Environmental Modelling & Software 84 (2016) 290–303

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong

Bing Gong*, Joaquín Ordieres-Meré

Department of Industrial Engineering, Business Administration and Statistic, E.T.S Industrial Engineering, Universidad Politécnica de Madrid, Calle de José Gutiérrez Abascal 2, 28006, Madrid, Spain

ARTICLE INFO

Article history:
Received 11 April 2016
Received in revised form
18 June 2016
Accepted 23 June 2016

Keywords:
Ozone level forecasting
Classification
Artificial intelligence
Re-sampling
Imbalanced data
Ensemble models

ABSTRACT

The objective of this study was to apply preprocessing and ensemble artificial intelligence classifiers to forecast daily maximum ozone threshold exceedances in the Hong Kong area. Preprocessing methods, including over-sampling, under-sampling, and the synthetic minority over-sampling technique, were employed to address the imbalance data problem. Ensemble algorithms are proposed to improve the classifier's accuracy. Moreover, a distance-based regional data set was generated to capture ozone transportation characteristics. The results show that a combination of preprocessing methods and ensemble algorithms can effectively forecast ozone threshold exceedances. Furthermore, this study advises on the relative importance of the different variables for ozone pollution prediction and confirms that regional data facilitate better forecasting. The results of this research can be promoted by the Hong Kong authorities for improving the existing forecasting tools. Moreover, the results can facilitate researchers' selection of the appropriate techniques in their future research.

1. Introduction

For many years, the air pollution problem has been attracting public attention, since it can cause serious health problems. The United States Environmental Protection Agency (EPA) set particle pollution (frequently referred to as particulate matter) (PM), ground-level ozone (O_3), carbon monoxide (CO), sulfur oxides (SO_2), nitrogen oxides (NO_x), and lead as 'critical pollutants'. Among these six pollutants, O_3, which has adverse effects on public health and agricultural yields, is one of the most dangerous pollutants (United States Environmental Protection Agency, 2015; Salazar-Ruiz et al., 2008; Khatibi et al., 2013; Lam et al., 2015).

To assist the control of O_3 pollution, the Hong Kong Special Administrative Region of the People 's Republic of China (HK) government established a maximum ozone threshold of 160 $\mu g/m^3$. Ozone levels in excess of this threshold may threaten public health. Therefore, accurate and prompt prediction of O_3 concentrations that exceed the threshold is of great importance to the management of the public pollution warning system.

Since the 1970*s*, regression and auto-regressive models, such as time series analysis, have been widely used in O_3 forecasting in many studies. However, the results suggested that traditional time series techniques fail to forecast O_3 accurately (Comrie, 1997; Chattopadhyay and Bandyopadhyay, 2007). As a replacement, artificial intelligence (AI) techniques emerged and proved to be more effective for ozone prediction (Robeson and Steyn, 1990; McCollister and Wilson, 1975). Of the AI techniques, the support vector machine (SVM) (Hájek and Olej, 2012; Salazar-Ruiz et al., 2008; Wang et al., 2008; Feng et al., 2011; Yu et al., 2012; Ortiz-García et al., 2010), artificial neural network (ANN) (Schlink et al., 2006; Salazar-Ruiz et al., 2008; Tsai et al., 2009), and decision tree (DT) (Zhang and Fan, 2008; Birant, 2011) are frequently employed in the domain of pollution forecasting. However, they often fail to accurately predict extreme concentrations and are of limited use because of the observational limitations (Zhang et al., 2012).

Since 2006, ensemble forecasting has begun to receive more attention, as ensemble algorithms can improve forecasting accuracy and enhance the generalization capability (Zhang et al., 2012). Among these, boosting, bagging, and stacking are the most popular

* Corresponding author.
E-mail addresses: bing.gong@alumnos.upm.es (B. Gong), j.ordieres@upm.es (J. Ordieres-Meré).

http://dx.doi.org/10.1016/j.envsoft.2016.06.020
1364-8152/© 2016 Elsevier Ltd. All rights reserved.

techniques and showed promising pollution forecasting results (Yang et al., 2010; Singh et al., 2013; Siwek and Osowski, 2012; Debry and Mallet, 2014). However, as the ensemble models' inherent limitation associated with the single AI models, that is, the learning and training process, still leads to a focus on and a bias toward to low level concentrations (majority instances), these studies failed to forecast the "few events," i.e., the threshold exceedances of O_3 concentrations, which constitutes an imbalance data problem.

To circumvent this imbalance problem, researchers in the domain of data mining proposed various methods, among which data re-sampling (Drummond and Holte, 2003; Chawla et al., 2004; Sun et al., 2009; Zhao et al., 2014), cost-sensitive learning (Lu and Wang, 2008; Tsai et al., 2009; Fontes et al., 2014), and algorithm modifications (López et al., 2012; Sun et al., 2009) are widely applied. These studies proved the effectiveness of cost-sensitive and algorithm modification methods, but preprocessing methods were never discussed. Nevertheless, the application of a preprocessing technique to ensemble learning algorithms has proved to be effective in the data mining domain (Galar et al., 2012).

