2000

Multivariate Correlation between Concentrations of Selected Herbicides and Derivatives in Outflows from Selected U.S. Midwestern Reservoirs

R. Tauler

D. Barcelo

E. Michael Thurman
U.S. Geological Survey

Multivariate Correlation between Concentrations of Selected Herbicides and Derivatives in Outflows from Selected U.S. Midwestern Reservoirs

R. TAULER,*,† D. BARCELO,‡ AND E. M. THURMAN§

Department of Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona 08028, Spain, Department of Environmental Chemistry, CID-CSIC, Jordi Girona Salgado 18-26, Barcelona 08034, Spain, and U.S. Geological Survey, 4821 Quail Crest Place, Lawrence, Kansas 66049-3839

Multivariate correlations between the concentrations of selected herbicides and herbicide derivatives in outflows from selected reservoirs in the Midwestern United States for April 1992 through September 1993 were investigated using principal component analysis (PCA) and multivariate curve resolution (MCR). Two independent sources for alachlor ethanesulfonic acid, one major source related to spring flush and seasonal runoff and another minor source related to groundwater, were identified using PCA. Results of MCR provided a semiquantitative interpretation of the environmental sources of the observed herbicide concentrations in reservoir outflows and allowed the examination of their temporal and geographical distributions. Samples with higher herbicide concentrations were collected from reservoirs in Indiana and Ohio, especially during the late spring and summer.

Introduction

Agricultural practices in the Midwestern United States have introduced herbicides into surface water and are a major concern for water quality (*1*−*3*). Approximately 75% of all premergent herbicides used in the United States are applied to row crops in an 11-state area, called the Corn Belt (*4*). Widespread detection of herbicides such as alachlor, atrazine, cyanazine, and metolachlor, all of them used extensively on corn, has occurred in monitoring and reconnaissance studies of surface water in the Mississippi River Basin area. Not only these four primary herbicides but also their derivatives and metabolites have been detected frequently in surface water.

Extensive sample analysis in the Midwestern United States area has been performed by the U.S. Geological Survey (USGS), and the results of this analysis provide an excellent data set for investigations related to temporal and geographical contribution of herbicides and herbicide derivatives in the Midwestern United States. Previous investigations using these data sets have been devoted to the study of the occurrence, formation, and transport of some of the herbicides and derivatives (*5*, *6*) and to the proposal of the deethylatrazine/atrazine ratio as a new indicator of the onset of the spring flush of herbicides into surface water (*7*). All these previous studies have been performed considering each of the measured variables (herbicide concentrations) individually and looking only at their pairwise correlations. When the number of variables is high and there is partial correlation between the different measured variables, the extraction of environmental information from this individual analysis of each variable becomes troublesome. In the present work, a deeper study of the multivariate correlations between the concentrations of the herbicides and their derivatives is proposed using two multivariate exploratory data analysis techniques, principal component analysis (PCA) (*8*) and multivariate curve resolution (MCR) (*9*, *10*). Using these two chemometric multivariate techniques, the extraction of hidden environmental information and especially the detection of multicomponent environmental sources of these herbicides are possible.

PCA is a frequently used multivariate technique that provides a powerful tool for data compression, exploration, and interpretation. PCA allows the investigation of the variance sources present in a multivariate data set using a reduced set of orthogonal variables or principal components (PCs), which are a linear combination of the original measured variables.

As it has been stated in previous works (*11*, *12*), the apportionment and environmental-source identification from environmental data sets are problems similar to the species resolution problem in spectrometric mixture analysis. In both cases, the goal of the analysis is the identification of the sources of data variance and the resolution of the profiles of these sources. The pure component spectra resolved in mixture spectrochemical analysis are analogous to the composition profiles resolved in environmental analysis. The concentration profiles in mixture spectrochemical analysis are analogous to the contribution profiles in environmental analysis. This analogy shows that methods such as MCR, which were developed initially for the analysis of chemical processes monitored spectroscopically, can be applied also to the resolution of environmental sources from environmental data sets such as those presented in this paper.

Whereas PCA performs a pure mathematical decomposition of the data imposing constraints, like orthogonality, not fulfilled by true data variance sources, MCR performs a similar data decomposition using natural constraints, like nonnegativity, fulfilled by the true data variance sources. In this paper, the two approaches are used in a complementary way, and the differences of their application are discussed.

Additionally, an interesting aspect of many environmental data sets is their three-way data structure (*13*). This means that the data sets can be ordered using three modes, ways, or orders of measurement that can be, for example, the measured variables in each sample (what constituent concentrations are measured), where these samples were measured, and when these samples were measured.

The goals of the work presented in this paper are the identification of the environmental sources causing the observed data variation in the data set under study and the determination of the composition profiles of these sources and their temporal and geographical contributions. In this way a semiquantitative apportionment of the source contributions for each of the analyzed samples may be possible.

* Corresponding author fax: 34 93 4021233; e-mail: roma@quimio.qui.ub.es.
† University of Barcelona.
‡ CID-CSIC.
§ U.S. Geological Survey.

