Atmospheric Pollution Research 6 (2015) 540-549

Atmospheric Pollution Research

www.atmospolres.com

Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales

Ana Russo [1], Pedro G. Lind [2,3], Frank Raischel [1], Ricardo Trigo [1], Manuel Mendes [4]

[1] *Instituto Dom Luiz, Faculdade de Ciências da Universidade de Lisboa, Campo Grande Edifício C8, Piso 3, 1749–016 Lisboa, Portugal*
[2] *TWIST – Turbulence, Wind energy and Stochastics, Institute of Physics, Carl–von–Ossietzky University of Oldenburg, DE–26111 Oldenburg, Germany*
[3] *ForWind – Center for Wind Energy Research, Institute of Physics, Carl–von–Ossietzky University of Oldenburg, DE–26111 Oldenburg, Germany*
[4] *Instituto Português do Mar e da Atmosfera, Rua C–Aeroporto,1749–077 Lisbon, Portugal*

ABSTRACT

We present a simple neural network and data pre–selection framework, discriminating the most essential input data for accurately forecasting the concentrations of PM_{10}, based on observations for the years between 2002 and 2006 in the metropolitan region of Lisbon, Portugal. Starting from a broad panoply of different data sets collected at several air quality and meteorological stations, a forward stepwise regression procedure is applied enabling to automatically identify the most important variables for predicting the pollutant and also to rank them in order of importance. The importance of this variable ranking is discussed, showing that it is very sensitive to the urban location where measurements are obtained. Additionally, the importance of Circulation Weather Types is highlighted, characterizing synoptic scale circulation patterns and the concentration of pollutants. We then quantify the performance of linear and non–linear neural network models when applied to PM_{10} concentrations. In the light of contradictory results of previous studies, our results show no clear superiority for the case studied of non–linear models over linear models. While all models show similar predictive performances, we find important differences in false alarm rates and demonstrate the importance of removing weekly cycles from input variables.

Keywords: *Pollution, PM_{10}, forward stepwise regression, circulation weather types, neural networks*

Corresponding Author:
Ana Russo
☎ : *+351-217-500-818*
🖷 : *+351-217-500-807*
✉ : *acrusso@fc.ul.pt*

Article History:
Received: 30 September 2014
Revised: 15 December 2014
Accepted: 15 December 2014

doi: 10.5094/APR.2015.060

1. Introduction

Air pollution is a global threat to public health and to the environment, particularly in urban areas (Kolehmainen et al., 2001; EEA, 2013). Urban air pollution is a complex mixture of toxic components, which may induce acute and chronic responses from sensitive groups, (Kolehmainen et al., 2001; Wong et al., 2002; Diaz et al., 2004). Therefore, forecasting air pollution concentrations in urban locations emerges as a priority for guaranteeing life and environmental quality (Kolehmainen et al., 2001; EEA, 2013).

Modeling air pollution allows describing the causal relationship between emissions, meteorology, atmospheric concentrations, deposition, and other factors, including the determination of the effectiveness of remediation strategies, and the simulation of future scenarios. Despite of the above mentioned advantages of pollution modeling, the choice for a certain modeling approach should be done with some parsimony. Particularly, the time–lag in which air pollution prediction is performed should allow effective alert procedures in urban centers.

Different methodologies have been applied to characterize and forecast the dispersion of air pollutants, from the most simple approaches, such as box models (Middleton, 1998), or persistence and regression models (Shi and Harrison, 1997), to the most complex dynamical model systems, such as CHIMERE (Monteiro et al., 2005), or the CMAQ–Community Multiscale Air Quality Model (Luecken et al., 2006; Arasa et al., 2010).

Simpler models are often used as they can provide a fast overview. However, they rely on significant simplifying assumptions and usually do not describe the complex processes and interactions that control the transport and chemical behavior of pollutants in the atmosphere (Luecken et al., 2006).

