Evaluating Urban Background Metal Concentration Clusters with Bayesian Networks: Southeastern Urban Centers in EPA Region 4

John F. Carriger[1], Robert G. Ford[2], Tim Frederick[3], Sydney Chan[3], Yuen-Chang Fung[4]

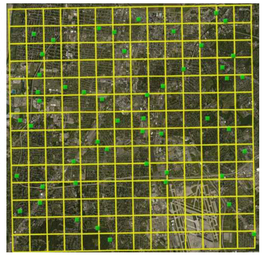
[1]U.S. Environmental Protection Agency, Office of Research and Development, Center for Environmental Solutions and Emergency Response, Land Remediation Technology Division, Environmental Decision Analytics Branch, Cincinnati, OH
[2]U.S. Environmental Protection Agency, Office of Research and Development, Center for Environmental Solutions and Emergency Response, Land Remediation Technology Division, Contaminated Sites and Sediments Branch, Cincinnati, OH
[3] U.S. Environmental Protection Agency, Region 4, Superfund and Emergency Management Division, Resource and Scientific Integrity Branch, Scientific Support Section, Atlanta, GA
[4]Tetra Tech, Inc.

John Carriger | carriger.john@epa.gov

www.epa.gov

Urban Background Study (UBS)- EPA Region 4/ORD

Soils in urban settings are likely to contain elevated levels of certain metals due to human activity, non-point source industrial operations, and from infrastructure materials along with natural background (non-anthropogenic) levels. Because these increased contaminant concentrations are due to anthropogenic urban activity and not site-related point source releases, they can be considered to represent urban background soil concentrations. The U.S. Environmental Protection Agency recently collected a comprehensive sampling of background surface soil concentrations to support risk assessment and risk management of urban locations in the Southeastern U.S.



Louisville, KY sampling sites

Sampling used a within-city strategy capturing variation of urban soil metal concentration within and across cities
Eight cities chosen in five states-

- Florida (Gainesville)
- Kentucky (Lexington, Louisville)
- North Carolina (Raleigh, Winston-Salem)
- South Carolina (Columbia)
- Tennessee (Chattanooga, Memphis)

Sampling strategy within cities

- 7x7 mile square with 0.5x0.5 mile cells
- Chattanooga had 5x5 mile grid
- 50 cells randomly chosen for sampling
- Field collectors chose sample locations within grid cells

Measurements at metals sampling sites

- **Surface soil metal concentrations**
- Coordinates
- Land use
- Surface type (e.g., grass)
- Landmarks
- Emission sources for metals
- Visual soil characteristics
- Other sampling-/site-related notes

Criteria for Selected Sample Locations

Qualify:

- *Areas that appear to be representative of the larger urban setting*
- *Locations that appear to be undisturbed by recent activity*
- *Public areas, such as along right-of-ways and within government-owned property*

Disqualify:

- *Private/residential yards*
- *Industrial properties or in obvious significant pollutant outfall area for nearby industry*
- *Areas of relatively recent development/redevelopment*
- *Low-lying areas that may be routinely subjected to flooding or inundation, such as wetlands and/or where surface runoff could accumulate*

From: Urban Background Study Webinar (https://www.epa.gov/research-states/regional-urban-background-study-webinar-archive)

Intended uses for urban background study results

- Differentiate site-specific contamination from background in urban areas
- Differentiate natural/anthropogenic background in urban areas
- Capture variability in a large-scale data set for background concentrations
- Support development of remediation goals for contaminated sites
- Compare background concentrations between cities
- Compare future changes to urban background concentrations
- Support human and ecological risk assessments

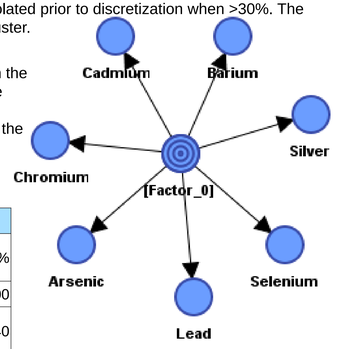
Chattanooga, TN sampling sites

Even with a comprehensive data set, setting threshold concentrations for metals for individual sites can be difficult, especially in urban settings, given the varying background and historical contributions to concentrations in different soils. Bayesian networks are useful for machine learning and discovering patterns in data. One machine learning tool that is especially useful for examining background levels is data clustering.

UBS Metals Clustering

Expectation maximization data clustering was used in Bayesialab 9.1. The clusters were based on concentration data from seven metals analyzed in all of the cities. Non-detects were substituted with zeroes and zeroes were isolated prior to discretization when >30%. The algorithm could seek clusters from 2 to 5 with a minimum of 85% purity for accepting a cluster.

Purity examines the fit of the samples in the case file to a corresponding cluster. The neighborhood provides the overlap with the data to other clusters. Purity is computed from the average probability of the clusters given each sample. Contingency table fit represents the percentage the data fit to the model lies between an unconnected model (0%) and a fully connected model (100%). Hypercube cells per state is an aligned measure that considers the size of the probability tables. The purity of the resulting clusters was high. However, the contingency table fit was low. The three clusters were interpretable in terms of higher concentrations (C2), lower concentrations (C1) and moderate concentrations (C3).



