SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 29, No. 1, pp. 64–72, 2010

Printed in the USA
DOI: 10.1002/etc.24

HIERARCHICAL CLUSTER ANALYSIS OF POLYCHLORINATED DIOXINS AND FURANS IN MICHIGAN, USA, SOILS: EVALUATION OF INDUSTRIAL AND BACKGROUND CONGENER PROFILES

TIMOTHY P. TOWEY,*†‡ SHU-CHI CHANG,†§ AVERY DEMOND,† DANIEL WRIGHT,† NOÉMI BARABÁS,‡ ALFRED FRANZBLAU,‖ DAVID H. GARABRANT,‖ BRENDA W. GILLESPIE,# JAMES LEPKOWSKI,†† WILLIAM LUKSEMBURG,‡‡ and PETER ADRIAENS†

†Department of Civil and Environmental Engineering, University of Michigan College of Engineering, Ann Arbor, Michigan 48109, USA

‡LimnoTech, 501 Avis Drive, Ann Arbor, Michigan 48108, USA

§Department of Environmental Engineering, College of Engineering, National Chung Hsing University, 250 Kuo Kuang Road, Taichung 40227, Taiwan

‖Department of Environmental Health Sciences, University of Michigan School of Public Health, Ann Arbor, Michigan 48109, USA

#Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan 48109, USA

††University of Michigan Institute for Social Research, Ann Arbor, Michigan 48109, USA

‡‡Vista Analytical Laboratory, 1100 Windfield Way, El Dorado Hills, California 95762, USA

(*Submitted* 28 *October* 2008; *Returned for Revision* 21 *March* 2009; *Accepted* 14 *July* 2009)

Abstract—As part of the University of Michigan Dioxin Exposure Study, soil samples were collected from 766 residential properties near the Tittabawassee River between Midland and Saginaw; near the Dow Chemical Facility in Midland; and, for comparison, in the other areas of Midland and Saginaw Counties and in Jackson and Calhoun Counties, all located in the state of Michigan, USA. A total of 2,081 soil samples were analyzed for 17 polychlorinated dibenzo-*p*-dioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs). In order to better understand the distribution and sources of the PCDD/F congeners in the study area, hierarchical cluster analysis (HCA) was used to statistically group samples with similar congener patterns. The analysis yielded a total of 13 clusters, including: 3 clusters among the soils impacted by contamination present in the Tittabawassee River sediments, a cluster comprised mainly of samples collected within the depositional area of the Dow incinerator complex, a small cluster of samples with elevated 2,3,7,8-tetrachlorinated dibenzo-*p*-dioxin (TCDD), and several clusters exhibiting background patterns. The clusters related to the Tittabawassee River floodplain contamination all contained elevated PCDF levels and were differentiated from one another primarily by their relative concentrations of higher-chlorinated PCDDs, a difference likely related to both extent and timing of impacts from Tittabawassee sediments. The background clusters appear to be related to combustion processes and are differentiated, in part, by their relative fractions of TCDD. Thus, HCA was useful for identifying congener profile characteristics in both contaminated and background soil samples. Environ. Toxicol. Chem. 2010;29:64–72. © 2009 SETAC

Keywords—Polychlorinated dibenzo-*p*-dioxins Polychlorinated dibenzofurans Cluster analysis Multivariate statistics Tittabawassee River

INTRODUCTION

The University of Michigan Dioxin Exposure Study (UMDES) was undertaken to evaluate the impact of the discharge of polychlorinated dibenzo-*p*-dioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) from the Dow Chemical Company facilities in Midland, Michigan, USA, on the residents' body burdens of these compounds [1]. The Dow Chemical Company began operations in Midland, Michigan in 1897 and continues to the present. Chemical processes at Dow that may have resulted in the historic discharge of PCDDs and PCDFs to the environment include: electrolysis processes in the 1910s [2]; chlorophenol production, which started in the late 1930s and continued until 1980 [3]; and the incineration and open burning of waste materials dating back to the 1930s [2]. To investigate the impact of contamination on human exposures, the study included participants from four populations in Midland, Saginaw, and part of Bay Counties (MI), and from a comparison population in Jackson and Calhoun Counties (MI) and comprised measurements of dioxin-like compounds in soil, household dust, and serum, as well as the administration of a questionnaire.

Previous studies of dioxin-like compounds in soils in the vicinity of the Tittabawassee River have found that toxic equivalent (TEQ) levels are elevated, and that PCDFs are the most significant contributor to the TEQ [4]. Emissions from the incinerator located at the Dow Chemical plant in Midland have been shown to resemble those from other hazardous waste combustion facilities [5] with PCDDs as the primary contributor to TEQ. Summary statistical analysis of the congener distributions found in the soils collected as part of the UMDES shows agreement with these general conclusions [6], yet the analysis was based on looking at the means of the primary contributors to the TEQ. In order to better understand the patterns of PCDDs and PCDFs in the environment to which the study participants

All Supplemental Data may be found in the online version of this article.

* To whom correspondence may be addressed (ttowey@limno.com).

Published online 23 September 2009 in Wiley InterScience (www.interscience.wiley.com).

may have been exposed, hierarchical cluster analysis (HCA) was applied to the soil dataset.

Cluster analysis is used to better understand datasets through the quantitative grouping of items with similar properties [7]. In the context of evaluating congener patterns of dioxin-like compounds, cluster analysis has been used to both group variables (congeners) [8,9] and observations (samples) [10–15]. Wenning et al. [10] used both principal components analysis (PCA) and cluster analysis to identify five distinct sources in 19 samples collected from Newark Bay, New Jersey, USA. No source identifications were made; however, they concluded that the congener patterns in Newark Bay samples were different from those collected from a nearby former 2,4,5-trichlorophenoxyacetic acid manufacturing plant. In a related study [11], the congener profiles of the Newark Bay samples were demonstrated to be similar to those from other industrialized waterways. Hagenmaier et al. [12] used HCA to evaluate sewage sludge samples from 30 wastewater plants in Germany. By grouping the profiles based on relative congener concentrations of the 2,3,7,8-substituted congeners with known sources, the researchers were able to infer linkages with aerial deposition, river sediments, automobile exhaust, and pentachlorophenol. Götz et al. [14] found HCA to be the method that yielded the most plausible results for evaluating potential PCDD/F sources in and near the River Elbe in Germany, based on a comparison of several multivariate statistical techniques. Their analysis suggests that the contamination in the River Elbe is related to the industrial center of Bitterfield, and that the contamination may be due to both chemical production and metallurgical processes.

