Atmospheric Pollution Research 7 (2016) 412–418

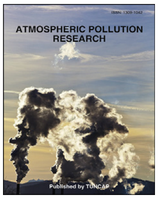
HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: http://www.journals.elsevier.com/locate/apr



Original article

# Reliable prediction of carbon monoxide using developed support vector machine

Saber Moazami [a], Roohollah Noori [b, *], Bahman Jabbarian Amiri [c], Bijan Yeganeh [d], Sadegh Partani [e], Salman Safavi [f]

[a] Department of Civil Engineering, College of Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr, Tehran, Iran
[b] Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Tehran, Iran
[c] Department of Fisheries and Environmental Sciences, College of Natural Resources, University of Tehran, Tehran, Iran
[d] International Laboratory for Air Quality and Health, Queensland University of Technology, Brisbane, Australia
[e] Department of Civil Engineering, Islamic Azad University-Central Tehran Branch (IAUCTB), Tehran, Iran
[f] Department of Civil Engineering, Faculty of Engineering, Malard Branch, Islamic Azad University, Malard, Iran

ARTICLE INFO

*Article history:*
Received 3 August 2015
Received in revised form
22 October 2015
Accepted 22 October 2015
Available online 21 November 2015

*Keywords:*
Carbon monoxide
Support vector machine
Uncertainty
Artificial intelligence

ABSTRACT

Air pollution modeling is always along with uncertainties which results in improper decision making and affects the health of the people exposed to the pollution. Therefore, the determination of model uncertainty can improve air pollution control strategies especially in critical conditions. This study aims to develop an appropriate methodology for determination of uncertainty in support vector regression (SVR) as a well-known modeling approach in atmospheric science. The methodology is based on running SVR model many times using different calibration datasets. The robustness of the proposed methodology was checked to predict the next day carbon monoxide (CO) concentrations in Tehran metropolitan. Thereafter, a comparison was carried out between the results of the present study and another research on uncertainty determination of adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN). Generally, the results showed that the SVR had less uncertainty in CO prediction than the ANN and ANFIS models. Moreover, repetition of SVR runs with different calibration datasets resulted in different SVR responses. Different SVR responses provided the required information to determine the band of uncertainty for predictions, using specific lower and upper percentiles. Besides, it is found that more than 75% and 78% of SVR predictions are located in the band of uncertainty determined by 2.5th–97.5th and 0.5th–99th percentiles, respectively.

## 1. Introduction

In the past decades, several approaches including regression based methods (Burrows et al., 1995; Sousa et al., 2007, 2009), deterministic models (Papakonstantinou et al., 2003; Duci et al., 2004; Kesarkar et al., 2007; Markakis et al., 2015) and artificial intelligence (AI) techniques (Nagendra and Khare, 2006; Grivas and Chaloulakou, 2006; Nagendra and Khare, 2008; Pai et al., 2011, 2013; Singh et al., 2012; Elangasinghe et al., 2014; Pai and Chen, 2015; Russo et al., 2015; Shahraiyni et al., 2015) were proved to be satisfactorily applied for air pollution forecasting. One of the important weaknesses of these approaches is high uncertainties, related to the calculations, inputs and complicated inherent of turbulence processes, in the atmosphere. This may have critical flaws in air pollution studies since the uncertainties associated with this type of predictions may result in improper decision making and affects the health of the people exposed to the pollution (Fisher and Ireland, 2001; Vardoulakis et al., 2002). Therefore, analyzing uncertainty is a significant aspect of air pollution modeling. In this regards, the present study aims to determine the uncertainty in support vector regression (SVR) models used for carbon monoxide (CO) prediction in the atmosphere of Tehran, Iran.

AI models are structurally and theoretically different from numerical ones. However, there are more differences among AI techniques compared to the differences among the numerical ones.

* Corresponding author. Tel.: +98 9128352488.
*E-mail address:* noor@ut.ac.ir (R. Noori).
Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

http://dx.doi.org/10.1016/j.apr.2015.10.022
1309-1042/Copyright © 2015 Turkish National Committee for Air Pollution Research and Control. Production and hosting by Elsevier B.V. All rights reserved.

All numerical models apply specific formulations i.e. advection–dispersion equation, along with a specific solving approach depending on the applied model, while there are important structural differences among the AI models. For example, adaptive neuro-fuzzy inference system (ANFIS) model is based on fuzzy theory, while artificial neural network (ANN) model is based on Boolean logic. Besides, numerical models are formulated in one, two or three dimensions applying specified advection–dispersion equation (physically based models), while AI models are practically black boxes (known as data-driven models) and lack underlying mathematical theory. Therefore, tuning the AI models is completely based on the data sampling patterns for calibration and verification purposes, and finding a logical relationship between independent and dependent variables through repetitive trial and error processes. Thus, AI models are more sensitive to input variables than numerical ones. As a result, AI models are subjected to more input-related uncertainty than numerical ones.

According to the above-mentioned facts, formulizing a structure to determine uncertainty of AI techniques may be more complicated than doing so for numerical models. Therefore, although there are numerous works to determine the uncertainty of numerical air quality models (Fox, 1984; Bergin et al., 1999; Dabberdt and Miller, 2000; Manomaiphiboon and Russell, 2004; Lumbreras et al., 2009; Rosa et al., 2011; Zhao et al., 2013), few works have been focused on determining the uncertainty of AI techniques. More specifically, in air pollution studies, although there are a lot of researches applying AI models, lack of those studies in uncertainty determination is still felt. So far, the uncertainty of both ANN and ANFIS models has only been considered for air pollutants estimations (Noori et al., 2010). However, despite the fact that in recent years SVR model has had an extensive application in air pollution studies (Lu and Wang, 2005; Singh et al., 2013; Noori et al., 2013b), the uncertainty of this model has never been developed in this field.

