(X'X)$, the independent variables should not have high relativity, because in this situation $(X'X)$ matrix can not become inverse and we will have more error in the data and calculations. To solve this problem, we should remove the multicolinearity between independent variables with PCA approach. In this research after removing the problem of multicolinearity on independent variables by the PCA, an appropriate model (PCA-MLR) was developed for predicting CO concentration. However, in the MLR calculation, stepwise algorithm has been used. In this method, entering the variables to the MLR is step by step condition, from the most important of them to the less important of them.

To achieve the best network structure for estimating CO concentration, various structures of FFMLP was investigated. Finally, a 13-22-1 architecture was selected for the best architecture of the network. Also, after removing the multicolinearity between independent variables, an appropriate PCA-MLR model was developed for prediction of CO concentration by stepwise algorithm.

In this step by performing PCA from 12 Principal Components (PCs), just 8 PCs were meaningful to enter the model. It estimates the CO concentration the regard to these new input variables. Finally, a PCA-MLR model is constructed that its equation is given below:

$$\begin{aligned} CO = {} & 4.92 + 0.60 \times (PC1) - 0.57 \times (PC9) + 0.35 \times (PC6) - 0.29 \times (PC5) - 0.24 \times (PC3) \\ & + 0.24 \times (PC11) + 0.16 \times (PC2) + 0.13 \times (PC7) \end{aligned}$$

For better judgment and selection of one on them, the Threshold Statistic (TS) index of testing step calculated and presented. For example this index shows that Absolute Relative Error (ARE) for 75% prediction of testing stage in ANN is 20%. This value (ARE) is 25% in the PCA-MLR model. However 90% of prediction of testing stage in ANN and PCA-MLR models are ARE equal to 41% and 53% respectively.

Finally, The use of FFMLP in prediction of average daily CO concentration in Tehran, is offered.

Key word: Principal component analysis, Artificial neural network, Carbon monoxide, Multivariate linear regression, Tehran

۱ مقدمه

آلودگی هوا با توجه به اثرات زیانبار آن بر انسان و محیط زیست در دهه اخیر، موضوعی بسیار با اهمیت است و توجه بسیاری از محققان را در این زمینه به خود جلب کرده است. این مشکل در شهر تهران به‌دلیل توپوگرافی خاص آن که با کوه‌های اطراف احاطه شده و همچنین حجم زیاد منابع آلودگی متحرک و ثابت، مشکلات سلامتی و اقتصادی زیادی را ایجاد ساخته است. کوه‌های اطراف تهران مانع مؤثری در مقابل نفوذ توده‌های گوناگون هوا به‌شمار می‌روند و به‌همین دلیل هوای تهران از آرامش و سکون بیشتری نسبت به مناطق مجاور برخوردار است. درضمن به‌دلیل وارونگی هوا در فصل‌های سرد سال، روزهای ناسالم این فصول هنگامی‌که ترافیک شهری سنگین‌تر است بروز می‌کند (بیات، ۱۳۸۳). از بین آلاینده‌های هوا در شهر تهران، کربن‌مونوکسید به‌دلیل حجم ترافیکی سنگین ناشی از ترابری، استفاده از خودروهای غیراستاندارد، مشکل احتراق ناقص سوخت‌های مورد استفاده در خودروها و بی‌توجهی که در طی سال‌های گذشته نسبت به آلودگی هوا در این شهر صورت گرفته، باعث مشکلات سلامتی و اقتصادی زیادی شده است. با توجه به اثرات مهلکی که کربن‌مونوکسید می‌تواند بر سلامتی انسان داشته باشد، اتخاذ تصمیمات لازم برای برنامه‌ریزی صحیح در مقابله با این معضل ضروری به‌نظر می‌رسد. پیش‌بینی غلظت آلاینده‌های شاخص گامی مؤثر در ایجاد تصمیمات لازم برای مقابله با آلودگی هوا است. برای این منظور می‌توان از مدل‌های در دسترس و روش‌های آماری استفاده کرد. تحقیقات صورت گرفته در زمینه پیش‌بینی کوتاه مدت آلاینده‌های هوا با استفاده از روش‌های آماری، سودمندی این روش‌ها را به اثبات رسانده است (گیلبرت، ۱۹۸۷؛ زانتی، ۱۹۹۰). این روش‌ها دارای مزیت‌هایی نسبت به روش‌های قطعی (deterministic methods) پیش‌بینی آلودگی هوا (مدل‌های پیش‌ساخته موجود) هستند. مثلاً در استفاده از این روش‌ها به اطلاعات انتشار و ضرایب انتشار آلاینده‌ها، که دسترسی به آنها در بیشتر موارد با مشکلاتی همراه است، نیازی نیست، زیرا اساس کار این روش‌ها به‌کارگیری اطلاعات مربوط به متغیرهای هواشناسی و آلودگی هوا است که دسترسی به آنها از طریق شبکه‌های سنجش آلودگی هوا به راحتی امکان‌پذیر است. از طرف دیگر ساختار مدل‌های آماری اغلب ساده‌تر از مدل‌های قطعی است، به‌طوری که استفاده از این مدل‌ها از سوی افرادی که تخصص زیادی نیز در زمینه آلودگی هوا ندارند، امکان‌پذیر است (نانری و همکاران، ۲۰۰۴). به هر جهت مدل‌های آماری ساخته شده برای یک منطقه، فقط در همان منطقه خاص قابل استفاده‌اند و در مناطق دیگر نمی‌توان از آنها استفاده کرد، زیرا با استفاده از اطلاعات همان منطقه ساخته و کالیبره شده‌اند.

روش‌های آماری متعددی برای پیش‌بینی غلظت آلاینده‌های هوا وجود دارند که از بین آنها تا به حال شبکه‌های عصبی (ANNs, Artificial Neural Networks)، مدل‌های رگرسیونی خطی و غیرخطی در پژوهش‌های مربوط به آلودگی هوا، به‌طور موفقیت‌آمیزی مورد استفاده قرار گرفته‌اند (بزنار و همکاران، ۱۹۹۳؛ فینزی و همکاران، ۱۹۹۸؛ گاردنر و درلینگ، ۱۹۹۸؛ نانری و همکاران، ۱۹۹۸؛ نانری و همکاران، ۲۰۰۱). شبکه عصبی از اوایل دهه ۱۹۹۰ در زمینه پیش‌بینی آلاینده‌های هوا، مورد استفاده قرار گرفت و اولین‌بار بزنار و همکارن (۱۹۹۳) برای پیش‌بینی غلظت گوگرددی‌اکسید، در نواحی صنعتی آلوده کشور اسلوونی از آن استفاده کردند. در ادامه سعی شده‌است اطلاعات کاملی در زمینه استفاده از شبکه عصبی و مدل‌های رگرسیونی که در تحقیقات قبلی مورد استفاده قرار گرفته‌اند، ارائه شود.

مسهلم و همکاران (۱۹۹۶) در تحقیقی، رابطه بین پارامترهای ترافیکی و غلظت کربن‌مونوکسید در یک چهارراه را که با ساختمان‌هایی از وزش باد محافظت شده بود، با استفاده از شبکه عصبی محاسبه کردند. درزدوویچ و همکاران (۱۹۹۷) با استفاده از مدل شبکه عصبی اقدام به پیش‌بینی غلظت ساعتی کربن‌مونوکسید در نواحی شهری روساریوی ایتالیا کردند. ناجندرا و خار (۲۰۰۴) نیز با توسعه مدل شبکه عصبی، غلظت کربن‌مونوکسید در یک بزرگ‌راه شهری را پیش‌بینی کردند. در این تحقیق از شش مشخصه ترافیکی و متغیرهای هواشناسی برای ساخت مدل و پیش‌بینی غلظت کربن‌مونوکسید استفاده شد.