These studies utilized samples collected primarily from contaminated areas, and the number of samples analyzed ranged from 19 to 407. In contrast, the soil samples obtained as part of UMDES included those contaminated by the flooding of the Tittabawassee River, by the aerial deposition from incinerators, as well as samples taken from areas with no known industrial impact. In addition, a total of 766 residential properties were sampled, resulting in a dataset of 2,081 samples. The present study examines the utility of hierarchical cluster analysis in evaluating a large dataset that includes samples collected from dispersed geographic regions, and in evaluating trends in both background samples and those impacted by industrial sources. Cluster centroid analysis is used to infer source attributes from characteristic patterns in each cluster, and a visualization technique to display the results of cluster analysis for large environmental datasets is proposed.

METHODS

Study populations

Five populations in Midland, Saginaw, Bay, Jackson, and Calhoun Counties, Michigan were sampled using a two-stage area probability household sample design [1]. The five populations were: (Tittabawassee River) Floodplain (203 properties with soil samples), Near Floodplain (164 properties), Midland Plume (37 properties), Other Midland/Saginaw (168 properties), and Jackson/Calhoun (194 properties). The Floodplain population included respondents whose property was partially or wholly within the Federal Emergency Management Agency (FEMA) defined 100-year floodplain and respondents outside of the FEMA floodplain who indicated that their property had been flooded by the Tittabawassee River. The Near Floodplain population was defined as respondents who lived in census blocks that included areas in the 100-year FEMA floodplain, but whose properties themselves were outside of the FEMA floodplain and had not, to the respondents' knowledge, been flooded by the Tittabawassee River. The Plume represents the residents in the city of Midland whose properties were likely impacted by the historical discharges from incinerators at the Dow facility, as delineated using a paired atmospheric-transport and geostatistical model [16]. The Other Midland/Saginaw population represented areas of Midland, Saginaw, and part of Bay Counties outside of the Floodplain, Near Floodplain, and Plume. A map showing the location of the four Midland and Saginaw County study populations is included as Figure 1. Jackson and Calhoun counties were selected for comparison because they are demographically similar to Midland and Saginaw, but are located approximately 200 km from Midland and are not impacted by emissions from Dow. In order to be eligible for soil sampling in UMDES, subjects were required to have lived in their residence at least five years and had to be the owner of their residence and property. A detailed description of the subject selection methodology is available in Garabrant et al. [1].

Soil sample collection and analysis

Each selected property was sampled in multiple locations from the surface to a depth of 15 cm (6 inches). Up to three sets of samples were collected from each property: around the perimeter of the residence (house perimeter), from the gardens where skin contact was likely (garden), and, for those properties located in the Floodplain, near the river (near river). The cores from the house perimeter and near river sets were separated and composited into two strata: 0 to 2.5 cm (1 inch) and 1 to 6 inches (1–6 inches). Samples from the garden set were composited in their entirety from 0 to 15 cm. From each property, up to five composited soil samples were produced (2 house perimeter, 1 soil contact, 2 near river). Not all composited samples were submitted for analysis. The house perimeter 0 to 2.5 cm sample and, if collected, the garden 0 to 15 cm sample were analyzed from every residence. Additional samples from some residences were submitted based on the population from which the sample was collected and the results of the house perimeter 0 to 2.5 cm sample [6]. A total of 2,081 soil samples from 766 properties were analyzed for the World Health Organization (WHO) designated 29 PCDD, PCDF, and PCB Congeners by Vista Analytical Laboratory (El Dorado Hills, CA, USA) using internal modifications of U.S. Environmental Protection Agency (U.S. EPA) Method 1688A [17] and U.S. EPA Method 8290 [18]. Eighty-three blind duplicate samples were submitted. The average relative percent difference between duplicate samples for all congeners was 19.6%. Additionally, eight samples from the soils analyzed as part of the Ninth International Intercalibration Study [19] (http://www.intercal.se/documents/Final_Report_9th_round_2004.pdf) were submitted as reference material. The average relative percent difference between the results from the samples submitted as part of the present study and the median of the Intercalibration Study results was 17.1%. Toxic equivalent values were calculated using the 2005 WHO toxic equivalency factors for PCDDs and

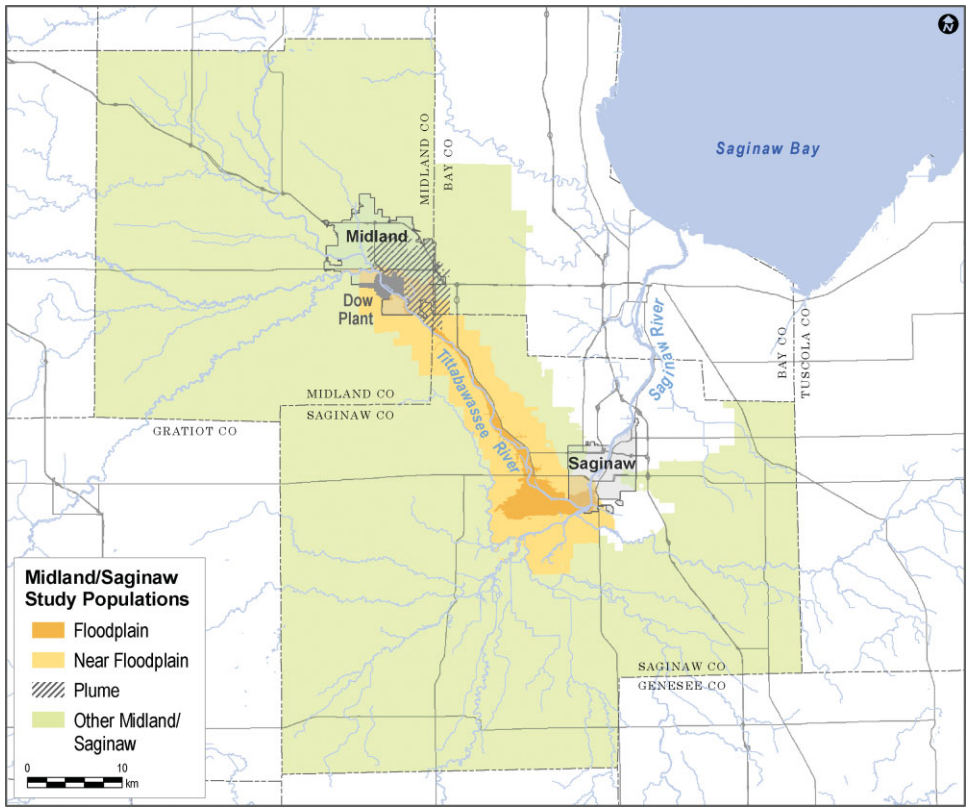

Fig. 1. Map showing locations of the University of Michigan Dioxin Exposure Study populations in Midland and Saginaw Counties, Michigan, USA.