10.1021/es000884m CCC: $19.00
Published on Web 07/08/2000

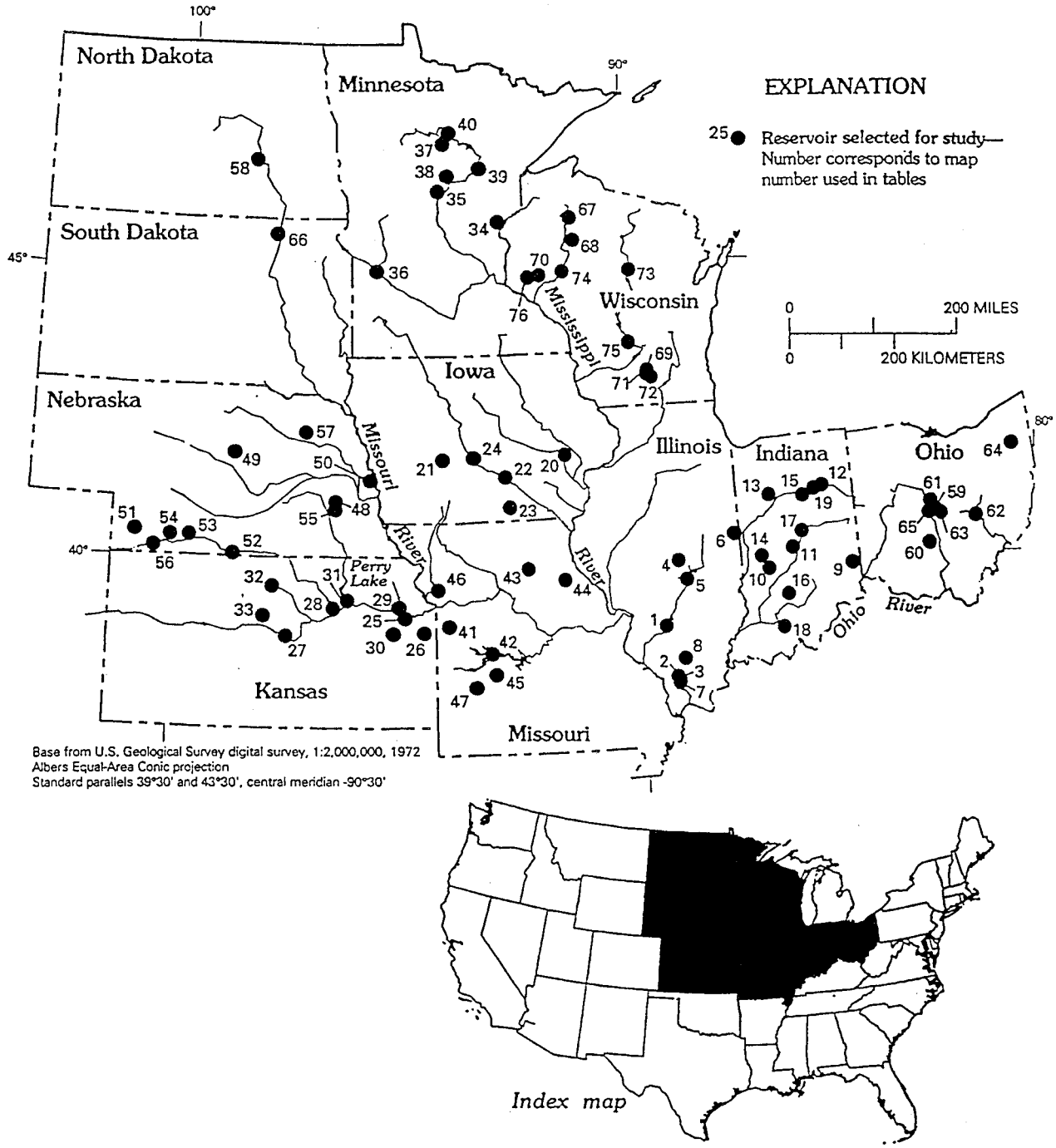

FIGURE 1. Location of study area and collected reservoirs in the Midwestern United States from which outflow samples were selected from April 1992 through September 1993.

To achieve these goals, PCA and MCR methods were applied and extended to the analysis of three-way data.

Experimental Data

The study area (Figure 1) comprises about 720 000 km^2 of land in 11 states (Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin) that drain to the Ohio, Upper Mississippi, and Lower Missouri Rivers. Outflows from 76 reservoirs were sampled eight times (approximately bimonthly) from April 1992 through September 1993. The timing and frequency of these samples made it possible to determine approximately when maximum and minimum concentrations of herbicides occurred in the reservoir outflow. Samples collected at the beginning of the study (late April or early May 1992) were collected before significant postplanting reservoir discharge occurred. Samples collected during June or early July 1992 were collected after significant postplanting runoff and flushing of the reservoirs had occurred. Samples collected in September 1993 were collected following the 1993 flood. Further details about selection of reservoirs, sample collection methods, and sample preparation are given elsewhere (*14*, *15*).

Herbicide samples were analyzed at the USGS laboratory in Lawrence, KS. The analysis included nine herbicides (alachlor, ametryn, atrazine, cyanazine, metolachlor, metribuzin, prometon, propazine, and simazine) as well as two atrazine derivatives (deethylatrazine and deisopropylatrazine) and three cyanazine derivatives (cyanazine amide, deethylcyanazine, and deethylcyanazine amide). The meth-

