Applied Machine-Learning Tools to Enhance PFAS Analytics and Support More Effective Site Management

RemTech 2024

Paul Hurst M.Sc., P.Eng (paul.hurst@wsp.com)

Skyler Sorsby, M.Sc., P.G. (skyler.sorsby@wsp.com)

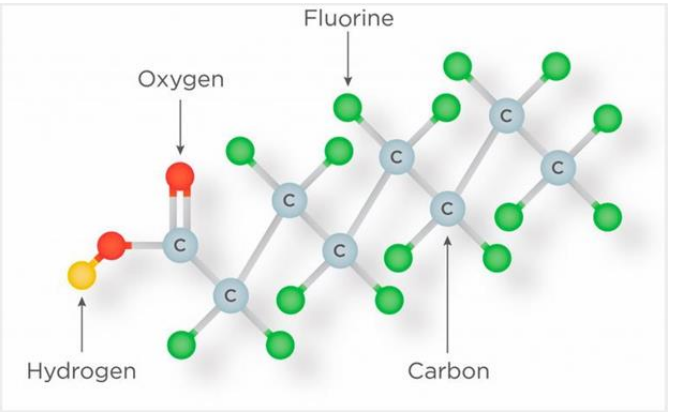
Stefano Marconetto, M.Sc., P.Eng (stefano.marconetto@wsp.com)

Agenda

- (Some of the) Challenges with PFAS
- Why do I need advanced analytics?
- Dive into selected techniques
  - HCA / GMM Clustering For Pattern Identification and Estimating Ambient Levels
  - Partitioning the Site
  - PCA Bi Plots
  - Identifying Sources
- Key Takeaways
- Q&A

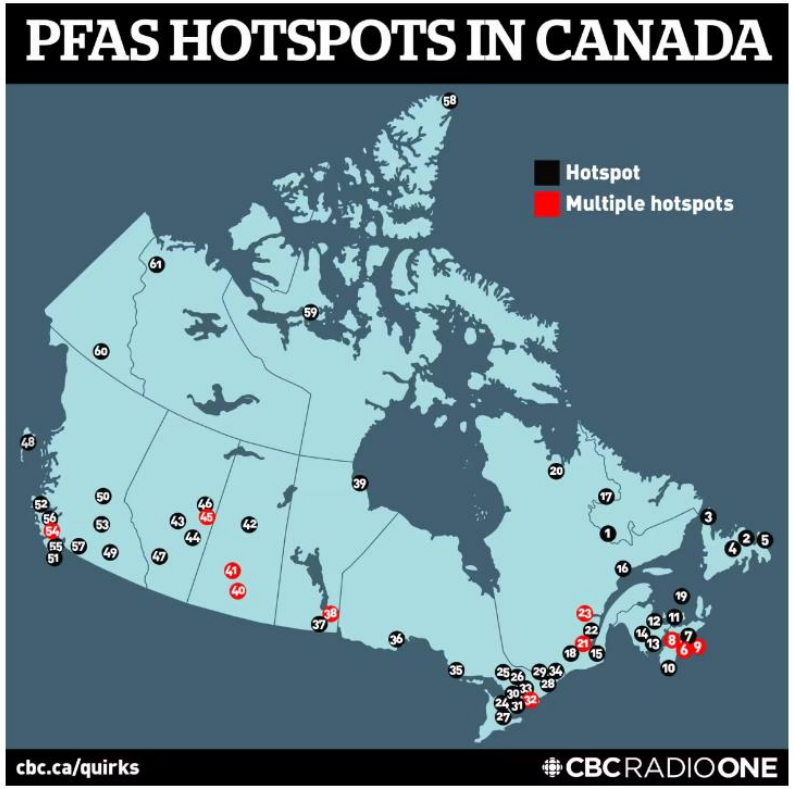
(Some of the) Challenges with PFAS

PFOA
perfluorooctanoic acid



Georgia Department of Natural Resources

- Very low drinking water standards
  - Canada: 30 ng/L (Sum of 25 specific PFAS)
  - USA: PFOA / PFOS: 4 ng/L, PFNA/PFHxS: 10 ng/L
- Numerous sources (industries, airports, water waste treatment plants, more)
- Concentrations in ambient background
  - Up to 11 ng/L (PFOA) and 14 ng/L (PFOS) **in precipitation alone**[1] (Study on Great Lakes)

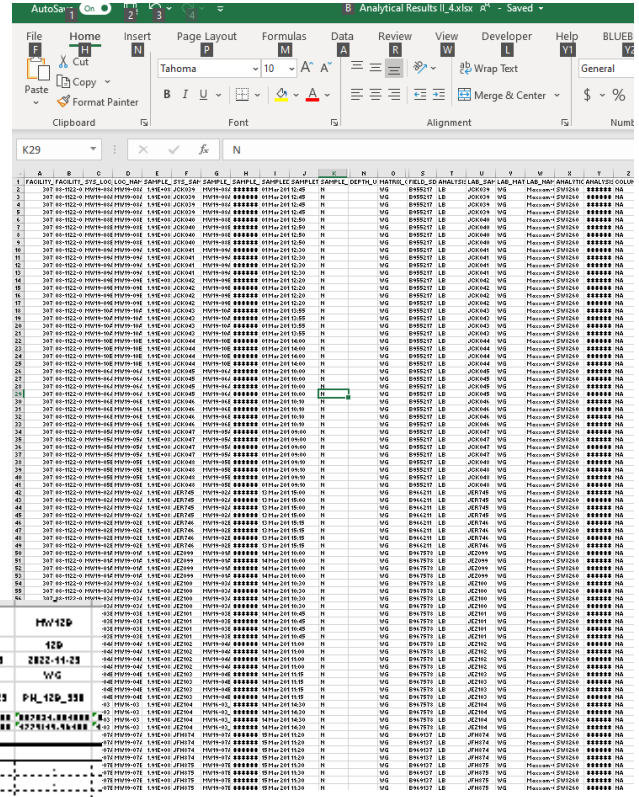


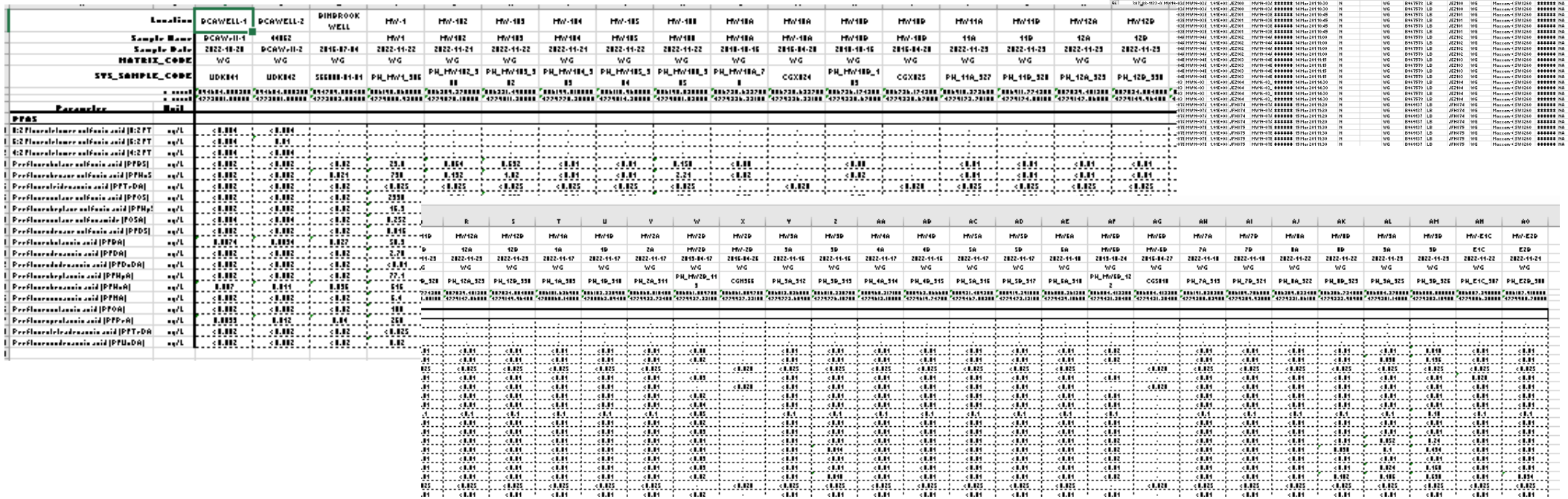
https://www.cbc.ca/radio/quirks/nov-7-fast-radio-bursts-in-our-galaxy-monkeys-with-a-puberty-switch-and-more-1.5789388/forever-chemicals-can-have-far-reaching-consequences-need-more-regulation-in-canada-scientists-say-1.5789395

[1]Gewurtz et al. (2019), [2]Salvatore et al. (2022)

Why do I need advanced analytics for PFAS?

