Atmospheric Environment 92 (2014) 19–30

Contents lists available at ScienceDirect

Atmospheric Environment

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

Review

Explorative forecasting of air pollution

D. Domańska [a,*], M. Wojtylak [b]

[a] Institute of Computer Science, University of Silesia, Bedzińska 39, 41-200 Sosnowiec, Poland
[b] Institute of Meteorology and Water Management (IMGW), Bratków 10, 40-045 Katowice, Poland

HIGHLIGHTS

- We propose method of explorative forecasting of air pollution.
- The method to forecast uses additional information about stations and wind direction.
- We use real meteorological data (from past to present) from areas in Poland.
- The method can be used to forecast hourly and daily average values of pollution.
- Explorative forecasting has competitive results to those in the literature.

ARTICLE INFO

Article history:
Received 14 August 2013
Received in revised form
17 March 2014
Accepted 21 March 2014
Available online 25 March 2014

Keywords:
Prediction
Air pollution forecasting
Fuzzy numbers
Data mining
Expert system

ABSTRACT

In the paper a model to predict immission concentrations of PM_{10}, SO_2, O_3 for a selected number of forward time steps is proposed. The proposed model (e-APFM) is an extension of the Air Pollution Forecasting Model (APFM). APFM requires historical data for a large number of points in time, particularly weather forecast, meteorological and pollution data. e-APFM additionally requires information about the wind direction in sectors and meteorological station. This information also permits pollution at meteorological stations for which we do not have the necessary data (in particular the data about pollution) to be forecast. The experimental verification of the proposed model was conducted on the data from the Institute of Meteorology and Water Management in Poland over a period of two years (between January 2011 and December 2012). Experiments show that the e-APFM method has lower deviations between the measured and predicted concentrations compared to the APFM method for the first day and similar deviations for the next two days (for hourly values) and for the first day and mostly worse for the second and third day (for daily values).

1. Introduction

The methodology of forecast creation is a very complex process and some of the steps are labour-consuming, e.g., data preparation. In addition to the data preparation, the steps in forecast creation are: the determination of the prediction aim and the prognostic premises, the qualification of the forecasts methods, the determination of the forecast, the use of the forecast, the monitoring and evaluation of the forecast.

In the work we first concentrate on the qualification of the forecast methods, and next on improving the forecast determination method. Forecast methods can be classified according to many criteria, e.g., automation criterion, the level of the complication of the method or type of the forecast. The first criterion divides the forecast methods into automatic and non-automatic methods. Automatic methods do not require any intervention by humans, in contrast to the non-automatic methods. The second criterion for classification is the level of the complication of the method. We distinguish between simple and complex methods. The third and final criterion divides forecast methods according to their type.

Several types of forecasting methods have been discussed in the literature (Chatfield, 2000; Armstrong, 2001a,b). They are fundamentally divided into two main streams: judgemental methods and statistical methods (in which cases forecasts can be given without specific knowledge about emission). Other methods, i.e., those based on physical formulas, e.g., using Navier–Stokes equation of motion or focused on creating own physical model (in cases in which the knowledge about emission is important) also exist. In the paper we present methods that do not use the values of

* Corresponding author. Modelling and Computer Graphics, University of Silesia, Będzińska 39, 41-200 Sosnowiec, Poland.
E-mail addresses: ddomanska@poczta.onet.pl (D. Domańska), monitoring.katowice@imgw.pl (M. Wojtylak).

http://dx.doi.org/10.1016/j.atmosenv.2014.03.041
1352-2310/

emission to forecast so we focus on describing the first two types of forecasting methods. The judgemental methods are the basis for forecasting methods that use experts' opinions. In many situations, the first step is to ask the experts (Harvey, 2001). Sometimes this is enough; as experts may make excellent forecasts. An expert's opinion is, however, subject to biases and shortcomings.

On the other hand, one can descend forecasting methods from the univariate statistical methods by using extrapolation methods. They consist of methods that use the values of a series to predict other values (Armstrong, 2001a,b; Remus and O'Connor, 2001). For a given variable, the univariate methods are based only on the past values of that variable.

Pure extrapolation of a time series assumes that all that we need to know is contained in the historical values of the series that is being forecast. For cross-sectional extrapolations, it is assumed that the evidence from one set of data can be generalised to another set.

One can descend the forecasting method by analogies from the methods that use expert opinions and one can use extrapolation models to descend the forecasting method by analogies. These methods are used when theory does not deliver a fully causal explanation of the phenomenon and the empirical research indicates the variability of the cause and method of their interaction. Formally, the use of analogies can help in the forecasting of experts. It can reduce systematic errors that are caused by the optimism and non-realistic seeing of someone's capabilities.

