Available online at www.sciencedirect.com
SCIENCE DIRECT®
Environmental Modelling & Software 20 (2005) 1263–1271
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations

Sabah A. Abdul-Wahab[a,*], Charles S. Bakheit[b], Saleh M. Al-Alawi[a]

[a] *Sultan Qaboos University, College of Engineering, Mechanical and Industrial Engineering Department, P.O. Box 33, Al Khod, Postal code 123, Muscat, Oman*
[b] *Sultan Qaboos University, Department of Mathematics and Statistics, College of Science, Al Khod, Postal code 123, Muscat, Oman*

Received 5 June 2003; received in revised form 16 June 2004; accepted 1 September 2004

Abstract

Data on the concentrations of seven environmental pollutants (CH_4, NMHC, CO, CO_2, NO, NO_2 and SO_2) and meteorological variables (wind speed and direction, air temperature, relative humidity and solar radiation) were employed to predict the concentration of ozone in the atmosphere using both multiple linear and principal component regression methods. Separate analyses were carried out for day light and night time periods. For both periods the pollutants were highly correlated, but were all negatively correlated with ozone. Multiple regression analysis was used to fit the ozone data using the pollutant and meteorological variables as predictors. A variable selection method based on high loadings of varimax rotated principal components was used to obtain subsets of the predictor variables to be included in the regression model of the logarithm of the ozone data. It was found that while high temperature and high solar energy tended to increase the day time ozone concentrations, the pollutants NO and SO_2 being emitted to the atmosphere were being depleted. Night time ozone concentrations were influenced predominantly by the nitrogen oxides ($NO+NO_2$), with the meteorological variables playing no significant role. However, the model did not predict the night time ozone concentrations as accurately as it did for the day time. This could be due to other factors that were not explicitly considered in this study.

Keywords: Statistical analysis; Principal component analysis; Regression analysis; Variable selection methods

1. Introduction

It has been reported that tropospheric ozone (O_3) is the principal index substance of photochemical smog. It has been recognized as one of the principal pollutants that degrades air quality (Van Eijkeren et al., 2002; Xu and Zhu, 1994). It is a key precursor of the hydroxyl radical (OH) which controls the oxidizing power of the atmosphere (Logan et al., 1981; Thompson, 1992). The concentration of OH in turn influences the concentrations of many trace species such as CH_4, CO and SO_2 (Poulida et al., 1991). In addition, high ozone levels not only play a role in damage to plant species, various natural materials and manufactured goods, but also lead to the damage of lung tissues in humans (Wang and Georgopoulos, 2001). There are no significant primary emissions of ozone into the atmosphere and all the ozone found has been formed by chemical reactions that occur in the air (WHO, 1976). Ozone, therefore, is a secondary photochemical pollutant that is not polluting in its own right. It is produced from anthropogenic precursors that include industrial and

* Corresponding author. Tel.: +968 515 360; fax: +968 513 416.
E-mail address: sabah1@squ.edu.om (S.A. Abdul-Wahab).

1364-8152/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2004.09.001

vehicular emissions of volatile organic compounds (VOC) and oxides of nitrogen (NO_x). This is the main reason why ozone is such a serious environmental problem that is difficult to control and predict.

Ozone is produced when the primary pollutants NO_x and VOCs (often called non-methane hydrocarbons, NMHC) interact under the action of sunlight. Collectively, NO_x and VOCs are referred to as ozone precursors. Additional mechanisms for the formation of tropospheric ozone include stratospheric injection and processes that influence the abundance of NO_2. The net result is that O_3 production is limited by the supply of CO, CH_4, NO, peroxyl radicals, and other hydrocarbons (Blankinship, 1996).

The destruction of ozone takes place through a number of pathways, the most important of which is surface deposition. Additional pathways for ozone consumption are available through the oxidation of SO_2 in liquid phase reactions. The rate of these reactions, in addition to rates of the competing transport and scavenging processes, vary widely in accordance with the meteorological and photolytic conditions. Scavenging processes dominate the removal of O_3 and odd nitrogen species ($NO_y = NO + NO_2 + HNO_3 +$ organic nitrates) such that their lifetimes are much lower in the continental or marine boundary layer than in the upper troposphere (Blankinship, 1996).

In addition to being affected by the non-linear nature of photochemistry, the relationship between ozone and its precursors is complicated by the fact that meteorological and chemical processes can interact over a very wide range of temporal and spatial scales. For example, chemical reaction rates range from very fast to very slow. Fast reactions have a direct impact in the locality of the emissions and can be strongly affected by atmospheric mixing. On the other hand, slow reactions are relatively insensitive to local mixing and affect a wider, regional or global spatial area (Georgopoulos, 1995; Wang and Georgopoulos, 2001).

Such relationships between meteorological conditions, and ozone concentrations have been examined in several studies which have used a combination of statistical regression, graphical analysis, fuzzy logic based method, and cluster analysis (Abdul-Wahab et al., 1996, 2000; Abdul-Wahab, 2001; Abdul-Wahab and Al-Alawi, 2002; Blankinship, 1996; Buhr et al., 1995; Clark and Karl, 1982; Cox and Chu, 1991; Lavecchia et al., 1996; Peton et al., 2000).

Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several independent (predictor) variables. In spite of its evident success in many applications, however, the regression approach can face serious difficulties when the independent variables are correlated with each other (McAdams et al., 2000). Multicollinearity, or high correlation between the independent variables in a regression equation, can make it difficult to correctly identify the most important contributors to a physical process. One method for removing such multicollinearity and redundant independent variables is to use multivariate data analysis (MDA) techniques. MDA have been used for analyzing voluminous environmental data (Buhr et al., 1992, 1995; Chang et al., 1988; Sanchez et al., 1986; Statheropoulos et al., 1998).

One such method is Principal component analysis (PCA), which has been employed in air-quality studies (Maenhaut et al., 1989; Statheropoulos et al., 1998; Shi and Harrison, 1997; Tian et al., 1989; Vaidya et al., 2000) to separate interrelationships into statistically independent basic components. They are equally useful in regression analysis for mitigating the problem of multicollinearity and in exploring the relations among the independent variables, particularly if it is not obvious which of the variables should be the predictors. The new variables from the PCA become ideal to use as predictors in a regression equation since they optimize spatial patterns and remove possible complications caused by multicollinearity.

2. Principal component analysis

Essentially, PCA maximizes the correlation between the original variables to form new variables that are mutually orthogonal, or uncorrelated (Tsunami, 1999). It is a special case of factor analysis which transforms the original set of inter-correlated variables into a new set of an equal number of independent uncorrelated variables or principal components (PCs) that are linear combinations of the original variables. The principal components are ordered in such a way that the first PC explains most of the variance in the data, and each subsequent one accounts for the largest proportion of variability that has not been accounted for by its predecessors. Although the number of PCs equals the number of independent original variables, generally, most of the variation in the data set can be explained by the first few principal components that can be used to represent the original observations.

Principal component methods are also used for selecting subsets of variables for a regression equation. One such application is to obtain a varimax rotation of the principal components, and to retain a subset of the original variables associated with each of the first few components, which are then used as predictors in the regression. Varimax rotation ensures that each variable is maximally correlated with only one principal component and a near zero association with the other components. More details on these and other methods can be found elsewhere (Jolliffe, 1986; Malinowski, 1991; Statheropoulos et al., 1998).