Integrated Application of RSIMCA, t-SNE, and PVA for Chemometric Mapping of PCDD/F Congeners

Mark J. Cejas* and Robert C. Barrick

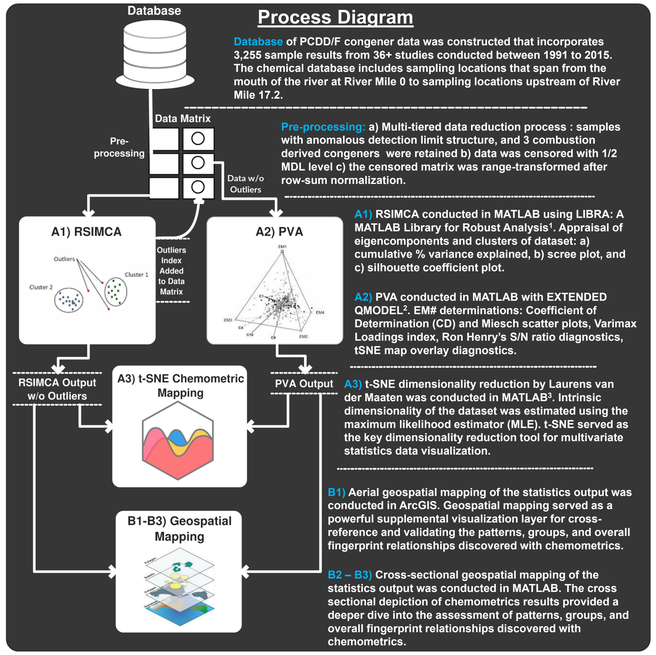
GROUP 1

* Now with Pace Analytical® Energy Services™

Abstract

Background/Objectives. Unique distributions of the 17 toxic polychlorinated dibenzodioxins and dibenzofuran (PCDD/F) congeners used in risk assessments are the basis for distinguishing major sources in sediment environments. Studies to date typically have approached data assessments using either sample classifications to assess relationships among sample PCDD/F distributions or linear mixing models to distinguish and apportion endmember fingerprints in samples. Both approaches have advantages but our objective was to improve interpretations by integrating these techniques in a tiered assessment of data. **Approach.** Robust Independent Modeling of Class Analogy (RSIMCA) is a robust algorithm that classifies sample fingerprints into groups or clusters, while minimizing the impact of outliers on classical statistical estimates RSIMCA was applied to classify over 2,800 river sediment samples into "clusters" or as outliers to all clusters after an initial screening of a 3,255 sample dataset. This multivariate statistical output is compressed from seven latent dimensions into two interpretable dimensions using t-Distributed Stochastic Neighbor Embedding (tSNE) . Polytopic Vector Analysis (PVA) is a model for the discovery of pure source compositional features (EMs), and their proportional contributions in samples. PVA was used to identify distinct source end-members based on PCDD/F characteristics of the classified samples.

Process Diagram



Database of PCDD/F congener data was constructed that incorporates 3,255 sample results from 36+ studies conducted between 1991 to 2015. The chemical database includes sampling locations that span from the mouth of the river at River Mile 0 to sampling locations upstream of River Mile 17.2.

Pre-processing: a) Multi-tiered data reduction process : samples with anomalous detection limit structure, and 3 combustion derived congeners were retained b) data was censored with 1/2 MDL level c) the censored matrix was range-transformed after row-sum normalization.

A1) RSIMCA conducted in MATLAB using LIBRA: A MATLAB Library for Robust Analysis[1]. Appraisal of eigencomponents and clusters of dataset: a) cumulative % variance explained, b) scree plot, and c) silhouette coefficient plot.

A2) PVA conducted in MATLAB with EXTENDED QMODEL[2]. EM# determinations: Coefficient of Determination (CD) and Miesch scatter plots, Varimax Loadings index, Ron Henry's S/N ratio diagnostics, tSNE map overlay diagnostics.

