Environ. Sci. Technol. 1993, 27, 2430-2434

Source Identification for Multiple Chemical Exposure Using Pattern Recognition and Classification Techniques

Barton P. Simmons* and Robert C. Spear

Center for Occupational and Environmental Health, School of Public Health, University of California, Berkeley, California 94720

To characterize sources of exposure to organic solvent mixtures, breathing zone air samples were collected from workers in a printing/bookbinding plant. The analysis of air samples revealed complex patterns of exposure to organic solvents. Principal component analysis (PCA) plus classification and regression tree (CART) analysis identified sources of exposure and verified that exposure classification of workers by job type corresponded closely to exposures measured using personal air monitoring. The variable loadings for most principal components of air exposure matched identifiable combinations of solvent mixtures used in the workplace. PCA and CART models both accurately described the sources of multiple chemical exposures.

Introduction

Many important occupational and environmental problems are caused by exposure to chemical mixtures. Examples include petroleum products, synthetic chemicals such as polychlorinated biphenyls, various natural products such as lignins and tannins, industrial wastes, and combustion products. There now exist several techniques for the analysis of chemical mixtures, such as gas chromatography, high-performance liquid chromatography, ion chromatography, and inductively coupled plasma atomic emission spectroscopy. However, much of the data generated from these techniques is wasted, or languishes, for the lack of adequate methods for coupling the multivariate nature of these data to human exposure in meaningful ways.

A first step in the analysis of complex mixtures is to resolve the exposure into source contributions with known properties. Pattern recognition techniques have been successfully applied to many chemical mixture problems (*1*). The entire field of "source–receptor analysis" has developed to solve the problem of identification and quantitation of sources of ambient air particulates (*2*). This paper describes the use of pattern recognition techniques for characterizing occupational exposure to organic solvents. The advantages of this approach over conventional monitoring techniques are both an improved identification of the sources of exposure and an improved characterization of how these sources interact at the point of the individual.

Experimental Section

Description of the Mixed-Solvent Study. In a pilot study, principal component analysis was done on existing data from an earlier University of Quebec study on solvent exposure and health effects in printers (*3*). The purpose of this earlier study was to measure the worker exposure to solvents and to correlate the exposure to color vision loss. We showed that principal component analysis could successfully be used to improve the characterization of individual worker exposure. Based on the results of this preliminary work, a new study was planned.

The follow-up study was conducted in a university printing/reproduction facility which included shops for printing, photolithography, bookbinding, and polycopy, plus areas remote from these shops which served as control areas. All workers in the printing/reproduction areas were invited to participate in the study. Eighteen workers volunteered to participate. In addition, five workers in other parts of the same building participated in the study as controls. Each worker completed an initial work history questionnaire.

The air samples which were collected in this study are summarized in Table I.

Table I. Air Samples

work area	no. of workers	no. of air samples
polycopy	5	23
bookbinding	8	33
photolithography	2	9
printing	3	15
total	18	80

Table II. Generic Solvents by Work Area

product name	primary use
Varn 253 (Varn Products Co.)	print shop
Blanket Wash (Ernest Green & Son, Ltd.)	print shop, polycopy
Blankrola (AM International, Inc.)	print shop, polycopy
Deglazing (Multigraphics)	photocopy
Electrostatic (Abdick)	photolithography

Bulk Samples. Bulk solvent samples were collected for comparison with air samples. The generic solvents of greatest use are listed in Table II. Although the areas of primary use are also listed, it was common practice to allow workers the choice of solvents for a particular task. Therefore, the use of particular solvent products cannot be strictly assigned by work area.

To characterize potential sources, selected bulk samples were analyzed by gas chromatography/mass spectroscopy (GC/MS) using EPA method 8270 (*4*), with a DB-5 wide-bore capillary column.

Air Samples. Breathing zone samples were collected during normal working hours on each day for 1 week, using passive samplers (SKC 530 Anasorb CA) and charcoal tubes. The badges were removed and capped if the workers left the work area for lunch. The sampling rates provided by the manufacturer (SKC) were as follows: toluene 9.05 mL/min; perchloroethene 8.62 mL/min. Other solvents were assigned the toluene sampling rate of 9.05 mL/min. The adsorbent was extracted with carbon disulfide desorption, and the extracts were analyzed by gas chroma-

* Address correspondence to this author at his present address: Hazardous Materials Laboratory, California Department of Health Services, 2151 Berkeley Way, Berkeley, California 94704.

