Self Organizing Maps for Analysis of Polycyclic Aromatic Hydrocarbons 3-Way Data from Spilled Oils

ARTICLE *in* ANALYTICAL CHEMISTRY · APRIL 2010

Impact Factor: 5.64 · DOI: 10.1021/ac100706j · Source: PubMed

Anal. Chem. **2010,** *82,* 4264–4271

Self Organizing Maps for Analysis of Polycyclic Aromatic Hydrocarbons 3-Way Data from Spilled Oils

R. Fernández-Varela,[†] M. P. Gómez-Carracedo,[†] D. Ballabio,[‡] J. M. Andrade,*[,†] V. Consonni,[‡] and R. Todeschini[‡]

Department of Analytical Chemistry, University of A Coruña; Campus da Zapateira s/n, E-15071, A Coruña, Spain, and Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1-20126 Milano, Italy

In this paper, the application of a new method based on self-organizing maps (SOM; termed MOLMAP, molecular map of atom-level properties) to handle 3-way data generated in a monitoring environmental study is presented. The study comprised 50 polycyclic aromatic hydrocarbons (PAHs) analyzed in samples derived from the weathering of six oil products (four crude oils and two fuel oils) spilled under controlled conditions for about 4 months. MOLMAP yielded useful information on each mode of the data cube: weathering samples, spilled oil products, and PAHs. Thus, the different behaviors of the six oils were ascertained, along with their particular evolution on time, and their weathering patterns were studied in terms of the original PAHs. Thus, the two heaviest products (two fuel oils) were characterized by two neurons whose more relevant weights were associated to heavy PAHs, as C_1-fluoranthene, C_2-fluoranthene, benzo(ghi)perylene, and dibenz(ah)anthracene. The six spilled products were projected on different regions on both the MOLMAP-SOM and a subsequent principal components analysis (PCA) scatter plot, developed using the so-called MOLMAP-scores. Besides, it was possible to further differentiate between unweathered, or slightly weathered, samples and the most weathered ones. The more relevant PAHs characterizing those samples were assessed studying the weights of the neurons in which the samples got projected.

There is a worldwide increasing claim for the analysis and monitoring of harmful chemicals originated in industrial activities reaching the environment. A good example of this is the huge concern raised among U.S. and European citizens after recent successive oil spillages caused by ship tankers *Erika* (SW France, 1999, 31 000 tons), *Prestige* (NW Spain, 2002, 63 000 tons), *Bouchard* (Buzzards Bay, Massachusetts, USA, 2003, 320 tons), *Athos 1* (Delaware River, USA, 2004, 860 tons), *Selendang Ayu* (Unalaska Island, Alaska, 2004, 18 000 tons), *Cosco Busan* (San Francisco Bay, 2007, 220 tons), *Hebei Spirit* (South Corea, 2007, ca. 10 000 tons), and *Full City* (South of Langesund, Norway, 2009, 200 tons). Despite that some light and volatile chemicals leave the oils after some hours (e.g., light aliphatic hydrocarbons, polycyclic aromatic hydrocarbons with one or two benzene rings, and alkylnaphthalenes), others will last for a long time in the environment. This is the case for most polycyclic aromatic hydrocarbons (PAHs), some of which have been recognized as carcinogenic and mutagenic by the U.S. EPA and European Safety Agency.[1] Environmental monitoring programs deployed at the areas affected by such accidents require both powerful analytical methods and reliable data analysis procedures to extract information capable of evaluating the evolution of the pollution event and, likely, determine which compounds will remain in the environment on a medium-long term.

Current developments in analytical methodologies lead to very low limits of detection and increased sensitivity and allow for the elucidation of the different species of the same contaminant. Particularly, gas chromatography coupled to mass spectrometry detection (GC/MS) proved the most suitable analytical methodology to detect the pollutants considered here, PAHs, including a suite of biomarkers as hopanes, steranes, diasteranes, triaromatic steroids, and diamandoids.[2] Nevertheless, this powerful analytical technique requires application of various data treatment tools (often, multivariate pattern recognition methods) to get the most out of the raw information present within the bulk of data obtained after an environmental study. Further, as environmental monitoring calls for deploying several sampling campaigns on time, the simultaneous analysis of various data sets is a must. Such data can be arranged intuitively as a "data cube", that is samples (rows) × variables (columns) × samplings (slices). There have been important efforts to study those data cubes (in general, *N*-way data arrangements), and several 3-way techniques were employed in the last years in the environmental field.

Several recent studies can be mentioned just to focus the discussions. The temporal evolution of river water physicochemical quality and soil pollution by road traffic were studied by parallel factor analysis (PARAFAC), matrix augmented princi-

* To whom correspondence should be addressed. E-mail: andrade@udc.es. Fax: +34981167065.
[†] University of A Coruña.
[‡] University of Milano-Bicocca.

(1) Harvey, R. G. *Polycyclic Aromatic Hydrocarbons: Chemistry and Carcinogenicity*; Cambridge University Press: Cambridge, U.K., 1991.
(2) Wang, Z.; Stout, S.; Fingas, M. *Environ. Forensics* **2006**, *7*, 105–146.