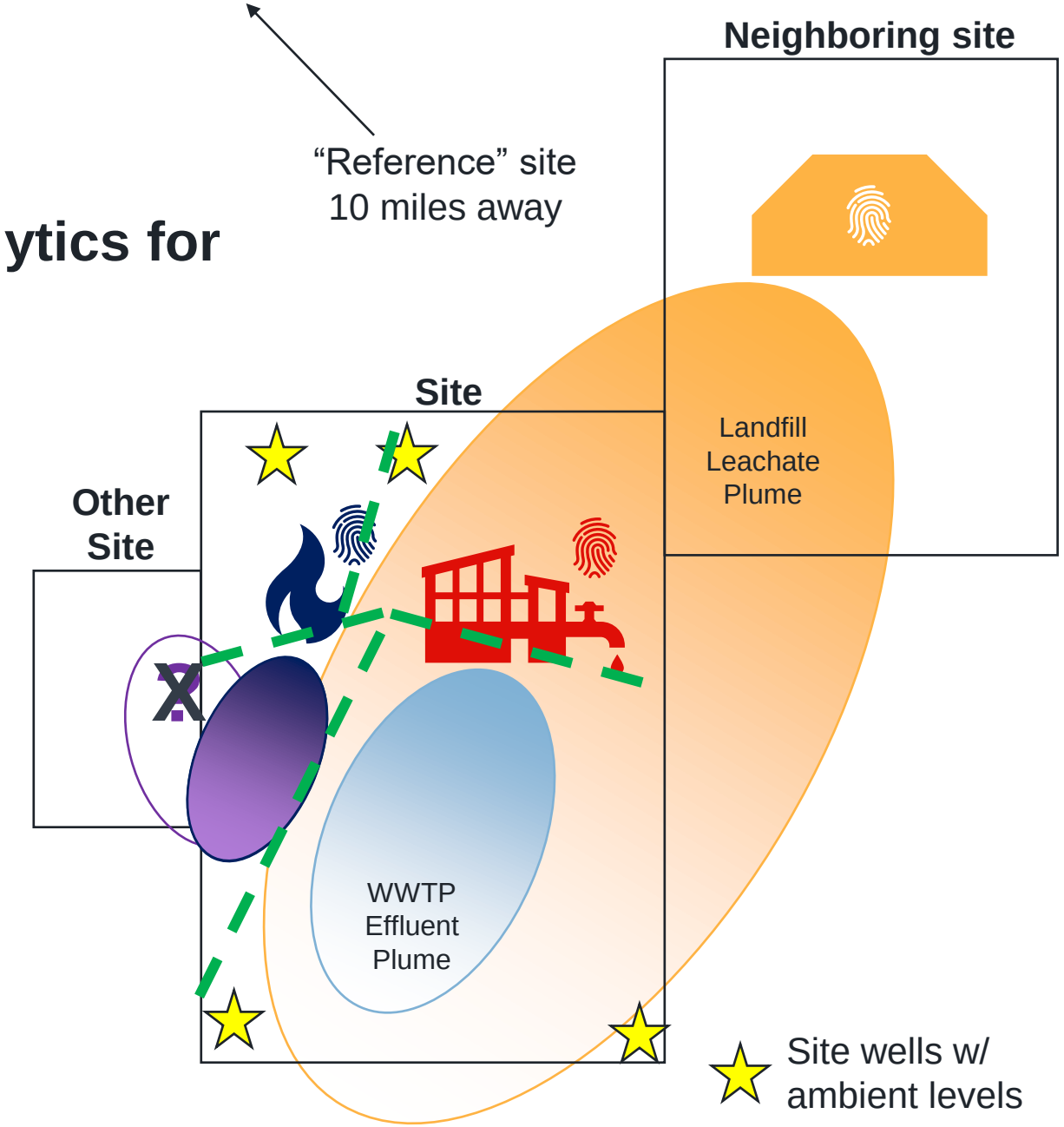
- Large datasets are **hard to interpret and manage!**





Why do I need advanced analytics for PFAS?

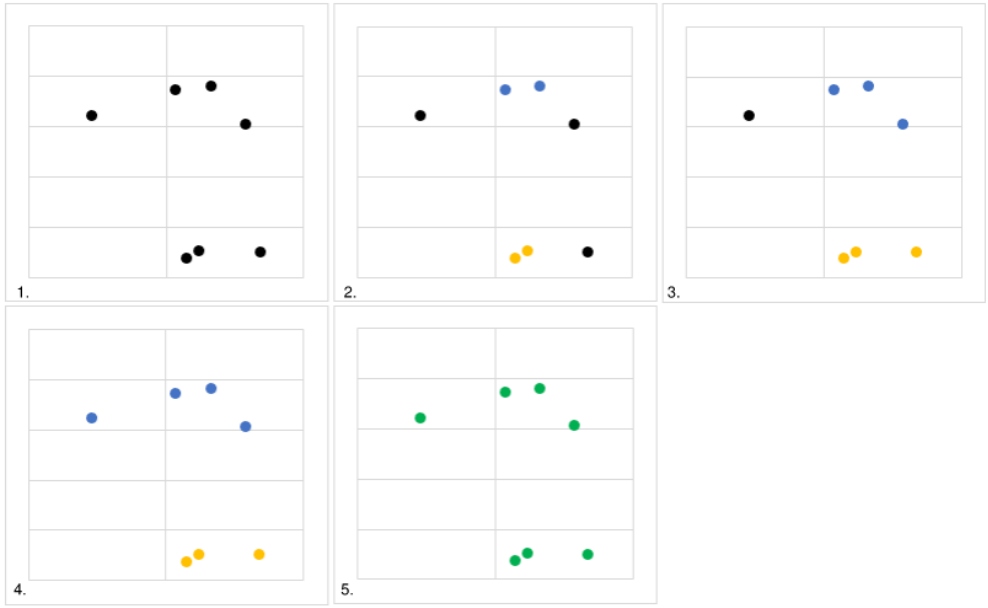
- Locate multiple **source areas**
  - Onsite vs. neighboring site
  - Varying onsite sources (i.e., multiple generations of AFFF)
- Distinguish sources from **transport patterns**
- Identify potential **precursor contributions**
- Define site-specific **ambient levels**
- **Data-driven site partitioning** for investigation / remediation

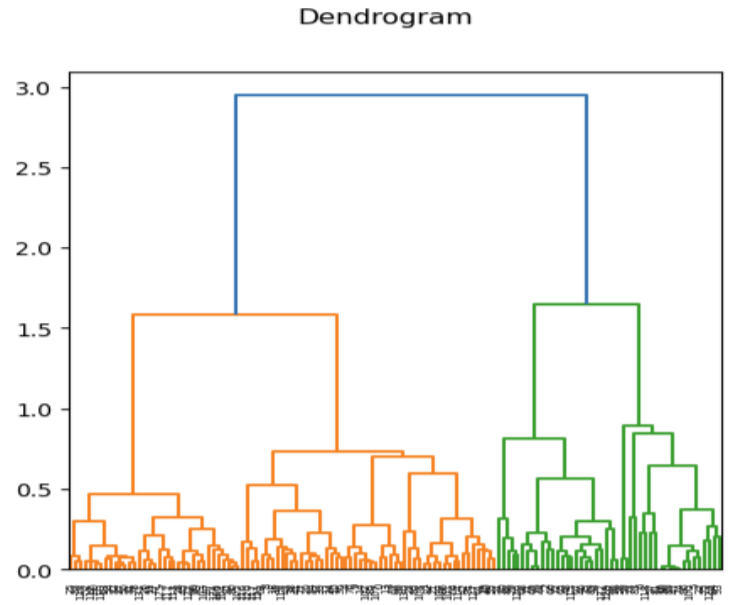


Hierarchical Clustering (HCA)

