Engineering Applications of Artificial Intelligence 32 (2014) 1–9

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A cluster multivariate statistical method for environmental quality management

Qinliang Tan [a,b,*], Yongmei Wei [a], Minnan Wang [a], Yuan Liu [a]

[a] College of Economy and Management, North China Electric Power University, Beijing 102206, PR China
[b] Research Center for Beijing Energy Development, Beijing 102206, PR China

ARTICLE INFO

Article history:
Received 25 December 2012
Received in revised form
22 December 2013
Accepted 7 February 2014
Available online 19 March 2014

Keywords:
Air quality management
Forecasting
Multivariate statistics
Regression

ABSTRACT

This study advances a Linear-regression-based stepwise cluster analysis (LSCA) and a quadratic-regression-based stepwise cluster analysis (QSCA) model. The models have the advantages of (1) being independent of complex atmospheric, meteorological and topographical information, (2) dealing with continuous and discrete variables, as well as nonlinear relationships among the variables, (3) facilitating finer analysis of within-cluster variations for the stepwise cluster analysis (SCA) outputs, leading to an improved forecasting accuracy, and (4) providing a reasonable result–interpretation since any variation of the explanatory variable will lead to the corresponding change of the response level. The models are then applied to the city of Xiamen in China for forecasting in air quality management. They are also compared with several alternative statistical models: SCA, decision trees (DT) and quadratic regression (QR). The results show that, among the five prediction models, the QSCA has the best forecasting performance followed by the LSCA. Since the diverse and potentially nontechnical user communities' interests are embraced by the approaches, they could be robust in terms of the varying levels of knowledge backgrounds, application capabilities, and data availability.

1. Introduction

Air quality forecasting is important for developing regional environmental management plans. It is used to assess impacts of regional economic development on the atmospheric environment. During past years, a large number of environmental quality management tools have been developed (Chock and Winkler, 1997; Kumbaroglu, 1997; Kim, 2004; Cai, 2006). Among them, statistical inference has been widely-used for air quality management when no knowledge about the causes of pollutant emission and dispersion is available. It has proved to be useful in quantifying a relation between air quality and its impact factors (Park et al., 2010). During past decades, many research efforts were made to develop statistical algorithms to support air quality predictions and management. Such algorithms could be divided into simplified linear (Slini et al., 2002; Kapetanios et al., 2008; Eynard et al., 2011) and nonlinear models (Wang et al., 2003; Michelle et al., 2005; Semenov, 2003; Pal and Mather, 2003; Rahideh et al., 2012).

Gardner and Dorling (1998) gave a general introduction and discussion regarding recent applications of artificial neural network to atmospheric quality management. Kalapanidas and Avouris (2001) present an air quality monitoring framework for management of urban areas where the task of short-term prediction of key-pollutants concentrations is a daily activity of major importance. They further discussed an NEMO approach to support short-term prediction of NO_2 maximum concentration levels in Athens, Greece; the NEMO performance was compared with that of a back propagating neural network and a decision tree; results showed that the overall performance of NEMO can be a good candidate supporting environmental management experts in operational conditions. Wang et al. (2001) developed a nested air quality predicting modeling system (NAQPMS) to investigate the various processes that govern the loading of chemical species and aerosols at various scales of atmospheric emotions in urban and regional scales. Hutchison et al., 2004 present some new approaches to integrated data into a real-time prediction methodology to support operational air quality forecasts in environmental management; the approaches were demonstrated based on remotely sensed satellite observations when the transient pollution can be isolated from local sources. Hoek et al. (2008) proposed the land-use regression method to assess outdoor air pollution; the method combines monitoring of air pollution at

* Corresponding author at: College of Economy and Management, North China Electric Power University, Beijing 102206, PR China. Tel.: +86 10 61772914.
E-mail address: tql@ncepu.edu.cn (Q. Tan).

http://dx.doi.org/10.1016/j.engappai.2014.02.007
0952-1976

typically 20–100 locations, spread over the study area, and development of stochastic models using predictor variables usually obtained through geographic information systems (GIS); it was applied to model annual mean concentrations of NO_2, NO_x, $PM_{2.5}$, the soot content of $PM_{2.5}$ and VOCs in different settings, including European and North-American cities; it was found that the performance of the method in urban areas is typically better or equivalent to geo-statistical methods, such as kriging, and dispersion models.

Of particular interest, Huang (1992) developed a model based on stepwise cluster analysis (SCA), which is used to deal with discrete and nonlinear systems in an air quality forecasting framework. The algorithm was effective in reflecting complex natures of the atmospheric environment with an improved accuracy level in forecasting air quality. Due to its advantages over the conventional multivariate statistical algorithms, the algorithm was applied to multiple types of environmental systems (Breiman, 2001; Iorgulescu and Beven, 2004; Refsgaard et al., 2003). For example, based on the SCA algorithm, Huang (2004) established a set of nonlinear, discrete predictors for estimating groundwater quality at a petroleum-contaminated site. The predictors were later integrated into an optimization framework for helping determine the optimal remediation strategies. Similar algorithms could also be found in many other studies (Breiman, 2001; Liu and Ranjithan, 2010), with the procedure of cluster mergence being neglected. Iorgulescu and Beven (2004) presented an algorithm similar to SCA for the analysis of rainfall–runoff relationships in catchments. A regression-tree algorithm was proposed by Yang et al. (2003) to investigate the hydrological relationships in watershed systems. Moreover, several decision-tree algorithms were developed for medical diagnosis and changes in freshwater acidification trends (Tsumoto, 2004; Thomas et al., 2002) and pattern cognitions in industry-oriented systems (Ezzedine et al., 2005; Kurgan and Musilek, 2006; Shahbaz et al., 2009).