In the last decades, significant progress has been made in air–quality dispersion models (Arasa et al., 2010). However, being highly non–linear, they require large amounts of accurate input data and are computationally expensive (Dutot et al., 2007; Elangasinghe et al., 2014).

Statistical models, such as Artificial Neural Networks (NN), have been shown to constitute a promising alternative to deterministic models (Yi and Prybutok, 1996; Cobourn et al., 2000; Gardner and Dorling, 2000a; Hooyberghs et al., 2005; Dutot et al., 2007; Papanastasiou et al., 2007; Lal and Tripathy, 2012; Nejadkoorki and Baroutian, 2012; Elangasinghe et al., 2014). These models are often regarded as a good compromise between simplicity and effectiveness, being capable of modeling the effect of non–linearities and fluctuations.

Although NN models may involve greater uncertainty, the input data requirements are less strict. Several NN models have been tested comparing the potential of different approaches when applied to different pollutants and prediction time lags (Yi and Prybutok, 1996; Gardner and Dorling, 2000a; Kukkonen et al., 2003; Hooyberghs et al., 2005). Other authors have proven better

forecasting results of NN over multiple linear regression (MLR) (Kukkonen et al., 2003; Agirre–Basurko et al., 2006). More recently, Russo et al. (2013) showed that, combining NN models and stochastic data analysis, allows diminishing the requirement of large training data sets often appearing when constructing a NN model.

Despite these improvements, forecasting NN models still present some caveats that need to be properly addressed (Lal and Tripathy, 2012). The construction of the best NN structure and the choice of input parameters constitutes a challenge (Chaloulakou et al., 2003; Hooyberghs et al., 2005; Perez and Reyes, 2006; Lal and Tripathy, 2012), as any set of input data can be fed into any NN architecture for training and evaluation, but not all possible combinations can be realistically tested. Comrie (1997) and Cobourn et al. (2000) have performed comparison studies between NN and regression models to forecast ozone concentrations, both showing that NN outcomes are only equal or slightly better than regression. In contrast, Gardner and Dorling (2000b) showed that there is a significant increase in performance when using non–linear models. For PM_{10}, the results are different to some extent, and it is possible to find in the literature different applications where NN can perform well, depending on input parameters (Chaloulakou et al., 2003; Perez and Reyes, 2006; Nejadkoorki and Baroutian, 2012). Comparison statistics between linear and nonlinear models presented by Chaloulakou et al. (2003) and Perez and Reyes (2006) indicate that the NN approach has an edge over linear models, expressed both in terms of prediction error and of episodic prediction ability, demonstrating that NN models, if properly trained and formed, can provide adequate solutions to particulate pollution prognostic demands. Thus, a good choice of input variables appears to be very important (Chaloulakou et al., 2003; Perez and Reyes, 2006; Hooyberghs et al., 2005) and should be performed with parsimony. Even though several studies revealed that certain weather parameters are relevant to model air pollutant concentrations (e.g. temperature, wind speed and direction, humidity) (Hooyberghs et al., 2005; Demuzere et al., 2009), the majority of the research focused on individual meteorological variables and non–automated procedures of variables' selection. Moreover, several studies have been published establishing important links between synoptic scale circulation patterns, usually named Circulation Weather Types (CWT), and air pollution (Dayan and Levy, 2002; Demuzere et al., 2009; Saavedra et al., 2012; Russo et al., 2014), relating a particular air mass to dispersion conditions and also to the mesoscale and local meteorological behavior (Dayan and Levy, 2005). Nevertheless, to the best of our knowledge, there are no studies in the literature focusing on the application over the Iberian Peninsula of objective automatic classification procedures of CWT as a predictor for air quality modeling.