PCDFs [20] and are denoted as TEQ_{DF2005}. Further details regarding soil sample collection and analysis, along with summary statistics of the soil results, are presented in Demond et al. [6]. The WHO-29 includes only 2,3,7,8-substituted PCDD/F congeners. Although a full set of tri- through octa-chlorinated congeners can be useful for source identification, many published source profiles are based only on the 17 PCDD/F congeners included in the WHO-29 [21,22].

Data treatment and transformation

The results presented here are based on the 17 PCDDs and PCDFs because they are the more significant contributors to both soil and serum TEQ levels in the target populations. Preliminary analysis suggested that the inclusion of PCBs obscured findings related to PCDDs and PCDFs. Also, the use of only the PCDD/F congeners allowed for comparison with published source profiles. If the concentration of a particular congener was below the limit of detection (LOD), the concentration was recorded as $LOD/\sqrt{2}$ [23]. Since the congener data exhibited log-normal distributions, a natural logarithm transformation of $\ln(x + 1)$ was undertaken. The addition of 1 prevents the variability of very low concentrations from unduly influencing the results. A constant-row-sum transformation was used, in which the sum of each row was converted to unity and the natural-logarithm-transformed concentration value of each congener in each sample was converted to a fraction of unity. Finally, a range transformation, as described by Johnson et al. [24], was applied to each congener. The natural-logarithm and constant-row-sum transformations reduce the influence of samples with high concentrations, while the range transformation reduces the influence of congeners with high variability. The goal of the present study was to identify large-scale trends in both industrially impacted and background soils in order to better understand potential sources of exposure. These transformations increase the formation of large clusters and decrease the formation of clusters with only a few samples.

Hierarchical cluster analysis

As a data reduction step and to make preliminary inference about inter-congener relationships, principal components analysis was performed on the correlation matrix of the transformed data using Minitab15 [25] software. The principal components that accounted for 95% of the cumulative variance were selected for further use in the HCA.

Hierarchical cluster analysis was performed using the Minitab15 Cluster Observations utility by grouping similar soil samples based on their principal component scores. This is an agglomerative clustering tool, meaning that the process starts with all of the samples as separate clusters and then merges the two most similar clusters in each step. The similarity of clusters is determined based on their positions in multidimensional space (in this case, their positions based on a plot of principal component scores). Choice of linkage method (between which part of the clusters similarity is measured) and model size (number of clusters) may affect how the clusters are agglomerated, an exploratory analysis was performed to determine the appropriate linkage method (average, centroid, complete, median, or single) and model size (7–14 clusters). As was previously noted, the goal of the analysis was to identify

large-scale trends in the dataset. Therefore, the selection of linkage method was based on the numbers of clusters formed that contained large numbers of samples. The decision regarding model size was evaluated by applying a knee-of-the-curve, or elbow criterion, on a plot of similarity (calculated as the ratio of the minimum distance at that agglomeration step to the maximum interobservation distance in the dataset [26]) as a function of the number of clusters.

Visualization of results

Minitab15 software allows for the creation of a distance to cluster-centroid matrix. The sample within each cluster that was closest to the centroid was selected to represent that cluster. A congener pattern for each cluster centroid was produced using the original congener-specific soil concentrations from the selected sample. Both original concentration, as a fraction of total PCDD/Fs, and contribution to TEQ_{DF2005} patterns were produced. To allow for the visualization of the congener pattern of the large dataset (a matrix of 2,081 samples by 17 congeners), a heatmap was used. This is a common technique in genetic microarray studies to represent the results of cluster analysis [27,28]. The constant-row-sum transformed data (not range transformed) were sorted according to cluster membership and then by TEQ. Using the sorted data, a heatmap was generated using a Visual Basic code in Microsoft Excel 2008 that colors a worksheet cell based on the magnitude of each congener concentration of each sample in the dataset.

RESULTS AND DISCUSSION

Cluster process results

Seven principal components were determined to explain 95% of the cumulative variance in the dataset (Supplemental Data, Fig. S1) and were selected to be used in the HCA. The contribution of each congener to each principal component (PC) is presented in Table 1. Principal component 1 is characterized by large positive contributions from most of the PDCF congeners; PC 2 is characterized by high negative contributions from most of the PCDD congeners, and PC 3 has large positive contributions from the higher-chlorinated PCDFs. The other PCs contain a mix of contributions from both PCDDs and PCDFs.

To evaluate the linkage methods, models with seven clusters were evaluated, corresponding with the number of principal components used in the analysis. All of the linkage methods, except for complete and average linkage, created a single large cluster and six small clusters with fewer than 30 samples each. Complete linkage created the most clusters with at least 100 samples and the fewest clusters with less than 30 samples, and was, therefore, selected as the appropriate method. Complete linkage was used in the evaluation of model size. The application of an elbow criterion on a plot of similarity as a function of model size (Supplemental Data, Fig. S2) was used to determine the appropriate number of clusters. Although a number of reasonable choices for the elbow exist, a clear increase in similarity occurs when moving from 12 to 13 clusters; additional increases by including more clusters are smaller. This additional information gained from moving from 12 to 13 clusters was evaluated in the context of the study populations. The additional cluster split a portion of elevated TEQ samples into two groups, one of which was more prevalent in the Floodplain population and the other more prevalent in the Near Floodplain population. Since the split was related to elevated TEQ samples and suggested impacts that varied geographically, 13 clusters were retained.

The clustering of samples using complete linkage and 13 clusters is illustrated in Figure 2 and Supplemental Data, Figure S3. Figure 2 shows a 3-D score plot of the four clusters that included samples with TEQ_{DF2005} values greater than the state of Michigan direct contact soil criteria of 90 pg/g; only these four clusters are included for ease of visualization. Clusters 3, 5, and 8, particularly cluster 8, have high scores for PC 1, the PC with high contributions from PCDFs. Cluster 6 has a large negative contribution from PC 2, which has high negative contributions from the PCDDs, indicating a positive contribution from PCDDs in cluster 6. Supplemental Data, Figure S3 is a matrix of 2-D score plots for all seven PCs used in the HCA and includes all 13 clusters.