The authors introduce an explorative forecast as an extension of the forecasting method using analogies. The diagram of the explorative forecast procedure is presented in Fig. 1. The forecast factors are data that are related to the future (forecasts, for the meteorological data we have weather forecasts), which coexist with historical data, i.e., data from the present and the past, and that are correlated with them. The idea of an explorative forecast is the forecasting of data basing on the similarities between the historical data and the data that are related to the future. The aim of the prediction is a time series of the selected phenomenon.

Air pollution is a fundamental problem in many parts of the world. It is closely connected with the emission and immission of pollution. The basic process through which the emission of pollution into the atmospheric air is the burning of fossil fuels in power stations, thermal-electric power stations, individual domestic furnaces, transport, natural disasters etc. The health of human is one of the short-term effects of air pollution, mainly in metropolises (Tabaku et al., 2011). The consequences that are connected with air pollution have brought about an increase in the social consciousness about the necessity to be concerned about air quality in both developing and developed countries (Klejnowski et al., 2006; Ośródka et al., 2006).

Statistical models, methods based on reasoning rules, neural networks, filtering of time series, support vector machines, clustering analysis, etc. are often used to forecast pollution concentrations (Hooyberghs et al., 2005; Perez and Reyes, 2006; Siwek and Osowski, 2012; Yu et al., 2012). The methodology of forecasting using the neural networks, genetic algorithm and linear regression model is presented in Grivas and Chaloulakou (2006). The methods that are used to forecast pollution concentrations in which artificial neural networks and support vector machines are presented in Benvenuto and Marani (2000), Božnar et al. (2004), Perez et al. (2000). We find an approach to the pollution forecasting with NO_2, O_3, CO and PM_{10} (hourly forecast for 1 day) based on neural networks and families of univariate regression models in Hrust et al. (2009). The error measures MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) were used to check the valuation of the methods. A method to forecast hourly values of O_3 for a 24 h horizon using neural networks was described in the paper of (Coman et al., 2008). A method to forecast an average day value of O_3 using neural networks is described by Sousa et al. (2006). A method of pollution forecasting for three days (as average day value) based on plain model using the chosen meteorological values and the pollution concentrations was described by Kurt and Oktay (2010). Further methods in this work are based on geographic models: the single-site neighbourhood model, the two-site neighbourhood model, and the distance-based model. A method to forecast values of PM_{10}, SO_2 and NO_2 as an average day value using a statistical method based on time series analysis, multiple regression model and based on a multiple layer perceptron network was presented by Jiang et al. (2004). A review of forecasting methods that are used to predict the values of pollution is presented in Zhang et al. (2012a, 2012b).

The APFM (Air Pollution Forecasting Model) method presented in Domańska and Wojtylak (2012a) is fed with weather forecasts, meteorological parameters and pollution concentrations. The extended APFM method, which will be presented in this paper, additionally contains information about wind direction and synoptic stations.

The paper is organised in the following way. In Section 2, we present an air pollution forecasting model and describe its steps in details. In Section 3, we describe the experimental setup: the dataset used in the air pollution forecasting model, error measure, station similarities and the experiments. The proposed method was tested on the data from the Silesian voivodeship. Weather forecasts from four areas were computed. There are four synoptic stations and sixteen meteorological stations (not all are active). We also discuss the results that were obtained. In Section 4, we give some conclusions.

2. The APFM model including the division into stations and areas

The proposed model is an extension of the APFM method (Domańska and Wojtylak, 2012a) and is called e-APFM (extended Air Pollution Forecasting Model). The model forecasts pollution concentrations, e.g., particulate matter PM_{10}, SO_2, O_3 for any selected day (usually the next day) or hours on the condition that the weather forecasts for that day are available. The basic terms that are used in the pollution concentrations forecasting model are described in Domańska and Wojtylak (2012a).

Let us define some terms used in the model:

- time horizon T – number of terms for which the forecast will be computed,
- time step Δt, e.g., 10 min, 1 h, 1 day,
- weather forecast – forecast of meteorological parameters; forecast for a selected day that is computed from the physico-chemical models for the area and the selected attributes, e.g., temperature, wind speed, pressure etc. for each $i \cdot \Delta t$, where $i = 1,\dots,T$,
- meteorological parameters – meteorological data that was measured at the synoptic station; we measure selected

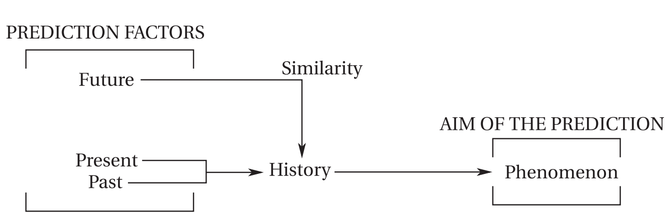

Fig. 1. The explorative forecast procedure diagram.