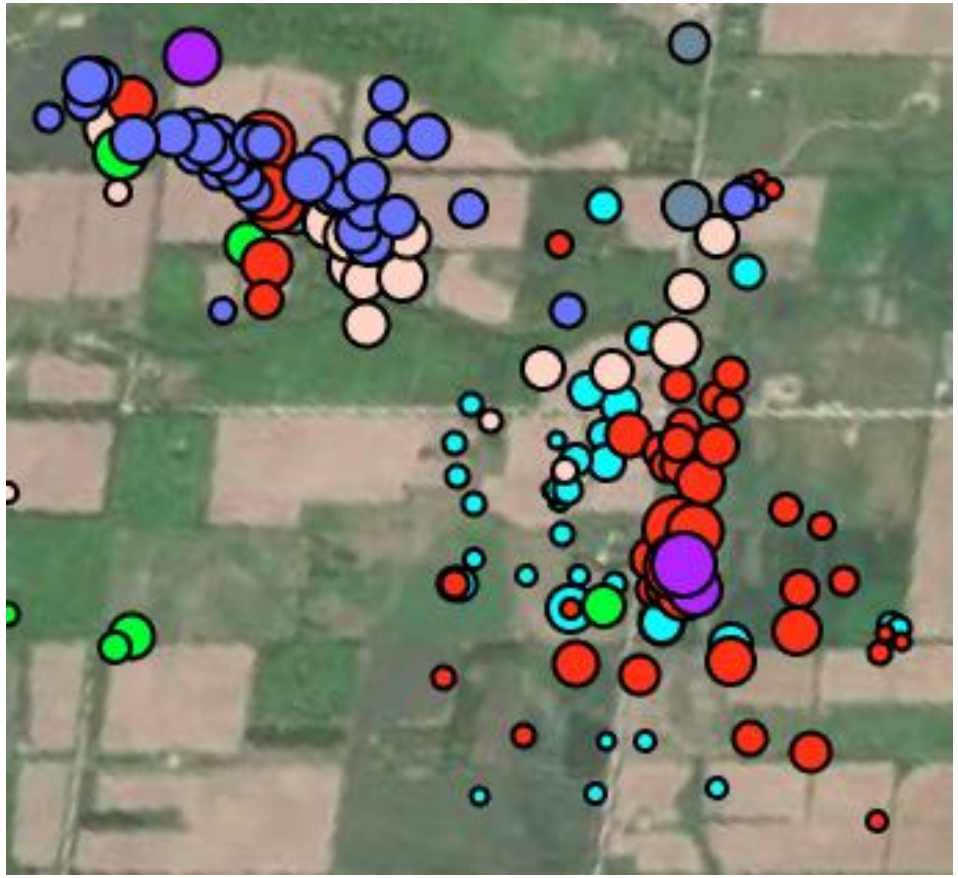
What is HCA Clustering?

- Linking algorithm that iteratively groups similar samples based on their similarly to each other.
- HCA works by taking the two most similar points and makes them one cluster, and, continues searching for the next closest point to convert into clusters until no points remain.
- A hierarchy of clusters is created and a user can then select how many clusters to work with for data analysis.

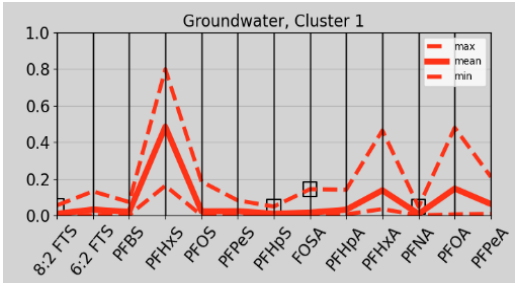


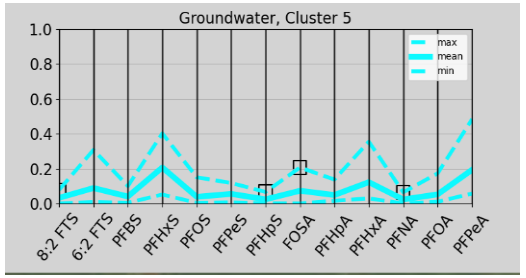


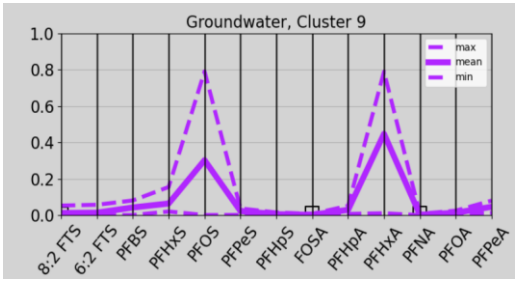
Hierarchical Clustering Example

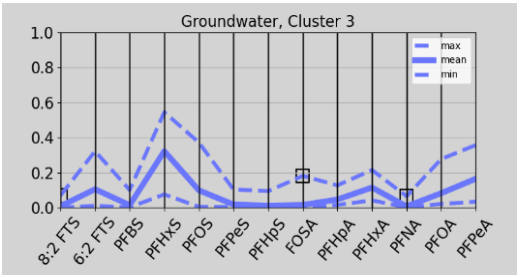


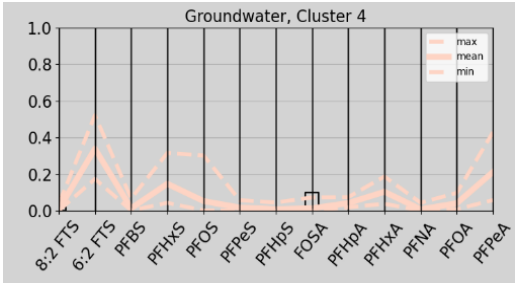
Google Earth anonymous aerial used











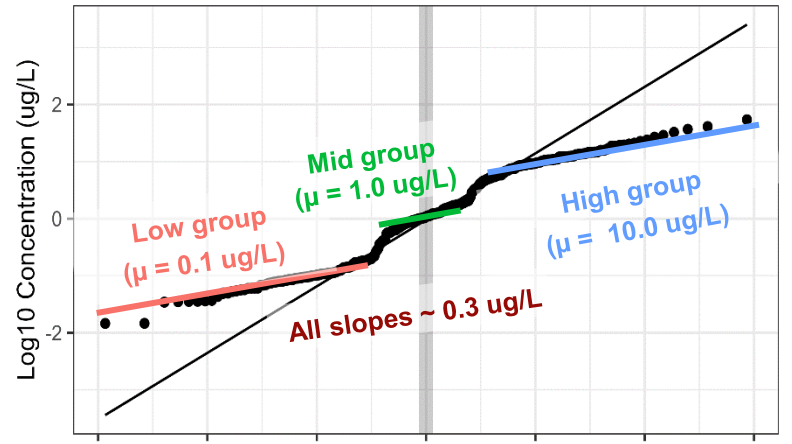
Gaussian Mixture Model (GMM) Clustering

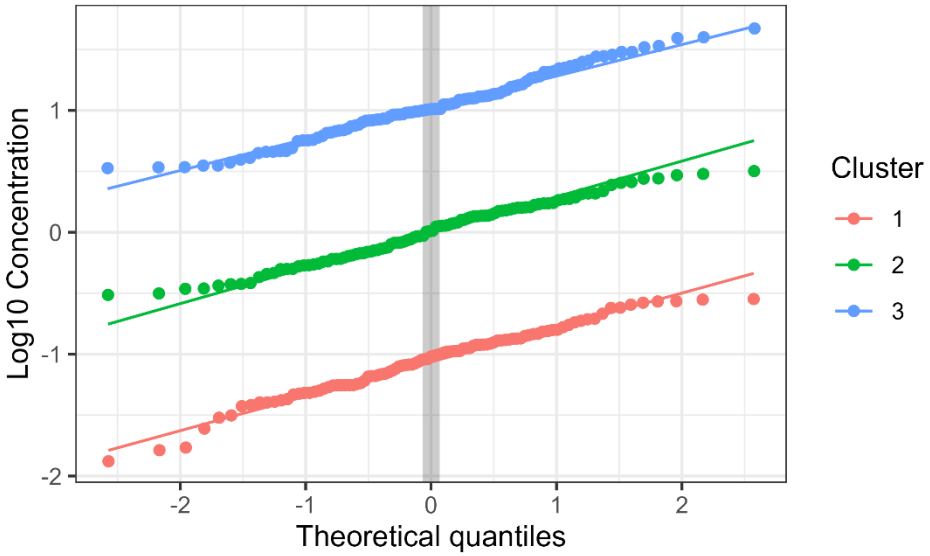
Problem: site-specific ambient levels may exceed regulatory criteria. Reference locations may not adequately represent full range of site-specific levels.