A3) t-SNE dimensionality reduction by Laurens van der Maaten was conducted in MATLAB[3]. Intrinsic dimensionality of the dataset was estimated using the maximum likelihood estimator (MLE). t-SNE served as the key dimensionality reduction tool for multivariate statistics data visualization.

B1) Aerial geospatial mapping of the statistics output was conducted in ArcGIS. Geospatial mapping served as a powerful supplemental visualization layer for cross-reference and validating the patterns, groups, and overall fingerprint relationships discovered with chemometrics.

B2 – B3) Cross-sectional geospatial mapping of the statistics output was conducted in MATLAB. The cross sectional depiction of chemometrics results provided a deeper dive into the assessment of patterns, groups, and overall fingerprint relationships discovered with chemometrics.

Key References

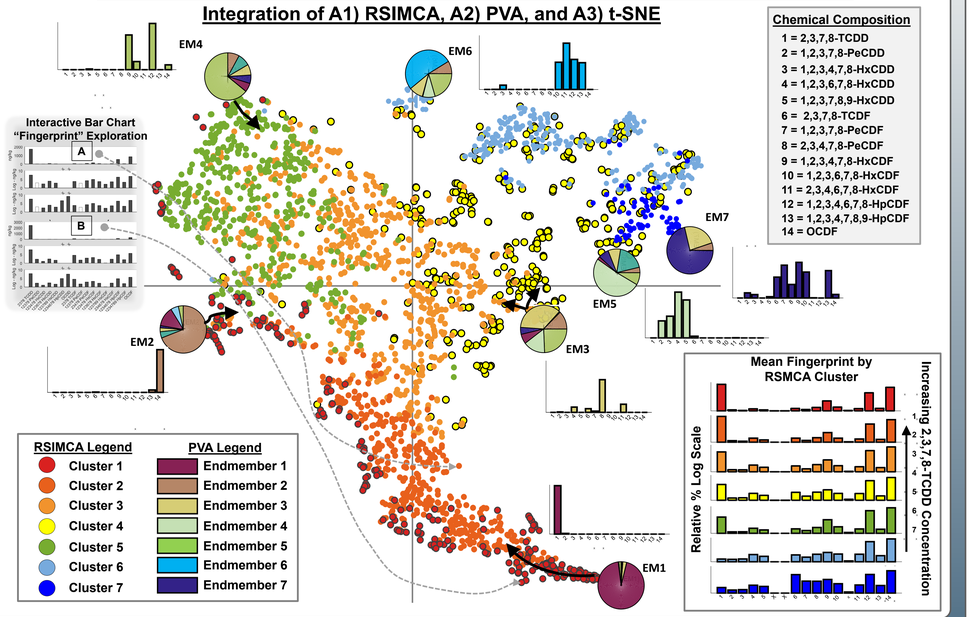
1) Branden, K. Vanden, and Mia Hubert. "Robust classification in high dimensions based on the SIMCA method." *Chemometrics and Intelligent Laboratory Systems* 79.1-2 (2005): 10-21.

2) Full, William E., Robert Ehrlich, and J. E. Klovan. "EXTENDED QMODEL—Objective definition of external end members in the analysis of mixtures." *Journal of the International Association for Mathematical Geology* 13.4 (1981): 331-344.

3) Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.