0013-936X/93/0927-2430$04.00/0

Table III. Results of Bulk Solvents GC/MS Analysis

solvent mixture	major components
Varn 253	C_7 aliphatic and aromatic hydrocarbons, including methylhexane, heptane, methylcyclohexane, and toluene
Blanket Wash	C_8–C_{10} hydrocarbon mixture, primarily substituted aromatics, e.g., trimethylbenzene
Blankrola	About 30% tetrachloroethene and 70% mixture of C_9–C_{11} aliphatic hydrocarbons, e.g., nonane and decane
Deglazing	20% tetrachloroethene and 80% dichloromethane
Electrostatic	2-propanol and 1,1,1-trichloroethane

tography with flame ionization detection (GC-FID) according to NIOSH method 1501 (*5*), using a 0.53-mm-diameter Supelco SPB-1 capillary column with the following conditions: injector temperature, 175 °C; column temperature, 35 °C for 5 min, 8 °C/min to 100 °C, hold for 5 min; detector temperature, 200 °C; nitrogen flow rate, 4 mL/min; makeup gas, 40 mL/min nitrogen. Analysis of backup windows showed that the sampling rate was 96% of expected. Bulk solvent samples were diluted with carbon disulfide and analyzed in the same manner as badge samples.

Data Analysis. Principal component analysis was done with SIMCA (*6*) using v-fold cross-validation to calculate statistical significance. Only statistically significant components were retained. Classification of sources was done using SIMCA plus classification and regression trees (CART) (*7*). CART analysis was done with n-fold cross-validation to estimate the misclassification rate.

Results

Bulk Solvents. A summary of GC/MS results is shown in Table III.

Air Samples. The total air concentration for each sample was calculated by summing the total of all peak concentrations. The limit of quantitation was 12.6 $\mu g/m^3$, calculated as toluene. Since the data were approximately log-normally distributed, values below the quantitation limit were substituted with $LOD/(2)^{1/2}$ (*8*); 12.6 $\mu g/m^3/(2)^{1/2} = 8.93\ \mu g/m^3$ was used for values below the quantiation limit. The analysis of badge samples produced data with up to 50 peaks measured in any one sample. This resulted in a data matrix of 50 peak concentrations and 105 samples. Results for control workers were consistently low, as expected, and were excluded from further analysis. The total air concentrations were approximately log-normally distributed, with a geometric mean of 1160 $\mu g/m^3$ and a geometric standard deviation of 3.40. Thus the data are consistent with the model in which the distribution of individual exposures is distributed log-normally and the distribution of mean concentrations for individuals is also distributed log-normally. This model has been used for the analysis of benzene exposures in the petroleum refining industry (*9*). Data were log-transformed prior to calculation of principal components.

Principal Component Analysis of Air Results. Six statistically significant principal components were calculated, which explained 75% of the total variance. Figure 1 shows a plot of the first two principal components (PCs), which explained 61% of the total variance. Samples for

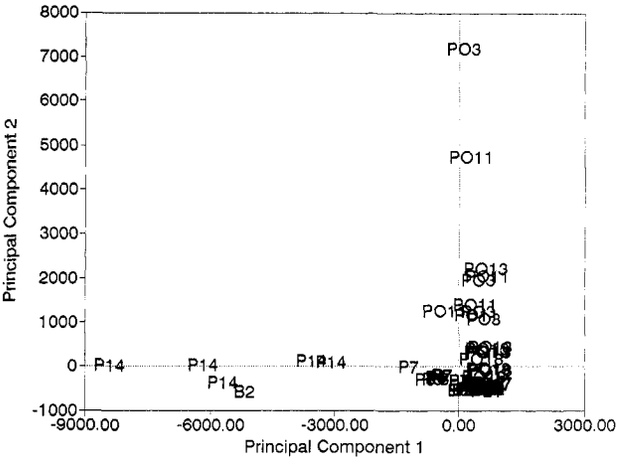

Figure 1. Sample scores for first two principal components: P, printing; PO polycopy; B, bookbinding; PL, photolithography.