Solution: GMM uses familiar statistical practices to extract low vs high concentrations reflective of ambient levels and impacts.

Benefit: Holistic approach to identify patterns and potential ambient reducing the need to assess location by location

Sample	PFOS	PFOS	PFHxS	…
1	↕	↕	↕	↕
2				
3				
…				





Estimating ambient levels

Gaussian Mixture Model (GMM) clustering

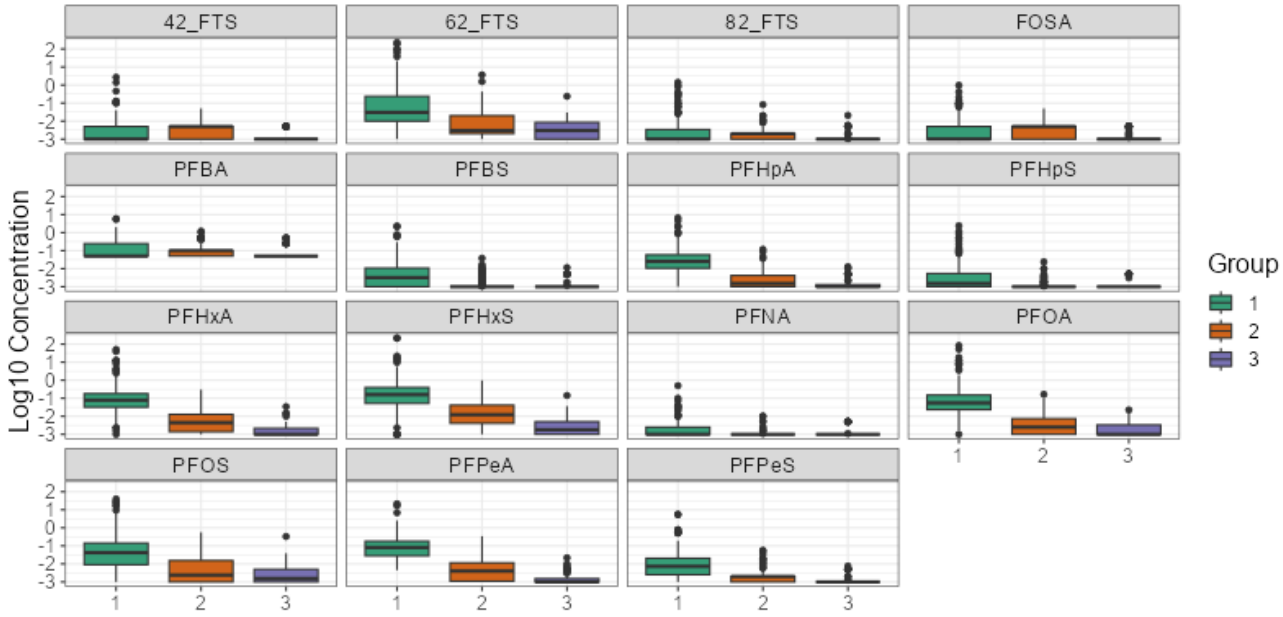
Clear difference in concentration levels between groups.

Cluster 1 = potential point impacts / sources.

Cluster 2 = potential diffuse impacts.

Cluster 3 = potential ambient levels.

Percentage detected | Box-and-whisker plot | Classification plot



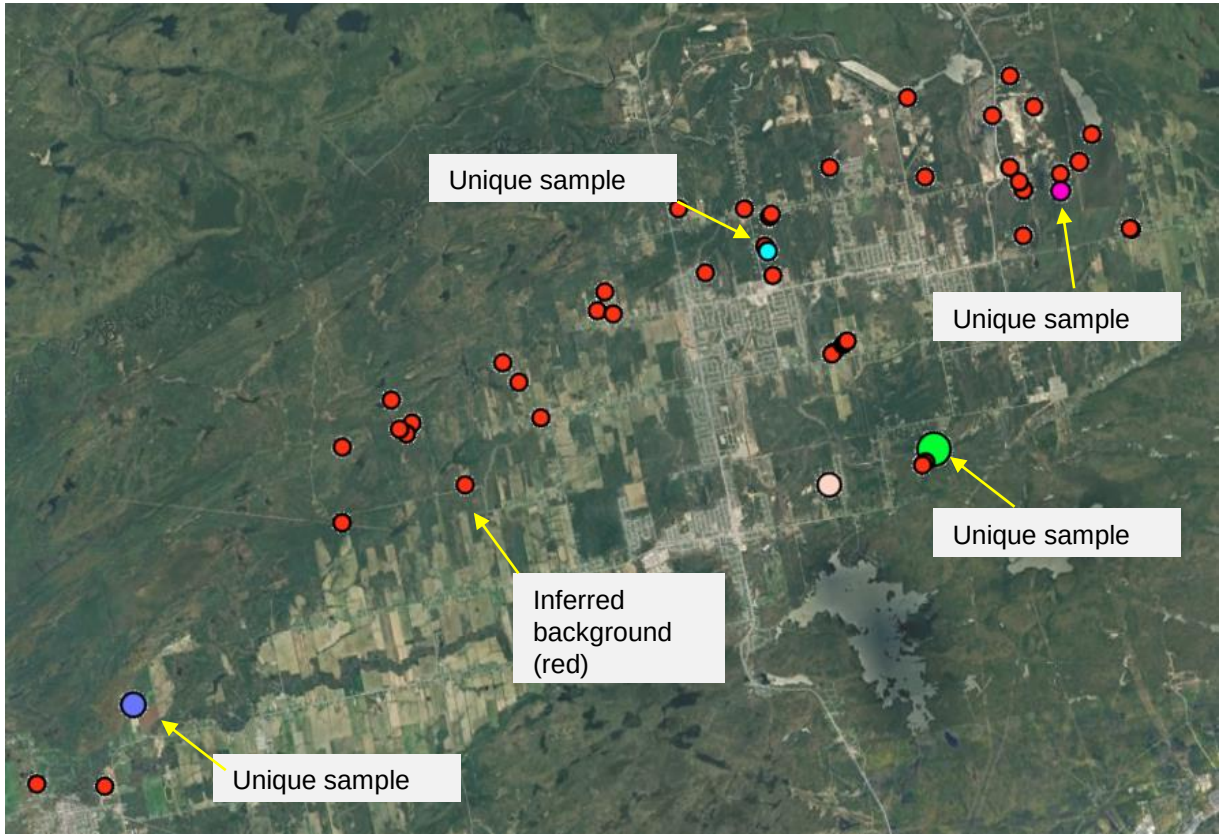
Show 25 entries Search: PFOS

CLASS	CHEMICAL_NAME	Minimum	Median	Mean	Maximum
1	PFOS	0.001	0.04180	1.02000	39.600
2	PFOS	0.001	0.00240	0.01810	0.564
3	PFOS	0.001	0.00155	0.00752	0.329
CLASS	CHEMICAL_NAME	Minimum	Median	Mean	Maximum

Showing 1 to 3 of 3 entries (filtered from 45 total entries) Previous 1 Next

Clustering Example

- GMM used to review potential PFAS impacts in drinking water
- Generally similar parameters observed across clusters
- Quickly identifed **potential and unique drinking** water samples
- Red cluster identifed as **ambient / background**
- Allowed **focused investigation** to assess are identifed samples due to potential other PFAS sources, elevated detection limits, etc.