A) Chemometric Mapping of Compositional Fingerprints

Integration of A1) RSIMCA, A2) PVA, and A3) t-SNE



2,3,7,8-TCDD Focused Discussion

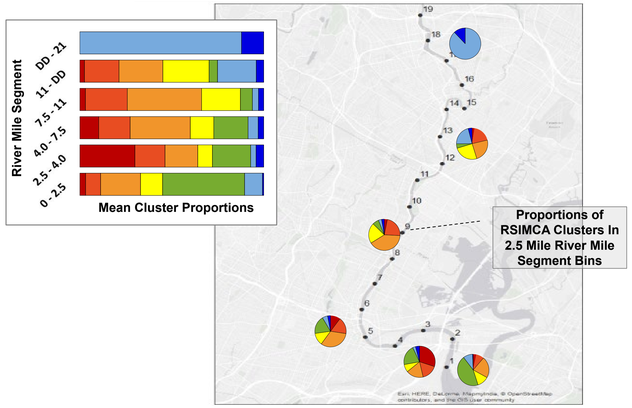
A) This figure is a snapshot of our integrated approach to data visualization. In addition to integrating data visualizations of 3 chemometrics methodologies, the data-viz maps were designed as an interactive platform for fingerprint bar chart data exploration and review. This presentation introduces the model development groundwork of 3,255 in-river samples. The preponderance of evidence that emerged from chemical fingerprint dynamics in-river, and model applications to the study of source candidate upland samples is the focus of our current research. A set of preliminary findings will be presented tomorrow in the *platform session A6*. The mixing evolution of chemical source fingerprints and background influences was a dominant characteristic of the PCDD/F chemistry in the river sediments. For brevity, this poster addresses only a key aspect of the 2,3,7,8-TCDD driven RSIMCA clusters, PVA endmembers, and geospatial patterns. **A1)**. RSIMCA Cluster 1 contains samples with the highest expression of 2,3,7,8-TCDD concentrations in the river sediments. Furthermore, Cluster 1 contained samples with the highest relative expression of PVA EM-1. Cluster 1 is an assemblage of samples with significant representation of EM-1 dominated samples. **A2)** PVA identified the dominant source fingerprint attributes of the river sediments, and modeled their mixing confluence in the river sediment system. EM-1 represents the dominant 2,3,7,8-TCDD source fingerprint attribute in the river sediments. **A3)** t-SNE is an emergent "Big Data" visualization technique that served as the data visualization map for RSIMCA and PVA results. With t-SNE, the high dimensional structure of complex data is accurately reflected in 2D. **B)** Geospatial analysis of the chemometrics results helped base the notable fingerprint groups and pattern in a river system context. **B1 & B2)** Geospatial rendering of the RSIMCA results show a correspondence between spatial changes in 2,3,7,8-TCDD concentration and spatial prevalence of Cluster 1. The positive relationship between concentration and fingerprint parameters highlights a dominant 2,3,7,8-TCDD source region between River Mile 2 and 6. **B3)** Geospatial rendering of the PVA results was generally consistent with the RSIMCA interpretation, as the maximum proportional contribution of EM-1 was localized in the highest 2,3,7,8-TCDD concentration zone between River Mile 2 and 6.

Conclusion

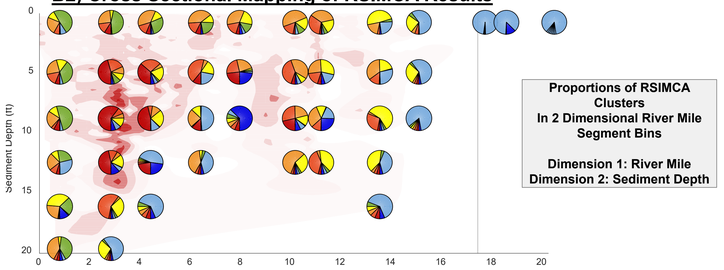
An integrated chemometrics strategy was designed to explore and understand (PCDD/F) chemical fingerprint dynamics in a river system. The novel approach led to insights supporting a conclusion that the dynamics of 2,3,7,8-TCDD occurrence in a large segment of the river sediments is likely related to a dominant source zone between River Mile 2 and 6.

B) Geospatial Mapping of Compositional Fingerprints

B1) Aerial Mapping of RSIMCA Results



B2) Cross Sectional Mapping of RSIMCA Results



B3) Cross Sectional Mapping of PVA Results (EM-1 Focus)