10.1021/ac100706j
Published on Web 04/23/2010

pal components (MA-PCA),[3] and Procrustes rotation.[4] The distribution of metals in river sediments was evaluated employing Tucker3[5] whereas the fractionation patterns of metals in soils, fish, sediments, and water were assessed by principal component analysis (PCA), MA-PCA, 3-PCA, PARAFAC, MCR-ALS (multivariate curve resolution-alternating least squares), and TUCKER3.[6,7] Closely related to the present study, weathered diesel oils and oil spills were studied by PARAFAC[8−10] while MA-PCA, PR, and PARAFAC were employed to treat GC/MS and mid-infrared (IR) spectrometry after a fuel spillage on the sea.[11,12] Besides, 3-way partial least-squares (PLS) and PARAFAC were applied to analyze PAHs in water,[13] and a screening method combining excitation−emission fluorescence and PARAFAC was presented for lubricant oils, crude oils, and heavy and light fuel oils.[14] Finally, several practical aspects of oil spill fingerprinting employing PCA and PARAFAC were reviewed recently.[15]

These chemometric methods yield satisfactory results, but they are complex both from a conceptual and a mathematical standpoint. Following, there is room for other approaches to study how the samples behave and to ascertain whether they present patterns. Hence, the objective of this work is to investigate the use of a new method (termed originally molecular map of atom-level properties, MOLMAP) to handle data generated in a monitoring environmental study where 50 PAHs were analyzed in samples derived from the weathering of six oil products spilled under controlled conditions for ca. 4 months. MOLMAP was developed initially for the study of disparate molecule chemical information organized into a 3-way data structure. Its main idea consists of rearranging the 3-way data cube so that a classical Kohonen neural network (or self-organizing map, SOM) can be used to look for similarities among the samples and extract sample patterns.[16,17] Accordingly, MOLMAP could be considered as a multiway analysis technique complementing other more complex multiway models (PARAFAC, Tucker3, MCR-PLS, etc), much in the same way as common PCA and cluster analysis complement each other. To the best of our knowledge, this is the first time that the three modes of the 3-way data cube (samples, oil products, and variables) are studied simultaneously using MOLMAP. In addition, the temporal evolution of each oil in the internal MOLMAP-SOM was assessed directly in terms of the experimental variables causing them.

EXPERIMENTAL SECTION

Samples. Six oil products were spilled in special containers, and their weathering processes were monitored for ca. 4 months. They were four original crude oils (Maya, Ashtart, Brent, and Sahara Blend), a Marine fuel oil (briefly, IFO), and the original fuel oil of the tanker *Prestige* sunk off the Galician coast (NW Spain on November 2002). The crude oils were chosen to cover the different types of petroleum types as the Maya oil was very heavy (high specific gravity), Ashtart was intermediate, and Brent and Sahara Blend were very light crude oils. The Prestige and IFO fuel oils are heavy residues of crude oil distillation processes. All products were released (500 mL) on a two-compartment special metallic container filled with seawater (60−70 L) and weathered under atmospheric conditions. A closed-circuit water pump continuously agitated the system by pouring water over the oil−water surface so that washing, emulsion, wave agitation, and sea movement were simulated as close as possible (more details can be found elsewhere[18,19]). Sample aliquots were withdrawn at preset intervals from the oil lumps. Following current practices (see, e.g., ref 20), aliquots were drawn more frequently during the first stages of weathering as the evaporation processes occur most intensely (initially they are the most relevant ones). After this, the weathering processes slow down and, so, the aliquots were more spaced on time. In total, 17 aliquots of the oil lumps were taken on time, and they are referred to as "weathering samples", which is a short for "samples taken along the weathering process of each spilled oil". The spillages of the crude oils were monitored from June 27th to October 17th (summer 2005, approximately 14 h of total solar irradiation/day, temperatures ranged from 18 to 28 °C) at 0, 5, 10, 24, 36, 48, 72, and 90 h and 7, 10, 15, 21, 29, 42, 57, 72, and 114 days. The fuel oils were spilled from January 9th to April 17th (winter 2006, approximately 8 h of solar irradiation/day, temperatures ranged from 7 to 16 °C) to mimic the release from the Prestige tanker. The fuel oil aliquots were sampled at 0, 5, 10, 24, 36, 48, 60, 72, and 84 h, and 7, 11, 14, 21, 28, 42, 56, and 101 days. The last samples of the fuels could not match those of the crude oils because of a lack of oil in the containers.

Note that, accordingly, each weathering sample is made of a $J \times K$ matrix, where each jth row corresponds to one of the oil products spilled into the containers and each kth column corresponds to the kth measured PAH. Each J x K matrix is therein referred to as the "input submatrix" of the ith weathering (Figure 1).

The organic phase was transferred to 50 mL Pyrex centrifuge tubes where around 1 g of anhydrous sodium sulfate (Merck, 99.0%, Damstard, Germany) was added, and the tubes were centrifuged at 3000 rpm for 30 min. Whether emulsions appeared, NaCl (Panreac, 99.5%, Barcelona, Spain) and another gram of

(3) Felipe-Sotelo, M.; Andrade, J. M.; Carlosena, A.; Tauler, R. *Anal. Chim. Acta* **2007**, *583*, 128–137.
(4) Andrade, J. M.; Kubista, M.; Carlosena, A.; Prada, D. *Anal. Chim. Acta* **2007**, *603*, 20–29.
(5) Singh, K. P.; Malik, A.; Basant, N.; Singh, V. K.; Basant, A. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 185–193.
(6) Peré-Trepat, E.; Ginebreda, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 69–83.
(7) Pardo, R.; Vega, M.; Debán, L.; Cazurro, C.; Carretero, C. *Anal. Chim. Acta* **2008**, *606*, 26–36.
(8) Ebrahimi, D.; Li, J.; Hibbert, D. B. *J. Chromatogr., A* **2007**, *1166*, 163–170.
(9) Ebrahimi, D.; Hibbert, D. B. *J. Chromatogr., A* **2008**, *1198−1199*, 181–187.
(10) Gaganis, V.; Pasadakis, N. *Anal. Chim. Acta* **2006**, *573−574*, 328–332.
(11) Andrade, J. M.; Fresco, P.; Muniategui, S.; Prada, D. *Talanta* **2008**, *77*, 863–869.
(12) Grueiro-Noche, G.; Andrade, J. M.; Muniategui-Lorenzo, S.; López-Mahía, P.; Prada-Rodríguez, D. *Environ. Pollut.* **2010**, *158*, 207–214.
(13) Beltrán, J. L.; Guiteras, J.; Ferrer, R. *Anal. Chem.* **1998**, *70*, 1949–1955.
(14) Christensen, J. H.; Hansen, A. B.; Mortensen, J.; Andersen, O. *Anal. Chem.* **2005**, *77*, 2210–2217.
(15) Christensen, J. H.; Tomasi, G. *J. Chromatogr., A* **2007**, *1169*, 1–22.
(16) Zhang, Q.-Y.; Aires-de-Sousa, J. *J. Chem. Inf. Model.* **2005**, *45*, 1775–1783.
(17) Zhang, Q.-Y.; Aires-de-Sousa, J. *J. Chem. Inf. Model.* **2007**, *47*, 1–8.
(18) Fernández-Varela, R.; Suárez-Rodríguez, D.; Gómez-Carracedo, M. P.; Andrade, J. M.; Fernández, E.; Muniategui, S.; Prada, D. *Talanta* **2005**, *68*, 116–125.
(19) Fernández-Varela, R.; Gómez-Carracedo, M. P.; Fresco-Rivera, P.; Andrade, J. M.; Muniategui, S.; Prada, D. *Talanta* **2006**, *69*, 409–417.
(20) Li, J.; Fuller, S.; Cattle, J.; Way, C. P.; Hibbert, D. B. *Anal. Chim. Acta* **2004**, *514*, 51–56.