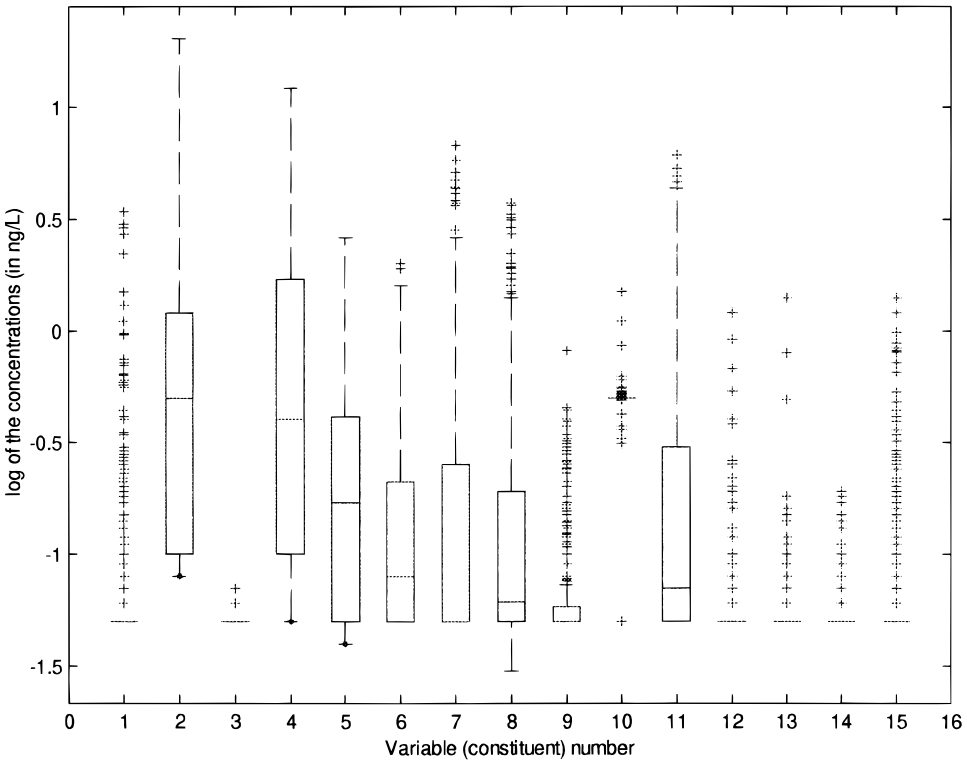

FIGURE 2. Box plots of the log of the measured variables or concentrations of the constituents (herbicides and derivatives). Each variable has 608 measured values (total number of measured samples). Identification of variables: 1, alachlor (ala); 2, alachlor ethanesulfonic acid (alachlor ESA); 3, ametryn; 4, atrazine (atraz); 5, deethylatrazine (DEA); 6, deisopropylatrazine (DIA); 7, cyanazine; 8, cyanazine amide; 9, deethylcyanazine; 10, deethylcyanazine amide; 11, metolachlor (metol); 12, metribuzin; 13, prometon; 14, propazine; 15, simazine. For each variable, the box has lines at the lower quartile (≤25%), median (≤50%), and upper quartile (≤75%) values. The whiskers are the lines extending from each end of the box to show the extent of the data up to 1.5 times the interquartile range (IRQ). Outliers are marked with + symbols.

ods of Thurman et al. (*16*) were used for herbicide GC/MS analysis. In addition, the ethanesulfonic acid derivative of alachlor (alachlor ethanesulfonic acid, ESA) was isolated by solid-phase extraction (SPE) and analyzed by enzyme-linked immunosorbent assay ELISA (*17*). Further details about analytical determinations, quality assurance, and analytical results are given elsewhere (*14*, *15*).

The whole data set consisted of the concentrations (μg/L) of 15 herbicides and herbicide derivatives in samples from 76 reservoir outflows throughout the Midwestern Corn Belt (Figure 1). The whole data set had 9120 entries or constituent concentrations that could be arranged in different ways. Box plots of the concentrations measured for each constituent in samples from reservoir outflows at different collection times are given in Figure 2. Larger concentration variations were observed for alachlor ESA, atrazine, deethylatrazine (DEA), desisopropylatrazine (DIA), cyanazine, cyanazine-amide, and metholachlor (variable numbers 2, 4−8, and 11, respectively, in Figure 2). As there were three different identification indexes for each concentration value (constituent, reservoir, and collection time), the data could be ordered in a three-way data structure (data cube) as shown in Figure 3.

Methods

Data Pretreatment. In the work presented in this paper, data were arranged to include one data set or data matrix per reservoir, with 8 row samples (sample collection times) and 15 column variables (constituents). This gave 76 data sets or data matrixes (76 reservoirs). These data structure may be arranged in a three-way data cube structure of dimensions 76 × 88 × 15, or they could be arranged in a column-wise augmented matrix of dimensions 608 × 15, i.e., with 608 row samples and 15 column variables (Figure 3). Since both PCA and MCR data decompositions are able only to work with two-way augmented data matrixes, the second data arrangement was used in this paper.

Two problems considered before the multivariate data analysis was begun were the large number of values less than the limit of detection and the presence of missing values. Values less than the limit of detection were assumed to be positive and equal to either the limit of detection or to zero. Approximately one-half of the entries in the original data were less than the limit of detection (approximately 52%). In the calculations, these values were set equal to the limit of detection values for that variable (constituent concentration) or to zero. Both strategies were tested and gave similar results. Values less than the limit of detection were not distributed equally among variables. For variables 1, 3−9, and 11−15, the limit of detection was equal to 0.05 μg/L. For variable 2 (alachlor ESA), the limit of detection was 0.1 μg/L, and for variable 10 (deethylcyanazine amide), the limit of detection was 0.5 μg/L. Variable 10 was specially problematic because it had a large number of missing values and a large number of values less than the limit of detection (96%). Only 26 of the 608 values were higher than the limit of detection of this variable. Variable 3 (ametryn) was also very problematic because only two samples contained concentrations greater than its limit of detection. Other variables with few values greater than limit of detection were variables 12−15.

Missing values were handled using the MATLAB function "missdat" of the MATLAB PLS Toolbox (*18*). Using this method, missing values were set to zero, and a PCA model was calculated for the whole data set. The missing values

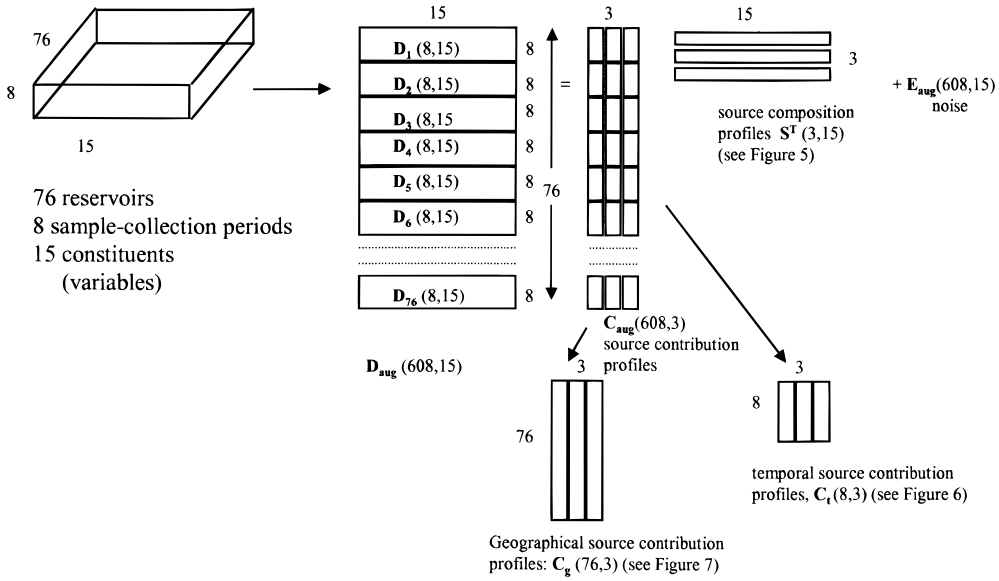

FIGURE 3. Three- and two-way augmented data arrangements. Concentration of 15 constituents (variables) were measured in 608 samples collected at 76 reservoir outflows from 11 states (see Figure 1) during eight sample collection time periods.