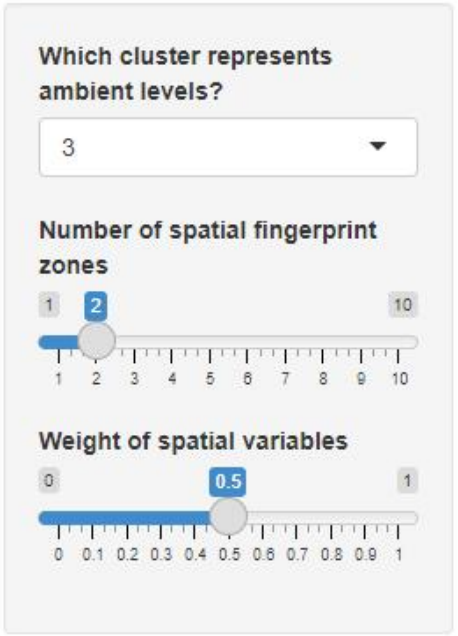
Google Earth anonymous aerial used

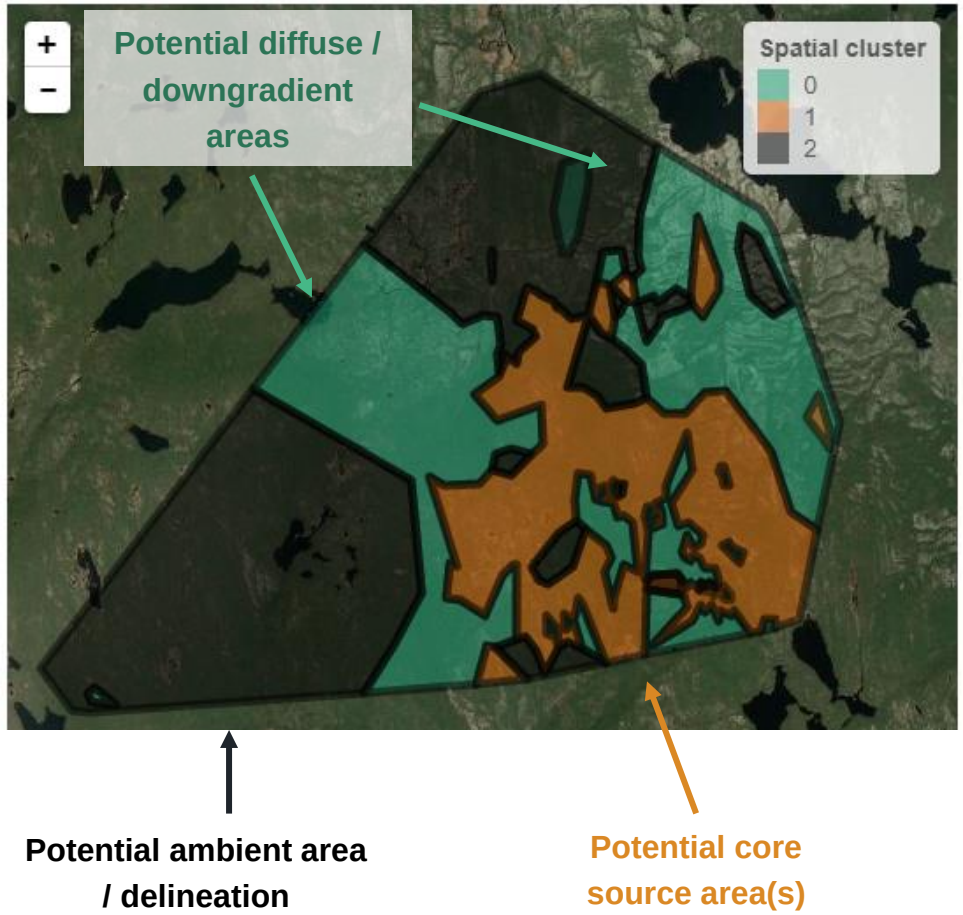
Site partitioning

Problem: **Challenging to partition areas** for detailed investigation and management at complex sites.

Solution: Spatial clustering incorporates fingerprints, ambient levels, and spatial relationships to **define chemically similar zones**.

Benefit: **Data-driven site partitioning** for further investigation or remediation, incorporating knowledge of ambient levels.

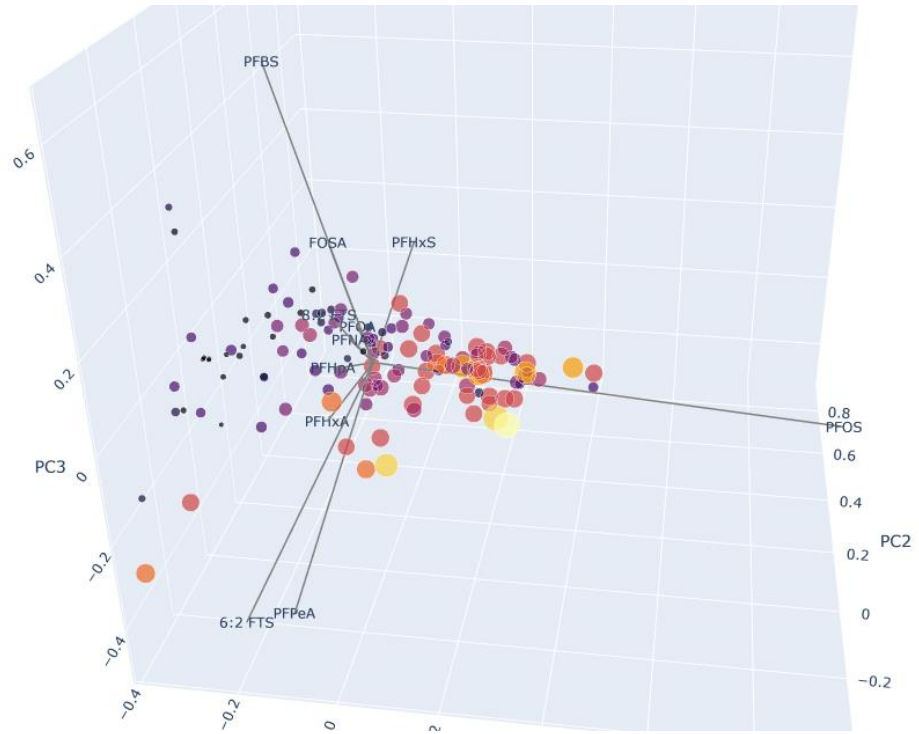




Google Earth anonymous aerial used

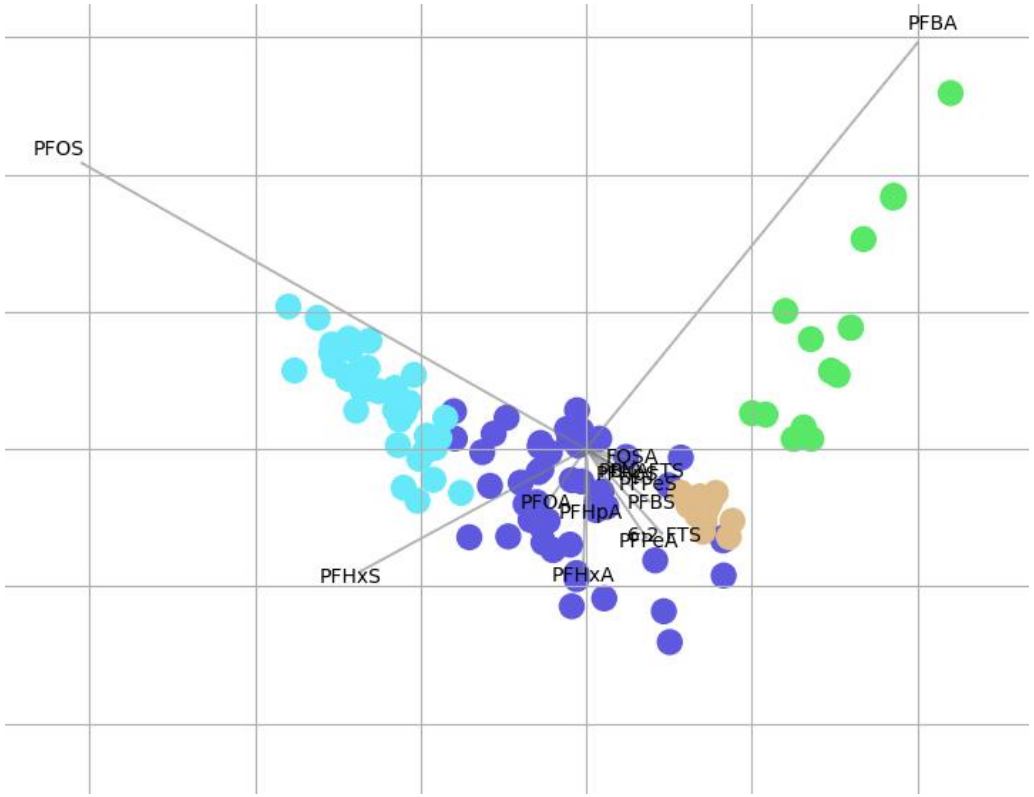
PCA Bi-Plots

- Linear dimensionality reduction technique that transforms the data into a new coordinate system (termed Principal Components, or PC)
- PCA creates orthogonal (right angle) PC to explain the maximum amount of variance per PC.
- **Clustering of Samples**: The proximity of points to each other would suggests that those samples have similar profiles
- **Direction and Length of Vectors**: The length and direction of each indicate the variable's significance in data variability
- **Correlation between Variables**: Variables close to each other or pointing in the same direction are positively correlated