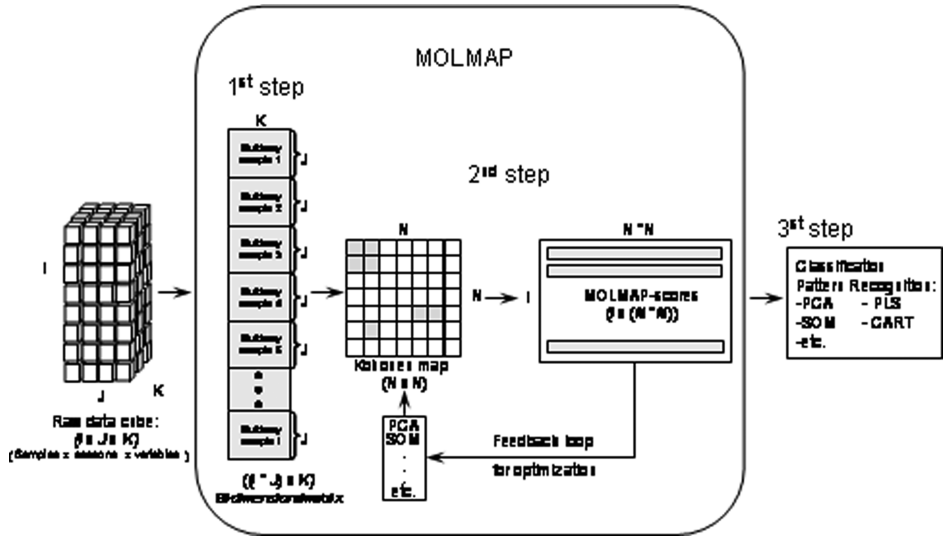

Figure 1. General scheme of the MOLMAP approach to handle 3-way data arrays. First step: unfolding of the 3-way data cube; second step: calculation of the "MOLMAP-scores" using an internal SOM; third step: usage of a pattern recognition or classification technique to unravel the main patterns of the data set.

sodium sulfate were added until they were broken down. To avoid excessive handling of stable emulsions, it was found satisfactory to thermostatize them at 50−60 °C (±1 °C) while centrifuging at 3000 rpm for 30−40 min (Seta Oil Test Centrifuge, Surrey, United Kingdom).[21]

Analytical Procedure. The analytical procedure was implemented following SINTEF recommendations,[22] after a detailed study of the robustness of the analytical conditions utilizing Plackett−Burman experimental designs. Critical variables were monitored daily employing a T^2 multivariate control chart. In total, 50 PAHs were analyzed; they were decahydronaphthalene (decalin, DE) and its C_1-, C_2-, C_3-, and C_4-homologous alkylated series (DE1, DE2, DE3, and DE4), naphthalene (N) and its C_1-, C_2-, C_3-, and C_4-homologous alkylated series (N1, N2, N3, and N4), benzothiophene (BT) and its C_1-, C_2-, C_3-, and C_4-homologous alkylated series (BT1, BT2, BT3, and BT4), biphenyl (B), acenaphthylene (ANY), acenaphthene (ANA), dibenzofuran (DBF), fluorene (F) and its C_1-, C_2-, and C_3-homologous alkylated series (F1, F2, and F3), dibenzothiophene (D) and its C_1-, C_2-, C_3-, and C_4- homologous alkylated series (D1, D2, D3, and D4), phenanthrene (P) plus anthracene (A) and its C_1-, C_2-, C_3-, and C_4-homologous alkylated series (P1, P2, P3, and P4), fluoranthene (FL) and its C_1-, C_2-, and C_3-homologous alkylated series (FL1, FL2, and FL3), pyrene (PY), benz(a)anthracene (BA) plus chrysene (C) and its C_1-, C_2-, C_3-, and C_4-homologous alkylated series (C1, C2, C3, and C4), benzo(b)fluoranthene (BbF) plus benzo(k)fluoranthene (BkF), benzo(e)pyrene (BEP), benzo(a)pyrene (BAP), perylene (PER), indene(123-cd)pyrene (IN), dibenz(ah)anthracene (DBA), and benzo(ghi)perylene (BPE). The RIS-PAH (deuterated acenaphthene, fluorene, and perylene) standard was used to quantify PAHs.

The analytical figures of merit were evaluated as well. Recoveries were assessed with the surrogate internal standard SIS-PAH (naphthalene-d_{10}, phenanthrene-d_{10}, chrysene-d_{12}) (Chiron Laboratory, Norway). *N*-Hexane (Merck, Damstard, Germany) was used as diluent. They ranged 70−100% for all compounds, except for DE (66%, because it is the lightest molecule) and for P (120%, due to its coelution with A). Repeatability and reproducibility were calculated using 0.5 $\mu g \cdot mL^{-1}$ standards, performing four replicates/day and ten replicates/three days, respectively. The values ranged from 0.05 to 0.27 $\mu g\ mL^{-1}$ for repeatability and from 0.05 to 0.33 $\mu g \cdot mL^{-1}$ for reproducibility. The method detection limits for the PAHs ranged from 0.06 to 4.60 $ng \cdot mL^{-1}$, for BT and BAP, respectively. Trueness was evaluated using the SRM 1582 certified material (Chiron Laboratory, Norway). More specific details can be found elsewhere.[23] To perform the chemometric studies presented here, all areas were normalized to C_{30}-hopane.