TABLE 1. Correlation between Nine Important Variables[a]

	alachlor (1)	alachlor ESA (2)	atrazine (4)	DEA (5)	DIA (6)	cyanazine (7)	cyanazine amide (8)	deethylcyanazine amide (9)	metolachlor (11)
alachlor (1)	1.00								
alachlor ESA (2)	0.69	1.00							
atrazine (4)	0.70	0.68	1.00						
DEA (5)	0.61	0.73	0.92	1.00					
DIA (6)	0.63	0.73	0.86	0.92	1.00				
cyanazine (7)	0.55	0.57	0.69	0.67	0.78	1.00			
cyanazineamide (8)	0.52	0.62	0.69	0.71	0.80	0.85	1.00		
deethylcyanazine amide (9)	0.45	0.47	0.60	0.50	0.66	0.84	0.83	1.00	
metolachlor(11)	0.81	0.73	0.84	0.80	0.81	0.65	0.64	0.51	1.00

[a] Variables with most of the values larger than detection limits. In parentheses, the identification number of the variables (see caption of Figure 2) is given.

then are replaced with those values that were most consistent with the PCA model. A new model was then recalculated, and the process was repeated until the estimates of the missing data converged. From the 9120 entries in data matrix $\mathbf{D}_{aug}$, 284 were missing (3.1% of the total). The contribution of these missing values was not uniform. For some variables, like variables 8−10 (that is cyanazine derivatives), the number of missing values was very high; 94 missing values from 608 values for these three variables (~15% of the total) were found. For cyanazine (variable 7) and metolachlor (variable 11), only one missing value was present. Looking at the original data, these values usually corresponded to the measurements in samples collected in early winter (January) or summer (June or July), but all the reservoir outflows were not the same. The missing values were evaluated using a PCA with a model of five components, which explained practically all the experimental data variance (99.9%).

It is obvious from this preliminary study that the more reliable variables are numbers 1, 2, 4−9, and 11 (see caption of Figure 2). Accordingly, the whole data analysis was performed either using all variables (data matrix **D**) or using only the nine more informative variables (data matrix $\mathbf{D}_r$). As all the constituent concentrations were measured in the same scale units and apportionment of source contributions was intended, initial data analysis was performed without any scaling or mean centering. Comparison of the results obtained in this way with those obtained using data scaling was performed also. Because a large number of the data values were near the detection limit, log transformation of the concentrations also was examined, and the results were compared.

The pairwise correlations between two variables were also preliminary investigated to see the relationships between the variations of the different constituent concentrations in the different samples analyzed. This was accomplished by calculating the correlation coefficients between all the values corresponding to two selected variables (Table 1).

Linear Model and Principal Component Analysis. The basic assumption in this study was that most of the observed data variance followed a linear model with a reduced number of components or environmental sources. Each source was defined by a particular composition profile describing the relative amounts of the different correlated variables (herbicide and derivative concentrations). This assumption is analogous to the usual assumption in spectrometric mixture analysis made on the basis of Beer's absorption law, where the absorption measured at different wavelengths is additive (linear) and highly correlated.

The following model and equations were assumed:

$$\mathbf{D}_i = \mathbf{C}_i\mathbf{S}^T + \mathbf{E}_i \quad i = 1, ..., 76 \tag{1}$$

where $\mathbf{D}_i$ is one of the individual data matrixes obtained when the 15 herbicide and derivative concentrations (variables) were measured in the eight (time) collection samples of water from reservoir i outflow. There are a total of 76 $\mathbf{D}_i$ matrixes of dimensions 8 × 15. $\mathbf{C}_i$ is the matrix of the temporal source

contribution profiles to reservoir i. This matrix has the dimensions of $8 \times N$, where N is the number of sources detected during data analysis. Thus, this model assumes that there are also a total of 76 $\mathbf{C}_i$ matrixes, one for each reservoir; that is, the temporal source contributions are not exactly equal in all the samples of reservoir outflow. Matrix $\mathbf{S}^T$ gives the composition of the N detected sources, and it has dimensions of $N \times 15$. The model assumes that the composition of each source is unique, i.e., that a particular source is defined by a unique composition of herbicide and derivative concentrations. Finally, matrix $\mathbf{E}_i$ has the residual data variations not modeled by the N detected sources, and it has the same dimensions as $\mathbf{D}_i$. In Figure 3, a detailed description of the data structue and linear model are given for the case that the number of resolved components is equal to three, $N = 3$. To ensure that the $\mathbf{S}^T$ matrix in eq 1 was common for all $\mathbf{D}_i$ data sets, the analysis was performed simultaneously for all the 76 $\mathbf{D}_i$ matrixes using:

$$\begin{pmatrix} D_1 \\ D_2 \\ \cdots \\ \cdots \\ D_{72} \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \\ \cdots \\ \cdots \\ C_{72} \end{pmatrix} S^T + \begin{pmatrix} E_1 \\ E_2 \\ \cdots \\ \cdots \\ E_{72} \end{pmatrix} \quad (2)$$

or in a more compact way:

$$\mathbf{D}_{aug} = \mathbf{C}_{aug}\mathbf{S}^T + \mathbf{E}_{aug} \quad (3)$$

where $\mathbf{D}_{aug}$ and $\mathbf{C}_{aug}$ are the corresponding column-wise augmented data and source contribution matrixes for the 76 samples of reservoir outflow corresponding to the data arrangement given in Figure 3.

The mathematical problem stated by model eqs 1−3 and Figure 3 can be summarized in the following way. Given data matrixes for each reservoir $\mathbf{D}_i$, find the temporal source contributions defined in matrixes $\mathbf{C}_i$ and the source composition profiles defined in matrix $\mathbf{S}^T$. First, the number of significant contributions to the whole data variance, N, needs to be estimated. Obviously, the analysis will focus on the major and distinct sources of data variance and not on the small contributions coming from multiple minor sources of data variation. Hence, for a model with a particular number of contributions, N, the residual matrixes, $\mathbf{E}_i$, will have still a substantial percentage of unexplained data variance coming from these multiple minor, unknown sources. This situation is clearly different from the situation usually encountered in the mixture analysis of spectrometric data (*9*, *10*) where most of the data variance can be explained by the selected model. In this paper, the number of resolved sources, N, was obtained from the amount of data variance they explain. Only those components explaining an appreciable amount of data variance (approximately >2%) were considered.

