Prediction of ozone concentration in ambient air using multivariate methods

A. Lengyel a,b,*, K. Héberger c, L. Paksy a, O. Bánhidi a,d, R. Rajkó e

a Department of Analytical Chemistry, Institute of Chemistry, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary
b Co-operation Center of Material Sciences and Mechatronics, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary
c Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, H-1025 Budapest, Pusztaszárd ut 59–67, Hungary
d DAM STEEL Co., Miskolc, Vasgyári út 43, H-3340, Hungary
e Department of Unit Operations and Environmental Engineering, University of Szeged College Faculty of Food Engineering, P.O. Box 433, Szeged H-6701, Hungary

Received 23 October 2003; received in revised form 7 July 2004; accepted 19 July 2004

Abstract

Multivariate statistical methods including pattern recognition (Principal Component Analysis—PCA) and modeling (Multiple Linear Regression—MLR, Partial Least Squares—PLS, as well as Principal Component Regression—PCR) methods were carried out to evaluate the state of ambient air in Miskolc (second largest city in Hungary). Samples were taken from near the ground at a place with an extremely heavy traffic. Although PCA is not able to determine the significance of variables, it can uncover their similarities and classify the cases. PCA revealed that it is worth to separate day and night data because different factors influence the ozone concentrations during all day. Ozone concentration was modeled by MLR and PCR with the same efficiency if the conditions of meteorological parameters were not changed (i.e. morning and afternoon). Without night data, PCR and PLS suggest that the main process is not a photochemical but a chemical one.

Keywords: Principal component analysis (PCA); Principal component regression (PCR); Partial least squares (PLS); Environmental data; Ambient air; Modeling of ozone concentration

1. Introduction

Multivariate techniques as Principal Component Analysis (PCA), Multiple Linear Regression (MLR),
Principal Component Regression (PCR) and Partial Least Squares (PLS) are powerful tools to handle several problems (Jolliffe, 1982; Wold et al., 1987; Næs and Martens, 1988; Vogt, 1989; Marbach and Heise, 1990; Bakken et al., 1997; Rius et al., 1997; Héberger, 1999; Otto, 1999). PCR has become a popular method in the last decades, mainly because it can provide good predictive models and can deal with multivariate data sets with highly correlated variables (Jolliffe, 1982; Næs and Martens, 1988; Vogt, 1989; Marbach and Heise, 1990; Davies, 1996; Bakken et al., 1997; Rius et al., 1997).

* Corresponding author. Address: Department of Analytical Chemistry, Institute of Chemistry, University of Miskolc, H-3515 Miskolc-Egyetemváros, Hungary. Tel.: +36 46 565 372; fax: +36 46 565 371.
E-mail address: akmla@gold.uni-miskolc.hu (A. Lengyel).
Moreover, PCA can be used for classification of the data, finding similarities, detecting outliers, etc. Finding physical significance to principal components was also possible in some cases (Héberger, 1999; Héberger and Görgényi, 1999). In the last decade, the application of these chemometric methods has been widely spread in the field of environmental sciences, especially in the field of water source and air protection.

In most cases, the following problems are to be tested as simple tasks during environmental investigations:

- establishing the type of distribution of measured data,
- proving independence of two distributions,
- determining equivalency of two distributions,
- filtering of outliers,
- establishing of correlation between two data sets (to detect global and/or local trends).

In a complex study, the tasks are more complicated, because we would like (i) to estimate and predict the influence of the pollutants on the environment, (ii) to build an adequate model for e.g. forecasting pollution situations. Similarly, prediction of seasonal variation in concentrations of various pollutants is an important goal (Geladi et al., 1999; Paterson et al., 1999; Alvarez et al., 2000; Lengyel et al., 2000; Pissimanis et al., 2000; Trainer et al., 2000; Klaus et al., 2001).