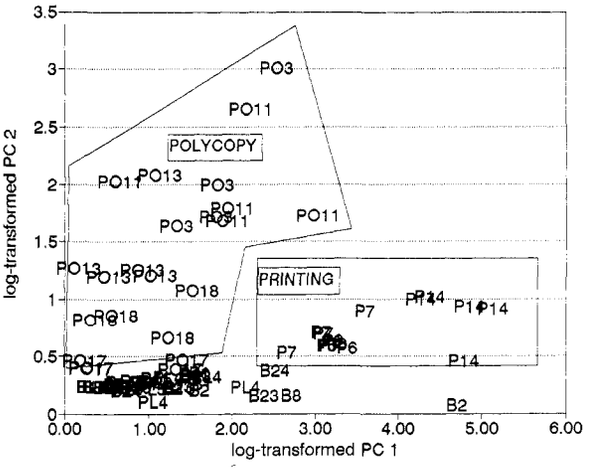

Figure 2. Log-transformed plot of first two principal components: P, printing; PO, polycopy; B, bookbinding; PL, photolithography.

printer 14 (labeled "P14"), and one bookbinding sample, B2, are separated from other samples by PC 1. In contrast, samples for polycopy workers tend to be separated from other samples by PC 2. Figure 2 shows the plot of log-transformed scores for PC 1 and PC 2, in order to resolve points near the origin. The separation of printer samples and polycopy samples is apparent.

The peak loadings were examined to explain the clustering of air samples. The peak loadings are the coefficients (also called β coefficients) which determine what weight will be given to each variable in the principal component. For example, peak 9 has a loading of −0.54 in PC 1, as shown in Figure 3. Figure 3 is a plot of the loadings for the first two principal components. A comparison of the worker chromatograms with the chromatograms for bulk solvents showed that peaks with large negative β 1 values represnt major components of Blanket Wash, a commercial solvent mixture. The plot shows that a negative PC 1 score is largely a measure of exposure to the components of Blanket Wash. In contrast, PC 2 is affected primarily by peak 19, which corresponds to perchloroethane (PCE, perchloroethylene), which was present in two of the solvent mixtures, Blankrola and Deglazing.

Looking at the loading plot in Figure 3 and the PC plots in Figures 1 and 2, it is apparent that exposure to Blanket Wash is a part of the exposure for printers and one B2 bookbinding sample. The chromatogram for the one

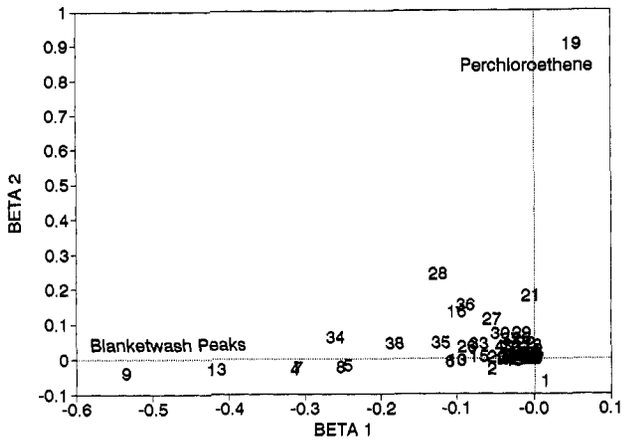

Figure 3. Peak loadings for first two principal components.

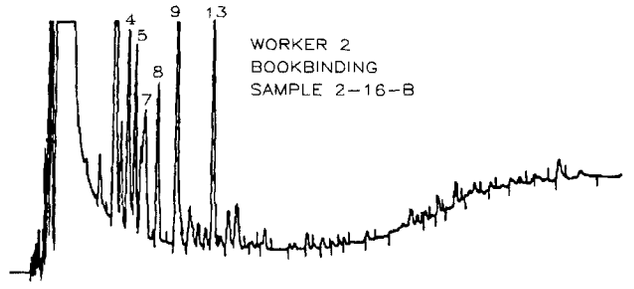

Figure 4. Chromatogram of anomolous bookbinding sample.