MOLMAP Theoretical Background. Artificial neural networks (ANNs) based on the Kohonen approach (Kohonen maps or SOMs) are self-organizing systems which are capable of solving the unsupervised rather than the supervised problems. In Kohonen maps, similar input objects are linked to the topological close neurons in the network; i.e., neurons that are located close to each other have similar reactions to similar inputs, while the neurons that are far apart have different reactions to similar inputs. Basically, in the Kohonen approach, the neurons learn to identify the location in the ANN that is most similar to the input vectors.

The Kohonen map is usually characterized by being a squared toroidal space that consists of a grid of neurons. Each neuron contains as many elements (weights) as the number of input variables. The weights of each neuron are randomly initialized between 0 and 1 and updated on the basis of the input vectors (i.e., samples), for a certain number of times (called training epochs). Both the number of neurons and epochs to be used to

(21) Fresco-Rivera, P.; Fernández-Varela, R.; Gómez-Carracedo, M. P.; Ramírez-Villalobos, R.; Prada, D.; Muniategui, S.; Andrade, J. M. *Talanta* **2007**, *74*, 163–175.

(22) Daling, P. S.; Fakness, L. G. *Laboratory and reporting instructions for the CEN/BT/TF 120 Oil Spill Identification-Round Robin Test-May, 2001*. Technical report, SINTEF report STF66-A02027, 2002.

(23) Fernández-Varela, R.; Andrade, J. M.; Muniategui, S.; Prada, D.; Ramírez-Villalobos, F. *Water Res.* **2009**, *43*, 1015–1026.

train the map must be defined by the user; more technical details can be found in the literature.[24−28]

SOMs possess outstanding capabilities to group (order) samples in an unsupervised nonhierarchical way.[29−31] This boostered research into new algorithms and, particularly, into supervised classification SOMs. The foundations of the SOMs have been already described, and therefore, they will not be repeated here. So far, SOMs have been restricted to the study of traditional 2-way data sets (typical data matrices structured as samples × variables), and only recently, some few attempts were made on chemoinformatic studies to handle 3-way data cubes.[16,17,32,33] Then, the feasibility of SOMs to cope with 3-way chemical data with classification purposes has been explored.[34]

Originally developed to cope with disparate data generated from molecular studies, MOLMAP stands for molecular map of atom-level properties, and it requires two major steps: (a) generation of the so-called "MOLMAP-scores" by means of a dedicated SOM and (b) development of a pattern recognition model using the MOLMAP-scores as independent variables. Noteworthy, the 3-way data structure is managed in the first step by reorganizing it so that the 3-way (or multiway) samples can be related using a classical SOM and, thus, their patterns visualized. Despite that traditional SOMs can be optimized by themselves, the MOLMAP-SOM should be optimized by evaluating the performance of a combined pattern recognition method (such as PCA, CART,[34] or Random Forest[16,17,32,33]) which uses the MOLMAP-scores to build a multivariate model (Figure 1).

To generate the MOLMAP-scores, we first need to unfold the $(I \times J \times K)$ 3-way data set in a new $((I^*J) \times K)$ bidimensional matrix whose (I^*J) rows are used to train the SOM.

Afterward, it is possible to calculate the MOLMAP scores on the basis of the obtained map. A representation of the samples of the multiway data set can be obtained by projecting them onto the trained map, one at a time. The pattern of activated neurons can be seen as a fingerprint of the sample and constitutes its MOLMAP score. As proposed in the literature,[16,17,27,34] each neuron gets a value equal to the number of times it is activated, and a value of 0.3 is added to each neuron multiplied by the number of times a neighbor is activated.

Finally, the map is transformed into a vector by concatenation of the neurons of the SOM. If the SOM is squared and constituted by N neurons on each side, the MOLMAP score of each sample has dimension N^*N and the entire MOLMAP score matrix has dimension $I \times N^*N$, where I is the number of samples in the original 3-way array. Basically, the MOLMAP score matrix is a two-way matrix where information of the original multiway data set is compressed by codifying the input vector positions in the SOM (Figure 1). In other words, each sample is described by new variables that encode information both on the second and third mode of the original data set. An extended explanation of applications of the MOLMAP approach to analytical data can be found in the literature.[34]

The architecture of the MOLMAP-SOM has to be optimized by varying the number of neurons and iterations of the SOM. The criterion to look for the best architecture is given by pattern recognition procedures applied to the MOLMAP-scores (e.g., number of groups, number of correct assignments to a group, total number of errors, etc., which should be stated initially by the scientist). Once the MOLMAP-SOM has been optimized, each 3-way sample of the training set is projected in turn onto it and the final sample-related MOLMAP-score is calculated. Finally, the MOLMAP-scores can be used as input to any of the different pattern recognition or classification techniques available in the literature (Figure 1). Here, we resorted to a traditional PCA, as we only searched for weathering patterns in the spilled oil products.

RESULTS AND DISCUSSION

Obtention of the MOLMAP-Scores. The 3-way raw data was arranged as a $17 \times 6 \times 50$ cube (weathering samples × oil products × PAHs). Four weathering samples were left apart to be projected on the final models and, thus, evaluate their predictive ability. Those samples were selected systematically so that each represented a part of the weathering process. All of them corresponded to weathering stages in the middle of the overall weathering period, neither at the very beginning of the aging nor at its end. Thus, the learning set was of dimension $(13 \times 6 \times 50)$ whereas the validation set was composed of four weathering samples $(4 \times 6 \times 50)$. They were the 3rd, 7th, 11th, and 14th weathering samples, which corresponded to 24 h, 5 days, 21 days, and 57 days (the crude oils) and to 24 h, 3 days, 14 days, and 42 days (the fuel oils), respectively.