For this purpose, modeling is the most suitable tool (e.g. using MLR, PLS, PCR etc.) E.g., P. Geladi and his colleagues investigated the mechanism of ozone reactions with nitrogen oxides close to the ground surface. They found NO, NO\textsubscript{2}, the temperature and UV radiation to be significant factors (Geladi et al., 1999). Generally, models of ozone arising near the ground level use analogous photochemical reactions as models for the stratosphere (Heicklen, 1976; Finlayson-Pitt and Pitt, 1986).

Ozone is a secondary air pollutant that has received extensive attention in the literature, mainly because it can cause harmful effects in vegetation during the growing season. Meteorological conditions strongly influence the efficiency of photochemical processes leading to ozone formation and destruction. Many papers (Paterson et al., 1999; Alvarez et al., 2000; Pissimanis et al., 2000; Trainer et al., 2000; Klaus et al., 2001) describe investigations in which the main goal was to model the air quality. It is ozone that specifically varies depending on the pollutant type and quantity. Several investigations (Paterson et al., 1999; Alvarez et al., 2000; Pissimanis et al., 2000; Klaus et al., 2001) used Principal Components Analysis to classify the spatial and temporal variations of ozone and to identify the important factors influencing ozone concentration.

Positive matrix factorization (PMF) produces non-negative factors for better interpretation of factors (Pata-tero and Tapper, 1994). This technique identifies the source and apportionment—important steps for the development of air pollution control strategies. Carbon monoxide, particulate matter, peroxyacetyl nitrate (PAN), isoprene, temperature, and ozone have been measured. Three physically interpretable factors were an isoprene-dominated factor, a local source factor, and a long-range transport factor (Paterson et al., 1999).

Field measurements of the concentrations of ozone, its precursors, the photochemical intermediates, and other photochemical products, as well as other atmospheric parameters were analyzed. The apportionment of volatile organic compounds (VOCs), carbon monoxide (CO) and the oxides of nitrogen (NO\textsubscript{x} = NO + NO\textsubscript{2}) among the various anthropogenic sources was established. Similarly, the importance of natural VOCs/CO and NO\textsubscript{x} relative to anthropogenic VOCs/CO and NO\textsubscript{x} in photochemical ozone production were determined. The mechanism of the photochemical ozone production, the efficiency of tropospheric ozone production relative to the NO\textsubscript{x} concentrations, and the importance of NO\textsubscript{x} relative to VOCs and CO in the observed ozone levels were also determined (Trainer et al., 2000).

The variation in ozone concentration is related to some meteorological variables (temperature, surface wind direction, speed, and global radiation) (Klaus et al., 2001).

Multivariate methods are frequently used for handling of monitoring data. Our aim was to analyze air quality data of a well-known but complicated structure. PCA will uncover the hidden structure of the data and show similarities among variables and cases (i.e. in the measurement time). Finally, we would like to model the ozone concentration using various chemometric techniques MLR, PLS and PCR.

2. Experimental data

The original data were arranged in a matrix form (see Table 1).

The following variables were used as columns: concentration of nitrogen monoxide (NO), of nitrogen diox-
The samples were arranged in the rows as were taken during time. All of these properties were continuously measured by a monitoring station. The date was chosen when the weather was fine not only at the chosen day, but also during several days before. The monitoring station was located at the most frequented square of Miskolc.
(Hungary), where two main roads with heavy traffic cross. Bus and coach stations are situated nearby this crossing. One of the market places of the city works in this area. Consequently, the streets are very crowded and considerably polluted.

As it can be seen in Table 1, the frequency of sampling was every half an hour (30 min). Samplings and analysis were continuous. Three cases (samples) were missing (No. 7, 43, 44) because of sampling error. In this paper, we named the variables with the short chemical formula if the parameter was a chemical compound. Otherwise, we used a maximum four character-long sign for the meteorological parameters (see bottom of Table 1).