PCA Bi-Plots – Example Interpretation

- **Green Points**: These points are clustered together and separate from the turquoise and purple points. They are also aligned more closely with the PFBA vector and not much with PFHxS or PFOS vectors, suggesting these samples have higher levels of PFBA and lower levels of PFHxS and PFOS.
- **Turquoise Points**: These points are a more diffused cloud and moderately associated with both PFOS and PFHxS, indicating mixed or average levels of these chemicals in these samples.
- **Purple Points**: These points are also a diffused and are scattered between PFOS, PFBA, and PFHxS, without a strong association with any, suggesting these samples have varied but generally moderate levels of all three chemicals.
- **Potential Clustering**: Given that there is no very distinct group(s) of points in the PCA bi-plot however, the results are likely show more partitioning rather than true clusters.



Identifying Unique Signals

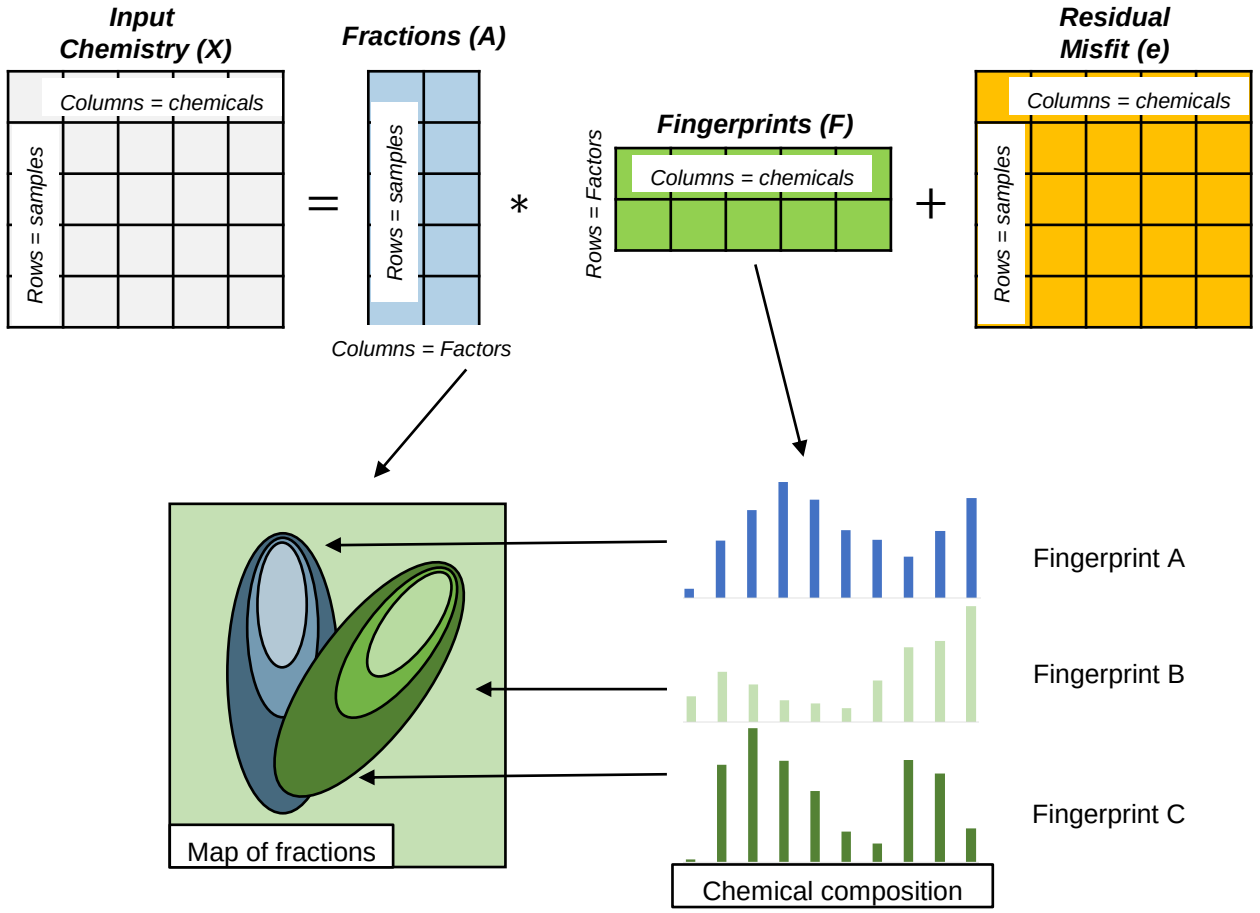
Sample	PFOS	PFOS	PFHxS	…
1	⟷			
2	⟷			
3	⟷			
…	⟷			

Blind Source Separation

Problem: **sources difficult to resolve** by traditional graphical methods.

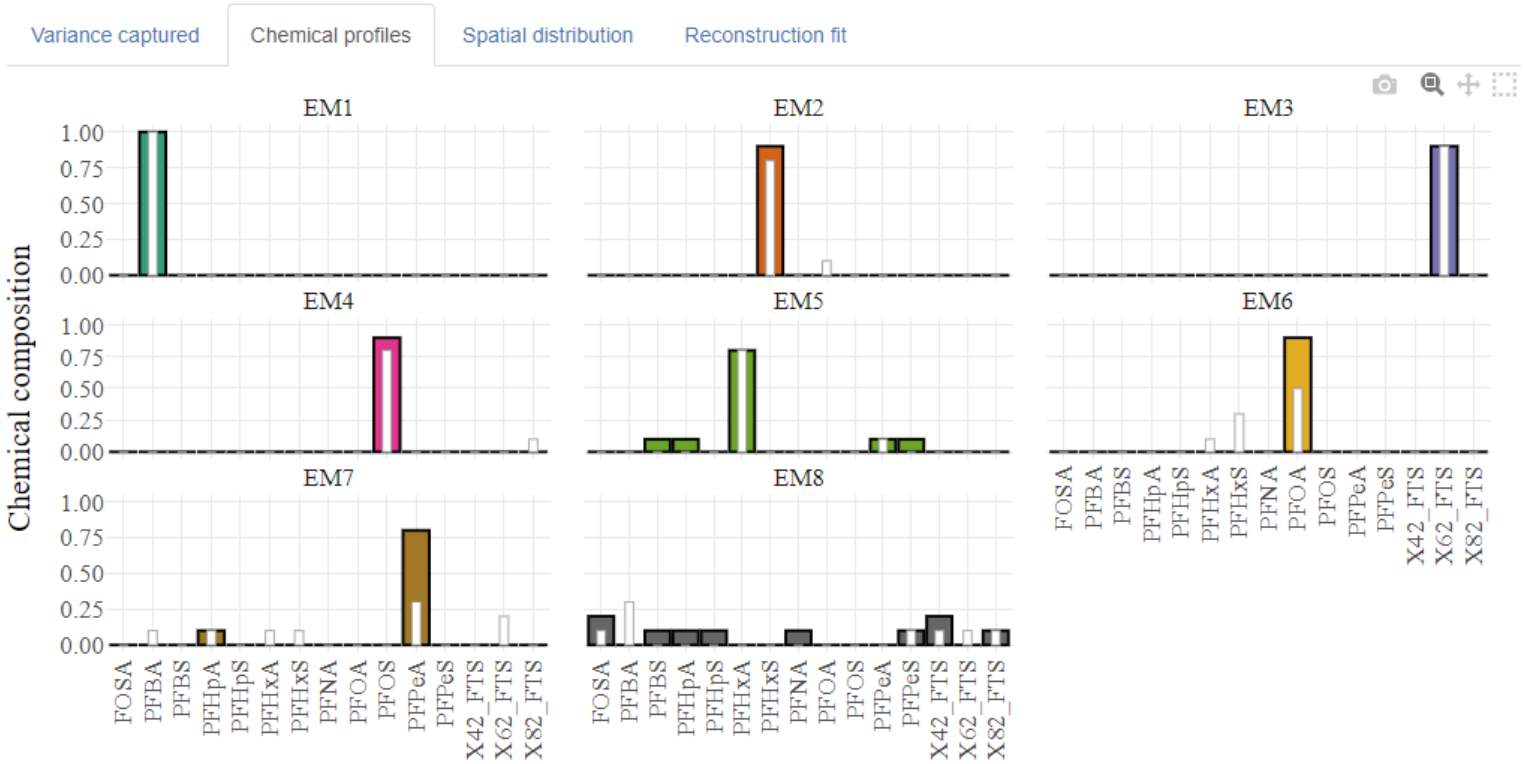
Solution: Matrix factorization performs **“blind source separation”** without prior knowledge.