One limitation of the above algorithms is the interpretability of resulting clusters. Within each cluster, variations of an explanatory variable would not result in any change of the response variables. In many atmospheric systems, in fact, the response variables are sensitive to any variation of an explanatory variable. Obviously, such interrelations can hardly be specified through the cluster-analysis algorithm. Another limitation is the piecewise nature of the generated cluster trees (Refsgaard et al., 2003; He et al., 2008b; David and Barbara, 2010). All the responses classified into each cluster are assigned to an identical prediction value. This is based on an assumption that, after the clusters are finalized, the impacts of explanatory variables on each response become insignificant. This assumption may lead to prediction errors. Thus, consideration of variations within each cluster would help improve the prediction accuracy.

This study aims to advance a linear-regression-based SCA (LSCA) and a quadratic-regression-based SCA (QSCA) modeling approach that may effectively address the differences between and within the clusters. A case study in the city of Xiamen, China, will then be provided for demonstrating the approaches' performance in urban air quality forecasting. The paper is organized as follows: Section 2 describes the methodology. Section 3 gives the overview, results and discussion of a case study. Section 4 provides conclusions.

2. Methodology

2.1. Modeling formulation

The modeling and solution process is shown in Fig. 1. Assume a sample consists of m pollution causes and p pollutants, denoted as

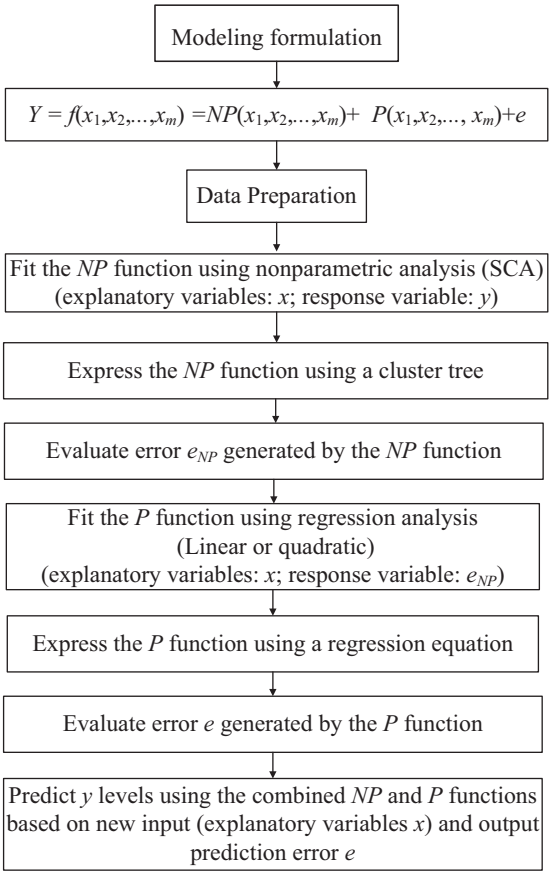

Fig. 1. Flowchart of the modeling formulation and the solution algorithm.

$x = (x_1, x_2, \ldots, x_m)$ and $y = (y_1, y_2, \ldots, y_p)$, respectively. To predict y through x_1, x_2,…, x_m, a mathematical model is formulated as follows:

$$y_1 = f_1(x_1, x_2, \ldots, x_m) = NP_1(x_1, x_2, \ldots, x_m) + P_1(x_1, x_2, \ldots, x_m) + e_1$$
$$y_2 = f_2(x_1, x_2, \ldots, x_m) = NP_2(x_1, x_2, \ldots, x_m) + P_2(x_1, x_2, \ldots, x_m) + e_2$$
$$\cdots$$
$$y_p = f_p(x_1, x_2, \ldots, x_m) = NP_p(x_1, x_2, \ldots, x_m) + P_p(x_1, x_2, \ldots, x_m) + e_p$$

where f_1, f_2,…, f_p are the functions reflecting the relationships between response and explanatory variables; NP_1, NP_2, …, NP_p denote the nonparametric functions; P_1, P_2, …, P_p denote the parametric functions.

2.2. Solution algorithm

The algorithm solving the model is divided into four stages: criterion establishment for clusters splitting and mergence, stepwise cluster analysis, regression analysis and example forecasting. The essence of this approach is, based on a given statistical criterion, to divide the fitting samples into a set of clusters that have significant differences; and then each cluster is assigned to a polynomial regression equation (i.e., regressor) representing a type of underlying relationship between variables.

[Stage 1]: This stage is to provide a criterion for determining whether the samples can be divided into two clusters and whether the two clusters can be merged into one. If n samples are used, one can have two matrixes $X = (x_{rt})_{m\times n}$ and $Y = (y_{kt})_{p\times n}$, where $r = 1, 2, \ldots, m$, $k = 1, 2, \ldots, p$, and $t = 1, 2, \ldots, n$. Let cluster h, which contains n_h samples, be cut into two sub-clusters e and f (e and f contain n_e and n_f samples, respectively,