Atmospheric Pollution Research xxx (2016) 1–9

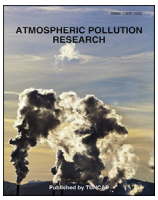
HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Atmospheric Pollution Research

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



Prediction of airborne nanoparticles at roadside location using a feed–forward artificial neural network

Abdullah N. Al–Dabbous [a, *], Prashant Kumar [b, c], Abdul Rehman Khan [d]

[a] Crises Decision Support Program, Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, P.O. Box 24885, 13109, Safat, Kuwait
[b] Department of Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences (FEPS), University of Surrey, Guildford, GU2 7XH, United Kingdom
[c] Environmental Flow (EnFlo) Research Centre, FEPS, University of Surrey, Guildford, GU2 7XH, United Kingdom
[d] Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, P.O. Box 24885, 13109, Safat, Kuwait

ARTICLE INFO

Article history:
Received 3 October 2016
Received in revised form
21 November 2016
Accepted 22 November 2016
Available online xxx

Keywords:
Artificial neural network
Prediction
Nanoparticles
Comparative evaluation
Multi–layer feed–forward network
Back–propagation training algorithm

ABSTRACT

Accurate prediction of nanoparticles is essential to provide adequate mitigation strategies for air quality management. On the contrary to PM_{10}, SO_2, O_3, NO_x and CO, nanoparticles are not routinely–monitored by environmental agencies as they are not regulated yet. Therefore, a prognostic supervised machine learning technique, namely feed–forward artificial neural network (ANN), has been used with a back–propagation algorithm, to stochastically predict PNCs in three size ranges ($N_{5–30}$, $N_{30–100}$ and $N_{100–300}$ nm). Seven models, covering a total of 525 simulations, were considered using different combinations of the routinely–measured meteorological and five pollutants variables as covariates. Each model included different numbers of hidden layers and neurons per layer in each simulation. Results of simulations were evaluated to achieve the optimum correspondence between the measured and predicted PNCs in each model (namely Models, M_1–M_7). The best prediction ability was provided by M_1 when all the covariate variables were used. The model, M_2, provided the lowest prediction performance since all the meteorological variables were omitted in this model. Models, M_3–M_7, that omitted one pollutant covariate, showed prediction ability similar to M_1. The results were within a factor of 2 from the measured values, and provided adequate solutions to PNCs' prognostic demands. These models are useful, particularly for the studied site where no nanoparticles measurement equipment exist, for determining the levels of particles in various size ranges. The model could be further used for other locations in Kuwait and elsewhere after adequate long–term measurements and training based on the routinely–monitored environmental data.

1. Introduction

Regulatory bodies worldwide have not reached a consensus regarding a legal threshold to control particle number concentrations (PNCs) in the ambient air (Kumar et al., 2014). The air quality standard for particles is based on mass concentrations of particulate matter less than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$). However, these standards do not regulate PNCs due to their negligible mass compared with PM_{10} and $PM_{2.5}$ (Heal et al., 2012). Consequently, most air quality monitors do not have nanoparticles monitors distributed in monitoring networks (Kumar et al., 2011b), which are highly expensive and very perceptive. Therefore, any relavent information about PNCs will be expedient using any consistent predictive model as a function of most commonly monitored pollutants in the ambient air.

The modelling of air pollutants usually fits into two modelling approaches: deterministic (i.e., dispersion models) and stochastic (i.e., statistical models) models that can be used in accurate modelling purposes (Mølgaard et al., 2012; Reggente et al., 2014). Artificial intelligence (AI), which was initially introduced by Robbins and Monro (1951), has wide applications in stochastic

* Corresponding author.
E-mail address: adabbous@kisr.edu.kw (A.N. Al–Dabbous).
Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

http://dx.doi.org/10.1016/j.apr.2016.11.004

Please cite this article in press as: Al–Dabbous, A.N., et al., Prediction of airborne nanoparticles at roadside location using a feed–forward artificial neural network, Atmospheric Pollution Research (2016), http://dx.doi.org/10.1016/j.apr.2016.11.004

prediction models, where no equations are required to describe the physical processes in the model. The most commonly−used AI application in prediction is the artificial neural network (ANN). In these models, a set of training data is used to derive a statistical description (i.e., automatically developed by ANN) of the relation between inputs (covariates) and outputs (targets) that can make predictions of the output data from unseen (i.e., new) input data within the bounds of the training covariates range. This statistical description is considered as a black box, where unknown simultaneous computational process is applied to map the relation between covariates and targets, which is one of the drawbacks of the ANN approach.

The targetted concentrations (i.e., PNCs) collected at a receptor site are mainly from various known and unknown sources making varying contributions. Deterministic models require the knowledge of sources and concentrations of nanoparticles, and the associated transformation and dispersion processes (i.e., Gaussian and Eleurian models) that are yet not fully understood (Kumar et al., 2011a). Therefore, stochastic models are preferred to overcome the limitations of the deterministic models.