Table 1. Congener coefficients for the seven principal components retained for use in hierarchical cluster analysis[a]

Congener	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
2,3,7,8-TCDD	0.00	−0.30	−0.43	−0.72	−0.20	0.07	−0.31
1,2,3,7,8-PeCDD	0.01	−0.45	−0.23	−0.08	0.04	−0.07	0.51
1,2,3,4,7,8-HxCDD	−0.06	−.45	−0.07	0.20	0.12	−0.03	0.44
1,2,3,6,7,8-HxCDD	−0.13	−0.40	0.13	0.19	0.19	−0.09	−0.53
1,2,3,7,8,9-HxCDD	−0.16	−0.38	−0.04	0.32	0.04	−0.28	−0.26
1,2,3,4,6,7,8-HpCDD	−0.32	0.09	0.01	0.16	−0.15	0.07	−0.09
OCDD	−0.29	0.19	−0.10	0.13	−0.26	0.14	0.12
2,3,7,8-TCDF	0.29	0.14	−0.22	0.04	0.28	−0.22	−0.03
1,2,3,7,8-PeCDF	0.31	0.10	−0.12	0.02	0.17	−0.13	0.00
2,3,4,7,8-PeCDF	0.31	0.09	0.00	0.03	0.03	−0.31	−0.10
1,2,3,4,7,8-HxCDF	0.31	0.03	−0.02	0.01	0.26	−0.03	−0.02
1,2,3,6,7,8-HxCDF	0.24	−0.12	0.38	−0.10	−0.51	−0.21	0.17
1,2,3,7,8,9-HxCDF	0.31	−0.04	0.04	0.07	0.03	0.36	0.01
2,3,4,6,7,8-HxCDF	0.23	−0.18	0.42	−0.01	−0.33	−0.06	−0.12
1,2,3,4,6,7,8-HpCDF	−0.21	0.01	0.48	−0.44	0.38	−0.33	0.14
1,2,3,4,7,8,9-HpCDF	0.23	−0.25	0.27	0.00	0.21	0.64	−0.06
OCDF	−0.28	0.09	0.23	−0.21	0.29	0.17	0.04

[a] PCs = principal components; T = tetra; Pe = penta; Hx = hexa; Hp = hepta; O = octa; CDD = chlorinated dibenzo-*p*-dioxin; CDF = chlorinated dibenzofuran.

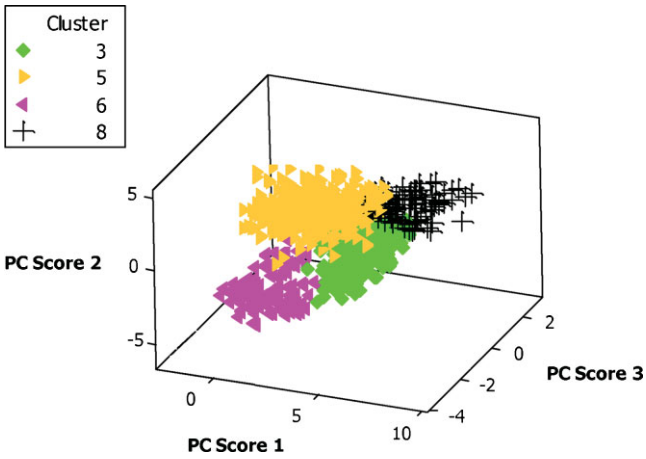

Fig. 2. Principal component (PC) score plot of first three PCs for the four clusters which contain samples with TEQ_{DF2005} (toxic equivalent based on 17 polychlorinated dibenzo-*p*-dioxins and polychlorinated dibenzofurans using the World Health Organization 2005 toxic equivalency factors) concentrations >90 pg/g.

Cluster summary statistics

To aid in the interpretation of the 13 clusters, their characteristics are shown in Table 2, including the number of samples, the mean TEQ_{DF2005}, and percent contributions to the TEQ_{DF2005} from PCDDs and PCDFs (note: the cluster numbering is based on the order that the clusters are formed in Minitab). Four clusters (3, 5, 6, and 8) include samples with TEQ_{DF2005} values above 90 pg/g. The PCDFs contribute a high percentage of the TEQ_{DF2005} in clusters 3, 5, and 8; in contrast, the PCDDs contribute a high percentage to the TEQ_{DF2005} in cluster 6. Clusters 1, 2, 4, and 7 include at least 100 samples, contain no samples above 90 pg/g TEQ_{DF2005}, and have mean TEQ_{DF2005} values of less than 10 pg/g; consequently, they seem to represent the background.

Figure 3 shows the distribution of cluster membership for each of the study populations. Both the clusters with elevated TEQ values and the background groupings vary by study population. The clusters with elevated PCDFs are found primarily in the Floodplain and Near Floodplain soils, and the cluster with elevated PCDDs is strongly associated with the Plume, comprising 84% of the soils from that population. Clusters 1 and 7, associated with the background, are primarily associated with the Jackson/Calhoun population; cluster 4 samples are found primarily in the Midland/Saginaw populations, with the exception of the Plume; and cluster 2 includes samples from all of the study populations, again with the exception of the Plume. Thus, the cluster separation allows for interpretation of the congener profiles in the context of region, and thus implies a differentiation of sources such as incineration, discharge of industrial process waste, deposition from long-range atmospheric transport, and emissions associated with small-scale processes.

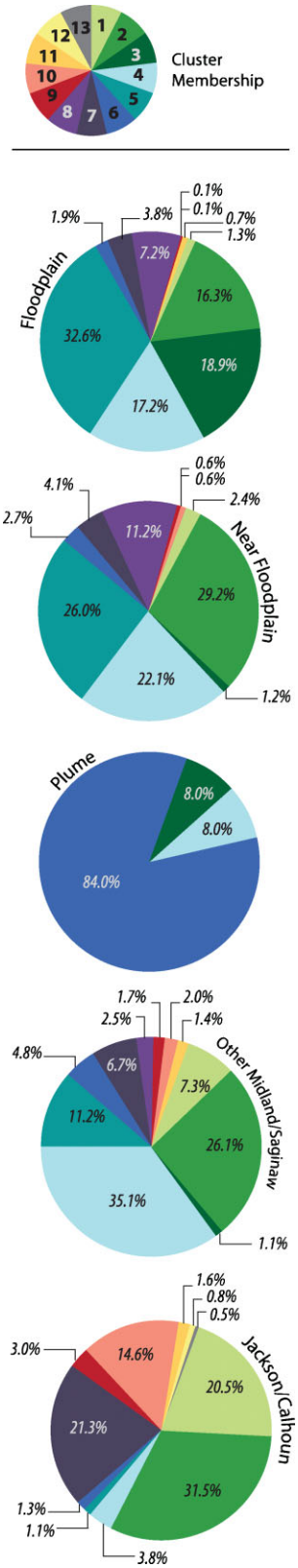

Fig. 3. Pie charts showing distribution of cluster membership for soil samples collected from each study population.