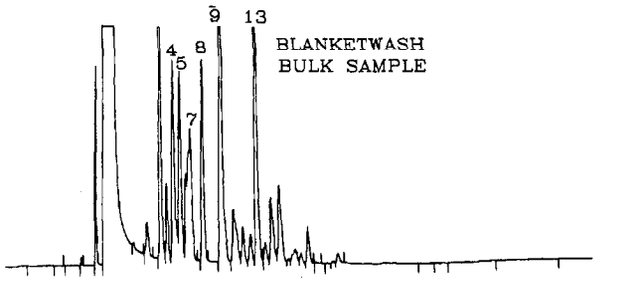

Figure 5. Chromatogram of Blanket Wash bulk sample.

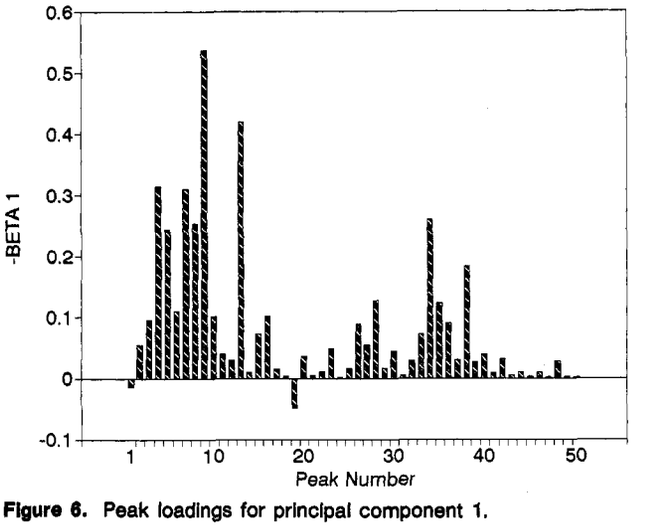

Figure 6. Peak loadings for principal component 1.

anomalous B2 sample is shown in Figure 4, and a chromatogram for the Blanket Wash bulk sample is shown in Figure 5. The comparison of the chromatogram for worker B2 and the chromatogram of the bulk Blanket Wash sample confirms that the pattern is undoubtedly from exposure to Blanket Wash vapors. In addition, the chromatogram for Blanket Wash confirms that the largest

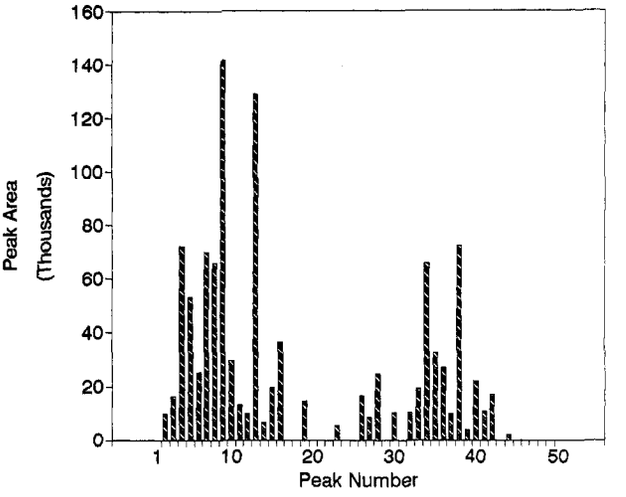

Figure 7. Combination of bulk solvents: Blanket Wash + 0.6 (Varn 253).

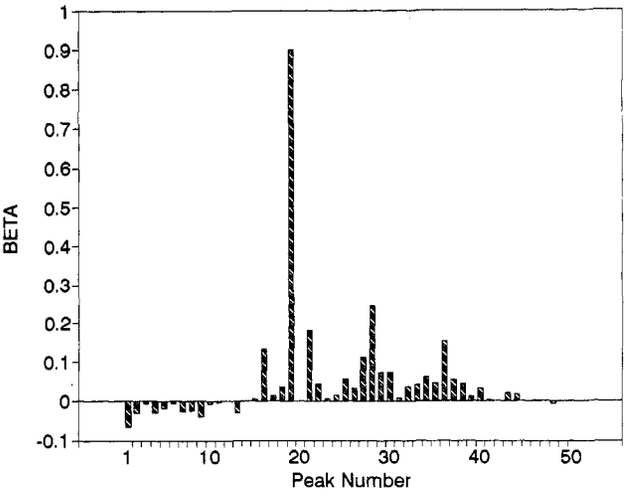

Figure 8. Peak loadings for principal component 2.

component are weighted strongly in PC 1. Figure 6 shows the loadings for all peaks in PC 1. As indicated above, Blanket Wash peaks are the highest weighted in PC 1. Another pattern is suggested and compares with the composition of Varn 253. The compositions of Blanket Wash and Varn 253 were mathematically combined in the ratio of 1:0.6 and produced the pattern in Figure 7. Thus PC 1 can be interpreted as a typical printer exposure to a mixture of solvent vapors from Blanket Wash and Varn 253, in the ratio of approximately 1:0.6.

