SENSITIVITY OF SOIL DIOXIN CONGENER PATTERN CLUSTER ANALYSIS FOR A COMMUNITY IN MICHIGAN, USA

Towey T[1], Adriaens P[2], Demond A[2], Chang, S-C[2], Gillespie B[3], Franzblau A[4], Garabrant D[4]

[1]LimnoTech, 501 Avis Dr, Ann Arbor, MI 48108 [2]Department of Civil and Environmental Engineering, University of Michigan College of Engineering, 1351 Beal, Ann Arbor, MI 48109; [2]Department of Biostatistics, University of Michigan School of Public Health, 109 Observatory, Ann Arbor, MI 48109; [3]Department of Environmental Health Sciences, University of Michigan School of Public Health, 109 Observatory, Ann Arbor, MI 48109

Keywords: Soil, Environmental samples, North America, PCDD/PCDF

Abstract

Sources of exposure to dioxins can be reflected in blood serum and often exhibit a regional character depending on the pathways deemed relevant to their contribution. With soil as a potential direct or indirect pathway of exposure, the sensitivity of hierarchical cluster analysis (HCA) of PCDD/PCDF concentrations in soil samples was investigated through comparisons with an initial understanding of PCDD/PCDF sources. Specifically, the sensitivity of HCA results to linkage method and cluster quantity was evaluated. For the soil samples collected as part of the University of Michigan Dioxin Exposure Study (UMDES), complete linkage best reflected the initial understanding of the data. Increasing the number of clusters offered additional resolution of potentially contributing PCDD/PCDF congener patterns. This analysis helped constrain the areas where variable sources of exposure are relevant, and may play a role in better defining exposure models for the UMDES population.

Introduction

The University of Michigan Dioxin Exposure Study was undertaken in response to concerns among the population of Midland and Saginaw Counties that the discharge of dioxin-like compounds from the Dow Chemical Company facilities in Midland have resulted in contamination of soils in the Tittabawassee River flood plain and areas of the City of Midland. There is concern that people's body burdens of polychlorinated dibenzodioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), and polychlorinated biphenyls (PCBs) may be elevated because of environmental contamination. A central goal of the UMDES was to determine the factors that explain variation in serum congener levels of PCDDs, PCDFs, and PCBs, and to quantify how much variation each factor explains.

One potential exposure source of particular interest to the population of Midland and Saginaw counties is residential soil. In order to better understand the distribution of PCDD/Fs in the soil of UMDES participants, an analysis of congener patterns in soil samples was performed using multivariate chemometric methods. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) were performed on the complete UMDES soil data set (2081 samples). Previous analysis indicated that the dominant source patterns were region-bound, and reflected incineration, other combustion, and direct industrial process effluent streams. The data further reflected substantial overlap between regions and sources, particularly when concentrations are at background levels. Hence, this study investigation the impact of assumptions on the sensitivity of the HCA results to linkage methods and cluster quantity were investigated.

Materials and Methods

Respondent selection, as well as soil sampling, compositing, and analysis methods, are described elsewhere.[1,2]

Data Preparation: All data received from the analytical laboratory went through a data cleaning step to ensure data quality. All values below limit of detection (LOD) were replaced with the LOD divided by the square root of two ($LOD/\sqrt{2}$) to create the untransformed soil data set.

Data Transformation: Since congener data exhibited log-normal distributions, a natural logarithm transformation of ln(x+1) was performed. A constant-row-sum transformation was used to convert the sum of each row 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 was applied to each column of the dataset to ensure the variation within each congener would be similar. The range transformation kept the PCA from being driven by several congeners with extreme variation.[3]

Principal Component Analysis: PCA was performed using Minitab[4] software. The seven principal components that accounted for 95% of the cumulative variance were selected for further use in the HCA.[5,6] For the analysis presented in this short paper, the soil PCB values were not used in the PCA. A previous paper describes initial results including PCB data.[7]

Hierarchical Cluster Analysis: HCA was performed based on Euclidean distance between samples based on principal component scores. The sensitivity of HCA to different linkage methods and cluster numbers was evaluated. The linkage methods that were evaluated were: average, single, complete, centroid, and mean. Description of each of these linkage methods is found elsewhere.[4,6] The sensitivity of cluster quantity on HCA results was evaluated using seven to 14 clusters. Seven clusters were used in the linkage comparison to correspond with the number of principal components that accounted for 95% of the cumulative variance.