پیش‌بینی آلاینده‌های دیگر در هوا نیز با استفاده از شبکه عصبی در تحقیقات گذشته مورد بررسی قرار گرفته است. گاردنر و درلینگ (۱۹۹۹) با استفاده از شبکه عصبی پرسپترون چندلایه در بخش مرکزی لندن، اقدام به پیش‌بینی غلظت ساعتی اکسیدهای نیتروژن و نیتروژن دی‌اکسید کردند. نتایج این تحقیق در مقایسه با تحقیقات صورت گرفته قبلی با استفاده از مدل‌های رگرسیونی (شی و هاریسون، ۱۹۹۷)، برتری مدل شبکه عصبی را به اثبات رساند. چلانی و همکاران (۲۰۰۲) مدل پیش‌بینی غلظت گوگرددی‌اکسید در سه محل متفاوت از شهر دهلی را با شبکه عصبی سه‌لایه، با یک‌لایه پنهان، ارائه کردند. آنها این کار را با استفاده از مدل رگرسیونی چندمتغیره نیز عملی ساختند و نتایج به‌دست‌آمده از این دو مدل را برای پیش‌بینی غلظت گوگرددی‌اکسید، با هم مقایسه کردند. نتایج به‌دست‌آمده از بررسی آنها نشان‌دهنده برتری مدل شبکه عصبی بود. ساهین و همکاران (۲۰۰۵) در تحقیقی با استفاده از شبکه عصبی سه‌لایه و رگرسیون غیرخطی، مدلی برای غلظت گوگرددی‌اکسید در شهر استانبول ترکیه ارائه کردند. نتایج این تحقیق نیز نشان‌دهنده برتری شبکه عصبی نسبت به رگرسیون غیرخطی بود. همچنین کورانی (۲۰۰۵) برای پیش‌بینی غلظت میانگین روزانه ذرات معلق کوچک‌تر از ۱۰ میکرون در میلان ایتالیا، مدل‌هایی با استفاده از شبکه عصبی و رگرسیون خطی ایجاد کرد. همان‌طور که ذکر شد، نتایج به‌دست آمده از اکثر این بررسی‌ها، نشان‌دهنده برتری مدل‌های شبکه عصبی نسبت به مدل‌های رگرسیونی در پیش‌بینی غلظت آلاینده‌های هوا است. برای استفاده از مدل‌های رگرسیونی، عواملی مانند نرمال نبودن متغیر وابسته (متغیر هدف) و همبستگی زیاد متغیرهای مستقل (متغیرهای ورودی به مدل برای پیش‌بینی متغیر وابسته) باعث ناپایداری مدل به‌دست آمده می‌شود و صحت نتایج به‌دست آمده از این روش را زیر سئوال می‌برد.

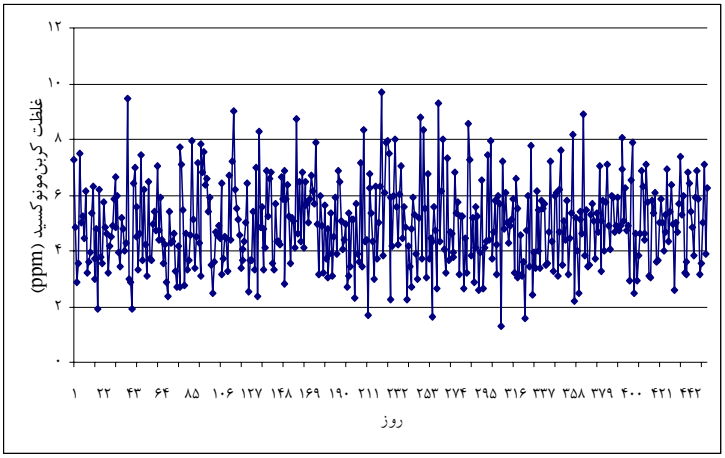
در این تحقیق، برای جلوگیری از مشکلات ذکرشده در مدل رگرسیونی، از روش تحلیل مؤلفه اصلی (principal component analysis) برای پردازش متغیرهای ورودی، حذف همبستگی بین متغیرهای مستقل و تفسیر بهتر نتایج مدل رگرسیون خطی چندمتغیره استفاده شده است. همچنین مدلی نیز با استفاده از شبکه عصبی برای این منظور ارائه و نتایج آن با نتایج به‌دست‌آمده از مدل رگرسیونی بر پایه تحلیل مؤلفه‌های اصلی مقایسه شده است. به این ترتیب در بخش دوم، روش تحقیق بیان می‌شود و در بخش سوم نتایج حاصل از کاربرد دو روش ذکرشده ارائه و باهم مقایسه می‌شوند. درنهایت در بخش چهارم با مجموعه‌ای از نکات اساسی به جمع‌بندی مقاله پرداخته شده است.

۲ روش تحقیق

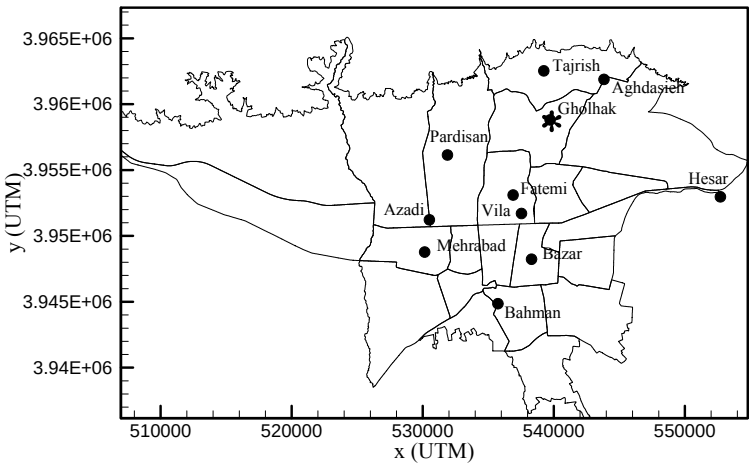
۲-۱ منطقه مورد بررسی و داده‌های مسئله

کلان‌شهر تهران، در کوهپایه‌های جنوبی رشته‌کوه البرز گسترده شده و حد فاصل طول جغرافیایی ۵۱ درجه و ۲ دقیقه شرقی تا ۵۱ درجه و ۳۶ دقیقه شرقی، به طول تقریبی

۵۰ کیلومتر، و عرض جغرافیایی ۳۵ درجه و ۳۴ دقیقه شمالی تا ۳۵ درجه و ۵۰ دقیقه شمالی، به عرض تقریبی ۳۰ کیلومتر را دربر گرفته است. ارتفاع این شهر در شمالی‌ترین نقطه حدود ۱۸۰۰ متر و جنوبی‌ترین نقطه ۹۵۰ متر از سطح دریا است. در حقیقت تهران در بین مناطق کوهستانی از شمال و نواحی کویری از جنوب احاطه شده است. در طی دوره‌ای سی ساله (از ۱۳۴۵ تا ۱۳۷۵) در این شهر، میانگین بارندگی ۲۴۰ میلی‌متر و میانگین تعداد روزهای یخ‌بندان ۴۸ روز در سال بوده است (بختیاری، ۱۳۸۵). جمعیت این شهر نیز مطابق با آخرین آمارگیری حدود ۸ میلیون نفر برآورد شده است (مرکز آمار ایران، ۱۳۸۵). نتایج به‌دست آمده از تحقیقی در مورد آلاینده‌های هوای این شهر، بیانگر این واقعیت است که ۹۰ درصد وزن کل آلاینده‌های هوای شهر تهران از خودروها منتشر می‌شود و ۱۰ درصد دیگر مربوط به منابع ثابت است (بیات، ۱۳۸۳). کربن‌مونوکسید نسبت به بقیه آلاینده‌های هوا در شهر تهران از اهمیت بیشتری برخوردار است، به‌طوری‌که بیش از سه‌چهارم وزن آلاینده‌های هوا در این شهر را کربن‌مونوکسید تشکیل می‌دهد (بیات، ۱۳۸۳). سری زمانی مربوط به میانگین غلظت روزانه این آلاینده در سال‌های ۱۳۸۳ و ۱۳۸۴ در شکل ۱ آمده است. در شهر تهران هم‌اکنون ۱۱ ایستگاه سنجش آلودگی هوا وجود دارد که کار اندازه‌گیری غلظت آلاینده‌های شاخص هوا را به انجام می‌رساند (شکل ۲). در این تحقیق به‌منظور پیش‌بینی میانگین غلظت روزانه کربن‌مونوکسید، از داده‌های هواشناسی شامل دما (temp)، رطوبت نسبی (hum)، فشار هوا (press)، سرعت باد (WS)، جهت باد (WD)، تابش خورشیدی (solar) و داده‌های آلودگی هوا شامل گوگرددی‌اکسید (SO_2)، کل هیدروکربن‌ها (THC)، ازن (O_3)، اکسیدهای نیتروژن (NO_X) و ذرات معلق با قطر کمتر از ۱۰ میکرون (PM_{10}) در ایستگاه قلهک واقع در شمال تهران، ثبت شده در سال‌های ۱۳۸۳ و ۱۳۸۴ استفاده شده است. برخی از مشخصات هر کدام از این متغیرها در جدول ۱ آورده شده است. به‌علت خاموشی و مشکلات فنی دستگاه سنجش آلودگی هوا در برخی از روزهای سال، اطلاعات برخی از روزها در طی این دو سال در دسترس نبودند و مجموعاً بعد از مرتب کردن اطلاعات ثبت شده، از اطلاعات ۴۵۴ روز ثبت شده در این ایستگاه در طی این دو سال استفاده شده است.



شکل ۱. سری زمانی میانگین غلظت روزانه برای کربن‌مونوکسید در سال‌های ۱۳۸۳ و ۱۳۸٤.



شکل ۲. موقعیت ایستگاه قلهک در بین ایستگاه‌های دیگر و شهر تهران با علامت ستاره (*) مشخص شده است.

جدول ۱. مقادیر میانگین، بیشینه، کمینه، میانه و انحراف معیار برای متغیرهای مورد استفاده در سال‌های ۱۳۸۳ و ۱۳۸٤.