Cluster centroid profiles

To better understand the origins of differences in the various clusters, cluster centroid profiles were extracted and compared. Figure 4 presents the centroid congener profiles of the four large background clusters (1, 2, 4, and 7) and four elevated TEQ clusters (3, 5, 6, and 8), generated using the sample closest to the center of the cluster. The fractional contribution of each congener to the total PCDD/Fs and the fractional contribution to the TEQ_{DF2005} are presented. The profiles are ordered to allow for

Table 2. Mean TEQ_{DF2005} concentrations and percent contributions to TEQ_{DF2005} from PCDDs and PCDFs for each cluster[a]

Cluster	No. of samples	Mean TEQ_{DF2005} (pg/g)	Mean PCDD contribution to TEQ_{DF2005} (%)	Mean PCDF contribution to TEQ_{DF2005} (%)
1	122	4.95	68	32
2	458	2.35	65	35
3	189	709	16	84
4	379	8.97	68	32
5	430	21.3	25	75
6	132	59	77	23
7	152	8.78	45	55
8	113	394	4	96
9	20	5.12	85	15
10	64	0.735	71	29
11	17	10.9	92	8
12	3	1.61	52	48
13	2	29.5	88	13

[a] TEQ_{DF2005} = toxic equivalent based on 17 PCDDs and PCDFs using the World Health Organization 2005 toxic equivalency factors; PCDDs = polychlorinated dibenzo-*p*-dioxins; PCDFs = polychlorinated dibenzofurans.

comparison between similar clusters. These profiles show that the background clusters appear to be very similar to one another and to profiles obtained in another background sampling study [22] (Supplemental Data, Fig. S4). The background clusters are differentiated primarily by their relative fractions of lesser-chlorinated PCDDs, particularly 2,3,7,8-TCDD. It is notable that cluster 4, which rarely occurs in Jackson/Calhoun, has a profile that more closely resembles samples found in the Plume (cluster 6). This suggests that the samples in cluster 4 may have been impacted by incineration at the Dow facility. For the house perimeter 0 to 2.5 cm samples in the Other Midland/Saginaw population, the average distance from the Dow facility in Midland is 17.5 km for cluster 4 samples, 27.1 km for cluster 2 samples, and 37.0 km for cluster 1 samples. Thus, proximity to Dow corresponds to an increased fraction of TCDD. Clusters 1 and 7, which are found frequently in Jackson/Calhoun but rarely in any of the Midland/Saginaw populations, have the lowest fraction of TCDD. Clusters 1 and 7 appear to have very similar profiles in terms of fractions of total PCDD/F; however, cluster 7 has a larger PCDF contribution to TEQ, particularly from 2,3,4,7,8-penta-CDF. The increased fraction of 2,3,4,7,8-penta-CDF is not present in the background clusters associated with the Midland and Saginaw populations, suggesting the presence of a PCDF source in or near Jackson and Calhoun Counties.

Of the clusters that include samples with elevated TEQ values, cluster 6, associated with the Plume population, is distinct in that PCDDs, particularly the lower chlorinated congeners, are the primary contributors to the TEQ. The profile of this cluster is similar to those of the background clusters. The pattern is comparable to a number of combustion-related patterns (including those of diesel fuel combustion, forest fires, and municipal waste incineration shown in Supplemental Data, Fig. S4) found in the U.S. EPA Inventory of Sources and Environmental Releases of Dioxin-Like Compounds [21] in that there is a larger fraction of PCDDs compared to PCDFs, and the proportion of the PCDD congener increases with increasing chlorination (e.g., OCDD is present in a larger proportion than 1,2,3,4,6,7,8-hepta-CDD). The resemblance of background samples with those found in the vicinity of an incinerator is consistent with the findings of Schuhmacher et al. [13].

The other three clusters (3, 5, and 8) with elevated TEQ values all have substantial PCDF contributions and appear to be related to sources in the Tittabawassee River Floodplain. These clusters are differentiated mainly by their relative fractions of PCDDs versus PCDFs, rather than by the distribution of the congeners within those families. Because cluster 5 has a much lower mean TEQ and a lower relative fraction of PCDFs relative to the other clusters, it appears that this cluster represents dilute-floodplain samples, or a mix of contributions from the Tittabawassee and background sources. However, this dilution effect does not differentiate clusters 3 and 8. The relative fraction of PCDFs in cluster 8 is higher than in cluster 3, but cluster 8 has a lower mean TEQ. Instead of dilution, the difference may be related to the relative impacts from separate industrial sources. The processes resulting in elevated PCDF levels in the Tittabawassee River floodplain are likely related to wastes from chlor-alkali production using graphite electrodes prior to the installation of a wastewater treatment system in the 1920s [2], as the profile in the Tittabawassee is dominated by PCDFs, similar to published graphite electrode sludge measurements [21]. Production of pentachlorophenol, which contains high levels of OCDD [21], occurred at the Dow facility in the period from 1937 to 1989 [3]. Cluster 8 samples may consist of sediments that were either deposited by, or moved from, the riverbed several decades ago, following the discharge of chlor-alkali related wastes but prior to the discharge and transport of pentachlorophenol-related wastes. The fact that the cluster with very low PCDD fractions is more prevalent in the Near Floodplain, where the contamination of soils may be the result of anthropogenic soil movement, rather than in the Floodplain, supports this hypothesis. Further investigation related to anthropogenic soil movement in the region is presented in Franzblau et al. [29].