The outcomes of the cluster analysis were evaluated by comparing the results to an initial understanding of the soil data. In the study region, two primary areas of soil contamination exist: the Tittabawassee River floodplain and the area in the plume of the former Dow incinerator in the City of Midland. The floodplain area contamination consists primarily of furans, likely the result of historic discharges associated with chloralkali operations at the Dow. The plume area soils contain more dioxins. The initial understanding of the data is that a distinct congener pattern is present in soils from the floodplain and a distinct congener pattern is present in the soils from the plume.

The inclusiveness and specificity of the HCA outcomes was evaluated with respect to the initial understanding of the data. To assist in the evaluation the inclusiveness of HCA results, a narrowly defined floodplain (NDFP) was delineated as soil samples from the floodplain region with a TEQ_{DF}-WHO_{05} greater than 100 pg/g. A narrowly defined plume (NDPL) was delineated as soil samples from the plume region with TEQs-WHO_{05} greater than 50 pg/g. To assist in the evaluation of the specificity of HCA results, a broadly defined floodplain (BDFP) was delineated as all soil samples from the floodplain population, which was determined by both the 100-year floodplain, as defined by the Federal Emergency Management Agency, and by self-identification by the study respondent as living on a property that has been flooded by the Tittabawassee River. A broadly defined plume (BDPL) was delineated as soil samples from the City of Midland.

Results and Discussion

The inclusiveness of the clusters was evaluated by comparing the percent of the NDFP and NDPL samples that were classified as belonging to the clusters associated with the floodplain and plume respectively. The specificity of the clusters was evaluated by comparing the percent of samples outside the BDFP and BDPL that were classified as belonging to the clusters associated with the floodplain and plume respectively. HCA methods that produced high percentages for inclusiveness and low percentages for specificity were considered more reflective of the initial understanding of the data. Results of the evaluation of cluster linkage methods using seven clusters are presented in Table 1.

Table 1. Evaluation of HCA linkage methods using seven clusters

	Average	Single	Complete	Centroid	Median
Percent of NDFP samples in FP cluster	99.6	100	99.6	95.3	95.3
Percent of NDPL samples in PL cluster	91.5	100	90.2	100	100
Percent outside BDFP in FP cluster	14.1	99.2	16.8	98.6	98.6
Percent outside BDPL in PL cluster	68.4	99.5	10.5	98.3	98.3

The complete linkage method best reproduces the initial understanding of the data. Single, centroid, and median linkage all tend to produce one large cluster with a few outlier samples classified as distinct clusters. For each of these three methods, samples from the floodplain and plume are grouped into the large cluster and are not differentiated from background soil samples. Average linkage produces a separate floodplain cluster, however the plume samples are grouped with background samples.

HCA is also sensitive to the quantity of clusters specified. Table 2 presents the evaluation of the cluster quantity using complete linkage.

Table 2. Evaluation of HCA cluster quantity using complete linkage

	Number of clusters							
	7	8	9	10	11	12	13	14
Percent of NDFP samples in FP cluster	99.6	99.6	94.1	94.1	94.1	94.1	75.1	75.1
Percent of NDPL samples in PL cluster	90.2	90.2	90.2	90.2	90.2	90.2	90.2	90.2
Percent outside BDFP in FP cluster	16.8	16.8	5.4	5.4	5.4	5.4	1.3	1.3
Percent outside BDPL in PL cluster	10.5	1.8	1.8	1.8	1.8	1.8	1.8	1.8

The results demonstrate that increasing the number of clusters improves the specificity of the groupings. Using nine clusters accurately reproduces the initial understanding of the data: a distinct floodplain cluster and a distinct plume cluster with only few samples outside of those regions that have similar congener distributions.