Metals cluster network. Factor_0 node contains cluster states

Overall Average Purity: 97%		
Cluster	Purity	Neighborhood
Cluster 1	98%	Cluster 3 2.1%
Cluster 3	97%	Cluster 2 2.6%
Cluster 2	90%	Cluster 3 9.5%

Performance Indices	
Contingency Table Fit	46%
Deviance	300
Hypercube Cells Per State	340

Posterior means for metal nodes from conditioning on cluster states			
Node	C1	C2	C3
Cadmium	0	0.80	0.28
Lead	55	390	61
Arsenic	2.1	11	6.0
Silver	0	0.18	0.051
Chromium	11	23	13
Barium	71	110	92
Selenium	0.011	0.071	0.035

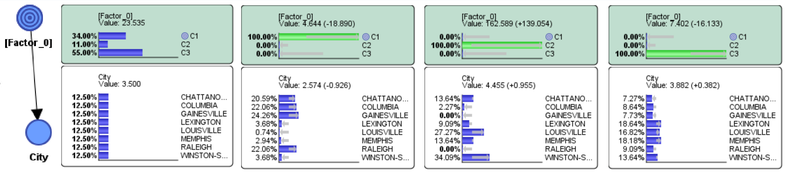
Marginal Probabilities for Factor_0 node	
Cluster 3	55%
Cluster 1	34%
Cluster 2	11%

Cluster 3 covers most of the samples (55%) and represents moderate concentrations. Cluster 2 represents higher concentrations but explains only 11% of the data set.

UBS Metals Clusters with Cities

Metals clusters with city node network. Factor_0 node contains cluster states from above built with metals data. The node was unattached from the metals and attached to the city node to examine the distribution of the cities for each cluster.

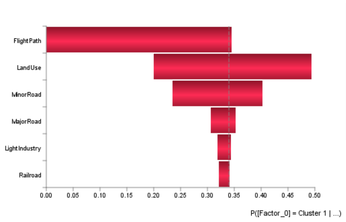
- Cluster 1 was primarily in Gainesville, Columbia, Raleigh and Chattanooga
- Cluster 2 was primarily in Louisville and Winston-Salem
- Cluster 3 was found throughout the cities but primarily Lexington, Louisville and Memphis



Cluster-city model marginal probabilities | Cluster-city model with conditioning on C1 | Cluster-city model with conditioning on C2 | Cluster-city model with conditioning on C3

UBS Metals Clusters with Land Use and Emissions

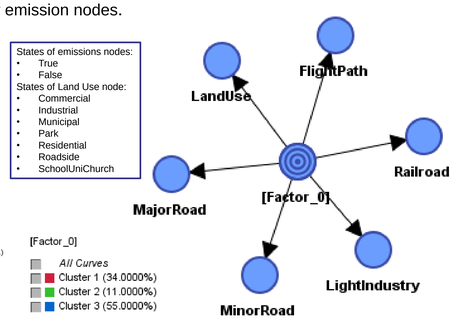
Nearby emission sources were also attached to the UBS Metals Factor node and their relationship to the metals clusters was examined with tornado diagrams that provide the probability range for each cluster from conditioning on the land use or emission nodes.

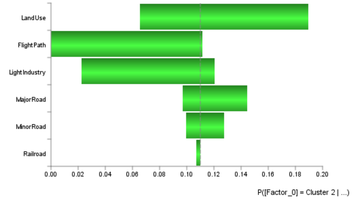


Cluster 1 probability ranges given emissions and land use node states

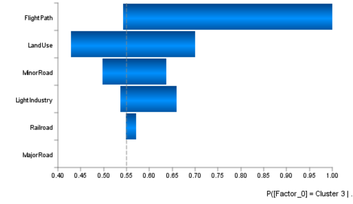
States of emissions nodes:
- True
- False

States of Land Use node:
- Commercial
- Industrial
- Municipal
- Park
- Residential
- Roadside
- SchoolUniChurch





Cluster 2 probability ranges given emissions and land use node states



Cluster 3 probability ranges given emissions and land use node states

Cluster 1 probability is decreased most by being near a flight path. It is increased most by residential land use.
Cluster 2 probability is decreased most by being near a flight path. It is increased most by roadside land use.
Cluster 3 probability is decreased most by residential land use. It is increased most by being near a flight path.

Conclusions

The clustering analysis can be useful for isolating, grouping, and/or comparing assumptions about background data when the clusters for the metals are homogeneous with respect to the data, stable, and interpretable with scientific knowledge of the differences in background concentrations. Once clusters are created for a background dataset of metals concentrations the clusters can be compared to location-specific identifiers and land use and nearby urban emission sources for further interpretability. Future work will examine additional data and node preparation approaches and assumptions for their influence on the cluster outcomes and fit of the data to the model.