2.1. Methods and procedures

The following part (Section 3) is divided according to the techniques used. First, a pattern recognition study was carried out to unravel similarities among variables and cases. To complete this pattern recognition, we applied Principal Component Analysis (PCA), i.e. PCA collects (classifies) the properties which have the same/similar effect on the investigated system. Secondly, modeling of ozone concentration was done using MLR, PCR and PLS. The model takes into account ozone concentrations near the ground surface at a place polluted with nitrogen oxides.

3. Results and discussion

3.1. Pattern recognition

PCA was used for classification of the data. Principal components (PCs) were calculated without rotation. The widely used “eigenvalues larger than one rule” did not suggest incorporating four principal components into the model, however this decision is supported with a screen plot (not shown here). PC1 to PC3 are common factors, whereas PC4 can be considered as a specific factor: the first PC correlates well with O₃, SO₂, HUM, TEMP and SUN, PC2 with monoxides (NO and CO); PC3 with wind velocity and direction, whereas PC4 with NO₂.

The loadings of PC1 and PC2 are plotted in Fig. 1. The point for humidity is an outlier; it is separated clearly from the other points and clusters. Two definite clusters can be observed: (i) pollutants as CO, NO, and (ii) SO₂ and sun exposure. The two clusters are completed with NO₃ and O₃ as well as TEMP, respectively, as it can be seen in Fig. 2.

The third cluster is composed of wind velocity and wind direction.

Four clusters can be seen from the third direction (Fig. 3).

The closest resemblance can be observed between NO and CO (cf. Fig. 1—their points were also close to each other in Fig. 1). Wind velocity and wind direction form another cluster as in Fig. 2. The presence of NO₂ in any
Cluster is arbitrary (apparent) only. It is well separated from the other factors in the fourth dimension. There is a strong difference between the daily and night temperature as well as between sun exposures during day and night. Consequently, humidity also changes during the day considerably. Score plots may show whether these changes become apparent (Figs. 4–6). The first PC separates day and night clearly. Two definite clusters can be observed in Fig. 4. Cluster 1 is on the right side (No. 45–48 and 1–16: from 22 h to 8:00). Cluster 2 is located on the right side (No. 17–42: from 08:30 to 21:30).

Cluster 1 can be decomposed into two sub-groups in Fig. 5. No. 8–16 is a new sub-group, which contains the data till 8 a.m. Sample 17 is an intermediate one, still, it is well distinguished.

It is absolutely clear that the night samples (1–6 and 45–48) and the early morning samples (8–16) create two separated groups confirmed by the observations presented in Fig. 5. Samples no. 17, 18 are well resolved from another direction (Fig. 6). This fact should be emphasized, because the PCA automatically separates data, which represent chemical conditions for the arising of ozone from the following reaction:

\[ \text{NO}_2 + \text{O}_2 = \text{NO} + \text{O}_3 \]  

(The reaction cannot be considered to be in equilibrium conditions.)

The data from 18 to 42 constitute a large group containing data measured during the day. This cluster will be investigated in the next part. As no ozone was formed during the night (cf. Table 1), it is not worth using all data in modeling the ozone formation and depletion.

The separation of the variables can clearly be seen in Fig. 7. Wind velocity and sun exposure form a cluster (they are very close), while ozone and temperature can be found in another cluster. It seems that CO, NO2 and humidity compose another cluster together with NO. Perhaps, NO can be related to SO2, but the wind direction is an individual variable. The point for temperature is the closest to the point for ozone concentration revealing its importance.

### 3.2. Modeling

Predictive models were built using MLR, PLS and PCR without intercepts for better interpretation.