For a particular number of components, the matrix decomposition using eqs 1−3 was not unique because there was rotational and scale freedom in the unconstrained solutions (*10*). This meant that there was an infinite number of possible solutions if no constraints were set during the linear data matrix decomposition formulated by these equations.

One of the more popular and used matrix data decompositions is PCA. In particular, the PCs identified by PCA are linear combinations of the original variables, which are orthonormal (orthogonal and normalized to unit length) and explain maximum variance. The goal of PCA is to represent the variation present in many variables using a small number of components or factors. A new row space is constructed in which to plot the samples by redefining the axes using factors rather than the original measured variables. The new axes, referred to as principal components or PCs, allow investigation of data matrixes with many variables and the viewing of the true multivariate nature of the data in a relatively small number of dimensions. With this new view, natural structures in the data can be identified.

When PCA is used, the matrix related to the source contributions, $\mathbf{C}_{aug}$, is called the scores matrix, and the matrix related to the source composition, $\mathbf{S}^T$, is called the loadings matrix. Under the two strong constraints of orthonormality and explained maximum variance, the matrix decomposition in eqs 1−3 is unique. However, the solutions for $\mathbf{C}_{aug}$ and $\mathbf{S}^T$ are pure mathematical solutions that do not correspond, in general, with physical solutions. Most of the work in the environmental literature concerning multivariate source apportionment is based on the PCA matrix decomposition followed by appropriate rotation and interpretation steps (*19*).

MCR-Alternating Least Squares. In this paper, a different approach is proposed. The new approach has already been applied to the input characterization of sedimentary, organic chemical markers in the north eastern Mediterranean Sea (*12*). To limit the number of possible solutions in the matrix decomposition proposed in eq 3 and to find solutions that were more easily interpretable from a physical point of view, the multivariate curve resolution (MCR) method, initially developed for the mixture analysis of spectrometric evolutionary processes (*9*, *10*), is proposed. This method decomposes the experimental data matrix using a constrained, alternating least-squares (ALS) algorithm that can be summarized in the following equations:

$$\mathbf{C}_{aug} = \mathbf{D}_{aug}(\mathbf{S}^T)^+ \quad (4)$$

and

$$\mathbf{S}^T = (\mathbf{C}_{aug})^+\mathbf{D}_{aug} \quad (5)$$

where $(\mathbf{S}^T)^+$ and $(\mathbf{C}_{aug})^+$ are the least-squares estimations of the pseudoinverse (*20*) of $\mathbf{S}^T$ and $\mathbf{C}_{aug}$ matrixes. Equations 4 and 5 are solved iteratively under nonnegativity constraints (*21*):

$$\mathbf{C}_{aug} \geq 0 \text{ and } \mathbf{S}^T \geq 0 \quad (6)$$

To start the iterative process, initial estimations are needed either for $\mathbf{S}^T$ or for $\mathbf{C}_{aug}$. In this paper, initial estimates for $\mathbf{S}^T$ were obtained from detection of pure or more selective variables or samples using a similar approach to that used in the SIMPLISMA method (*22*).

In contrast to PCA, the profiles obtained for $\mathbf{C}_{aug}$ and $\mathbf{S}^T$ were directly interpretable because they referred to physical values. $\mathbf{S}^T$ gave the source compositions (relative constituent concentrations in the sources), and $\mathbf{C}_{aug}$ gave the temporal and geographical source contribution for each sample. The conditions under which the ALS matrix decomposition using eqs 4−6 were unique depended on data selectivity (unique source compositions or source contributions, see ref *10*) and on local rank conditions (number of components needed to explain the variance of the different subsets of samples and (or) variables, see ref *23*). These conditions for unique solutions of two-way data decompositions have been analyzed in detail in previous works related to chromatographic and spectrometric mixture analysis data (*9*, *10*, *24*), and they can be also extended to environmental data.

Whereas the matrix of source composition $\mathbf{S}^T$ was directly interpretable giving the relative concentrations of each constituent (herbicide and derivatives) on each of the different ALS-resolved environmental sources, the matrix of the source contributions $\mathbf{C}_{aug}$ needed some rearrangement before its information be interpretable because this matrix

TABLE 2. PCA Results Percentage of Accumulated Explained Variance[a]

matrix	PC1	PC2	PC3
D (15 variables)[b]	84.6	94.4 (9.8)[c]	96.8 (2.4)
$\mathbf{D}_r$ (9 variables)[d]	85.8	95.7(9.9)	98.2 (2.4)
D scaled	99.5	99.8 (0.2)	99.8 (0.04)
$\mathbf{D}_r$ scaled	79.7	87.74(8.1)	91.7 (3.9)
D autoscaled	51.0	59.0 (8.1)	66.2 (7.1)
$\mathbf{D}_r$ autoscaled	73.6	84.6 (10.9)	89.8 (5.3)
D log transformed	91.4	97.0 (5.6)	97.9 (0.9)
$\mathbf{D}_r$ log transformed	89.5	95.6 (6.2)	97.2 (1.5)

[a] The percentage of explained data variance for a particular number of components is calculated using % var $= [\sum_{i,j}(d^e_{i,j} - d^c_{i,j})^2]/[\sum_{i,j} d^e_{i,j}{}^2] \times 100$, where $d^e_{i,j}$ is the experimental data values, and $d^c_{i,j}$ is the corresponding calculated values using N components in the PCA model. [b] Matrix **D** with all 15 variables. [c] In parentheses, the percentage of nonaccumulated explained variance for that particular component is given. [d] Matrix $\mathbf{D}_r$ (reduced) with 9 important variables (variable numbers 1, 2, 4, 5, 6, 7, 8, 9, and 11; see caption of Figure 2).

had the temporal and geographical information mixed. From the way the different data matrixes $\mathbf{D}_i$ were joined to build the augmented matrix $\mathbf{D}_{aug}$ (see Figure 3), every column of matrix $\mathbf{C}_{aug}$ corresponded to a resolved environmental source contribution and had 608 elements (8 sampling periods × 76 sampled reservoirs). Every one of the resolved contribution profiles in matrix $\mathbf{C}_{aug}$ could be folded and plotted in two different ways, showing the temporal or the geographical contribution. Average values showing the average temporal contributions and the average geographical contributions were also possible. As it is shown in Figure 3, the 76 contribution values obtained for each sample collection period may be averaged to give matrix $\mathbf{C}_t$ of temporal contribution source profiles. Conversely when the eight temporal contributions obtained for every reservoir outflow are averaged, matrix $\mathbf{C}_g$ of geographical contribution source profiles is obtained.