Heatmap representation

The heatmap presented in Figure 5 shows the fraction of each congener in each soil sample in the analysis. In order to evaluate which congener patterns are associated with elevated TEQ levels, a bar was placed adjacent to the samples to indicate those that exceeded 12.2 pg/g TEQ_{DF2005}, the 95th percentile of the Jackson/Calhoun house perimeter 0 to 2.5 cm samples, and those that exceeded 90 pg/g TEQ_{DF2005}, the direct soil contact criteria for the state of Michigan. The heatmap reinforces trends

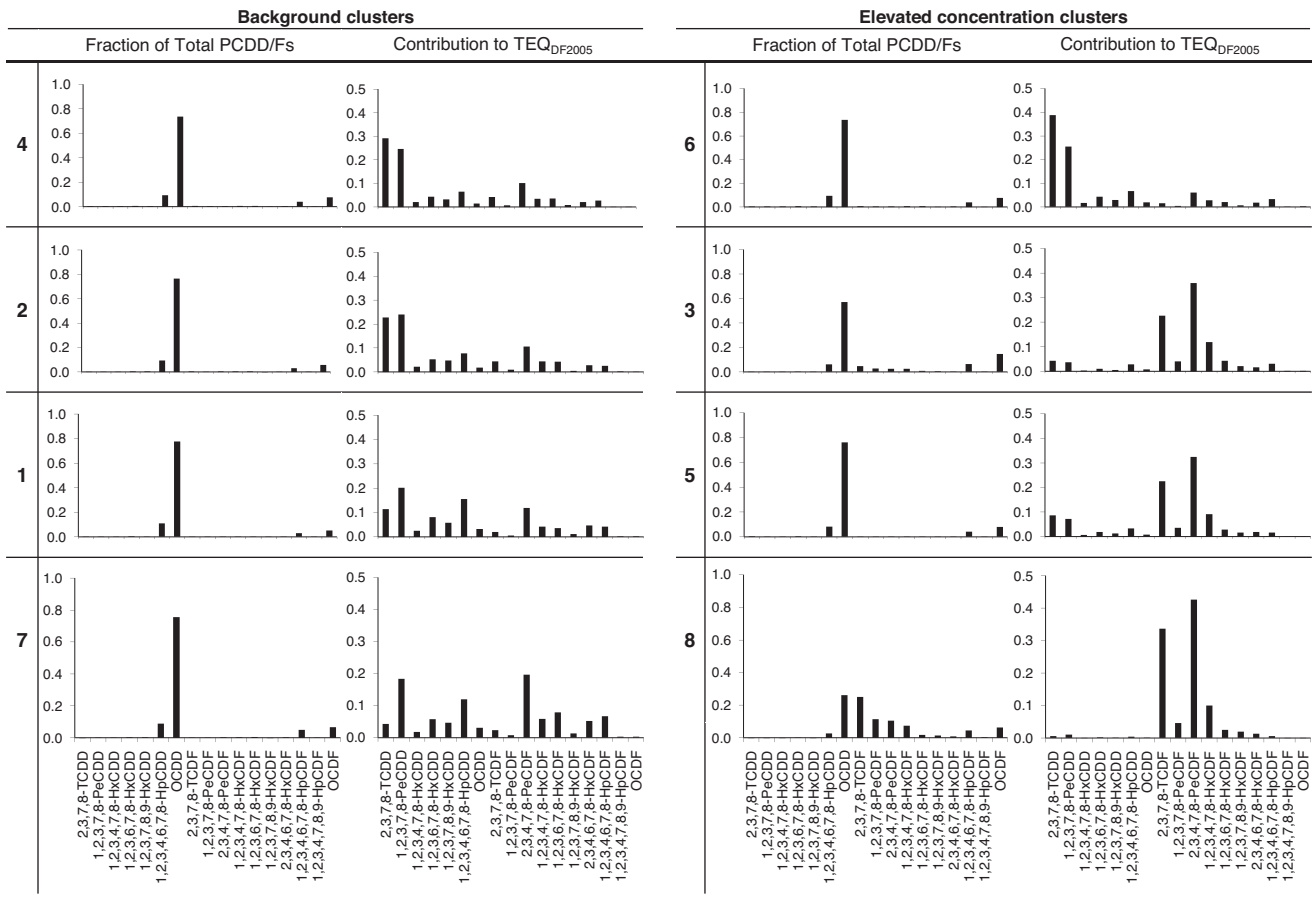

Fig. 4. Congener profiles of background and elevated concentration cluster centroids. Profiles ordered to facilitate comparison of similar profiles. T = tetra; Pe = penta; Hx = hexa; Hp = hepta; O = octa; CDD = chlorinated dibenzo-*p*-dioxi; CDF = chlorinated dibenzofuran; TEQ_{DF2005} = toxic equivalent based on 17 polychlorinated dibenzo-*p*-dioxins and polychlorinated dibenzofurans using the World Health Organization 2005 toxic equivalency factors.

identified in the cluster centroid analysis regarding the relative fractions of PCDDs and PCDFs associated with each cluster. For example, the larger fraction of PCDFs found in clusters 3, 5, and 8 is apparent from the dark red colors of the PCDF columns. This method of representation also allows an assessment of the homogeneity of the clusters. For example, it appears that the clusters with elevated TEQ values appear to have greater intra-cluster variability than the other clusters, perhaps because being impacted by a particular source differentiates samples significantly enough that they tend to form clusters even in cases when those samples contain large ranges of concentrations.

The heatmap also allows for examination of the smaller clusters, clusters 9 to 13, which were not evaluated using centroid congener profiles. One prominent feature is the higher fraction of TCDD in clusters 11 and 13. The mean TEQ_{DF2005} of the 17 samples in cluster 11 is 10.9 pg/g, which is only slightly elevated compared to the mean from the Jackson/Calhoun background clusters (4.95 pg/g and 8.97 pg/g). However, the mean contribution to TEQ_{DF2005} from TCDD is 77%, as compared to 19% in the largest background cluster, cluster 2. Cluster 11 includes samples from the Floodplain, Other Midland/Saginaw, and Jackson/Calhoun populations, so it does not have any clear geographic ties. Cluster 13 consists of two samples from the same property in the Jackson/Calhoun population. The mean TEQ_{DF2005} from those two samples is 29.5 pg/g with a mean contribution of 81% from TCDD. Some phenoxy-herbicides are known to contain TCDD [21]. The dispersed geographic distribution of the samples in these two clusters is consistent with the application of TCDD-containing herbicides as a potential source.

CONCLUSIONS

The results of HCA in evaluating the UMDES soils dataset suggest that the soils that have been impacted by Tittabawassee River sediments or the Dow incinerators can be differentiated from background soils based on their congener profiles, and that all of the UMDES soil samples above the Michigan 90 pg/g TEQ_{DF2005} direct contact criteria are likely related to those two sources. The analysis also shows that a subset of the samples impacted by the Tittabawassee River contamination, particularly among the soils of the Near Floodplain population, appear to have been impacted by processes that generated elevated PCDF levels, but not processes that generated higher-chlorinated PCDDs. Further, Dow incineration processes may have had a small but measurable effect on some soils from the Midland and Saginaw populations outside of the Plume population, based on the fact that their relative levels of TCDD look more similar to the samples in the Plume than to samples in Jackson and Calhoun Counties. Finally, a small number of residences exhibit congener profiles possibly related to the application of chlorinated pesticides.

