

Data Mining of the Relationship between Volatile Organic Components and Transient High Ozone Formation

Feng Gan and Philip K. Hopke[†]

(Department of Chemical Engineering, Clarkson University, Potsdam, NY 13699-5705, USA)

Abstract

The aim of this research is to identify the relationships between volatile organic components (VOCs) and transient high ozone formation in the Houston area. The ozone is not emitted to the atmosphere directly but is formed by chemical reactions in the atmosphere. In Houston, short-term (1 hour) sharp increases are observed followed by a rapid decrease back to typical concentrations. Automatic gas chromatographs (GC) are operated at several sites which cryogenically collect VOCs during an hour and then the compounds are flash evaporated into the GC for analysis. Chromatographic data for more than 65 VOCs are stored in analysis report text files. A program has been developed to read the amount of each component in the measurements such that a data set is generated that includes the concentrations of each VOC for each hourly sample. A subset of the data is selected that corresponds to the period of the positive ozone transient and these data are used in the data mining process.

Based on a chemical mass balance analysis, a linear model was established between the subset and the positive ozone transition. Non-negative least squares was used to calculate the regression coefficient of the VOCs that have the most significant positive relationship to the positive ozone transition. The results show that more attention might be paid to a lot of unknown VOCs, which have significant relationships to the transient high ozone formation.

Keywords: Data mining, Volatile organic components, Transient high ozone formation

[†]Correspondent author. Email: hopkepk@clarkson.edu

Introduction

The development of modern analysis instrument makes it easy to collect huge data in a short time. But, huge data collection of a long term measurement makes it not an easy thing to extract meaningful information either. Obtaining useful knowledge from the rapidly growing volumes of chemical data is being a challenge to chemists[1]. Although obtaining knowledge from a small data set or short term observation is still valid method, a new technology, named knowledge discovery in database (KDD) or data mining(DM) [2-5], is absorbing the attention of chemists [1, 6-9]. The aim of DM is to discover new knowledge from huge database. It is no doubt that it will play an important role in future chemistry. Details about DM can be found in references 2-5, we will not discuss it further here.

In this work, we will use DM method to environment data. The aim is to identify the relationships between volatile organic components (VOCs) and transient high ozone formation in the Houston area. Our interesting is in that the ozone is not emitted to the atmosphere directly but is formed by chemical reactions in the atmosphere. In Houston, short-term (1 hour) sharp increases are observed followed by a rapid decrease back to typical concentrations. Some components, such as ethylene, propylene etc., are thought to be the cause of the transient high ozone formation based on some measurements from automatic gas chromatographs (GC). But, in our opinion, it will be a good way to extract this information from historic GC data by using a reasonable method such as DM. As huge historic GC data is available, what we need to do is to design a DM process to extract the information from the data set.

Based on chemical mass balance, a linear model was established between the subset of the data and the positive ozone transition. Non-negative least squares [10-12] was used to calculate the regression coefficients of the VOCs that have the most significant positive relationship to the positive ozone transition. It will be a meaningful work in the understanding of the cause of transient high ozone formation.

Methodology

The O₃ formation reaction can be written as following:



where, R and P represent reactants and products respectively. r and p are the corresponding coefficients.

From chemical reaction theory, one knows that

$$O_3 \propto p_i P_i \quad \text{or} \quad O_3 = \alpha_i p_i P_i = \rho_i P_i \quad (2)$$

where, $\alpha_i > 0$ and $p_i > 0$, so $\rho_i > 0$.

Equation. (2) means that all other products can be represented in the form of O_3 . If we consider the chemical mass balance (CMB), eqn.(1) can be revised as

$$\sum_i r_i |\Delta R_i| = \Delta O_3 + \sum_i p_i \Delta P_i \quad (3)$$

According to eqn.(2), one gets

$$\sum_i r_i |\Delta R_i| = k \Delta O_3 \quad (4)$$

where,

$$k = \left(1 + \sum_i \frac{1}{p_i}\right) \quad (5)$$

In this work, eqn. (4) is used to data mining. As r and k are positive, non-negative least squares (NNLS) regression will be a proper method to calculate the regression coefficients, r_i/k . From these coefficients, the relationship of each component to the transient high ozone can be determined. Then the ones with significant t-test value will be the potential cause for the ozone formation.