مشخصه	PM_{10}	Hum	NO_X	Press	SO_2	THC	Temp	WD	WS	CH_4	O^3	solar	CO
واحد	ug/m^3	%RH	ppm	mBar	ppm	ppm	°C	Deg	m/s	ppm	ppm	KW/M^2	ppm
میانگین	9.81	49.17	0.18	845.97	0.03	4.03	18.93	182.28	0.84	1.67	0.01	0.25	4.93
بیشینه	190.27	94.29	0.64	903.30	0.07	8.73	38.29	282.73	20.71	4.73	0.04	0.59	9.67
کمینه	3.05	10.14	0.00	597.59	0.00	0.00	0.59	32.27	0.00	0.00	0.00	0.01	1.30
میانه	8.92	47.64	0.12	850.71	0.03	4.34	20.87	181.61	0.80	1.82	0.01	0.27	4.81
انحراف معیار	9.08	18.55	0.16	19.80	0.01	1.78	9.07	28.59	1.18	0.97	0.01	0.12	1.53

۲-۲ شبکه عصبی مصنوعی

فن شبکه عصبی مصنوعی را اولین‌بار مک‌کلاچ و پیتز (۱۹۴۳) ارائه کردند. با وجود به‌کارگیری یک ساختمان ساده از این مدل، سرعت و قدرت محاسباتی آن بشدت مورد توجه قرار گرفت. ANN ها مدل‌هایی محاسباتی هستند که قادرند رابطه بین ورودی‌ها و خروجی‌های یک دستگاه فیزیکی (هر چند پیچیده و غیرخطی) را با شبکه‌ای از گره‌ها که همگی با هم متصل‌اند، تعیین کنند. از مهم‌ترین عوامل تعریف ANNs نحوه معماری آن است. ساختار ANNs که معماری به آن اطلاق می‌شود، به‌شکلی است که نرون‌ها در دسته‌هایی که لایه نام دارند، مرتب می‌شوند. معماری معمول ANNs متشکل از سه‌لایه است، لایه ورودی (داده‌ها را در شبکه توزیع می‌کند)، لایه پنهان (داده‌ها را پردازش می‌کند) و لایه خروجی (نتایج را به ازای ورودی‌های مشخص، استخراج می‌کند). یک شبکه می‌تواند چندین لایه پنهان داشته باشد. به هر حال کارهای نظری صورت گرفته در این زمینه نشان داده‌اند که یک لایه پنهان برای این گونه مدل‌ها می‌تواند هر تابع پیچیده و غیرخطی را تقریب زند (سیبنکو، ۱۹۸۹؛ هورنیک و همکاران، ۱۹۸۹). هم‌چنین نتایج تجربی و عملی نیز این

موضوع را تایید می‌کنند (زانگ و همکاران، ۱۹۹۸).

تقسیم‌بندی متغیرها برای استفاده از آنها در شبکه عصبی به روش‌های متفاوتی صورت می‌گیرد. در این تحقیق از روش درستی‌سنجی موازی استفاده شده است. درستی‌سنجی موازی روشی است که به دفعات در مدل‌های شبکه عصبی مورد استفاده قرار گرفته است (استون، ۱۹۷۴). این روش برای تعیین زمان پایان آموزش و مقایسهٔ توانایی تعمیم مدل‌های متفاوت مورد استفاده قرار می‌گیرد (بوردن و همکاران، ۱۹۹۷). در روش درستی‌سنجی موازی از دستهٔ مستقلی از داده‌ها برای آزمایش قدرت تعمیم مدل، در مراحل متفاوت آموزش استفاده می‌شود. دستهٔ مستقل به این معناست که داده‌های آن نباید در دو دستهٔ آموزش و درستی‌یابی وجود داشته باشند. در نتیجه در این روش، داده‌ها به سه دسته تقسیم می‌شوند. دسته مربوط به آموزش شبکه، که با آنها وزن‌های شبکه تعیین می‌شوند، دسته نظارت بر آموزش شبکه که با بررسی خطای این دسته درحین آموزش شبکه نسبت به توقف محاسبات، تصمیم گرفته می‌شود و دسته درستی‌سنجی که توانایی شبکه پس از آموزش را بررسی می‌کند. درحین آموزش تا زمانی‌که خطای مربوط به سری داده‌های نظارت کاهش یابد، آموزش ادامه می‌یابد. هنگامی‌که خطای مربوط به داده‌های درستی‌یابی شروع به افزایش کند، آموزش متوقف می‌شود. با به‌کار بردن این روش، که روش توقف آموزش (stop training algorithm) نیز نامیده می‌شود، امکان استفاده از معماری‌های پیچیده‌تر در طراحی شبکه برای کاربر فراهم می‌شود، بدون اینکه مشکل فوقِ برازشی (overfitting) روی دهد و با قرار دادن پاره‌ای معیارها، به محض روی دادن این مشکل در شبکه، آموزش متوقف می‌شود. بدین ترتیب معیارهای مورد اشاره نقش مهمی در این روش ایفا می‌کنند (کولیبالی و همکاران، ۲۰۰۰).

با توجه به مطالب ذکرشده، دراین مقاله از شبکه عصبی پرسپترون چندلایه feed-forward با یک لایه پنهان و روش توقف آموزش، برای پیش‌بینی غلظت کربن‌مونوکسید استفاده شده است.

۲-۳ تحلیل مؤلفه‌های اصلی

تحلیل مؤلفه اصلی از روش‌های آماری چندمتغیره است که می‌توان از آن برای کاهش تعداد متغیرها و تفسیر بهتر اطلاعات استفاده کرد (کامدویرن و همکاران، ۲۰۰۵). با اعمال این روش، متغیرهای ورودی اولیه به مؤلفه‌های جدید بدون همبستگی تبدیل می‌شوند؛ به‌طوری‌که مؤلفه‌های ایجاد شده، ترکیبی خطی از متغیرهای ورودی‌اند (لو و همکاران، ۲۰۰۳). با استفاده از این روش، ترکیباتی از P متغیر $X_p , \dots , X_2 , X_1$ برای ایجاد P مؤلفه مستقل $Z_p , \dots , Z_2 , Z_1$ برقرار می‌شود. نبود همبستگی بین این مؤلفه‌ها یک ویژگی مفید است به این معنی که مؤلفه‌ها جنبه‌های متفاوتی از پارامترهای اصلی را نمایان می‌سازند (مانلی، ۱۹۸۶). در این روش اطلاعات پارامترهای اصلی با کمترین تلفات در مؤلفه‌های حاصل آورده می‌شود (جانسون و ویچرن، ۱۹۸۲). هر مؤلفه اصلی می‌تواند با دنباله زیر مشخص شود:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p \qquad (۱)$$

که درفرمول (۱)، Z_i معرف مؤلفه موردنظر، a_{ij} بردار ویژه مربوطه و X_i نیز متغیرهای اصلی است. این اطلاعات از حل معادله زیر به دست می‌آید (جانسون و ویچرن، ۱۹۸۲).

$$\left|R - \lambda I\right| = 0 \qquad (۲)$$

که در آن I ماتریس واحد، R ماتریس واریانس-کوواریانس متغیرهای ورودی و λ نیز مقادیر ویژه این ماتریس است. از این مقادیر ویژه، بردارهای ویژه به‌دست می‌آیند. برای دستیابی به این مقادیر مراحل زیر باید صورت گیرد:

۱- محاسبه فاکتور KMO: مقدار این فاکتور بین صفر تا یک متغیر است. این فاکتور به کمک ضرایب همبستگی ساده و ضرایب همبستگی جزیی طبق فرمول (۳) محاسبه می‌شود.

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \qquad i \neq j \qquad (۳)$$

در فرمول (۳)، r_{ij} ضریب همبستگی ساده بین متغیرهای i و j و a_{ij} نیز ضریب همبستگی جزئی متغیرهای i و j به شرط ثابت بودن سایر متغیرهاست.

با توجه به فرمول (۳) کمیت فاکتور KMO با ضرایب همبستگی ساده رابطه مستقیم و با ضرایب همبستگی جزیی رابطه معکوس دارد. مقادیر بالاتر KMO مستلزم کوچک بودن ضرایب همبستگی جزیی (که برآورد ضریب همبستگی جملات خطا) و بیانگر دقت محاسبات مربوطه با استفاده از روش تحلیل مؤلفه اصلی است. در صورتی‌که این فاکتور بزرگ‌تر از ۰/۵ به‌دست آید، این امر نشان‌دهنده امکان اجرای تحلیل مؤلفه اصلی بر متغیرهای ورودی است (سینگ و همکاران، ۲۰۰۴؛ شرستا و کازاما، ۲۰۰۷).