When the quantity of clusters is increased to 13, even fewer samples outside the broadly defined floodplain are included in the floodplain cluster. However, the inclusiveness of the floodplain cluster is also decreased. This is due to a second cluster being formed from the set of floodplain samples. An investigation of the congener distributions of the different clusters demonstrates that the two floodplain clusters do have distinct congener distributions. Table presents the TEQ_{DF}-WHO_{05} and contribution of the dioxins and furans to the TEQ_{DF}-WHO_{05} for select clusters from the analysis using complete linkage and 13 clusters. It is expected that this analysis will have an impact on the expected dioxin/furan profiles generated by the centroid samples resulting from the analysis, and on the spatial distribution of the extracted patterns.

Table 3. TEQ and contribution from dioxins and furans for select clusters using 13 clusters and complete linkage

Cluster	No. of samples	Mean TEQ_{DF}-WHO_{05}	Dioxin contribution to TEQ	Furan contribution to TEQ
Largest background cluster	458	2.35	64.6%	35.4%
Plume cluster	132	59.00	77.6%	22.4%
Floodplain cluster 1	189	709.23	10.0%	90.0%
Floodplain cluster 2	113	394.46	1.5%	98.5%

Both of the floodplain clusters have elevated TEQ levels, with high percentage contributions from the furans. However, floodplain cluster 2 has a particularly low (1.5%) contribution from the dioxins. This may indicate that the samples in floodplain cluster 2 are less impacted by pentachlorophenol, a secondary source of elevated dioxin levels in the Tittabawassee River floodplain. Evaluating the relative mass of two indicator congeners in the clusters supports this conclusion. The mass ratio of 1,2,3,4,6,7,8-HpCDD (indicative of pentachlorophenol contamination) to 2,3,4,7,8-PeCDF (indicative of chloralkali related furan contamination) is 2.6 for floodplain cluster 1 and 0.5 for floodplain cluster 2.

The timeline of contaminant discharge and geographic distribution of floodplain clusters may offer an explanation as to why floodplain cluster 2 samples are less impacted by pentachlorophenol, but still have high levels of other floodplain contaminants. The production of pentachlorophenol is a more recent process at the Dow facility than the chloralkali process. Floodplain cluster 2 samples may represent soils deposited before pentachlorophenol had been discharged or fully distributed throughout the river sediment. This is supported by the fact that samples in this cluster tend to be located further from the Tittabawassee River.

For the UMDES soils data, increasing the quantity of clusters used in HCA can offer improved resolution for assessing the distribution of congener patterns.

Acknowledgements
Financial support for this study comes from the Dow Chemical Company through an unrestricted grant to the University of Michigan. The authors acknowledge Ms. Sharyn Vantine for her continued assistance and Drs. Linda Birnbaum, Ron Hites, Paolo Boffetta and Marie Haring Sweeney for their guidance as members of our Scientific Advisory Board.

References
1. Olson, K, Garabrant, D, Franzblau, A, Adriaens, P, Gillespie, B, Lepkowski, J, Lohr-Ward, B, Ladronka, K, Sinibaldi, J, Chang, S-C, Chen, Q, Demond, A, Gwinn, D, Hedgeman, E, Hong, B, Knutson, K, Lee, S-Y, Sima, C, Towey, R, Wright, D, Zwica, L. *Organohalogen Comp* 2006.
2. Demond A, Towey T, Chang SC, Adriaens P, Luksemburg W, Maier M, Favaro K, Wenning R, Kennington B. *Organohalogen Comp* 2006.
3. Johnson GW, Ehrlich R. Environmental Forensics. 2002; 3, 59.
4. Minitab, Inc., State College, Pennsylvania, USA.
5. Jolliffe IT. *Principal Component Analysis*. Springer-Verlag, New York. 1986;
6. Hair Jr JF, Black WC. Cluster analysis. In: *Multivariate Data Analysis*, Hair Jr JF, Anderson RE, Tatham RL, Black WC. (ed.), Prentice-Hall Inc, New Jersey, 1998: 147.
7. Chang S-C, Adriaens P, Towey T, Wright D, Demond A, Gillespie B, Franzblau A, Garabrant D. *Organohalogen Comp* 2006.