Hence, this paper presents a pioneering effort in uncertainty analysis of AI models, studying the SVR model. In other words, as there is no proper reference on determining the uncertainty of SVR, this work can be considered as a novel study. Moreover, the method proposed by this research can be applied as a pattern to determine the uncertainty of SVR in air pollution researches in future.

## 2. Case study area and data

Tehran, capital of Iran, is surrounded by mountains to the north, west and east. The results of previous studies about air pollution of Tehran demonstrate that 90% by weight of total air pollutants are generated from traffic and only 10% from other sources (Bayat, 2005). In comparison with other air pollutants in Tehran, CO and particulate matter equal to or less than 10 μm in diameter ($PM_{10}$) are more than the others. In this study, the pollution data {$PM_{10}$, total hydrocarbons (THC), nitrogen oxides ($NO_x$), methane ($CH_4$), sulfur dioxide ($SO_2$) and ozone ($O_3$)} and meteorological variables {pressure (Press), temperature (Temp), wind direction (WD), wind speed (WS) and relative humidity (Hum)} from Gholhak station in north of Tehran was selected to be applied in predicting the daily CO concentration. Statistical characteristics of the used data are available in Noori et al. (2010).

## 3. Methodology

Here, only a brief description is presented, since the SVR theory has been described in detail, in numerous works (e.g., Vapnik, 1998; Abe, 2005; Lu and Wang, 2005). SVR is a supervised learning method, which estimates the dependent variable **y** on a set of independent variables **x** applying deterministic function $\mathbf{y} = \{\mathbf{w}^T \cdot \phi(\mathbf{x}) + b\} +$ noise. In SVR model the noise term is represented by error tolerance ($\varepsilon$), and **w** (vector of coefficients) and $b$ (constant) are the regression function parameters. Besides, $\phi$ is the kernel function to transform input data to a high-dimensional feature space, in which the input data become more separable, compared to the original input space. The task is then to find a functional form for $\{\mathbf{w}^T \cdot \phi(\mathbf{x}) + b\}$. This can be achieved by tuning the SVR model on a sample set, *i.e.*, calibration data. Then, **w** and $b$ are derived by minimizing the error function (Eq. (1)) subject to Eq. (2) (Goel and Pal, 2009; Pal et al., 2011):

$$\frac{1}{2}\mathbf{w}^T \cdot \mathbf{w} + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi_i^{\cdot} \tag{1}$$

$$\begin{aligned} &\mathbf{w}^T \cdot \phi(\mathbf{x}_i) + b - \mathbf{y}_i \le \varepsilon + \xi_i^{\cdot} \\ &\mathbf{y}_i - \mathbf{w}^T \cdot \phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \\ &\xi_i, \xi_i^{\cdot} \ge 0, \quad i = 1, \ldots, N \end{aligned} \tag{2}$$

where $C$ indicates a positive constant that determines the degree of penalized loss when a calibration error occurs, $N$ is the sample size, and $\xi_i$ and $\xi_i^{\cdot}$ are slack variables specifying the upper and lower calibration error subject to $\varepsilon$ (Pal and Goel, 2007). It is noted that all SVR codes were run in MATLAB software environment.

According to the above context regarding SVR model, the error function will be optimized with respect to the model's input data; in other words, the error function is directly related to the model's input data. Therefore, it is necessary to evaluate the model performance through different calibration patterns in order to assess the model uncertainty caused by changes in the input data. The detailed procedure of the uncertainty determination of the SVR model is shown in Fig. 1.

According to Fig. 1, the calibration pattern should be generated, and outputs should consequently be computed by the model. Moreover, this process should be repeated for many times. In this study, SVR and forward selection-SVR (FS-SVR) models were calibrated by a percentage of data as an alternative choice, and data sampling process for calibration of these two models were repeated in appropriate times (here, 1000 times). As a result, tuning the parameters of SVR and FS-SVR models were determined 1000 times for each model. Therefore, a range of outputs related to the uncertainty in the 1000 calibrated SVR and FS-SVR models would be determined by applying this massive computational technique. To evaluate the uncertainty of SVR and FS-SVR models, the percentage of measured data bracketed by $R$ percent predicted uncertainties ($R$PPU) can be applied for both calibration and verification steps of the models. The $R$PPU is computed by determining $\{0.5 \times (100 - R)\%\}$ $(X_L)$ and $\{R + 0.5 \times (100 - R)\%\}$ $(X_U)$ of normal distribution function obtained from 1000 times forecasting process as follows.

$$\text{Bracketed by } R\text{PPU} = \frac{1}{k}\text{Count}(k|X_L \le k \le X_U) \times 100 \tag{3}$$

where $k$ is the number of dataset in calibration or verification steps of the models. Moreover, $d$-factor parameter proposed by Abbaspour et al. (2007) can be applied to evaluate the average width of confidence interval band, as Eq. (4).

$$d - \text{factor} = \frac{\overline{\text{d}}_X}{\sigma_X} \tag{4}$$

In Eq. (4),$\sigma_X$ represents the standard deviation of the measured variable $X$ and $\overline{\text{d}}_X$ is the average distance between the upper and the lower band determined from Eq. (5) (Abbaspour et al., 2007; Aqil et al., 2007; Noori et al., 2013a):