۲- استاندارد کردن متغیرهای ورودی: در این مرحله متغیرهای ورودی مطابق فرمول (۴) به نحوی استاندارد می‌شوند که دارای میانگین صفر و انحراف‌معیار یک باشند.

$$Z = \frac{X - \mu}{\sigma} \qquad (۴)$$

در فرمول (۴)، Z معادل استاندارد شده مشاهدات، X مشاهدات خام ورودی، μ میانگین هر متغیر و σ نیز مقادیر انحراف معیار برای هر متغیر است.

۳- محاسبه ماتریس واریانس-کوواریانس برای متغیرها: این ماتریس، که ماتریسی متقارن است، میزان تغییرات در نمونه و میزان همبستگی P متغیر را با هم نشان می‌دهد. عضوهای روی قطر اصلی این ماتریس، واریانس متغیرهای ورودی و بقیه درایه‌های این ماتریس، کوواریانس بین متغیرهای ورودی است. چون برای تشکیل این ماتریس از داده‌های استاندارد شده استفاده شده است، به‌همین دلیل این ماتریس معادل ماتریس همبستگی بین متغیرهای ورودی است.

۴- اجرای چرخش مناسب روی ماتریس ضرایب مؤلفه‌ها: چون در بسیاری از موارد تعدادی از متغیرها به یک مؤلفه ویژه یا حتی به تعدادی از مؤلفه‌ها بستگی دارند، تفسیر مؤلفه‌ها مشکل خواهد بود. از این‌رو روش‌هایی پدید آمده است که بدون تغییر میزان اشتراک، باعث تفسیر ساده‌تر عوامل می‌شوند. این روش‌ها، همان دوران مؤلفه‌ها هستند و به دو نوع دوران عمود و دوران مایل تقسیم می‌شوند. در چرخش عمود استقلال بین مؤلفه‌ها حفظ می‌شود. یکی از روش‌های چرخش عمودی که بیشتر مورد استفاده قرار می‌گیرد، چرخش varimax نامیده می‌شود (وگا و همکاران، ۱۹۹۸؛ هلنا و همکاران، ۲۰۰۰؛ سیمونوف و همکاران، ۲۰۰۳). این روش نسبت به بقیه روش‌ها نتایج بهتری را درپی دارد و به‌مثابه چرخش استاندارد توصیه می‌شود (مانلی، ۱۹۸۶). جزئیات بیشتر در مورد تحلیل مؤلفه اصلی در منابع دیگر در دسترس است (نوری و همکاران ۱۳۸۶؛ دیویس، ۱۹۸۶؛ واکرناجل، ۱۹۹۵؛ تاباکنیک و فیدل، ۲۰۰۱؛ اویانگ، ۲۰۰۵).

۴-۲ مدل رگرسیون خطی چندمتغیره

مدل رگرسیونی به شکل ماتریسی را می‌توان به‌صورت معادله زیر نشان داد:

$$Y = X\beta + e \qquad (۵)$$

در فرمول (۵)، β ماتریس ضرایب رگرسیون، e ماتریس خطای برازش و Y نیز ماتریس پاسخ می‌باشد. با حل

معادله بالا بر حسب β خواهیم داشت:

$$\beta = (X'X)^{-1}(X'Y) \qquad (۶)$$

که در رابطه (۶)، X' ترانهاده ماتریس X است. برای محاسبه معکوس $(X'X)$، لازم است متغیرهای مستقل همبستگی زیادی نداشته باشند، زیرا در این صورت ماتریس $(X'X)$ را نمی‌توان معکوس کرد و باعث افزایش خطا در اثر گرد کردن داده‌ها و محاسبات می‌شود. برای رفع این مشکل باید قبل از ساخت مدل رگرسیونی، همبستگی بین متغیرهای مستقل را از بین برد. در این خصوص روش مناسب، استفاده از تحلیل مؤلفه‌های اصلی روی متغیرهای مستقل ورودی به مدل است. معیار قضاوت برای رفع این مشکل با اجرای تحلیل مؤلفه‌های اصلی روی متغیرهای ورودی، فاکتور تورم واریانس است. عدد ایدئال برای فاکتور تورم واریانس یک است و مقادیر بزرگ‌تر از ۱۰ برای تورم واریانس نشانه ناپایداری مدل رگرسیونی است (هاکینگ، ۲۰۰۳). در این تحقیق پس از رفع مشکل همبستگی در متغیرهای مستقل، مدلی مناسب با استفاده از فن رگرسیون خطی چندمتغیره برای پیش‌بینی غلظت روزانه کربن‌مونوکسید توسعه یافته و در محاسبات رگرسیونی از الگوریتم گام به گام (stepwise) استفاده شده است. در این روش ورود متغیرها به مدل رگرسیون به‌صورت مرحله‌ای، از مهم‌ترین متغیر تا کم اهمیت‌ترین آنها، صورت می‌گیرد. معیار میزان اهمیت متغیر در مدل، مقدار سطح معنی‌داری یا آماره F متناظر با آن در جدول آزمون معنی‌داری متغیرها است.

۲-۵ معیار ارزیابی اعتبار مدل‌های مورد استفاده

یکی از معیارهای مورد بررسی اعتبار نتایج به‌دست‌آمده از مدل رگرسیونی و شبکه عصبی، معیار برازندگی ضریب همبستگی (R)، است که مقدار آن بین ۱- و ۱+ تغییر می‌کند و مقادیر مطلق نزدیک به ۱ تطابق بهتر داده‌های مشاهده‌ای و برآوردشده را نشان می‌دهند. ولی با توجه به این مسئله که مقدار R تحت تاثیر داده‌های پرت است، باید از آن به اتفاق پارامترهای دیگر استفاده کرد (لگاتس و مک‌کاب، ۱۹۹۹). به‌همین دلیل در این تحقیق از معیارهای ریشه متوسط مربعات خطا (RMSE, Root Mean Squre Error)، و میانگین نسبی خطای مطلق (MARE, Mean Absulete Realative Error)، نیز برای بررسی اعتبار نتایج، استفاده شده است. نحوه محاسبه این دو معیار در روابط (۷) و (۸) آمده است.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \qquad (۷)$$

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\frac{|O_i - P_i|}{O_i} \qquad (۸)$$

در روابط (۷) و (۸)، o_i مقدار غلظت کربن‌مونوکسید مشاهده شده، P_i مقدار غلظت کربن‌مونوکسید پیش‌بینی شده و n تعداد مشاهدات است.

شاخص‌های معرفی شده در بالا، نمایه‌های آماری کلی‌اند و هیچ اطلاعاتی در مورد نحوه توزیع خطا ارائه نمی‌کنند. از این‌رو برای ارزیابی توانایی مدل‌ها، نیاز به شاخص‌های آماری است که نحوه توزیع خطا در مدل‌های ساخته شده را مشخص سازند. به‌همین منظور (علاوه بر سه پارامتر آماری ذکر شده) از نمودار پراکندگی قدرمطلق مقادیر خطای نسبی برای ارزیابی نهایی مدل‌ها استفاده شده است.

۳ بحث و نتایج

۳-۱ آنالیز حساسیت

در حالت کلی، شبکه عصبی اطلاعاتی در مورد اهمیت و میزان تأثیر متغیرهای ورودی بر خروجی نمی‌دهد. برای فهم درصد تأثیر هریک از متغیرهای ورودی بر خروجی می‌توان از آنالیز حساسیت استفاده کرد. با آنالیز

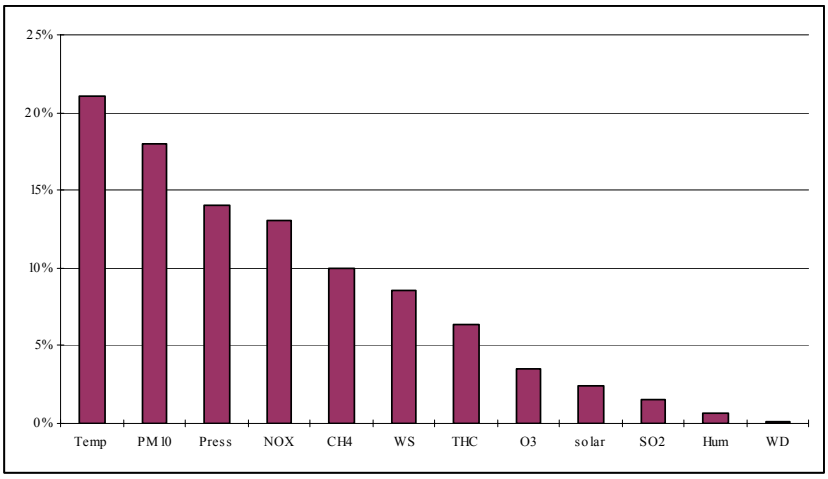
کردن حساسیت متغیرهای ورودی که بیشترین تغییرات را در خروجی به‌وجود می‌آورند، مشخص می‌شود. در نتیجه این متغیرها می‌بایست با دقت و حساسیت بالایی اندازه‌گیری و به‌کار برده شوند این متغیرها از درجه اهمیت بیشتری نسبت به بقیه متغیرها برخوردارند. در همین راستا در این تحقیق به‌منظور شناخت درصد تأثیر هر یک از پارامترهای ورودی بر غلظت روزانه کربن‌مونوکسید، آنالیز حساسیت صورت گرفت که نتیجه آن در شکل ۳ آمده است.