Few factors should be selected for the use of PCR. First, we calculated PCA using the data of the day (No. 18–42) excluded variable of ozone. The screen plot of the PCA suggests that four factors are relevant. We calculated PCR with cross-validation (leave-one-out) using the PLS_Toolbox 2.1 for MATLAB 6, Eigenvector Research, Inc., 830 Wapato Lake Road, Manson, WA 98831, www.eigenvector.com. Table 2 shows that four factors should be used for the correct model based on both \( X \) (all variables excluded ozone) and \( y \) (only the variable of ozone) blocks. The residual plots drawn for three and four factors also confirm that choosing four factors is correct.
The forward selection MLR in stepwise mode (without intercept) results the following equation:

\[ O_3 = 2.058 \times \text{TEMP} - 0.2537 \times \text{NO} + 1.479 \times \text{SO}_2 \\
- 4.547 \times \text{WIND} - 0.5505 \times \text{NO}_2 - 0.02820 \times \text{DIR} \]

\[ R = 0.99936; \quad F(6, 19) = 2461.5; \quad p < 0.00000; \]

\[ S = 1.690 \quad (2) \]

Table 2 compares the results of PCR, PLS and MLR methods.

A comparison suggests that the different kind of techniques provides various results considering interpretation. The coefficients for SO\(_2\), WIND and TEMP are all positive whereas the coefficients for NO and CO are negative. PCR and PLS coefficients are more realistic than the ones obtained by MLR, e.g. coefficient of NO\(_2\) is positive, which corresponds to the reaction (3). SUN (solar radiation) is not significant by MLR, small but positive by PCR and PLS. The coefficients of HUM and TEMP are both considerable by PCR and PLS.

Consequently, no clear-cut conclusions can be drawn. NO is an important variable according to the majority of methods, still it does not comply with the reaction:

\[ \text{NO}_2 + \text{O}_2 \rightarrow \text{NO} + \text{O}_3 \quad (3) \]

and the equation below:

\[ [\text{O}_3] = K \times [\text{NO}_2] \times [\text{O}_2]/[\text{NO}] \quad (4) \]

Our model does not take into account the effects of several other parameters, e.g., organic compounds (organic radicals) and meteorological variables. According to the reaction given above, the concentration of NO\(_2\) should increase (and at the same time the concentration of NO should decrease) the concentration of O\(_3\). PCR and PLS comply with this. The disadvantage of the
PCR and PLS is that they cannot provide physically relevant factors. However, they have a very important advantage: four factors have definite effect. The PCR and PLS models show a weak effect of SO\(_2\), and WIND against the MLR model (it is clear that e.g. WIND-effect should not be as strong as the MLR shows).

All methods, PCR, PLS and MLR are able to reproduce the measured data within 9 ppb.

4. Conclusions

PCA is able to classify both the variables and the cases (or events). The dominant pattern suggests that the day and night data are separated according to the different ozone-producing processes. In other words, PCA is a very useful tool for selection of properties (variables) and different quality processes. In this case, it seems that O\(_3\) is eliminated in different ways (with different reactions) during day and night.

Models for erasing of ozone were built using PCR, PLS and MLR. They can predict ozone concentration with an error below 5, 1 and 2 ppb level, respectively. Comparing these three methods, some differences can be observed:

- PCR and PLS show the effect of NO\(_2\) for the erasing of O\(_3\) according to the reaction NO\(_2\) + O\(_2\) = NO + O\(_3\) (see Table 3) correctly;
- MLR does not reveal effect of humidity (HUM), CO concentration and sun exposure (SUN);
- MLR, PLS and PCR do not mark the effect of CO and sun exposure (SUN);
- NO concentration and temperature (TEMP) are influential according to the three methods;
- MLR shows significant effect of NO\(_2\) concentration and temperature (TEMP), as well;
- altogether PCA and PCR or PLS are very useful tools for grouping the data and analyzing/evaluating environmental data without human assistance.

The ozone concentration near the ground level at a heavy traffic place can be described by a non-photothermal reaction model using PCA, PCR and PLS. In this simple chemical reaction phenomena NO, TEMP and humidity play important role as reactants and the temperature as a kinetic parameter.

Acknowledgment

This work was supported by the Hungarian Research Foundation: OTKA T 037684, T 046484 and KTM 16127.

References