Results and Discussion

PCA Results. Table 2 shows the results of PCA for the different data structures. First, the whole raw data set was analyzed. Although always recommended in PCA studies, for the particular data set used in this study, mean centering of the data had little effect on the results. The results and plots obtained, whether mean centering was applied or not, were very similar. The first PC (principal component) explained 84.6% of the total variance, the second PC explained 9.8%; and the third PC explained 2.4%.With two and three PCs, 94.4 and 96.8% of the total variance was already explained. The fourth PC explained a little more than 1% (1.5%), and the fifth PC explained less than 1% of the total variance.

Most of the data variance was explained by the first three PCs, indicating that most of the information provided by the 15 original measured variables is explained using only these three components. This allows an easier graphical representation and interpretation of the data variance. Figure 4 gives the PCA loadings for the three main components. The first PC had high loadings for variable 2 (alachlor ESA) and for variable 4 (atrazine). The second PC also had high loadings for the same variables 2 and 4 but in an inverse way. This was interpreted as if two independent sources of variables 2 and 4 were present. The first PC is the major source of alachlor ESA and atrazine and also for the other major detected herbicides and derivatives (alachlor, atrazine, cyanazine, and metolachlor). Probably, it corresponds to the major spring flush of these herbicides. The second PC is related to an independent source of some of these compounds such as groundwater as has been proposed in previous studies of the same geographical area (*3*). The third

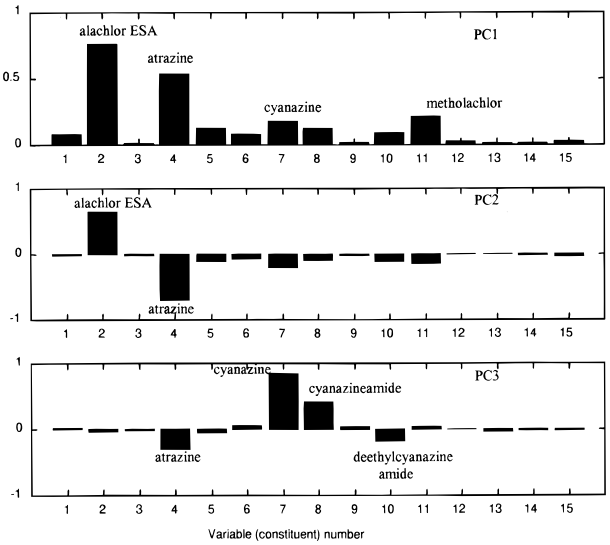

FIGURE 4. Principal component analysis (PCA) loadings. Variable number on the *x*-axis refers to constituent concentrations given in Figure 2 caption. Loadings (values in *y* axis) are normalized to unit length.

PC had high loadings for variables cyanazine (variable 7) and cyanazine amide (variable 8) but not for alachlor ESA and atrazine (variables 2 and 4). Metolachlor (variable 11) loads were more evident for PC1 than for PC2.

A large number of the samples have relatively low scores for PC1 and PC2 that account for most of the data variance (94.4%, Table 2). However, some samples have high scores for PC1 (for example, samples 12, 15, 17, and 19 from reservoirs in Indiana or samples 61 and 65 from reservoirs in Ohio, all of them collected in the summer) or high scores for PC2 (for example, samples from reservoir 10, corresponding to a sample collected during July from Cataract Lake, also in Indiana, and from reservoir 15, corresponding to a sample collected in Indiana during the summer). All these samples have high input concentrations of alachlor ESA and atrazine (variables 2 and 4). The fact that they have different scores with respect to PC2 (*y*-axis) is in agreement with the existence of two different environmental sources for alachlor ESA and atrazine as was previously proposed (*3*) and with the pattern shown by the PCA loadings (Figure 4). Scores for the third PC (high cyanazine) showed a similar trend for most of the samples, except for some individual samples (like the sample from reservoir 35, Lac Qui Parle Reservoir, Minnesota, collected during late June) with a much higher value for this third PC score (cyanazine concentration). When PCA was repeated considering only the nine more important variables (variables 1, 2, 4−9, and 11). PCA results obtained with this new decreased data set were similar to results just described (see Table 2). Although the elimination of six variables had a small effect on the PCA results when no transformation was applied to the data, decreasing the number of variables had a substantial effect on the scaled and autoscaled transformed data matrixes. This was due to the fact that the scaling of the variables where most of the constituent concentrations were close to the limit of detection gave very unreliable results. For instance, for the whole scaled and autoscaled data sets, less reliable constituents such as ametryn (variable 3) deethylcyanazineamide (variable 10), and variables 12−15 (see caption of Figure 2) gave high loadings for the first PC because when these variables were divided by their very small standard deviation in the scaling process, their values become very large. For the decreased variables data set, scaling and autoscaling pretreatment gave similar loading values. In this case, all the variables contributed similarly to the first PC with high loadings; alachlor,

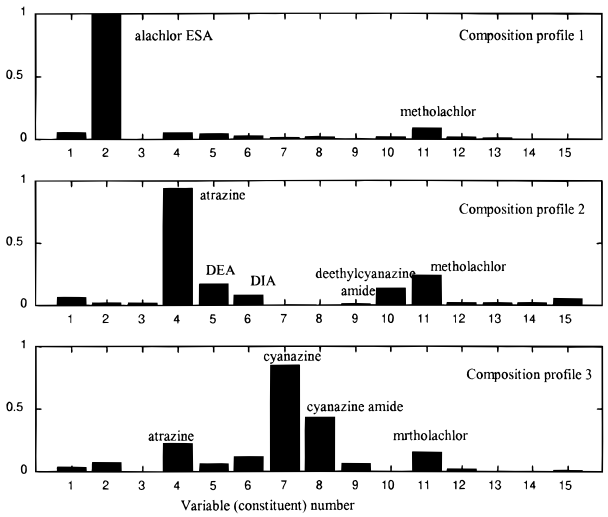

FIGURE 5. Source composition profiles resolved by multivariate curve resolution (MCR-ALS). Variable number in *x*-axis refers to constituent concentrations given in Figure 2 caption. The *y*-axis gives the relative constituent (herbicide and derivative) normalized concentrations obtained for each of the ALS-resolved environmental sources.