Benefit: **data-driven fingerprints + spatial distribution** to assess onsite / offsite sources & fate-and-transport patterns.



Identifying Unique Signals

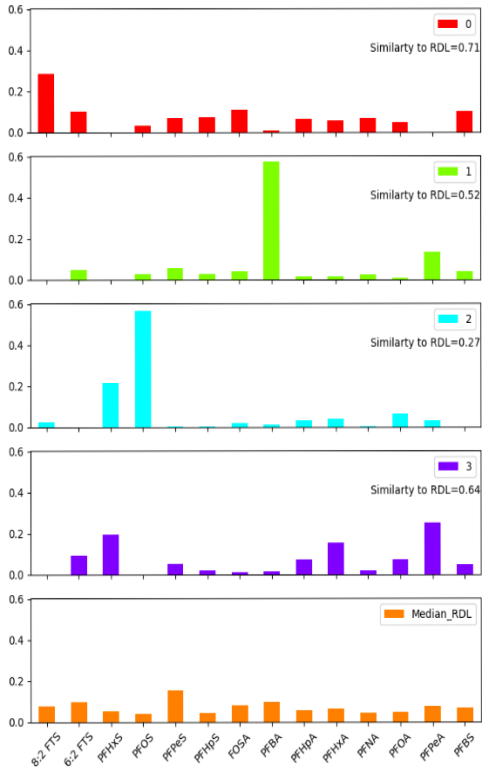
- Resolved fingerprints reflect “corners” of the compositional data space – the most unique profiles that **bracket the data**.
- Should (somewhat) **align with real samples** / established profiles.
- EM1 ~ an anomalous PFBA analytical signal
- EMs 2-7 potential source / F&T patterns
- EM8 ~ a profile of detection limits (background)

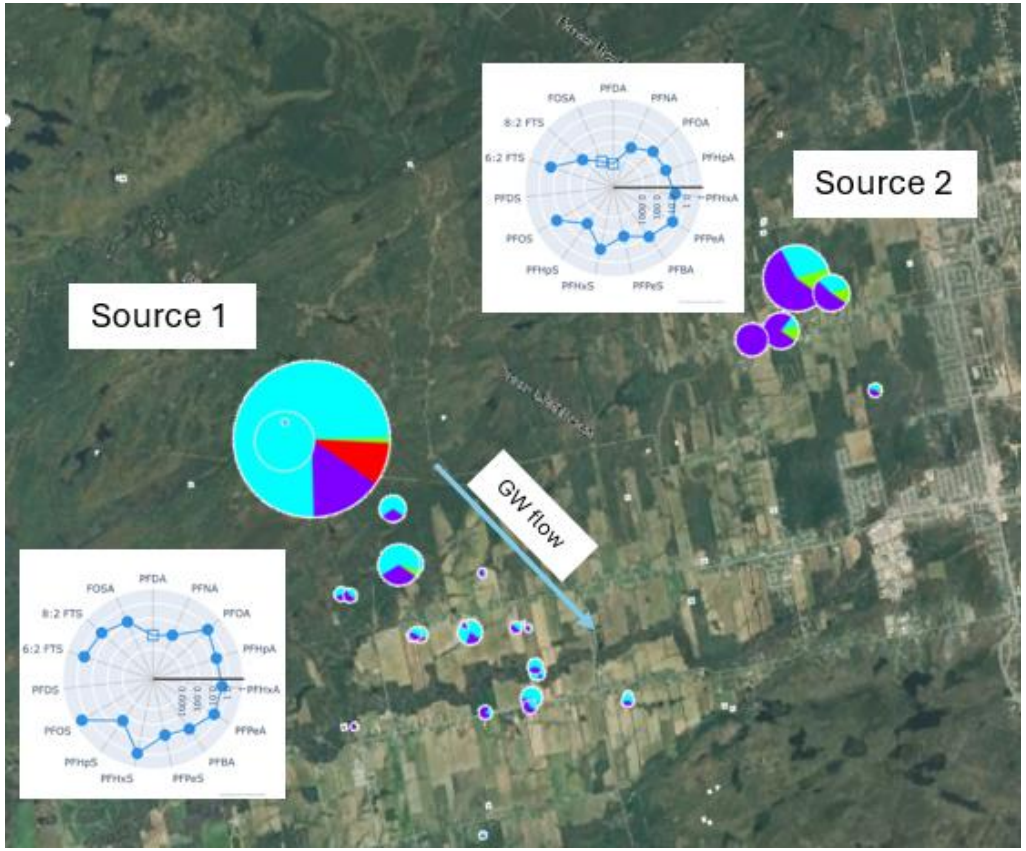


Identifying Unique Signals- Example

Unique Signal Results

- **Red:** Inferred background
- **Green:** PFBA dominated
- **Turquoise:** PFOS / PFHxS Dominated
- **Purple:** Primarily PFPeA, PFHpA, PFHxS





Google Earth anonymous aerial used

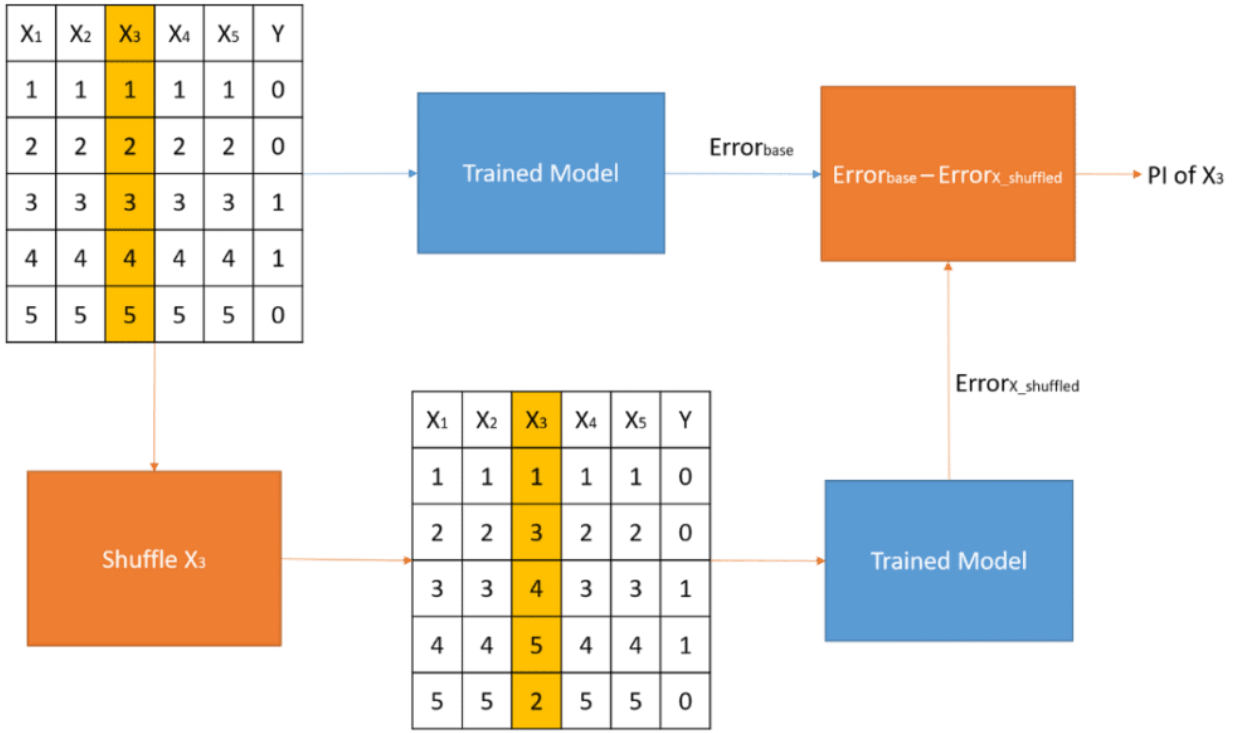
Explainable AI (XAI)