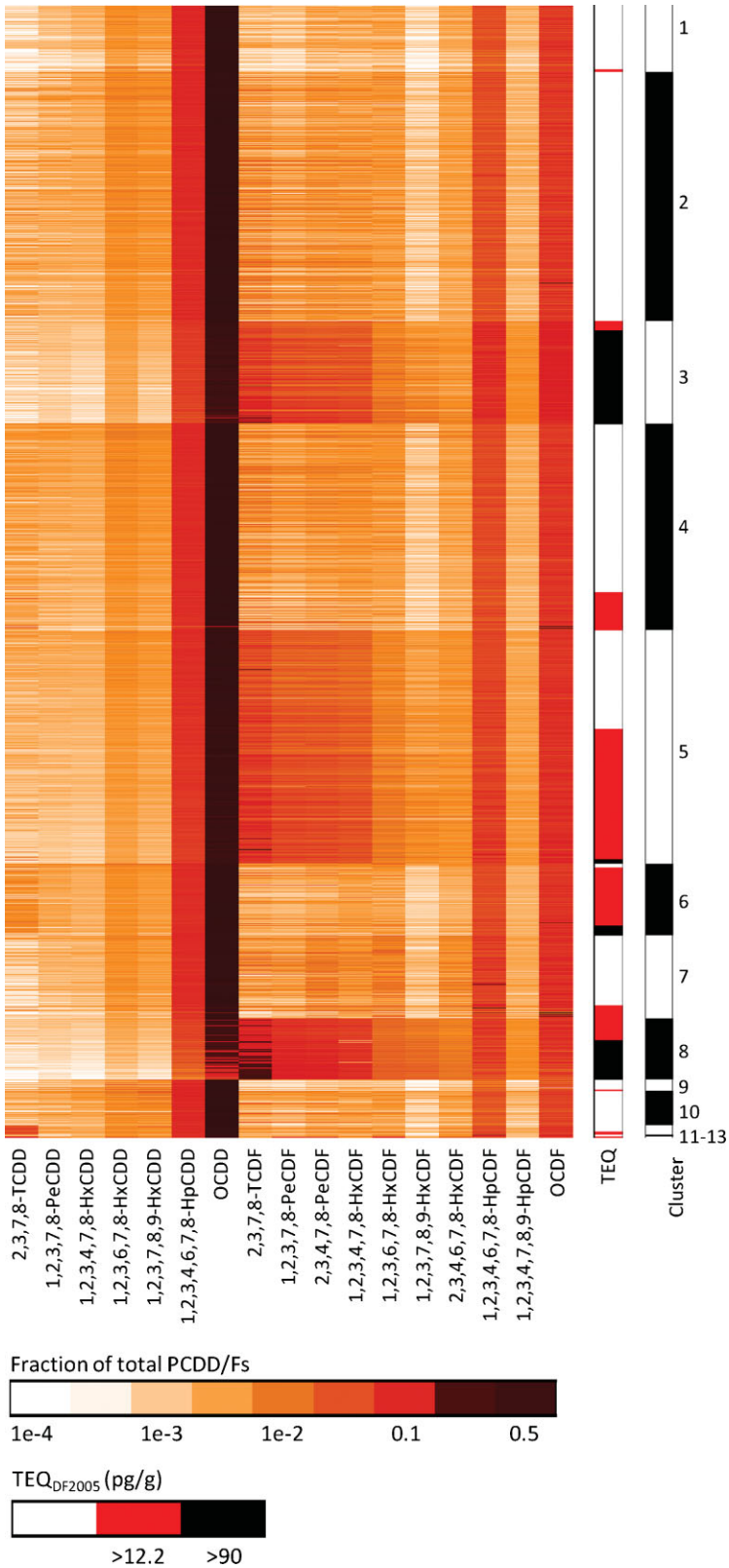

Fig. 5. Heatmap showing congener pattern of all soil samples collected as part of the University of Michigan Dioxin Exposure Study. The heatmap contains 2,081 rows, each of which represents a single sample. The column immediately to the right of the congener columns shows which samples have TEQ_{DF2005} concentrations >12.2 pg/g and 90 pg/g. The column to the far right shows the cluster break points. T = tetra; Pe = penta; Hx = hexa; Hp = hepta; O = octa; CDD = chlorinated dibenzo-*p*-dioxin; CDF = chlorinated dibenzofuran; TEQ_{DF2005} = toxic equivalent based on 17 polychlorinated dibenzo-*p*-dioxins and polychlorinated dibenzofurans using the World Health Organization 2005 toxic equivalency factors.

The present study demonstrates several benefits of the use of cluster analysis for the purpose of analyzing large and geographically dispersed datasets. The benefits include: the ability to make inferences about sources, the identification of smaller groups of samples with unusual congener patterns, the differentiation of samples that have been impacted by multiple sources to varying degrees, and the utility of HCA in creating visualizations to demonstrate results.

SUPPLEMENTAL DATA

Figure S1. Cumulative variance explained as a function of the number of principal components. Seven clusters retained using 0.95 threshold (indicated by dark dashed line).

Figure S2. Similarity as a function of the number of clusters. Thirteen clusters retained using elbow criterion.

Figure S3. Principal component (PC) score plots for seven components included in cluster analysis. Samples grouped by cluster membership.

Figure S4. Congener profiles of sources from U.S. Environmental Protection Agency (U.S. EPA) Inventory of Sources and Environmental Releases of Dioxin-Like Compounds and U.S. EPA Region 8 Denver Front Range study, including fraction of total analyzed polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/Fs) and the contribution to TEQ_{DF2005} (toxic equivalent based on 17 polychlorinated dibenzo-*p*-dioxins and polychlorinated dibenzofurans using the World Health Organization 2005 toxic equivalency factors). (567 KB PDF)

Acknowledgement—Financial support for the present study came from the Dow Chemical Company through an unrestricted grant to the University of Michigan. The authors thank Linda Birnbaum, Ronald A. Hites, Paolo Boffetta, Marie Haring Sweeney, and Sharyn Vantine.