همان‌طور که انتظار می‌رفت دما مؤثرترین عامل مؤثر بر میانگین غلظت روزانه کربن‌مونوکسید است، زیرا تغییرات دما باعث ناپایداری هوا و جابه‌جایی آن می‌شود. این عامل می‌تواند بر میزان کربن‌مونوکسید مشاهده شده تأثیرگذار باشد. دومین عامل مؤثر بر غلظت کربن‌مونوکسید، وجود ذرات معلق با قطر کوچک‌تر از ۱۰ میکرون است. با توجه به این مسئله که منبع اصلی انتشار این آلاینده نیز در شهر تهران همانند کربن‌مونوکسید، خودروها (منبع اصلی تولید کربن‌مونوکسید) هستند. انتظار می‌رود که به‌طور همزمان میزان این دو آلاینده کم یا زیاد شود. با توجه به شکل ۳ جهت باد کمترین تأثیر را نسبت به عامل‌های دیگر دارد و سرعت باد دارای تأثیر متوسطی است، زیرا امکان وزش بادهای شدید و یا بادهای دائم در تهران وجود ندارد. این مسئله از عوامل زیادی نظیر توپوگرافی و موقعیت جغرافیایی تهران و غیره تأثیر می‌گیرد که بررسی این عوامل هدف مقاله حاضر نیست و به تحقیقات جداگانه‌ای احتیاج دارد.

۳-۲ پیش پردازش متغیرهای ورودی به مدل رگرسیونی با PCA

بررسی اولیه نشان داد که بین متغیرهای ورودی مورد استفاده در این تحقیق همبستگی معنی‌داری وجود دارد که برای از بین بردن این مشکل، از روش تحلیل مؤلفه اصلی استفاده شد. مقدار KMO=۰/۷۲۱ نیز امکان اجرای تحلیل مؤلفه اصلی را تأیید کرد. برای اجرای این روش، پس از استاندارد کردن متغیرهای ورودی ماتریس متقارن همبستگی از مرتبه ۱۲ (معادل با تعداد متغیرهای ورودی) تشکیل شد، که نتایج آن در زیر آمده است.

$$\begin{bmatrix}
1.00 & -0.01 & -0.01 & -0.13 & 0.07 & 0.05 & 0.04 & 0.09 & 0.05 & 0.04 & -0.06 & 0.04 \\
-0.01 & 1.00 & 0.16 & 0.47 & -0.06 & 0.16 & -0.76 & -0.36 & 0.15 & 0.37 & -0.06 & -0.65 \\
-0.01 & 0.16 & 1.00 & 0.50 & -0.42 & 0.65 & -0.03 & 0.12 & 0.15 & 0.54 & 0.41 & -0.28 \\
-0.13 & 0.47 & 0.50 & 1.00 & -0.23 & 0.45 & -0.49 & -0.12 & 0.26 & 0.60 & 0.27 & -0.49 \\
0.07 & -0.06 & -0.42 & -0.23 & 1.00 & -0.48 & 0.00 & -0.11 & -0.07 & -0.47 & -0.22 & 0.21 \\
0.05 & 0.16 & 0.65 & 0.45 & -0.48 & 1.00 & -0.11 & 0.06 & 0.03 & 0.89 & 0.36 & -0.41 \\
0.04 & -0.76 & -0.03 & -0.49 & 0.00 & -0.11 & 1.00 & 0.32 & -0.05 & -0.33 & 0.18 & 0.74 \\
0.09 & -0.36 & 0.12 & -0.12 & -0.11 & 0.06 & 0.32 & 1.00 & 0.11 & -0.07 & 0.16 & 0.31 \\
0.05 & 0.15 & 0.15 & 0.26 & -0.07 & 0.03 & -0.05 & 0.11 & 1.00 & 0.06 & 0.11 & -0.05 \\
0.04 & 0.37 & 0.54 & 0.60 & -0.47 & 0.89 & -0.33 & -0.07 & 0.06 & 1.00 & 0.26 & -0.53 \\
-0.06 & -0.06 & 0.41 & 0.27 & -0.22 & 0.36 & 0.18 & 0.16 & 0.11 & 0.26 & 1.00 & 0.17 \\
0.04 & -0.65 & -0.28 & -0.49 & 0.21 & -0.41 & 0.74 & 0.31 & -0.05 & -0.53 & 0.17 & 1.00
\end{bmatrix}$$



شکل ۳. درصد تأثیر هر یک از متغیرها بر غلظت میانگین روزانه کربن‌مونوکسید.

از حل دستگاه معادلات (۲)، ۱۲ مقدار ویژه و به‌ازای هر مقدار ویژه ۱۲ بردار ویژه، حاصل می‌شود که با استفاده از آنها، ۱۲ مؤلفه از متغیرهای اولیه به‌دست می‌آید. مشخصات هر مؤلفه در بیان پراکندگی متغیرهای اولیه در جدول ۲ آورده شده است.

مقدار عددی هر مؤلفه با تقسیم مقادیر ویژه به‌دست آمده بر تعداد متغیرهای مورد استفاده، به‌دست می‌آید. درصد پراکندگی نیز از تقسیم مقدار عددی هر مؤلفه بر تعداد متغیرهای مورد استفاد، محاسبه می‌شود. همان‌طور که در جدول ۱ مشخص است به تعداد متغیرهای مورد استفاده، مؤلفه ایجاد شده است. مقادیر بردارهای ویژه که ضرایب هر مؤلفه را برای محاسبه آنها تعیین می‌کند، با استفاده از تحلیل مؤلفه اصلی، در جدول ۳ آمده است.

همان‌طور که در جدول ۲ نیز مشخص است، ۷ مؤلفه اول بیش از ۷۸ درصد کل پراکندگی داده‌های اصلی را بیان می‌کنند. جدول ۳ نیز ضرایب مربوط به هر متغیر اصلی (بردارهای ویژه) برای ایجاد مؤلفه‌ها را نشان می‌دهد.

۳-۳ ساخت مدل رگرسیونی اصلاح شده

به‌منظور استفاده از داده‌های مسئله در مدل رگرسیونی، لازم است که متغیر وابسته (غلظت روزانه مونوکسید کربن) دارای توزیع نرمال باشد. بدین منظور با استفاده از روش آماری کولموگوروف–اسمیرنوف (Kolmogorov-Smirnov) نرمال بودن داده‌های مورد استفاده، تأیید شد (مانلی، ۱۹۸۶). هم‌چنین با استفاده از روش تحلیل مؤلفه‌های اصلی مشکل همبستگی بین متغیرهای مستقل نیز رفع شد. پس از تأیید نرمال بودن متغیر وابسته و رفع مشکل همبستگی در متغیرهای مستقل، مدلی مناسب با استفاده از روش رگرسیون خطی چندمتغیره با الگوریتم گام به گام (Stepwise) برای پیش‌بینی غلظت روزانه کربن‌مونوکسید بسط یافت. که نتایج آن در جدول ۴ آمده است.

همان‌طور که در جدول ۴ مشخص است در مدل رگرسیونی با اجرای تحلیل مؤلفه‌های اصلی از ۱۲ مؤلفه، تنها ورود ۸ مؤلفه به مدل معنی‌دار بوده است و مدل مورد نظر غلظت کربن‌مونوکسید را با توجه به این مقادیر برآورد می‌کند. در جدول فوق مشاهده می‌شود که مدل رگرسیون حاصل از متغیرهای اولیه برای هر کدام از مؤلفه‌های ورودی، دارای مقادیر تورم واریانس نزدیک به یک (یعنی مقدار ایدئال) است. در نهایت پس از پردازش

جدول ۲. درصد پراکندگی که با هرمؤلفه با چرخش Varimax بیان می‌شود.