Therefore, the unfolding of the data cube yielded a $((13^*6) \times 50)$ or (78×50) bidimensional training matrix whose variables were range scaled. Then, each multiway sample of the unfolded matrix was used to optimize the architecture of the MOLMAP-SOM by means of the MOLMAP-scores. Following the explanations above, they were (1×49) vectors representing which neurons of the SOM were activated by each sample.

SOM Optimization. In this work, the architecture selected for the MOLMAP-SOM was that leading to a best discrimination of the oil products in the subsequent PCA developed on the MOLMAP-scores. This was measured by the total number of samples erroneously projected in the groups of oil products. Different trials were made varying the topology of the SOMs from 6×6 to 8×8 neurons and the number of iterations from 50 to 500, where from the best results were obtained for a 7×7 topology, trained during 100 epochs. Figure 2 shows the Kohonen top map with all the 78 training and 24 validation weathering samples. Recall that each point in this figure corresponds to a row of the unfolded matrix (i.e., to each of the weathered and unweathered samples withdrawn from each spilled oil product) and that the original 3-way test set $(4 \times 6 \times 50)$ was unfolded to

(24) Kohonen. T. *Self-Organizing Maps*; Springer: Berlin, 2001.
(25) Mukherjee, A. *J. Comput. Civil Eng.* **1997**, *11*, 74–77.
(26) Astel, A.; Tsakovski, S.; Barbieri, P.; Simeonov, V. *Water Res.* **2007**, *41*, 4566–4578.
(27) Fonseca, A. M.; Biscaya, J. L.; Aires-de-Sousa, J.; Lobo, A. M. *Anal. Chim. Acta* **2006**, *556*, 374–382.
(28) Wongravee, K.; Lloyd, G. R.; Silwood, C. J.; Grootveld, M.; Brereton, R. G. *Anal. Chem.* **2010**, *82*, 628–638.
(29) Kalteh, A. M.; Hjorth, P.; Berndtsson, R. *Environ. Model. Softw.* **2008**, *23*, 835–845.
(30) Marini, F. *Anal. Chim. Acta* **2009**, *635*, 121–131.
(31) Melssen, W.; Wehrens, R.; Buydens, L. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 99–113.
(32) Gupta, S.; Matthew, S.; Abreu, P. M.; Aires-de-Sousa, J. *Bioorg. Med. Chem.* **2006**, *14*, 1199–1206.
(33) Latino, D. A. R. S.; Aires-de-Sousa, J. *Angew. Chem., Int. Ed.* **2006**, *45*, 2066–2069.
(34) Ballabio, D.; Consonni, V.; Todeschini, R. *Anal. Chim. Acta* **2007**, *605*, 134–146.

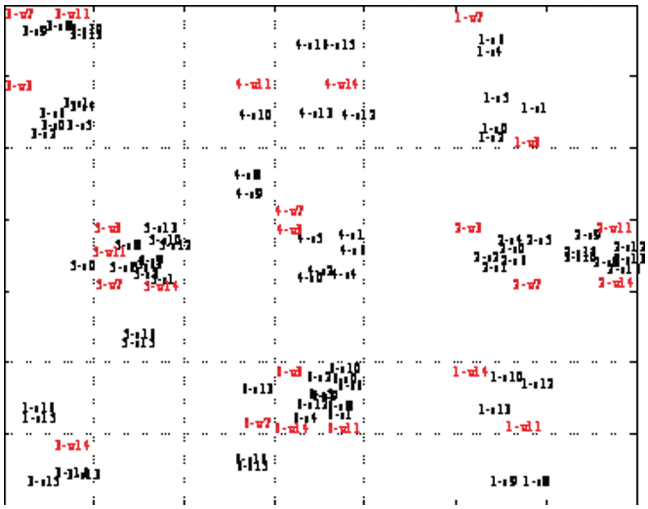

Figure 2. Kohonen top map of the MOLMAP model with the projection of the 78 training samples and 24 testing ones (s = training, v = validation). Labels represent the type of oil product to which the weathering sample belongs (1 = Ashtart, 2 = Brent, 3 = Maya, 4 = Sahara, 5 = IFO, and 6 = Prestige) and the extent of the weathering processs (0 = original unweathered sample, 16 = most weathered sample).

a (24 × 50) matrix). A good separation among the six products is observed. They become differentiated clearly in separate regions, except for the two most weathered samples of the Asthart crude oil, codes 1-s15 and 1-s16, which got mixed with the most weathered samples of the Maya crude oil, codes 3-s13, 3-s15, 3-s16, and 3-v14 at the bottom left corner of Figure 2, without a sound reason explaining it. It may be hypothesized that the heavy and intermediate oils reached a highly similar composition on the final stages of their weathering. Recall that the SOMs employed here were toroidal, which means that each edge of the Kohonen map has to be seen as connected with the opposite one. Hence, the neuron on the left bottom is connected to the neuron in the right top. The fact that the oil products lie in separate regions indicates that their weathering patterns are different.

This constitutes a very interesting result demonstrating that the SOM can resolve satisfactorily the confounding caused by the unfolding of the ($I \times J \times K$) data cube to a ((I*J) × K) matrix, which made the effects of the weathering samples and the type of oil product get mixed. Further, the SOM not only differentiated among the six oil products (even when they were weathered) but also modeled what occurred within each. In this way, the MOLMAP approach yields information on the oil products (2nd dimension of the data cube) and offers some clues on what is going on the first dimension, that is, the weathering samples.