In this paper, we address the issues previously mentioned, (1) aiming at developing daily forecast through the application of a circulation–to–environment approach based on the analysis of links between meteorological parameters, CWT and daily air quality measurements, and (2) introducing a simple framework for automatically ranking the set of variables used as input variables for training the NN model. To systematically develop a better air quality model, we apply both linear and non–linear NN models to predict PM_{10} daily average concentrations within the greater urban area of Lisbon, Portugal, based on historical air pollution and weather information. We choose to address only PM_{10}, that corresponds to inhalable particulate matter sized 10 µm or less, as this pollutant poses a major health risk (Stedman et al., 2002). Although pollutants' emissions in Europe have decreased over the last two decades, this did not lead to a corresponding reduction of concentrations of PM_{10} throughout Europe (EEA, 2011). Evidence has accumulated during the last years that there is a direct association between daily variations in the concentrations of airborne particles and a range of health indicators (Stedman et al., 2002; Wong et al., 2002; Diaz et al., 2004).

Despite the mitigating impact of the nearby Atlantic Ocean on the effects of aerosols and pollution (Almeida et al., 2013), Lisbon has been affected by several high pollution episodes in the last two decades, exceeding repeatedly the legal limits imposed for PM_{10} (APA, 2008; Russo et al., 2014). Those episodes are often related to the occurrence of synoptic patterns with an eastern component which results in an eastern/southeastern flow and advection of dryer continental air (Russo et al., 2014). Therefore, a good PM_{10} prediction procedure with a sufficiently large time–lag is needed to prevent the occurrence of exceeding concentrations.

The methodological approach here presented is very straightforward in terms of operational implementation and has low computational costs and thus can be relevant for daily surveillance and alert systems in the Lisbon area.

2. Data

2.1. Target data

We consider daily values of PM_{10} concentrations measured by twelve monitoring stations in the agglomeration of Lisbon (Figure 1), between 2002 and 2006, which record the atmospheric concentrations of major pollutants, such as gases (e.g. NO_2, NO and CO) and PM_{10}. This network is complemented by three meteorological monitoring stations, located near the stations of Avenida da Liberdade (AL), Lavradio (L) and Olivais (O).

A preliminary data analysis showed that it is difficult to identify a clear cycle in PM_{10}, cf. Figure S1 of the Supporting Material (SM). However, when analyzing Figure S1a, it is possible to identify higher values during winter and summer months and lower ones during autumn and spring. Nevertheless, the cyclic behavior is not as noticeable as it usually occurs with O_3 and NO_2, cf. Figure S1 of the SM.

Daily legal limits were often exceeded during the 2002–2006 period in all the monitoring stations (APA, 2008), but the number of days with exceeding values is especially impressive for AL and E (Entrecampos) stations. It is worth mentioning that, in both stations two types of exceedances occurred, as the daily legal limit (50 µg/m^3) was exceeded, but also the number of times that the daily limit can be exceeded per year (35 exceedances/year) was also surpassed (APA, 2008).

Thus, the target of the present work is to predict PM_{10} on day t+1 on each monitoring station based on measurements on day t of several input variables (Section 2.2.).

2.2. Input data for NN training

The 15 variables that are available as NN input data sets are shown in Table 1. Additionally to the pollutant's concentration measured on the previous day and at 00:00 UTC (Universal Time Coordinated), several available meteorological variables measured in the 3 monitoring stations were considered.

In order to include information regarding the atmospheric stability and circulation, which is an important factor for the accumulation of pollutants near the surface, two other variables were considered, namely the boundary layer and the daily CWT. Three boundary layer height (BLH) fields were retrieved from the ECMWF 40–years reanalysis (ECMWF, 2013) for the years 2002–2006: the 03:00 UTC (BLH5), 09:00 (BLH7) and 21:00 UTC (BLH11). The BLH varies along the day, and the anti–phase diurnal variations of PM mass concentrations and BLH indicate that the BLH is one of the important factors affecting air quality (Du et al., 2013). Thus, we decided to use 3 measures of the BLH, one during night time (BLH5), one during peak traffic hours (BLH7) and one after the normal work day ends (BLH11). The daily CWT classification was determined based on the Trigo and DaCamara (2000) approach,