مؤلفه	مقدار هر مؤلفه از ۱۲	درصد پراکندگی	درصد پراکندگی تجمعی
1	2.2	18.33	18.33
2	2.06	17.18	35.51
3	1.05	8.79	44.3
4	1.02	8.53	52.83
5	1.02	8.52	61.35
6	1.02	8.49	69.85
7	1.01	8.4	78.25
8	0.94	7.82	86.06
9	0.88	7.3	93.36
10	0.53	4.39	97.75
11	0.21	1.71	99.46
12	0.06	0.54	100

جدول ۳. مشخصات هر مؤلفه با چرخش Varimax.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
PM_{10}	0.01	0.03	-0.03	0.04	0.03	0.03	0.1	-0.01	-0.05	0.01	$2*10^{-3}$	$3^{*}10^{-5}$
Hum	**-0.94**	0.11	-0.01	-0.18	0.09	-0.02	-001	0.07	0.11	-0.03	0.21	0.01
NO_X	-0.04	0.37	0.21	0.07	0.07	-0.18	-0.01	0.86	0.18	-0.06	0.01	10^{-3}
Press	-0.33	0.31	0.14	-0.05	0.15	-0.06	-0.09	0.2	**0.83**	-0.09	-0.03	$2^{*}10^{-3}$
SO_2	0.01	-0.28	-0.07	-0.06	-0.03	**0.94**	0.04	-0.14	-0.05	0.04	$-2*10^{-3}$	$7*10^{-5}$
THC	-0.04	**0.91**	0.16	0.04	-0.01	-0.18	0.03	0.26	0.08	-0.11	$4*10^{-3}$	-0.17
Temp	**0.84**	-0.09	0.1	0.11	0.01	-0.02	0.02	0.06	-0.21	0.21	0.41	$-7*10^{-3}$
WD	0.21	0.8	0.07	0.97	0.07	-0.05	0.05	0.05	-0.03	0.06	0.01	$-2*10^{-3}$
WS	-0.06	0.3	0.05	0.06	**0.99**	-0.03	0.03	0.05	0.09	0	$3*10^{-3}$	$7*10^{-4}$
CH_4	-0.24	**0.88**	0.08	-0.04	0.02	-0.21	0.03	0.13	0.23	-0.08	-0.02	0.19
O_3	0.09	0.16	**0.96**	0.07	0.05	-0.07	-0.03	0.15	0.09	0.06	0.01	$-8*10^{-4}$
Solar	**0.57**	-0.32	0.16	0.15	-0.01	0.1	0.03	-0.11	-0.15	0.69	0.04	$3*10^{-4}$

جدول ٤. نتایج ورود هر مؤلفه به مدل رگرسیون خطی چندمتغیره.

پارامتر	ضریب هر مؤلفه	سطح معنی داری	فاکتور تورم واریانس
Constant	4.92	<0.01	
PC1	0.6	<0.01	1
PC9	-0.57	<0.01	1.01
PC6	0.35	<0.01	1
PC5	-0.29	<0.01	1
PC3	-0.24	<0.01	1
PC11	0.24	<0.01	1
PC2	0.16	0.01	1
PC7	0.13	0.04	1

اولیه روی متغیرهای ورودی، مدل رگرسیونی با اجرای تحلیل مؤلفه اصلی برای برآورد غلظت روزانه کربن‌مونوکسید تهیه شد که معادله آن در زیر آورده شده است.

$$CO = 4.92 + 0.60 \times (PC1) - 0.57 \times (PC9) + 0.35 \times (PC6) - 0.29 \times (PC5) - 0.24 \times (PC3) + 0.24 \times (PC11) + 0.16 \times (PC2) + 0.13 \times (PC7) \quad (۹)$$

مطابق با معادله فوق، مؤلفه اول (PC1)، که در تشکیل آن به ترتیب متغیرهای رطوبت نسبی، درجه حرارت و تابش خورشیدی بیشترین تأثیر را داشته‌اند، مهم‌ترین عامل تأثیرگذار بر غلظت روزانه کربن‌مونوکسید است. مؤلفه‌های موثر دیگر بر غلظت کربن‌مونوکسید، به ترتیب PC9، PC6، PC5، PC3، PC11، PC2 وPC7 هستند. در جدول ۳ متغیرهایی که بیشترین تأثیر را بر این مؤلفه‌ها داشته‌اند، با قلم مشکی پر نوشته شده‌اند. پس از ساخت مدل رگرسیونی اعتبار این مدل مورد بررسی قرار گرفت که نتایج مراحل آموزش و آزمون مدل در جدول ۵ و شکل‌های ۴ و ۵ آمده است.

همان‌طور که در جدول ۵ و شکل‌های ۴ و ۵ نیز مشخص است، نتایج مدل رگرسیونی در مقایسه با شبکه عصبی از دقت کمتری برخوردار است. و استفاده از این مدل برای برآورد غلظت روزانه کربن‌مونوکسید با خطای زیادی نسبت به مدل شبکه عصبی همراه است.

جدول ۵. نتایج مراحل آموزش و آزمون مدل رگرسیونی.

معیار ارزیابی	مرحله آموزش	مرحله آزمون
R	0.663	0.581
MARE	0.207	0.189
RMSE	1.181	1.138

۳-۴ شبکه عصبی

داده‌های مورد استفاده در مدل‌سازی با شبکه عصبی برای ارائه مدل پیش‌بینی غلظت کربن‌مونوکسید، به‌طور تصادفی به سه دسته تقسیم‌بندی شدند، که دسته اول برای ساخت مدل شامل اطلاعات ۳۱۹ روز، دسته دوم برای آزمون مدل ساخته شده شامل اطلاعات ۵۴ روز و دسته سوم نیز برای جلوگیری از مشکل فوق برازشی شبکه در هنگام ساخت مدل شامل اطلاعات ۸۱ روز بودند. پس از تقسیم‌بندی اطلاعات در سه دسته، شبکه عصبی Feed-Forward با یک لایه پنهان، به ازای تعداد نرون‌های متفاوت، برای ساخت بهترین مدل برآورد غلظت روزانه کربن‌مونوکسید مورد استفاده قرار گرفت. با توجه به معیارهای مورد بررسی R، MARE و RMSE ساختار شبکه شامل ۲۰ نرون درنقش معماری برتر شبکه برگزیده شد. همچنین TRAINLM درحکم تابع آموزش شبکه و TANSIG درحکم تابع انتقال مورد استفاده قرار گرفت. نتایج مراحل آموزش و آزمون شبکه در جدول ۶ و شکل‌های ۶ و ۷ آورده شده‌است.

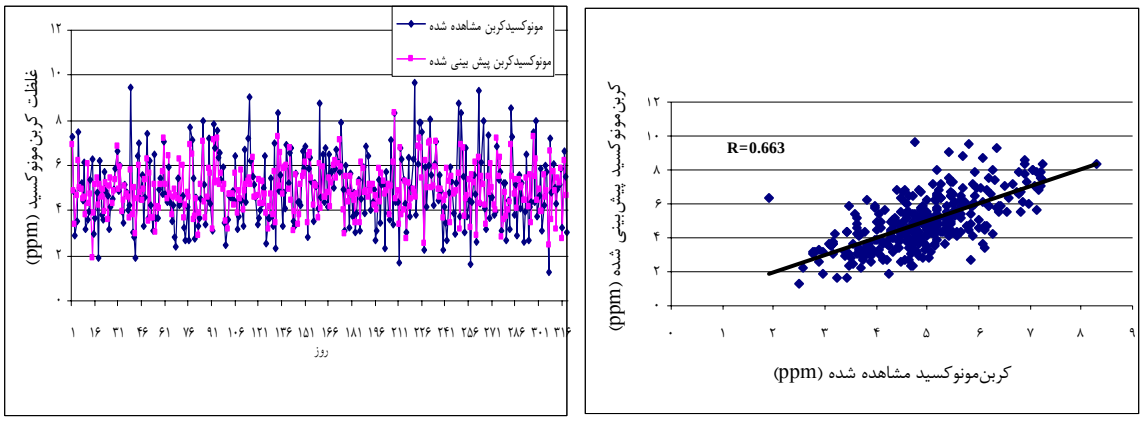
جدول ٦. نتایج مراحل آموزش و آزمون مدل شبکه عصبی.

معیار ارزیابی	مرحله آموزش	مرحله تست
R	0.805	0.716
MARE	0.164	0.158
RMSE	0.935	0.969

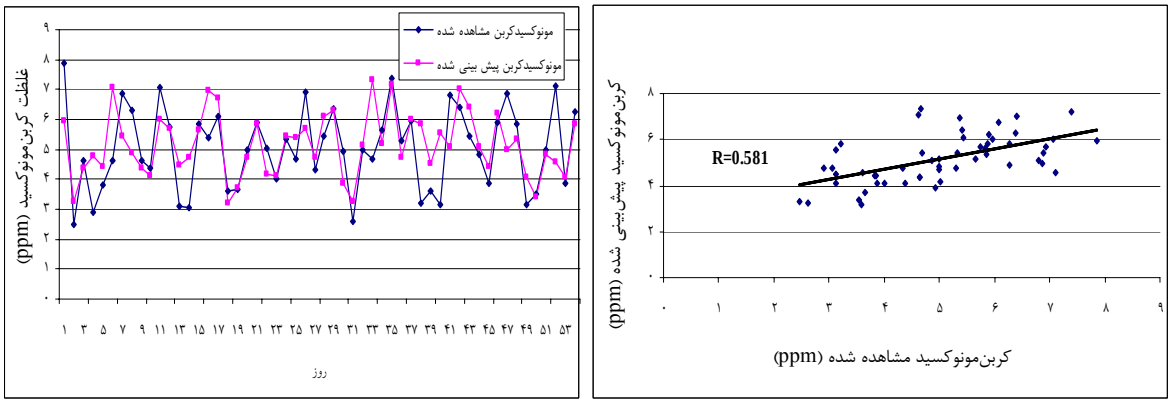
۳-۵ انتخاب مدل برتر

در این تحقیق برای گزینش مدل برتر از بین مدل ترکیبی رگرسیونی و PCA و مدل شبکه عصبی، از نمودار پراکندگی قدرمطلق مقادیر خطای نسبی استفاده شد. نتایج این تحلیل در شکل ۸ برای مرحله آموزش و در شکل ۹ برای مرحله آزمون شبکه آمده است. با توجه به این نمودار می‌توان تشخیص داد که خطای هر درصد از تعداد