In order to interpret the weathering samples, the (13 × 49) matrix of MOLMAP-scores has been explored by PCA. Figure 3a represents the 13 weathering samples of the training set in the PC1-PC2 subspace (explaining 87.7% of the total variance). It reveals a nice ordering of the 13 weathering stages, despite the first six ones could not be differentiated well (as expected) because of the small time frame they were obtained (i.e., 3 days, which it was thought it was not long enough to get a definite pattern in all products, mainly in the fuel oils, which had very little volatile compounds). Each point in the plot represents the behavior of the weathering samples taken on time for all the six oil products and, therefore, this plot suggests that the different stages of the weathering process can be differentiated. (Despite for Prestige and IFO, this might be not absolutely true, as detailed in the next studies.) The validation samples were projected in regions that can be logically associated to their age (see Figure 3a, samples coded with "v"). The weathering stage V7 became slightly displaced from the overall linear behavior, which may be justified by the complex chemical processes that occur with the varying atmospheric conditions during the ca. 4 months of the study. Indeed, around the third and fourth weathering days (samples #6 and #7) important changes on the weather conditions for the crude oils occurred (from rainy to sunny conditions). The "average" behavior of the validation aliquots numbered as 14 (V14 in Figure 3a) was not satisfactory, as they were projected slightly before the S13 samples. It was found that they involved quite different weathering periods (see Obtention of the MOLMAP-Scores). This, along with the differences in the weathering patterns (mostly due to the lower amount of volatiles in the fuel oils), would yield important differences

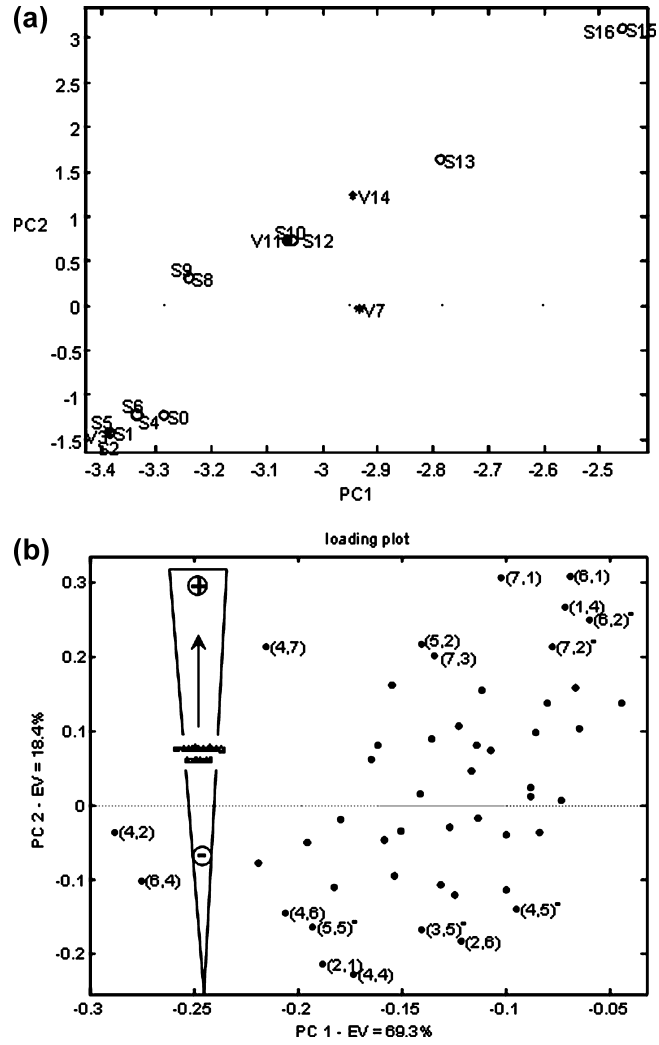

Figure 3. PCA study on the MOLMAP-scores considering the first two principal components: (a) score plot of the training samples (s) and projection of the validation samples (v) on the MOLMAP model. The codes represent the extent of the weathering process (0 = original unweathered sample, 16 = most weathered sample). (b) Loadings plot derived for the neurons of the MOLMAP-SOM. Neurons are identified by two numbers defining their position in the top map shown in Figure 2 (row, column). For an explanation on the neurons with an asterisk, see Interpretation of the MOLMAP SOM.

within the 14th aliquots that may have caused that wrong projection for the "average" behavior.

Interpretation of the MOLMAP SOM. Information on PAHs (3rd mode) should be linked to the findings of the other two modes (weathering samples and oil products), i.e., the MOLMAP-SOM and the MOLMAP-scores. We can start by the latter, as we are clearly interested in assessing which neurons determine the relative ordination of the weathering samples (i.e., how the weathering of the products goes on). To simplify explanations, the neurons will be identified by two numbers defining their position in the top map presented in Figure 2. Thus, the neuron (1,4) corresponds to the neuron located at the first row and fourth column. Considering the loadings obtained in the PCA analysis, Figure 3b shows that the most relevant neurons for the first principal component were (4,2) and (6,4). The second principal component was defined mostly by neurons (6,1), (7,1), (1,4), (4,7), (5,2), (7,3), with large positive loadings, and neurons (4,4), (2,1), (2,6), (4,6), and (6,4), with large negative loadings. Neurons denoted with an asterisk in the figure do not contain samples projected on them. Despite this, they received high loadings because of the way in which the scoring of the neurons was made to get the MOLMAP-scores. They are very close to neurons with several samples and, thus, they scored high but they are not of help to perform a chemical interpretation.

In the MOLMAP-SOM shown in Figure 2, neurons (4,2) and (6,4) contain almost all the samples of the Prestige and IFO fuel oils (but for the two oldest ones) and, so, they two characterize the heaviest products (fuels). The neurons with the largest negative loadings in the second factor (PC2), i. e., (2,1), (2,6), (4,2), (4,4), (4,6), and (6,4), contain the less weathered samples of the Maya, Ashtart, IFO, Sahara, Brent, and Prestige oils, respectively. On the contrary, neurons with the largest positive loadings contain the most weathered samples.