Data Preparation

The data come from several auto-GCs operated in the Houston area as part of their PAMS program. Table 1 lists the sites. The automatic GC operated by cryogenically collecting VOCs during an hour and then the flash evaporating the compounds into the GC for analysis. The normal processing of these chromatogram results in concentration values for a set of 65 compounds that have been preselected based on their likely ability to contribute to production of high ambient ozone concentrations. We have reanalyzed all of the chromatograms in order to identify all of the peaks that could be detected. Chromatographic data for more than 65 volatile

organic components are stored in analysis report text files. Programs have been developed to read the amount of each component in the measurements such that a data set is generated that includes the concentrations of each VOC for each hourly sample. A subset of the data, which corresponds to the periods of rapid ozone formation (≥ 40 ppb/hr), is selected and then used in the DM process.

It must note that there are a number of unknown components in the chromatographs along with the known 65 VOCs. All unknown components are marked as UNKNOWN in this work. We distinguish them by numbers according to their locations among the known 65 VOCs. It is often the case that several unknown components are in the consecutive location. In this situation, we take these unknowns as a single species and add their amounts. Appendix is a table of the known VOCs.

All programs are written by using Matlab 12.1 and run on PC (RAM 512 MB).

Results and Discussion

Table 2 to 4 are the results of the NNLS, with only results with significant t-test values being listed. From the tables one can see that some VOCs with high photochemical reactivity, such as isoprene, propylene and ethylene, are listed. More interesting is that a number of the UNKNOWNs also have significant t-test values. It means that further research work is needed to identify the UNKNOWNs. Particularly, UNKNOWN24 should be paid attention to because it appears at two locations.

From the results one can see that time period is also important in DM process. Long time periods will contain more information. That is the reason why DM process usually need huge data set. Table 4 shows the results of Aldine area. Because of rather short time period, only two components have significant t-test values. Although these two components have high t-test values, the reliability is lowered. Tables 2 and 3 show the results with relatively long time period where the reliability is relatively high. For example, ethylene shows significant t-test value at both Clinton and Deerpark.

Conclusion

The results of this work show that DM can extract useful information from historic environmental data. It appears that attention needs to be paid to these unknown VOCs because

they have significant relationship to the transient high ozone formation.

References

1. L.M.C. Buydens, T.H. Reijmers, M.L.M. Beckers, and R. Wehrens,, Chemom. Intell. Lab. Syst., 49, (1999) 121-133
2. U. Fayyad, R. Uthurusamy, Commun. ACM, 39 (1996) 27-34
3. C. Glymour, D. Madigan, D. Pregibon, P. Smyth, Commun. ACM, 39 (1996) 35-41
4. W. H. Inmon, Commun. ACM, 39 (1996) 49-50
5. U. Fayyad, D. Haussler, P. Stolorz, Commun. ACM, 39 (1996) 51-57
6. Y.Z. Liang and F. Gan, Anal. Chim. Acta, 446, (2001) 115-120
7. Q. Guo, W. Wu, D.L. Massart, C. Boucon and S. De Jong,, Chemom. Intell. Lab. Syst., 61, (2002) 123-132
8. K. Kafadar, Chemom. Intell. Lab. Syst., 60, (2002) 127-134
9. A. Inselberg, Chemom. Intell. Lab. Syst., 60, (2002) 147-159
10. C.L. Lawson and R. J. Hanson, Solving Least Squares Problems, pp.158-173, Prentic-Hall, Englewood Cliffs, NJ.
11. D. Wang and P.K. Hopke, Atmospheric Environment, 23, (1989) 2143-2150
12. F. Gan and Y.Z. Liang, Anal. Sci., 16, (2000) 603

Table 1. Data source and time period

Data Source	Time Period
Aldine	Aug., 2000 – Nov., 2000
Deerpark	Jan., 1997 — Feb., 2000
Clinton	Oct., 1995 — Sep., 2000