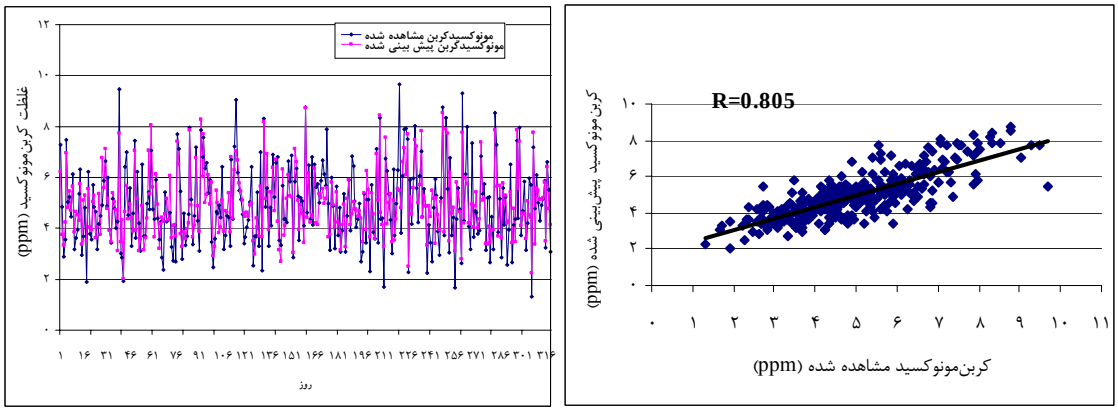
غلظت‌های پیش‌بینی شده کربن‌مونوکسید با دو مدل شبکه عصبی و رگرسیون، به چه میزان است. برای نمونه با توجه به شکل ۹ می‌توان گفت که قدرمطلق خطای نسبی برای آزمون مدل شبکه عصبی کمتر از ۵۳ درصد است، در حالی‌که این مقدار برای مدل رگرسیونی حدود ۸۳ درصد است. ۷۵ درصد پیش‌بینی‌های مدل شبکه عصبی درمرحله آزمون دارای خطای کمتر از ۲۰ درصدند، درحالی‌که در مدل رگرسیونی این خطا برابر ۲۵ درصد است. ۹۰ درصد از پیش‌بینی‌های مدل شبکه عصبی در مرحله آزمون دارای خطای کمتر از ۴۱ درصدند، درحالی‌که در مدل رگرسیونی این خطا برابر ۵۳ درصد است. همچنین نتایجی مشابه از شکل ۸ را می‌توان برای مراحل آموزش شبکه عصبی و مدل رگرسیونی به‌دست آورد. با مقایسه نتایج به‌دست آمده از مدل رگرسیونی و شبکه عصبی می‌توان دید که نتایج به‌دست آمده از شبکه عصبی بهتر از نتایج مدل رگرسیونی است. پس با توجه به این مطلب، مدل ایجاد شده با شبکه عصبی با یک لایه پنهان شامل ۲۰ نرون، به‌مثابه مدل برگزیده برای پیش‌بینی میانگین غلظت روزانه کربن‌مونوکسید در شهر تهران، انتخاب می‌شود.



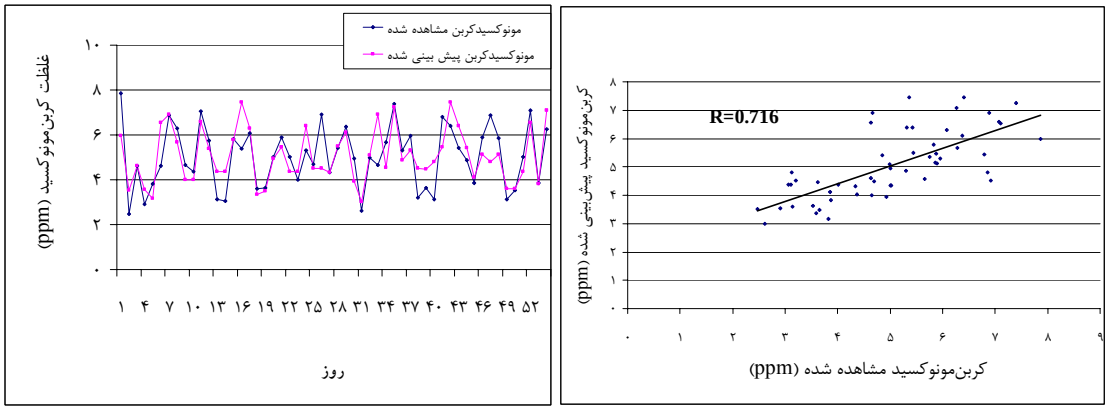
شکل ٤. نتایج مدل رگرسیونی در مرحله ساخت مدل.



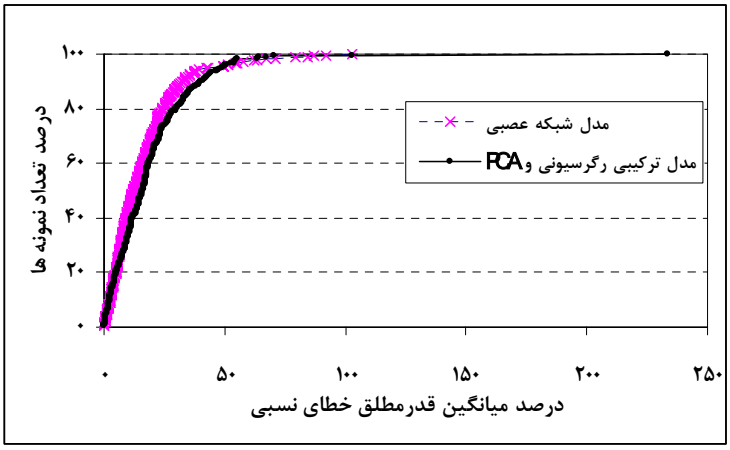
شکل ۵. نتایج مدل رگرسیونی در مرحله آزمون مدل.



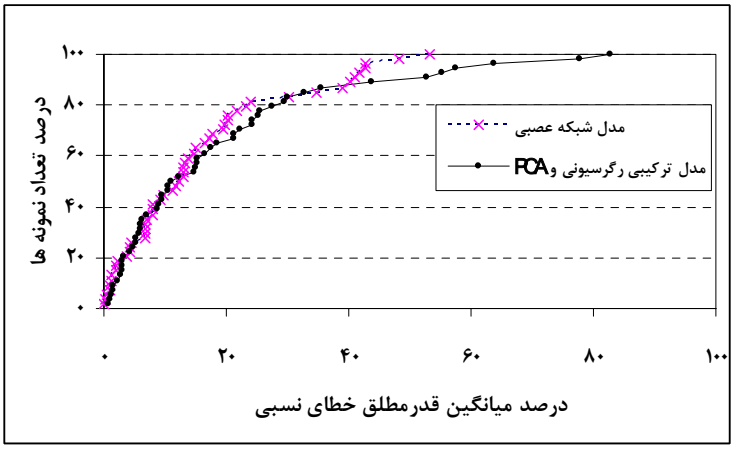
شکل ٦. نتایج شبکه عصبی در مرحله آموزش شبکه.



شکل ۷. نتایج شبکه عصبی در مرحله آزمون شبکه.



شکل ۸. نحوه پراکندگی مقادیر مطلق خطا نسبی در مرحله آموزش مدل‌ها.



شکل ۹. نحوه پراکندگی مقادیر مطلق خطا نسبی در مرحله آزمون مدل‌ها.

