Atmospheric Environment 79 (2013) 822–830

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Air quality prediction using optimal neural networks with stochastic variables

Ana Russo [a,*], Frank Raischel [b], Pedro G. Lind [b,c,d]

[a] Center for Geophysics, IDL, University of Lisbon, 1749-016 Lisboa, Portugal
[b] Center for Theoretical and Computational Physics, University of Lisbon, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
[c] TWIST – Turbulence, Wind Energy and Stochastics, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany
[d] ForWind – Center for Wind Energy Research, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany

HIGHLIGHTS

- Optimized variables selection on a multivariate system by stochastic data analysis.
- Selection of variables used as input for ANN air quality forecast models.
- Use of derived variables as input for ANN models maintains forecast capabilities.
- Use of derived variables as ANN's inputs reduces the amount of input variables.
- Methodology can be adapted to other ANN models in weather or geophysical forecast.

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form
26 July 2013
Accepted 29 July 2013

PACS:
[2010] 92.60.Sz
02.50.Ga
02.50.Ey

Keywords:
Pollutants
Neural networks
Stochastic systems
Environmental research

ABSTRACT

We apply recent methods in stochastic data analysis for discovering a set of few stochastic variables that represent the relevant information on a multivariate stochastic system, used as input for artificial neural network models for air quality forecast. We show that using these derived variables as input variables for training the neural networks it is possible to significantly reduce the amount of input variables necessary for the neural network model, without considerably changing the predictive power of the model. The reduced set of variables including these derived variables is therefore proposed as an optimal variable set for training neural network models in forecasting geophysical and weather properties. Finally, we briefly discuss other possible applications of such optimized neural network models.

1. Introduction

Urban air pollution is a complex mixture of toxic components with considerable impact on the inhabitants of urban regions, particularly those belonging to sensitive groups, such as children and people with previous heart and respiratory insufficiency (Kolehmainen et al., 2001). Therefore, forecasting the temporal evolution of air pollution concentrations in specific urban locations emerges as a priority for guaranteeing life quality in urban and metropolitan centers. With this aim, and in order to identify and predict in advance episodes of low air quality at regional and local scales, air quality forecasting models have been developed, considering the characteristics of atmospheric pollution and its consequent impact on people's health and life quality.

Straightforward approaches such as Box models (Middleton, 1997), Gaussian plume models (Reich et al., 1999), persistence and regression models (Shi and Harrison, 1999) are commonly applied to characterize and forecast air pollutants dispersion. These models are easy to implement and allow for the rapid calculation of forecasts. However, they include significant simplifications (Luecken et al., 2006) and usually do not describe the processes and interactions that control the transport and chemical

* Corresponding author.
E-mail addresses: ana_russo@yahoo.com, acrusso@fc.ul.pt (A. Russo).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.07.072

behavior of pollutants in the atmosphere (Luecken et al., 2006), important for instance for secondary pollutants (Sokhi et al., 2006). Improvements have been made with deterministic dispersion models and statistical-based approaches, which however, being highly nonlinear (Lal and Tripathy, 2012), require a large amount of accurate input data and are considerably expensive from the computational point of view (Dutot et al., 2007).

A promising alternative to all these models are artificial neural networks (ANN) (Lal and Tripathy, 2012; Nejadkoorki and Baroutian, 2012; Gardner and Dorling, 1998). Several ANN models have already been used for air quality forecast, in particular for forecasting hourly averages (Kolehmainen et al., 2001; Perez et al., 2000; Kukkonen et al., 2003) and daily maxima (Perez, 2001). Further, several authors compared already the potential of different approaches when applied to different pollutants and prediction time lags (Kukkonen et al., 2003; Yi and Prybutok, 2002; Gardner and Dorling, 2000; Hooyberghs et al., 2005). Still, though successful in many situations and having considerably less restrictions on the input data, large training data sets are usually required to improve accuracy and minimize uncertainty in the output data, which up to now has been a significant disadvantage of these models.

Recently, we applied methods from stochastic data analysis and statistical physics for deriving variables with reduced stochastic fluctuations (Vasconcelos et al., 2011) to empirical data in sets of NO_2 concentration measurements (Raischel et al, 2012). Such methods were introduced in the late nineties (Friedrich and Peinke, 1997; Friedrich et al., 2011) for analyzing measurements on complex stochastic processes, aiming for a quantitative estimation of drift and diffusion functions from sets of measurements that fully define the evolution equation of the underlying stochastic variables. The framework has already been applied successfully, for instance to describe turbulent flows (Friedrich and Peinke, 1997) and the evolution of climate indices (Lind et al., 2005, 2007), performance curves of wind turbines (Raischel et al, 2013), stock market indices (Friedrich et al., 2000), and oil prices (Ghasemi et al., 2007). At the same time, the basic method has been refined in particular for data with low sampling frequency (Kleinhans et al., 2005; Lade, 2009) and subjected to strong measurement noise (Boettcher et al., 2006; Lind et al., 2010; Carvalho et al., 2011).