REFERENCES

1. Garabrant DH, Franzblau A, Lepkowski J, Gillespie BW, Adriaens P, Demond A, Ward B, LaDronka K, Hedgeman E, Knutson K, Zwica L, Olson K, Towey T, Chen Q, Hong B. 2009. The University of Michigan Dioxin Exposure Study: Methods for an environmental exposure study of polychlorinated dioxins, furans, and biphenyls. *Environ Health Perspect* 117:803–810.
2. Ann Arbor, Technical Services. 2006. *Remedial Investigation Work Plan: Tittabawassee River and Upper Saginaw River*. Ann Arbor, MI, USA.
3. Collins JJ, Budinsky RA, Burns CJ, Lamparski LL, Carson ML, Martin GD, Wilken M. 2006. Serum dioxin levels in former chlorophenol workers. *J Exposure Sci Environ Epidemiol* 16:76–84.
4. Hilscherova K, Kannan K, Haruhiko N, Nobuyasu H, Nobuyoshi Y, Bradley PW, McCabe JM, Taylor AB, Giesy JP. 2003. Polychlorinated dibenzo-*p*-dioxin and dibenzofuran concentration profiles in sediments and flood-plain soils of the Tittabawassee River, Michigan. *Environ Sci Technol* 37:468–474.
5. U.S. Environmental Protection Agency. 1985. Soil screening survey at four midwestern sites. EPA 905/4-85-005. Environmental Services Division, Eastern District Office, Westlake, OH.
6. Demond A, Adriaens P, Towey T, Chang S-C, Hong B, Chen CW, Franzblau A, Garabrant D, Gillespie B, Hedgeman E, Knutson K, Lee CY, Lepkowski J, Olson K, Ward B, Zwica L, Luksemburg W, Maier M. 2008. Statistical comparison of residential soil concentrations of PCDDs, PCDFs, and PCBs from two communities in Michigan. *Environ Sci Technol* 42:5441–5448.
7. Massart DL, Kaufman L. 1989. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. Krieger, Malabar, FL, USA.
8. Pleil JD, Lorber MN. 2007. Relative congener scaling of polychlorinated dibenzo-p-dioxins and dibenzofurans to estimate building fire contributions in air, surface wipes, and dust samples. *Environ Sci Technol* 41: 7286–7293.
9. Antignac JP, Marchand CG, Gade C, Matayron G, Qannari EM, Bizec BL, Andre F. 2006. Studying variations in the PCDD/PCDF profile across various food products using multivariate statistical analysis. *Anal Bioanal Chem* 384:271–279.
10. Wenning RJ, Harris MA, Finley B, Paustenbach DJ, Bedbury H. 1993. Application of pattern recognition techniques to evaluate polychlorinated dibenzo-*p*-dioxin and dibenzofuran distributions in surficial sediments from the lower Passaic River and Newark Bay. *Ecotoxicol Environ Saf* 25:103–125.
11. Wenning RJ, Harris MA, Ungs MJ, Paustenbach DJ, Bedbury H. 1992. Chemometric comparisons of polychlorinated dibenzo-*p*-dioxin and dibenzofuran residues in surficial sediments from Newark Bay, New Jersey and other industrialized waterways. *Arch Environ Contam Toxicol* 22:397M–313.
12. Hagenmaier H, Lindig C, She J. 1994. Correlation of environmental occurrence of polychlorinated dibenzo-*p*-dioxins and dibenzofurans with possible sources. *Chemosphere* 29:2163–2174.
13. Schuhmacher M, Granero S, Xifro A, Domingo JL, Rivera J, Eljarrat E. 1998. Levels of PCDD/Fs in soil samples in the vicinity of a municipal solid waste incinerator. *Chemosphere* 37:2127–2137.
14. Götz R, Steiner B, Friesel P, Roch K, Walkow F, Maab V, Reincke H, Stachel B. 1998. Dioxin (PCDD/F) in the River Elbe—Investigations of their origin by multivariate statistical methods. *Chemosphere* 37:1987–2001.
15. Götz R, Lauer R. 2003. Analysis of sources of dioxin contamination in sediments and soils using multivariate statistical methods and neural networks. *Environ Sci Technol* 37:5559–5565.
16. Goovaerts P, Adriaens P, Demond A, Franzblau A, Garabrant D, Gillespie BW, Lepkowski J. 2008. Geostatistical modeling and spatial distribution of soil dioxins in the vicinity of an incinerator: 1. Theory and application to Midland, Michigan. *Environ Sci Technol* 42:3468–3654.
17. U.S. Environmental Protection Agency. 1999. Method 1668, Revision A: Chlorinated biphenyl congeners in water, soil, sediment, and tissue by high-resolution gas chromatography/high-resolution mass spectrometry (HRGC/HRMS). Washington, DC.
18. U.S. Environmental Protection Agency. 1994. Method 8290: Polychlorinated dibenzodioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) by high-resolution gas chromatography/high-resolution mass spectrometry (HRGC/HRMS). Washington, DC.
19. van Bavel B. 2004. Ninth Final Report, International Intercalibration Study. Workgroup International Intercalibration Studies, Orebro, Sweden.
20. Van den Berg M, Birnbaum LS, Denison M, De VM, Farland W, Feeley M, Fiedler H, Hakansson H, Hanberg A, Haws L, Rose M, Safe S, Schrenk D, Tohyama C, Tritscher A, Tuomisto J, Tysklind M, Walker N, Peterson RE. 2006. The 2005 World Health Organization reevaluation of human and mammalian toxic equivalency factors for dioxins and dioxin-like compounds. *Toxicol Sci* 93:223–241.
21. U.S. Environmental Protection Agency. 2006. An inventory of sources and environmental releases of dioxin-like compounds in the United States for the years 1987, 1995, and 2000. EPA/600/P-03/002F. Washington, DC.
22. U.S. Environmental Protection Agency. 2001. Denver front range study: Dioxins in surface soil. U.S. EPA Region 8, Denver, CO.
23. Hornung RW, Reed LD. 1990. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg* 5:46–51.
24. Johnson GW, Ehrlich R. 2002. State of the art report on multivariate chemometric methods in environmental forensics. *Environ Forensics* 3:59–79.
25. Minitab. 2006. MINITAB Statistical Software, Release 15 for Windows. State College, PA, USA.
26. Minitab. 2006. MINITAB Help. State College, PA, USA.
27. Kajiwara H, Nakamura M. 2008. Hierarchical cluster analyses and heat map analyses of silkworm tissues using R-statistics. *Jpn J Electrophor* 52:29–38.
28. Yi SG, Park T, Lee JK. 2008. Response projected clustering for direct association with physiological and clinical response data. *BMC Bioinformatics* 9:76–86.
29. Franzblau A, Demond A, Towey T, Adriaens P, Chang SC, Luksemburg W, Maier M, Garabrant D, Gillespie B, Lepkowski J, Chang WC, Chen QC, Hong B. 2008. Residences with anomalous soil concentrations of dioxin-like compounds in two communities in Michigan, USA: A case study. *Chemosphere* 74:395–403.