Furthermore, the ubiquity of local ozone results from the reaction of its precursors, such as NO_x, NO_2, and VOC, under certain meteorological conditions. These precursors are the consequences of anthropogenic emissions associated with transportation, industry activities, biomass burning, fossil fuel refinement, and distribution (Stojić et al., 2015). However, the ozone concentrations at local area are, probably, influenced by other factors, such as pollution transportation.

In such case, the use of both local and regional pollution data could benefit forecasting, which would help to enforce local pollution control actions and efficient abatement strategies in order to avoid a situation where all the cities are affected. Regional models, such as the Community Multiscale Air Quality (CMAQ) model and the urban atmospheric dispersion model (DAUMOD), were proposed in previous studies (Rojas, 2014; Liu et al., 2010). Ozone dispersion was considered; however, the computation cost of these models is high or they are not adequate for application in urban areas that present severe photochemical pollution conditions. AI, conversely, with the merits of fast computation and capacity to handle data that include complex photochemical reactions, has been applied in real-time and local ozone forecasting; nevertheless, the regional factor considered in previous studies was overlooked (Zhang et al., 2012).

In summary, in this study we determined whether a combination of preprocessing methods and ensemble AI techniques can solve the imbalance class problem and improve the classifier's accuracy in the field of forecasting daily maximum ozone threshold exceedances. Meanwhile, regional factors were considered by using a regional scale data set and assessing its empirical relevance.

The remaining parts of this paper are as follows. In the second part, the study area, data source, ozone characteristics, variables selection, re-sampling, and the AI methodologies used in this study are introduced. In the third part, the performances of the classifiers are compared, the importance of the variables is analyzed, the key factors influencing the ozone level are identified, and the ozone formation pattern in the selected area is specified. A comparison between regression models and classifier models is also presented. The final part summarizes the results of this study.

2. Research design

2.1. Study area and data collection

HK is located on the south coast of mainland China, close to the PRD area, which is one of the most developed in China.

In order to safeguard the health and well-being of the community and to build HK as a global and green city providing a high quality of life, the HK government implemented a wide range of measures to control local emissions from motor vehicles, shipping companies, power plants, and industrial and commercial processes. In addition, the government established a policy that the daily average ozone level should not exceed 160 $\mu g/m^3$ more than nine times per year.

The HK air pollution monitoring network contains 15 fixed monitoring stations: 12 general stations and 3 roadside stations (Fig. 1). The abbreviations of the stations can be found in Table 1. The hourly records of carbon monoxide (CO), fine suspended particulates (FSP), nitrogen dioxide (NO), nitrogen oxide (NO_x), ozone (O_3), respirable suspended particulates (RSP), and sulfur dioxide (SO_2) are provided by the Environmental Protection Department (EPD) of HK. The data can be downloaded free of charge from the EPD official Website (EPD, 2014). The HK Observatory (HKO) offers hourly meteorological data from 42 meteorological stations, such as temperature (TMP), relative humidity (RH), and wind direction (WDR). The meteorological data can be found at the Website of the HKO (HKO, 2014).

2.2. Characteristics of ozone concentrations

Fig. 2 illustrates that the number of industries is not consistent with the O_3 threshold exceedances at the local area (district) level in HK. For example, the number of O_3 exceedances in Tusen Wan (TW) and Kwai Chung (KC), which contain the highest number of manufacturing establishments, is relatively low. Conversely, the pollution levels in areas with a small number of manufacturers, such as Yue Long (YL) and Sha Tin (ST), are high.

Furthermore, HK is near one of the most developed areas, the Pearl River Delta (PRD), in China. The pollution from this remote emission source probably contributes to the local ozone formation. Fig. 3 presents the hourly O_3 concentration levels from October 5*th* to October 7*th*, 2008 at the TW and YL stations. YL is the nearest monitoring station to the PRD area. During this period, the wind direction at the YL station was between 270 and 315°. The distance between TW and YL is 12.5 km. The maximum wind speed was 11 m/s, the minimum was 0.7 m/s, and the average was around 5 m/s. In Fig. 3 it can be seen that the time of the O_3 peak level at the YL station is one to two hours in advance of that at the TW station. Therefore, the main responsibility for O_3 formation at TW is O_3 transportation rather than the local industrial activities at TW, which means that the pollution from industry in local areas in HK is not the only source for local O_3 formation. Thus, global factors, such as O_3 transportation, could play a significant role in O_3 pollution forecasting.

The diurnal variations in the average ozone level in July and October, at the TW, Tung Chung (TC), Kwai Chung (KC), and YL monitoring stations are given in Fig. 4. The ozone concentration levels at these stations demonstrate a consistent pattern in July, which is characterized by one peak in mid-afternoon, between 14:00 and 16:00, and two valleys between 06:00 and 08:00 and after 20:00, remaining stable at a low level from 00:00 to 05:00. In October, the ozone levels increased significantly at these four stations. However, the pattern observed during this period was different from that observed in July, when one peak appeared during the day. In October, two peaks appeared: from 03:00 to 06:00 and from 14:00 to 17:00.

Moreover, Fig. 5 shows the frequency of the days on which the daily maximum ozone level occurred at the corresponding time of the day. The x axis represents the time of day and the y axis represents the frequency at which the daily maximum ozone occurred at the corresponding time. Although the average ozone level