Table 2. Non-negative least squares results of Clinton ($t_{\infty,95}=1.653$)

Component name	NNLS Coefficient	t
ISOPRENE	62.186	4.2744
UNKNOWN47	40.0609	3.0208
ETHYLENE	38.0545	2.116
ETHANE	38.0784	2.0305
UNKNOWN24	24.7581	1.9995
UNKNOWN46	27.7411	1.8862
UNKNOWN62	24.9812	1.7641

Table 3. Non-negative least squares results of Deerpark ($t_{\infty,95}=1.653$)

Component name	NNLS Coefficient	t
ETHYLENE	35.8644	3.9329
ISO-BUTANE	47.8275	3.917
1-BUTENE	26.9956	2.5109
UNKNOWN24	25.458	2.4113
PROPYLENE	20.0395	1.7173

Table 4. Non-negative least squares results of Aldine ($t_{120,95}=1.658$)

Component name	NNLS Coefficient	t
ISO-PENTANE	63.8691	2.7023
UNKNOWN36	19.7665	1.7942

Appendix

Component name and retention time

No.	Component Name	Retention Time(min.)
1	METHANE	6.48
2	ETHANE	7.055
3	ETHYLENE	7.746
4	PROPANE	9.864
5	PROPYLENE	17.143
6	ISO-BUTANE	19.99
7	N-BUTANE	21.356
8	ACETYLENE	23.314
9	TRANS-2-BUTENE	27.234
10	1-BUTENE	27.734
11	CIS-2-BUTENE	28.638
12	CYCLOPENTANE	30.085
13	ISO-PENTANE	30.409
14	N-PENTANE	31.347
15	1,3-BUTADIENE	33.31
16	2-METHYL-2-BUTENE	34.19
17	CYCLOPENTENE	34.286
18	TRANS-2-PENTENE	34.483
19	3-METHYL-1-BUTENE	35.128
20	1-PENTENE	35.54
21	CIS-2-PENTENE	36.18
22	2,2-DIMETHYLBUTANE	37.11
23	2,3-DIMETHYLBUTANE	37.881
24	2-METHYLPENTANE	38.059
25	3-METHYLPENTANE	38.169

26	ISOPRENE	39.881
27	4-METHYL-1-PENTENE	42.235
28	2-METHYL-1-PENTENE	42.969
	(Another Column)	
29	N-HEXANE	11.245
30	TRANS-2-HEXENE	12.594
31	CIS-2-HEXENE	12.811
32	METHYLCYCLOPENTANE	13.476
33	2,4-DIMETHYLPENTANE	13.688
34	BENZENE	15.658
35	CYCLOHEXANE	16.583
36	2-METHYLHEXANE	17.354
37	2,3-DIMETHYLPENTANE	17.499
38	3-METHYLHEXANE	18.053
39	2,2,4-TRIMETHYLPENTA	19.266
40	N-HEPTANE	20.118
41	METHYLCYCLOHEXANE	21.784
42	2,3,4-TRIMETHYLPENTA	23.952
43	TOLUENE	24.259
44	2-METHYLHEPTANE	24.985
45	3-METHYLHEPTANE	25.501
46	N-OCTANE	27.163
47	ETHYLBENZENE	30.199
48	M&P-XYLENE	30.667
49	STYRENE	31.736
50	O-XYLENE	31.945
51	N-NONANE	32.703
52	ISO-PROPYLBENZENE	33.579

53	a-PINENE	34.130
54	N-PROPYLBENZENE	35.02
55	M-ETHYLTOLUENE	35.441
56	P-ETHYLTOLUENE	35.567
57	1,3,5-TRI-M-BENZENE	35.684
58	b-PINENE	35.724
59	O-ETHYLTOLUENE	36.209
60	1,2,4-TRI-M-BENZENE	36.936
61	1,2,3-TRI-M-BENZENE	37.307
62	N-DECANE	37.463
63	M-DIETHYLBENZENE	39.176
64	P-DIETHYLBENZENE	39.471
65	N-UNDECANE	41.216

* All UNKNOWN is represented by its location in the table. For example, UNKNOWN1 is at the location between METHANE and ETHANE..