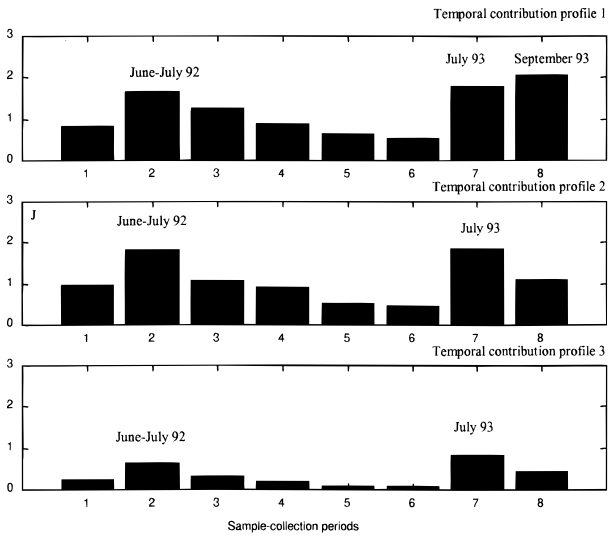

FIGURE 6. Temporal contribution profiles resolved by MCR-ALS (averaged for the 76 reservoir outflow sampling sites), matrix C_t in Figure 3. Identification of sample collection periods in *x*-axis: 1, late April−early May 1992; 2, late June−early July 1992; 3, late July−late August 1992; 4, end of September−mid-October 1992; 5, early January 1993; 6, mid-March 1993; 7, late July−late August 1993; and 8, September 1993 following the 1993 flood. The *y*-axis gives the relative temporal contributions obtained for each ALS-resolved environmental source.

cyanazine amide, and deethylcyanazine (variables 1, 8, and 9) were more important in the second PC, and alachlor, atrazine, and deisopropylatrazine (variables 1, 4, and 6) were more important in the third PC. Data interpretation in terms of source apportionment became more difficult when experimental data were scaled since the information and variance coming from input high concentration variables was mostly lost.

In Table 2, the results of PCA analysis for log-transformed data are given for both the 15- and 9-variable data sets. For the log-transformed data set, the effect of variable reduction in terms of explained variance was much less than for scaled and autoscaled data. No clear advantages were gained from the log-transformed data. In this case (log-transformed data), the first PC again had high loadings from most of the variables; the second PC had high loadings mostly from variables 2, 3, and 8; and the third PC had high loadings from variables 1, 2, 3, 6, 7, and 9. As the goal of the analysis was to distinguish the possible different sources of data variance, neither of the proposed data transformations gave more interpretable results that the nontransformed data matrix.

MCR-ALS Results. MCR-ALS results were interpreted in terms of resolution of the composition profiles of the three possible sources of data variation and in terms of the resolution of the temporal and geographical contributions of these sources. The composition profiles were not forced to be orthonormal or to explain maximum variance as in PCA but only to be nonnegative and to explain composited maximum variance (nonnegative least-squares solutions) (*21*). Whereas PCA profiles are pure abstract mathematical uncorrelated solutions, ALS profiles are solutions attempting to recover the real, physically correlated source profiles.

Figure 5 gives the three more important ALS-resolved source composition profiles when nonnegative constraints were used in the ALS decomposition of the whole augmented data matrix (eqs 1−6). Initial estimates for the ALS optimization were obtained from the detection of the purest samples using a procedure similar to the SIMPLISMA procedure (*21*). The three composition profiles selected for an initial estimation of matrix $\mathbf{S}^T$ were those from a sample with a high concentration of alachlor ESA (from reservoir 13 in Figure 1), from a sample with a high concentration of atrazine (from reservoir 40 in Figure 1), and from a sample with a high concentration of cyanazine (from reservoir 35 in Figure 1).

The percentage of explained variance achieved by MCR-ALS using three components and nonnegative constraints was 96.7%, which is similar to the variance obtained when PCA was applied (96.8%, see Table 2) for the same number of components.

Source composition profiles given in Figure 5 show that the first resolved profile accounted mostly for alachlor ESA concentrations; the second resolved profile accounted for atrazine and also for some metholachlor, deethylatrazine, and deethylcyanazine amide concentrations; and the third resolved composition profile accounted mostly for cyanazine, cyanazine amide, and some atrazine and metolachlor concentrations.

Source contributions (matrix $\mathbf{C}_{aug}$) were plotted in two different ways depending on whether the interest was to look at the temporal source contributions (matrix $\mathbf{C}_t$ in Figure 3) or to the geographical source contributions (matrix $\mathbf{C}_g$ in Figure 3). For a large number of reservoirs, the concentration contributions were very low, close to what seems to be the background contribution level. However, for a still significant number of profiles, there was a clear and repetitive (although not exactly equal) seasonal pattern, with two maxima during the summer of the 2 yr under study, 1992 and 1993.

From Figure 6, it is confirmed that the peak concentration of the herbicides and derivatives were obtained in the summer with a similar seasonal pattern for the three MCR-ALS-resolved profiles. Figure 7 gives the averaged geographical source contribution profiles of the three MCR-ALS-resolved environmental sources of herbicides and derivatives for the time periods under investigation. Reservoirs in Indiana (sites 12, 15, 17, and 19 in Figure 1) gave high values for the three resolved herbicide environmental sources. Also, some reservoirs in Ohio (sites 61, 63, and 65 in Figure 1) gave relatively high values for the first two resolved herbicide environmental sources. Reservoirs in the lower Missouri River Basin (sites 41−44, 46, and 48 in Figure 1) gave high values for the second source (mostly related to atrazine) but not for the first source (mostly related to alachlor ESA). The same results is evident for reservoir 1 in Illinois and reservoir 10 in Indiana.