The close match of bulk solvent chromatograms to air samples is due in part to the high volatility, or activity coefficients, of the solvents. Components with significantly lower volatility would be expected to distort this relationship. In this case, the close match of PC loadings to solvent composition confirms the physical interpretation of the principal components. Similarly, the PC 2 peak loadings (Figure 8) compare reasonably well with the chromatogram for Blankrola (Figure 9), but not with the chromatogram for any other solvent mixture. Perchloroethene, peak 19, dominates both the PC 2 loadings and the composition of Blankrola.

Examination of PC 3 revealed one previosly unidentified polycopy exposure which did not match any generic solvent patterns. PC 4 proved to be a measure of exposure to a photolithography solvent mixture. The probable sources of PCs are summarized in Table IV.

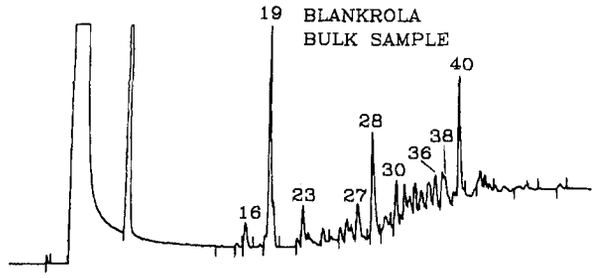

Figure 9. Chromatogram for Blankrola bulk solvent.

Table IV. Probable Sources of Principal Components

PC	weighted peaks[a]	probable source
1	9, 13, 4, 7, 8, 5	Blanket Wash
	34, 38, 28, 35, ...	Varn 253
2	19, 28, 21, 36, 16, ...	Blankrola
3	36, 37, 38, ...	unidentified
4	1, 2, 36, ...	unidentified (photolithography solvent)

[a] In order of loading.

Table V. Printing Samples Training Set for SIMCA (n = 15)

PC	variance	variance explained (%)	cumulative variance explained (%)
0	155 369		
1	13 970	141 399 (91.009)	141 399 (91.009)
2	5 761	8 209 (5.283)	149 608 (96.292)

Table VI. Results for SIMCA Classification of Samples

	actual work classif	
SIMCA classif	printer	nonprinter
printer	13	2
nonprinter	2	64

The initial analysis indicated that, with a few exceptions, the samples for each worker tended to cluster together. That is, the between-day pattern variability was less than the between-worker pattern variability. PCA was also performed on mean air concentrations for each worker, averaged over the work week, with similar results.

Classification of Exposures. Soft independent modeling of class analogy (SIMCA) classified samples with principal components calculated from objects of known class membership—a training set. Principal component analysis of the printing air samples produced two statistically significant principal components, as shown in Table V. As is shown, a model using two principal components can represent over 96% of the variance for the training set. By use of this model, the samples were classified as shown in Table VI. Thus the SIMCA classification accurately classified 77 of 81, or 95% of the samples.

Classification Trees. As an alternative to SIMCA analysis, the same air exposure data were analyzed using classification and regression tree (CART) analysis. In the CART nomenclature, each peak area is a variable and each sample is an observation. Two-thirds of the data set was used as a training set to grow classification trees, that is, to develop data-based rules that would classify a sample as belonging to a particular job category. The remaining data were then used as a test set to determine the relative misclassification for each tree. The tree with the minimum

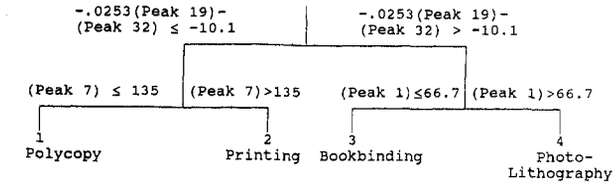

Figure 10. Classification tree from CART analysis.