Unlike linear multivariate statistical methods (e.g., ordinary least squares method, and partial least squares), ANN is able to model complex non−linear relationships between given parameters, without any assistance from the user, and to easily deal with high−dimensional data (Svozil et al., 1997). ANN showed remarkable performance and accuracy in capturing the complex non−linear associations within data, compared to traditional statistical models. For example, Chelani et al. (2002) also showed superior prediction ability of ANN (R^2 = 0.68, 0.72 and 0.63 for industrial, commercial and residential sites, respectively) against the multivariate regression models (R^2 = 0.57, 0.52 and 0.48 for industrial, commercial and residential sites, respectively) for SO_2 daily concentrations at three sites in Delhi, India. Furthermore, Kukkonen et al. (2003) demonstrated that ANN (R^2 = 0.71) outmatched the linear statistical model (R^2 = 0.47) and the deterministic modelling system (R^2 = 0.32), when predicting NO_2 hourly concentrations at two monitoring stations in central Helsinki, Finland, from 1996 to 1999. Likewise, ANN has been shown to perform better (R^2 = 0.65) than multi−linear regression method (R^2 = 0.60) for predicting PM_{10} daily concentrations in Athens, Greece (Chaloulakou et al., 2003). In Athens again, ANN displayed better predictions for hourly PM_{10} concentrations than linear regression models at four urban and suburban locations since the R^2 for ANN were in the 0.80−0.89 range compared with 0.29−0.35 for linear regression models (Grivas and Chaloulakou, 2006). Furthermore, Paschalidou et al. (2011) showed better predictions by ANN (R^2 = 0.65−0.76) than those given by principal component regression analysis (R^2 = 0.33−0.38) for hourly PM_{10} concentrations in four urban locations in Cyprus. In Kocaeli (Turkey), Özdemir and Taner (2014) reported that predictions of PM_{10} hourly concentrations by ANN (R^2 = 0.87 and 0.49 for urban and industrial sites, respectively) outperformed multi−linear regression (R^2 = 0.74 and 0.36 for urban and industrial sites, respectively), highlighting the more efficient predictions by ANN. Among the aforementioned studies, ANN has shown the highest predictive accuracy in term of their appealing adaptive nature and ability of modelling complex non-linear high-dimensional data, and is thereby considered a better predictive modelling tool.

Many fields have utilised ANN successfully. Some of them include air pollution (Moustris et al., 2010), waste management (Antanasijević et al., 2013), medicine (Lo et al., 2013), ecology (Larsen et al., 2012) and chemistry (Svozil et al., 1997). In terms of air pollution, ANN has predicted successfully the concentrations of PM_{10} (Paschalidou et al., 2011), SO_2 (Moustris et al., 2010), O_3 (Kandya et al., 2013), NO_2 (Nagendra and Khare, 2006), NO_x (Perez and Trier, 2001), CO (Moustris et al., 2010) and H_2S (Baawain and Al-Serihi, 2014), but application of this approach to the PNC predictions remain very limited (Table 1).

The current analysis presented as a part of this work differs from previous PNC studies (Table 1) in the following unique ways. Firstly, the PNC measurements were recorded, at a sampling rate of 10 Hz with a time response ($T_{90-10\%}$) as low as 200 ms, using one of the fastest available aerosol mobility size spectrometers, i.e., differential mobility spectrometer, DMS500 (Kumar et al., 2010). This has made a high resolution data available for the training and performance evaluation of our ANN models. Secondly, the sampling site is a representation of an urban location of an industrialised country of Arabian peninsula where petroleum and petrochemical products are the main source of revenue (Al-Dabbous and Kumar, 2015). Therefore, the model developed as a part of this work has a broad applicability after adequate training. Thirdly, unlike previous modelling efforts (Table 1), this is the first instance concerning the application of ANN for prediction of PNCs in the middle−east region. The novelty of the present ANN model is to relate simultaneously 3 targets, i.e., 5−30 nm (N_{5-30}; nucleation mode), 30−100 nm (N_{30-100}; Aitken mode), 100−300 nm ($N_{100-300}$; accumulation mode), to 5−7 (i.e., meteorological and pollutant) covariates in the best possible manner at a time resolution of 5 min, utilizing 525 simulations for evaluation. Lastly, other than the previous standard statistical modelling work of Reggente et al. (2014) and Sabaliauskas et al. (2012), most of studies have collected their data from an urban background locations (Table 1). Moreover, the work of Reggente et al. (2014) used a lower cut−off diameter of 25 nm; these are nucleation mode volatile particles and are cause of many uncertainties in nanoparticle models (Kumar et al., 2011a). Likewise, Sabaliauskas et al. (2012) considered a daily temporal resolution in their data analysis from urban locations. In this work, we use a lower cut−off diameter of 5 nm and an averaging time of 5 min, allowing to predict the nucleation mode particles that could contribute up to 77% of the total PNCs (Al-Dabbous and Kumar, 2014b; Kumar et al., 2009) and capture variability brought by nucleation mode particles to the total ambient PNCs, respectively.

In order to fill the above−noted research gaps, a supervised machine learning technique, namely multi−layer ANN, was applied to predict PNCs in three size ranges, i.e., N_{5-30}, N_{30-100} and $N_{100-300}$, using different combinations of seven routinely−measured meteorological (wind speed and temperature) and pollutant (PM_{10}, SO_2, O_3, NO_x and CO) variables as covariates.

2. Materials and methods

2.1. Multi−layer ANN

A multi−layer feed−forward ANN, trained with a supervised back−propagation training algorithm, is developed for the prediction of particles in three different sizes (Section 2.2). The general architecture of the network consists of input layer, hidden layers and output layer, as shown in Fig. 1. A single hidden layer is generally used in ANN prediction purposes (Hornik et al., 1989), however this practice is debated as more complex problems sometimes required more than one hidden layer (Chaloulakou et al., 2003). Therefore, networks with single, two and three hidden layers were assessed to choose optimum number of hidden layers than could yield acceptable prediction results. In these hidden layers, different numbers of hidden neurons were evaluated (described in Section 2.3). These layers are interconnected through a system of neurons by weights and output signals, which are originated from the neurons in input layer and fed forward towards the neurons in the following layer. The number of hidden neurons

Please cite this article in press as: Al−Dabbous, A.N., et al., Prediction of airborne nanoparticles at roadside location using a feed−forward artificial neural network, Atmospheric Pollution Research (2016), http://dx.doi.org/10.1016/j.apr.2016.11.004