From Figures 6 and 7, it is clear that alachlor ESA and atrazine are not always encountered simultaneously at high

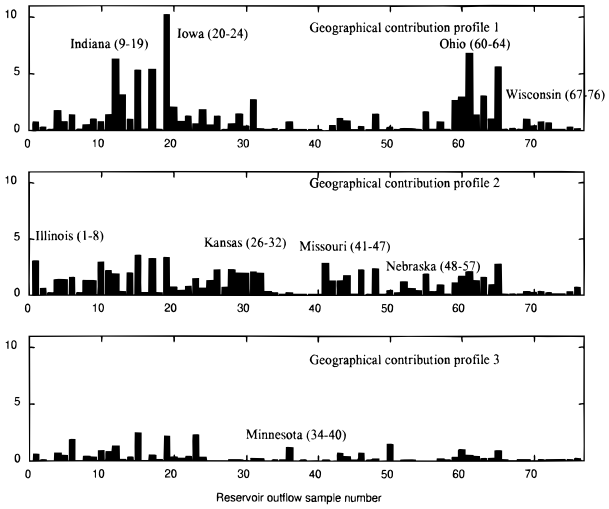

FIGURE 7. Geographical contribution profiles resolved by MCR-ALS (averaged for the eight sample collection periods), matrix C_g in Figure 3. Identification of reservoir outflow sample number is given in Figure 1. The *y*-axis gives the relative geographical contribution obtained for each ALS-resolved environmental source.

concentrations. Usually when the alachlor ESA concentration is high, then atrazine is also high, but the opposite is not true because in some locations atrazine is encountered at high concentrations but alachlor ESA is at low concentration. This agrees with the PCA results and again suggests two independent environmental sources of these two major herbicides in the area under study. Reservoirs 23 in Iowa, 36 in Minnesota, and 50 in Nebraska had high values for the third resolved source (mostly related with cyanazine) and low values for the other sources. In summary, looking in detail at the plots given in Figures 6 and 7, it is relatively easy to gain a rapid view of the temporal and geographical source contributions of the detected environmental sources of herbicides and their derivatives in the whole study area, giving source apportionment and environmental interpretation.

Finally, when the MCR-ALS analysis was applied to the reduced number of variables, the same results were obtained. When data pretreatment methods such as log transformation were applied, the resolved profiles did not provide any additional information to improve the interpretation of the observed data variance.

Acknowledgments

This is Comisión de Intercambio Cultural, Educativo y Científico entre España y los EEUU Project HNCCT 98148. The use of firm, trade, or brand names in this paper is for identification purposes only and does not constitute endorsement by the U.S. Government.

Literature Cited

(1) Humenik, F. J.; Smolen M. D.; Dressing, S. A. *Environ. Sci. Technol.* **1987**, *21*, 737−742.
(2) Thurman, E. M.; Goolsby, D. A.; Meyer, M. T.; Kolpin, D. W. *Environ. Sci. Technol.* **1991**, *25*, 1794−1796.
(3) Thurman, E. M.; Goolsby, D. A.; Meyer, M. T.; Mills, M. S.; Pomes, M. L.; Kolpin, D. W. *Environ. Sci. Technol.* **1992**, *26*, 2440−2447.
(4) Gianessi, L. P.; Puffer C. M. *Use of selected pesticides for agricultural crop production in the United States, 1982−1985*; National Technical Information Service: Springfield, VA, 1986.
(5) Thurman, E. M.; Meyer, M. T.; Mills, M. S.; Zimmerman, L. R.; Perry, C. A.; Goolsby, D. A. *Environ. Sci. Technol.* **1994**, *28*, 2267−2277.
(6) Thurman, E. M.; Goolsby D. A.; Aga, D. S.; Pomes M. L.; Meyer, M. T. *Environ. Sci. Technol.* **1996**, *30*, 569−574.
(7) Thurman, E. M.; Fallon, J. D. *Int. J. Environ. Anal. Chem.* **1996**, *65*, 203−214.
(8) Smeyers-Verbeke, J.; Den Hartog, W. H.; Dekker, J. C.; Coomans, D.; Buydens, L.; Massart, D. L. *Atmos. Environ.* **1984**, *18*, 2471−2478.
(9) Tauler, R.; Izquierdo-Ridorsa, A.; Casassas E. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 293−300.
(10) Tauler, R.; Smilde, A.; Kowalski, B. R. *J. Chemom.* **1995**, *9*, 31−58.
(11) Hopke, P. K. *Chemom. Intell. Lab. Syst.* **1991**, *10*, 21−43.
(12) Salou, J. S.; Tauler, R.; Bayona, J.; Tolosa, I. *Environ. Sci. Technol.* **1997**, *31*, 3482−3490.
(13) Zeng, Y.; Hopke, P. K. *J. Chemom.* **1992**, *6*, 65−83.
(14) Thurman, E. M.; Meyer, M. T.; Mills, M. S.; Zimmerman, L. R.; Perry, C. A.; Goolsby, D. A. *Environ. Sci. Technol.* **1994**, *28*, 2267−2277.
(15) Scribner, E. A.; Goolsby, D. A.; Thurman, E. M.; Meyer, M. T.; Battaglin W. A. *Open-File Rep.—U.S. Geol. Surv.* **1996**, *No. 96-393*.
(16) Thurman, E. M.; Meyer, M.; Pomes, M.; Perry, C. A.; Schwab, P. *Anal. Chem.* **1990**, *62*, 2043−2048.
(17) Aga D. S.; Thurman, E. M.; Yockel M.; Williams, T. *Environ. Sci. Technol.* **1996**, *30*, 592−597.
(18) Eigenvector Research. *PLS_Toolbox version 2.0*; Manson, WA, 1998.
(19) Golub, G. H.; Van Loan, Ch. F. *Matrix computations;* The Johns Hopkins University Press: London, 1989.
(20) Forina, M.; Lanteri, S.; Leardi, R. *Trends Anal. Chem.* **1987**, *6*, 250−251.
(21) Lawson, C. L.; Hanson, R. J. *Solving least squares problems*; Prentice-Hall: London, 1974.
(22) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425−1432.
(23) Manne, R. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 89−94.
(24) Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133−146.

Received for review January 7, 2000. Revised manuscript received May 4, 2000. Accepted May 10, 2000.

ES000884M