Table VII. Misclassification by Class with CART Analysis

	test sample		learning sample	
class	no. of cases	no. misclassif	no. of cases	no. misclassif
printing	5	1	10	0
polycopy	9	0	14	0
photolith	5	3	5	2
bookbinding	4	0	29	1
total	23	4	58	3

misclassification was identified by CART as the best classification tree. The tree shown in Figure 10, with three nodes, had the minimum misclassification. The splitting rules were based on linear combinations of peak concentrations. The rules are listed at their respective nodes in Figure 10.

A remarkable feature of the classification tree is the parsimonious use of data. Of the 50 peaks available, the classification rules used only four peaks, including one linear combination of two peaks, although a large number of linear combinations were possible (If only binary combinations are considered, for example, the number of possible combinations is 1225).

Table VII compares the classification of workers by their area of work and the classification by CART. In addition to the test sample technique, a v-fold cross-validation technique was also used for tree construction and evaluation, with similar results. Overall, the tree classified 91% of the cases correctly. For those cases which were assigned to other classes, an examination of the chromatograms revealed some interesting features. Only one printing sample was misclassified. An examination of its peak pattern showed that the exposure was relatively low overall, and several peaks typical of printing exposure, including peak 32, were not detected.

The misclassification of 5 of 10 of the photolithography samples was apparently due to the absence in some samples of peak 1, which was used for classification. This result is similar to the PCA result, which did not distinguish some photolithography samples from bookbinding samples.

One bookbinding sample, 2-16-B, was classified as a printing sample. This confirms the principal component analysis, which found that 2-16-B clustered with printing samples because of the Blanket Wash exposure in that sample.

Thus, although SIMCA and CART used very different approaches, the results were similar, both in the classification of samples and the peaks used for classification.

Conclusion

The conclusions from this analysis are as follows: (1) pattern recognition techniques can identify characteristic patterns of mixed chemical exposure; (2) because these

patterns of exposure are based on actual multiple measurements of exposure, they provide a more accurate classification of exposure than does job classification; (3) classification becomes less effective as exposures approach the limits of sampling and analysis. It should be noted that more sensitive analytical techniques, e.g., adsorbent tubes with thermal desorption, can provide sufficient data for classification of low-level (nonoccupational) exposure; (4) these techniques can successfully reduce multivariate exposure measurement to a few summary variables.

Acknowledgments

We thank Donna Mergler, Suzanne Bélanger, Luc Dallaire, and Chantal Jetté of the University of Quebec at Montreal for assistance in field work and Liza Finley for assistance in laboratory work. Laboratory work was conducted at the University of California Environmental Engineering and Health Sciences Laboratory. Supported by National Institute for Occupational Safety and Health Grant R03 OH02555-01.

Literature Cited

(1) Jurs, P. C. *Science* **1986**, *232*, 1219–1224.
(2) Gordon, G. E. *Environ. Sci. Technol.* **1988**, *22*, 1132–1142.
(3) Mergler, D.; Bélanger, S.; de Grosbois, S.; Vachon, N. *Toxicology* **1988**, *49*, 341–348.
(4) *Test methods for Evaluation of Solid Waste*; *Physical/Chemical Methods*, 3rd ed; SW-846; U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response, U.S. Government Printing Office: Washington, DC, 1986.
(5) *Methods of Analysis*, 3rd ed.; U.S. National Institute for Occupational Safety and Health, U.S. Government Printing Office: Washington, DC, 1986.
(6) Wold, S.; et al. Multivariate Data Analysis in Chemistry. In *Proceedings NATO Advanced Study Institute on Chemometrics*; Nowalshi, B. R., Ed.; Cosenza, Italy, September 1983, Reidel: Dordrecht, Holland, 194; pp 17–95.
(7) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C.; *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
(8) Hornung, R. W.; Reed, L. D. *App. Occup. Environ. Hyg.* **1990**, *5* (1), 46–51.
(9) Spear, R. C.; et al. *Appl. Ind. Hyg.* **1987**, *2* (4), 155–163.

Received for review January 14, 1993. Revised manuscript received June 1, 1993. Accepted June 14, 1993.