In this paper we present an important application of such variables: using them as input for training ANN enables one to reduce considerably the amount of input data needed for achieving a given accuracy. We argue that this reduction in the number of input variables is possible because the derived variables incorporate temporal correlations between independent and spatially separated monitoring stations. Moreover, as we quantitatively show below, when using this reduced amount of information that includes the derived variables, the predictive power of the ANN is not significantly changed, which is a major advantage when working with observational data which might include missing values. Combining a faster ANN training with the same predictive power may improve the ability and capability of alert system for air quality in large urban centers. We start in Sec. 2 by briefly describing ANN models as well as the main points of the stochastic data analysis procedure used. In Sec. 3 the empirical data is described, comprising two different data sets of NO_2 concentration measures in the city of Lisbon, Portugal (see Fig. 1). In Sec. 4 the results are discussed in the light of predictive power measures, and Sec. 5 concludes the paper.

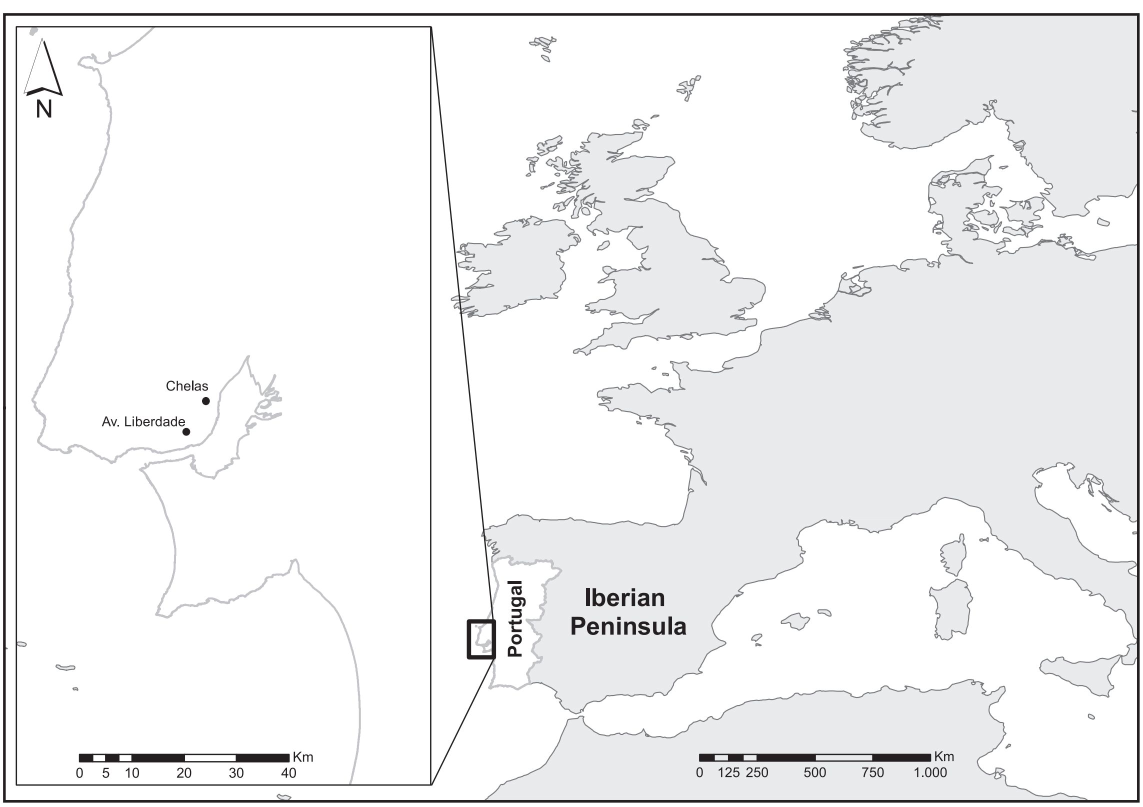

Fig. 1. NO_2 measurement stations in the region of Lisbon (Portugal) at the Southwestern coast of Europe. In this paper we focus on the set of measurements taken at the stations of Chelas and Avenida da Liberdade, approximately 4 km distant apart, with approximately $3x10^4$ data points. Each data set is extracted within the period between 2002 and 2006, with a frequency of 1 h^{-1}.