It was mentioned at the beginning of this section that we can draw information about the behavior of the PAHs from the MOLMAP-SOM. Once the most relevant neurons have been identified as indicated above, it is straightforward to inspect the profile of the weights of each relevant neuron of the SOM and to study how they relate to the raw data. Since each weight is associated with a variable, a pattern should emerge easily and we only explain the PCA-relevant ones. In the next paragraphs, the ordering of the weathering samples of each product and the chemical interpretation of the associated weights is given.

Group 1 (Ashtart) was projected on neurons (1,6), (2,6), (6,6), and (7,6). As mentioned above regarding PC2, neuron (2,6) clustered the less weathered samples (codes from 0 to 5) whereas neuron (6,6) grouped highly weathered samples (codes from 10 to 14), except for the most weathered ones (located with the two oldest Maya samples, which mixed with them in neuron (6,1), as noted above). The profiles of the weights of these four neurons, particularly of neuron (2,6), showed a clear predominance of two original PAHs containing sulfur; namely, C_2-benzothiophene and C_4-dibenzothiophene (PAHs #13 and #28, Figure 4a), although neuron (6,6) was more related to PAH #14 (C_3-benzothiophene) than to variable #13, which can be due to the Maya samples included on it.

Comparing the four weight profiles and scrutinizing the PAHs associated to each, it was observed that there is a clear difference

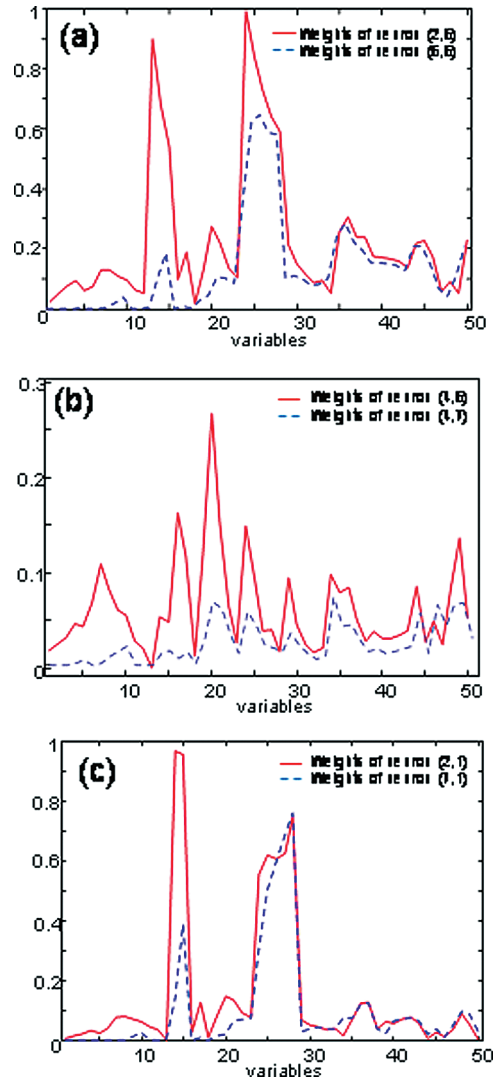

Figure 4. Profiles of the weights associated to the experimental variables of selected neurons. The weights are plotted against the number of the PAH (see text for details). (a) Ashtart, (b) Brent, and (c) Maya.

between the less-weathered samples (codes 0−3, clustered at neuron (2,6)) and the most weathered samples of the Ashtart oil (codes 10−14, grouped at neuron (6,6)), as PAHs containing sulfur (C_2-, C_3-, and C_4-benzothiophene, variables #13, 14, and 15) disappeared dramatically from the latter samples. This is apparent from the weights. Other weights changed also with time, although not so extensively, particularly, dibenzothiophene and C_1-dibenzothiophene (variables #24 and #25), Figure 4a.

Group 2 (Brent) got associated to neurons (4,6), the least weathered samples, and (4,7), the most weathered samples, both relevant in PC2 above. They had characteristic small weights for almost all variables (Figure 4b), of which only PAH #20 (fluorene) had a weight larger than 0.2 (recall that all weights range from 0 to 1). This reflected the characteristic low content of PAHs of this oil product, as it was a light crude oil. Neuron (4,7) showed a distinctive decrease of the weights associated to naphthalene and their alkylated homologous values (PAHs #6−10), as seen in Figure 4b. This fact reflected also what occurred with the concentrations of the PAHs measured on the Brent samples.

The Maya crude oil (Group 3) was related to neurons (1,1), (2,1), and (7,1) whose more important weights are those of PAHs #14, 15, and 28 (C_3-benzothiophene, C_4-benzothiophene, and C_4-dibenzothiophene, respectively). In neuron (7,1), the latter variable was much more important than the former two, probably because C_4-dibenzothiophene was much less prone to be degraded and, thus, characterize the samples with a stronger weathering (Figure 4c), which explains its importance in PC2. A relevant issue is the disappearance of the weight linked to PAH #14 (C_3-benzothiophene) and the large decrease of that associated to PAH #15 (C_4-benzothiophene) when moving from neuron (2,1) to (7,1). As mentioned above, these species contain sulfur and are relatively volatile. Neuron (1,1) had an intermediate behavior, and it characterized samples with intermediate weathering.

With regard to the lightest oil product (Sahara Blend, labeled as Group 4), its samples spread from neuron (4,4), least weathered samples, to neuron (2,4), intermediate weathered samples, and neuron (1,4), most weathered aliquots. These neurons showed important weights for many original PAHs (e.g., PAHs #8, 16, and 36; i.e., C_2-naphthalene, biphenyl, and C_2-fluoranthene, respectively) although, noteworthy, showed very low weights for PAHs #13 (C_2-benzothiophene, studied above as characteristic of the Ashtart oil), #18, and #48 (acenaphthene and indeno(123-cd)pyrene, see Figure 5a). This might be due to the large and fast depletion most PAHs suffered during the weathering of this very light crude oil. This dramatic diminution represents how PAHs disappeared with aging. For instance, C_3-decahydronaphthalene, C_4-decahydronaphthalene, C_1-naphthalene, C_2-naphthalene, and C_3-naphthalene, respectively (variables #4−9) had large weights on neuron (4,4), which almost disappeared in neuron (1,4), Figure 5a. Hence, a variable can be used not only to characterize the aging of this product but also to characterize a suite of them. The residue that went on when lightest variables disappeared was characterized by larger and heavier PAHs, as dibenzofuran (variable #19), C_1- and C_2-fluorene (#21 and #22), and fluoranthene (#34) whose weights in neuron (1,4) increased their importance with respect to neuron (4,4).