۴ نتیجه‌گیری

با توجه به این‌که بیش از سه‌چهارم وزن آلاینده‌های هوا در شهر تهران را کربن‌مونوکسید تشکیل می‌دهد، پیش‌بینی غلظت روزانه این آلاینده در ایجاد تصمیمات لازم برای مقابله با اثرات زیان‌بار آن در شهر تهران از اهمیت به‌سزایی برخوردار است، لذا هدف از این تحقیق ارائه مدل مناسبی برای برآورد غلظت روزانه این آلاینده بود. در این تحقیق اولین‌بار به فرض‌های ایجاد مدل رگرسیونی برای پیش‌بینی غلظت کربن‌مونوکسید توجه و مدل مناسب برای این منظور ساخته شد. نتایج به‌دست آمده از این تحقیق نشان می‌دهد که با توجه به سری زمانی غلظت روزانه کربن‌مونوکسید برای شهر تهران (شکل ۱) که دارای نوسانات زیادی است روش رگرسیون خطی چندمتغیره توانایی ارائه مدلی که بتواند این نوسانات را در نظر بگیرد، ندارد و در این موارد استفاده از شبکه عصبی که قادر به پیش‌بینی روابط غیرخطی و پیچیده بین ورودی‌ها و خروجی باشد راه‌حل مناسبی برای جایگزینی با رگرسیون خطی چندمتغیره است. علاوه بر آن در استفاده از مدل رگرسیونی خطی چندمتغیره فرض‌های زیادی وجود دارد که استفاده از آن را در مسائل عملی مشکل می‌سازد، به‌طوری‌که در بسیاری از تحقیقات صورت گرفته به دلیل اینکه، این فرض‌ها مباحث آماری پیچیده‌ای‌اند، از دید محققین مخفی مانده و به آنها توجهی نشده است. در نتیجه مدل ارائه شده آنها دارای دقت لازم نیست. در نهایت با توجه به مدت زمان اندک دوره آماری استفاده شده در این تحقیق (سال‌های ۱۳۸۳ و ۱۳۸۴)، توصیه می‌شود که بررسی‌های بیشتری در این زمینه با مدت زمان آماری طولانی‌تر و همچنین با استفاده از دیگر مدل‌ها نیز صورت پذیرد و نتایج آن با نتایج به‌دست آمده در این تحقیق مقایسه شود. این مهم نیز در دست اقدام است.

۵ تشکر و قدردانی

در پایان نویسندگان مقاله بر خود لازم می‌دانند از زحمات آقای دکتر حسین گنجی‌دوست، استاد گروه محیط‌زیست دانشگاه تربیت مدرس تهران، تشکر و قدردانی کنند.

۶ منابع

بختیاری، س.، ۱۳۸۵، اطلس کامل تهران. مؤسسه جغرافیایی و کارتوگرافی گیتاشناسی، تهران.

بیات، ر.، ۱۳۸۳، سهم‌بندی منابع تولید آلودگی هوای شهر تهران.، پایان‌نامه کارشناسی ارشد، مهندسی محیط‌زیست، دانشکده فنی، دانشگاه صنعتی شریف.

نوری، ر.، کراچیان، ر.، خدادادی، ا.، و شکیبایی نیا، ا.، ۱۳۸۶، ارزیابی اهمیت ایستگاه‌های پایش کیفی رودخانه‌ها با استفاده از آنالیزهای مؤلفه و فاکتور اصلی، مطالعه موردی: رودخانه کارون. مجله علمی- پژوهشی آب و فاضلاب، شماره ۶۳، صفحات ۶۰-۶۹.

Boznar, M., Lesjak, M., and Mlakar, P., 1993, A neural network based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmos. Environ., **27B**(2), 221-230.

Burden, F. R., Brereton, R. G., and Walsh, P. T., 1997, Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. Analyst, **122** (10), 1015-1022.

Camdevyren, H., Demyr, N., Kanik, A., and Keskyn, S., 2005, Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecol. Modell., **181**, 581-589.

Chelani, A. B., Chalapati Rao, C. V., Phadke, K. M., and Hasan, M. Z., 2002, Prediction of sulphur dioxide concentration using artificial neural networks, Environ. Modell. Softw., **17**, 161-168.

Corani, G., 2005, Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model., **185**, 513-529.

Coulibaly, P., Anctil, F., and Bobee, B., 2000, Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol., **230**, 244-257.

Cybenko, G., 1989, Approximation by superposition of a sigmoidal function. Math. Control Signal Syst., **2**, 303-314.

Davis, J. C., 1986, Statistical and data analysis in geology, second ed. John Wiley & Sons. New York. USA.

Dorzdowicz, B., Benz, S. J., Sonta, A. S. M., and Scenna, N. J., 1997, A neural network based model for the analysis of carbon monoxide concentration in the urban area of Rosario. In: Power, H., Tirabassis, T., Brebbia, C. A. (Eds.), Air Pollution, vol. V. Computational Mechanics Inc., Southampton, Boston, pp. 677-685.

Finzi, G., Volta, M., Nucifora, A., and Nunnari, G., 1998, Real-time ozone episode forecast: a comparison between neural network and grey-box models. In: Proceedings of the International ICSC/ IFAC Symposium on Neural Computation-NC'98, Vienna.

Gardner, M. W., and Dorling, S. R., 1998, Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences, Atmos. Environ., **32** (14/15), 2627-2636.

Gardner, M. W., and Dorling, S. R., 1999, Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos. Environ., **33**(5), 709-719.

Gilbert, R. O., 1987, Statistical methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York. USA.

Helena, B., Pardo, R., Vega, M., Barrado, E., Ferna´ ndez, J. M., and Fernandez, L., 2000, Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. Water Res., **34**, 807-816.

Hocking, R. R., 2003, Methods and application of linear models regression and analysis of variance, Wiley, Newjersey. USA.

Hornik, K., Stinchcombe, M., and White, H., 1989, Multilayer feedforward networks are universal approximators. Neural Networks, **2**(5), 359-366.

Johnson, R. A., and Wichern, D. W., 1982, Applied multivariate statistical analysis. Prentice-Hall Inc., Englewood Cliffs, SA, 590 pp.

http://www.sci.org.ir/portal/faces/public/census85/census85.natayej/census85.agetownship.

Legates, D. R., and McCabe, G. J., 1999, Evaluating the use of "Goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resour. Res., **35**, 233-241.

Lu, W. Z., Wang, W. J., Wang, X. K., Xu, Z. B., and Leung, A. Y. T., 2003, Using improved neural network to analyze RSP, NOX and NO2 levels in urban air in Mong Kok, Hong Kong. Environ. Monit. Assess., **87**, 235-254.

Manly, B. F. J., 1986, Multivariate Statistical Methods: A Primer. Chapman & Hall, London. UK.

McCulloch, W. S., and Pitts, W., 1943, A logical calculus of the ideas imminent in nervous

activity. B. Math. Biophys., **8**, 115-133.

Moseholm, L., Silva, J., and Larson, T. C., 1996, Forecasting carbon monoxide concentration near a sheltered intersections using video traffic surveillance and neural networks. Transport. Res., **D1**, 15-28.

Nagendra, S. M. S., and Khare, M., 2004, Artificial neural network based line source models for vehicular exhaust emission predictions of an urban roadway. J. Transport Environ, **9**(3), 199-208.

Nunnari, G., Bertucco, L., and Milio, D., 2001, Predicting daily average SO2 concentrations in the industrial area of Syracuse (Italy). In: Proceedings of ICANNGA 2001 Fifth International Conference on Artificial Neural Networks and Genetic Algorithm, Prague, Czech Republic, 22-25 April.

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R., and Chatterton, T., 2004, Modelling SO2 concentration at a point with statistical approaches, Environ. Modell. Softw., **19**, 887-905.

Nunnari, G., Nucifora, A., and Randieri, C., 1998, The application of neural techniques to the modelling of time series of atmospheric pollution data. Ecol. Modell., **111**, 187-205.

Ouyang, Y., 2005, Application of principal component and factor analysis to evaluate surface water quality monitoring network. Water Res., **39**, 2621-2635.

Sahin, U., Ucan, O. N., Bayat, C., and Oztorun, N., 2005, Modeling of SO2 distribution in Istanbul using artificial neural networks. Environ. Modell Assess., **10**, 135-142.

Shi, J. P., and Harrison, R. M., 1997, Regression modelling of hourly NOx and NO2 concentration in urban air in London. Atmos. Environ., **31** (24), 4081-4094.

Shrestha, S., and Kazama, F., 2007, Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environ. Modell & Softw., **22**, 464-475.

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M., and Kouimtzis, Th., 2003, Assessment of the surface water qualityin Northern Greece. Water Res., **37**, 4119-4124.

Singh, K. P., Malik, A., Mohan, D., and Sinha, S., 2004, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water qualityof Gomti River (India)-a case study. Water Res., **38**, 3980-3992.

Stone, M., 1974, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society B **36**, 111-147.

Tabachnick, B. G., and Fidell, L. S., 2001, Using Multivariate Statistics. Allyn and Bacon, Boston, London. UK.

Vega, M., Pardo, R., Barrado, E., and Deban, L., 1998, Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res., **32**, 3581-3592.

Wackernagel, H., 1995, Multivariate Geostatistics. An Introduction With Applications. Springer, New York and London. UK.

Zannetti, P., 1990, Air Pollution Modelling, Theories, Computational Methods and Available Software. Van Nostrand Reinhold, New York. USA.

Zhang, G., Patuwo, B. E., and Hu, M. Y., 1998, Forecasting with artificial neural networks: the state of the art. Int. J. Forecasting, **14**, 35-62.