Group 5 (IFO) was restricted to neurons (4,2) and (5,2), the latter containing the last two samples of the aging process. The distinctive pattern of their weights shows a very negligible importance of variables #1−25 and more relevance for the heaviest molecules, see Figure 5b (C_2-fluoranthene, benzo(ghi)perylene, and dibenz(ah)anthracene, i.e., PAHs #36, 47, and 49, respectively). This is coherent with the fact that the IFO product was a heavy fuel oil, with low quantities of volatile species, and was hard to degrade. The most relevant evolution during its weathering was a decrease on the weights associated to PAHs #7−10 (C_1- to C_4-naphthalene), #13−15 (C_2- to C_4-benzothiophene), #17 (acenaphthylene), and #20 (fluorene) when moving from neuron (4,2) to neuron (5,2), most weathered samples. Only variable #34 (fluoranthene) increased with IFO's aging, without a sound reason to explain it.

Group 6 (Prestige) was located essentially in neurons (6,4) and (7,3), the two oldest samples. Similar to the IFO product, these neurons are weighted almost only by the heaviest PAHs (Figure 5c), although the absolute values of the weights were almost twice those of the IFO (so that both products got differentiated). The

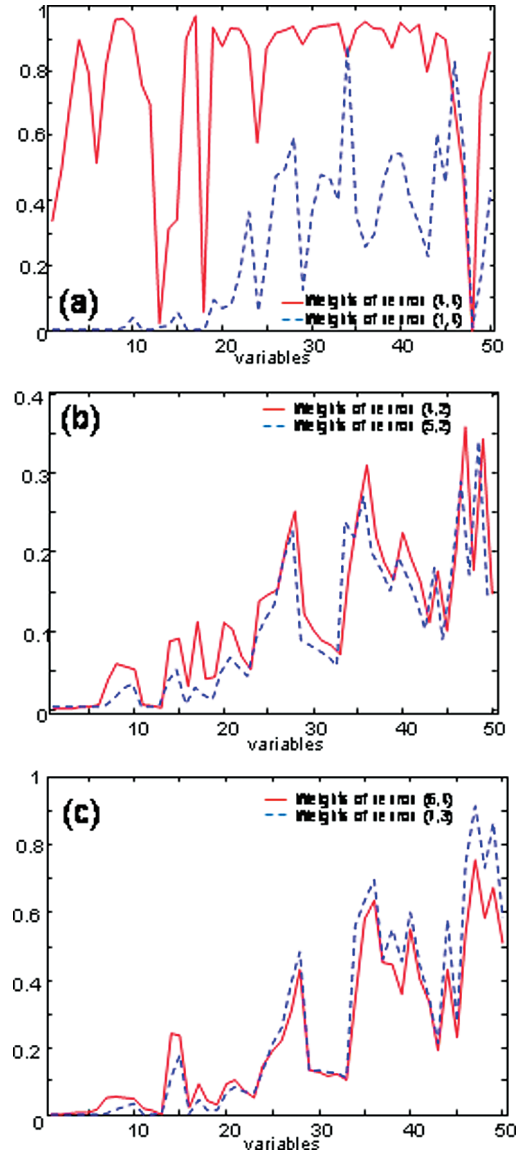

Figure 5. Profiles of the weights associated to the experimental variables of selected neurons. The weights are plotted against the number of the PAH (see text for details). (a) Sahara, (b) IFO, and (c) Prestige.

most relevant PAHs were fluoranthene, C_1-fluoranthene, C_2-fluoranthene, benzo(ghi)perylene, and dibenz(ah)anthracene; variables #34, 35, 36, 47, and 49, respectively. The aging process can be resumed by a decrease of the weights of the lightest PAHs, #6−10 (naphthalene to C_4-naphthalene), #17 (acenaphthylene), #21 (C_1-fluorene), #14 to 15 (C_3- and C_4-benzothiophene), and an increase on the importance of the heaviest PAHs #34 (fluoranthene), #36 (C_2-fluoranthene), #38 (C_4-fluoranthene), #40 (C_1-chrysene), #25−28 (C_1- to C_4-dibenzothiophene), #44 (benzo(b)fluoranthene and benzo(k)fluoranthene), #47 (perylene), and #49 (dibenz(ah)anthracene), as expected.

Finally, it is worth noting that all validation samples were projected in neurons corresponding to the correct oil products and weathering stage. Also, recall that some of the PAHs characterizing the most weathered samples in almost all six oil products were the heaviest ones, some of which (e.g., benzo(a)pyrene, variable #46) are recognized to be carcinogenic species.[1]

As a matter of conclusion, it can be stated that the MOLMAP approach resulted in a useful method to handle environmental 3-way data. Information was obtained for each mode of the data cube (samples, products, and variables); thus, the different behavior of the six studied oil products was observed, along with their particular evolution on time, and the weathering patterns were studied in terms of the original PAHs.

ACKNOWLEDGMENT

M.P.G.-C. acknowledges an "Ángeles Alvariño" research contract (PGIDIT, Xunta de Galicia (Galician Government)), as well as postdoc grants from the Xunta de Galicia and the Ministry of Education and Science ("José Castillejo" grants) to stay at the Universidade Nova de Lisboa and the Università degli Studi di Milano-Bicocca. The Galician Research Program (07MDS031103PR) is acknowledged for its support.

Received for review March 18, 2010. Accepted April 14, 2010.

AC100706J