Feature importance

Problem: Surface/subsurface conditions can **modify PFAS signatures** (i.e., geochemistry, precursor transformations, fate & transport)

Solution: Feature importance uses data resampling to **identify environmental attributes** that help explain signatures.

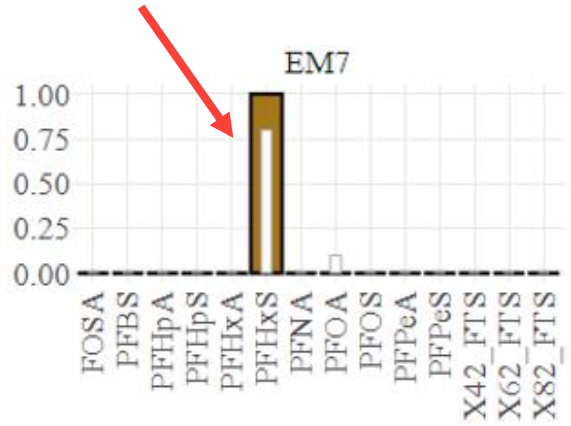
Benefit: Simplify complex data relations and **identify key geochemical / precursor drivers**, guide remedial design.

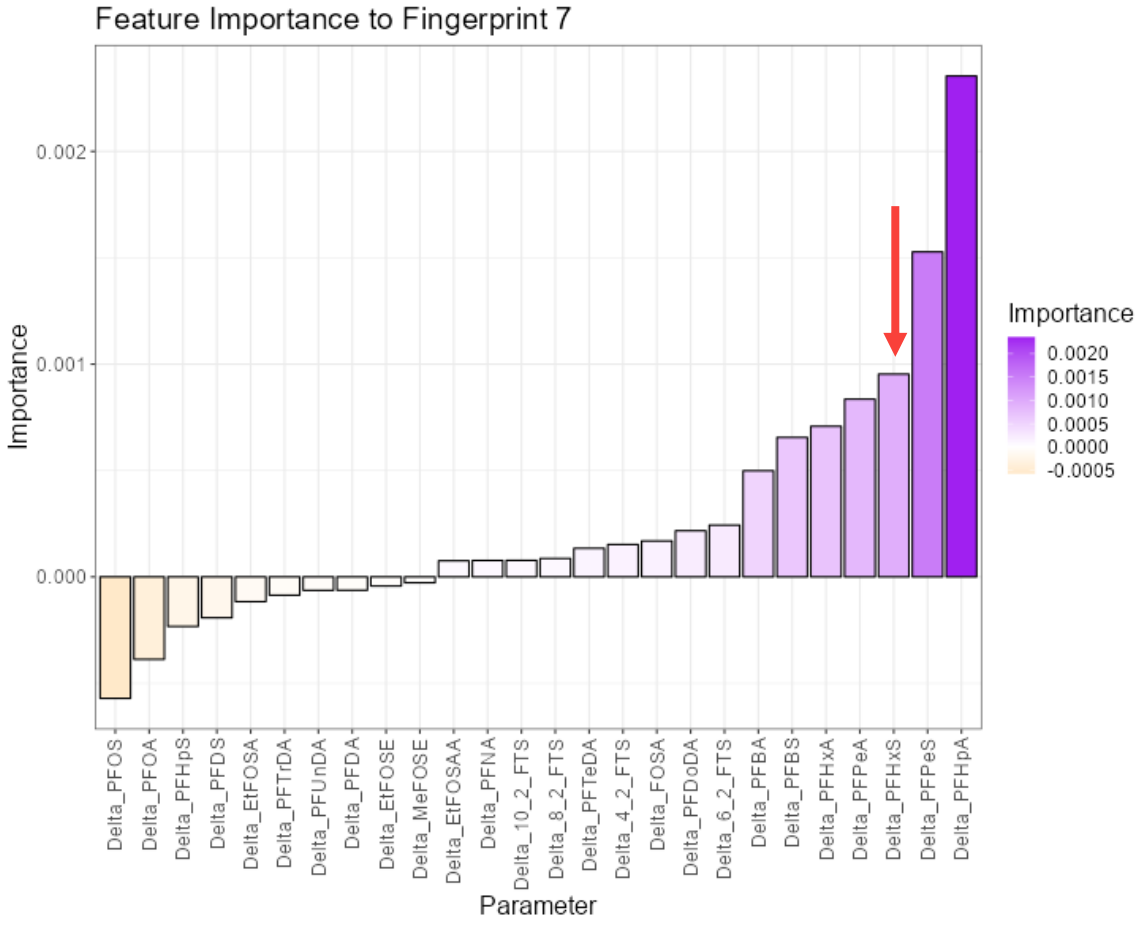


https://www.aporia.com/learn/feature-importance/explain-ml-models-with-permutation-importance/

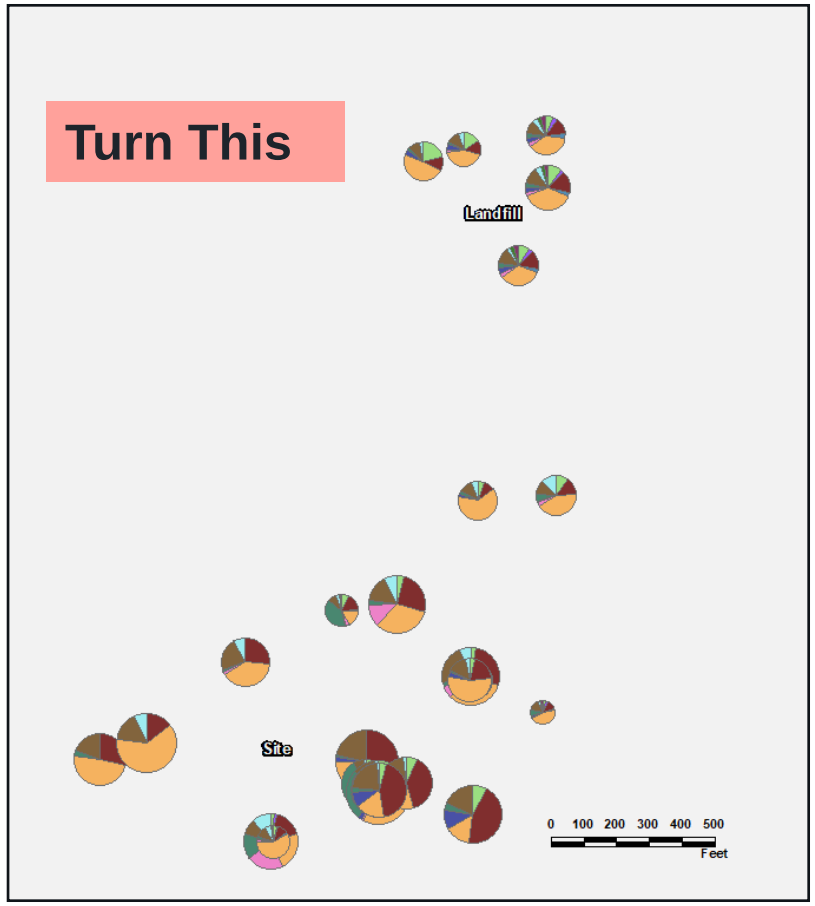
Explainable AI (XAI) Example

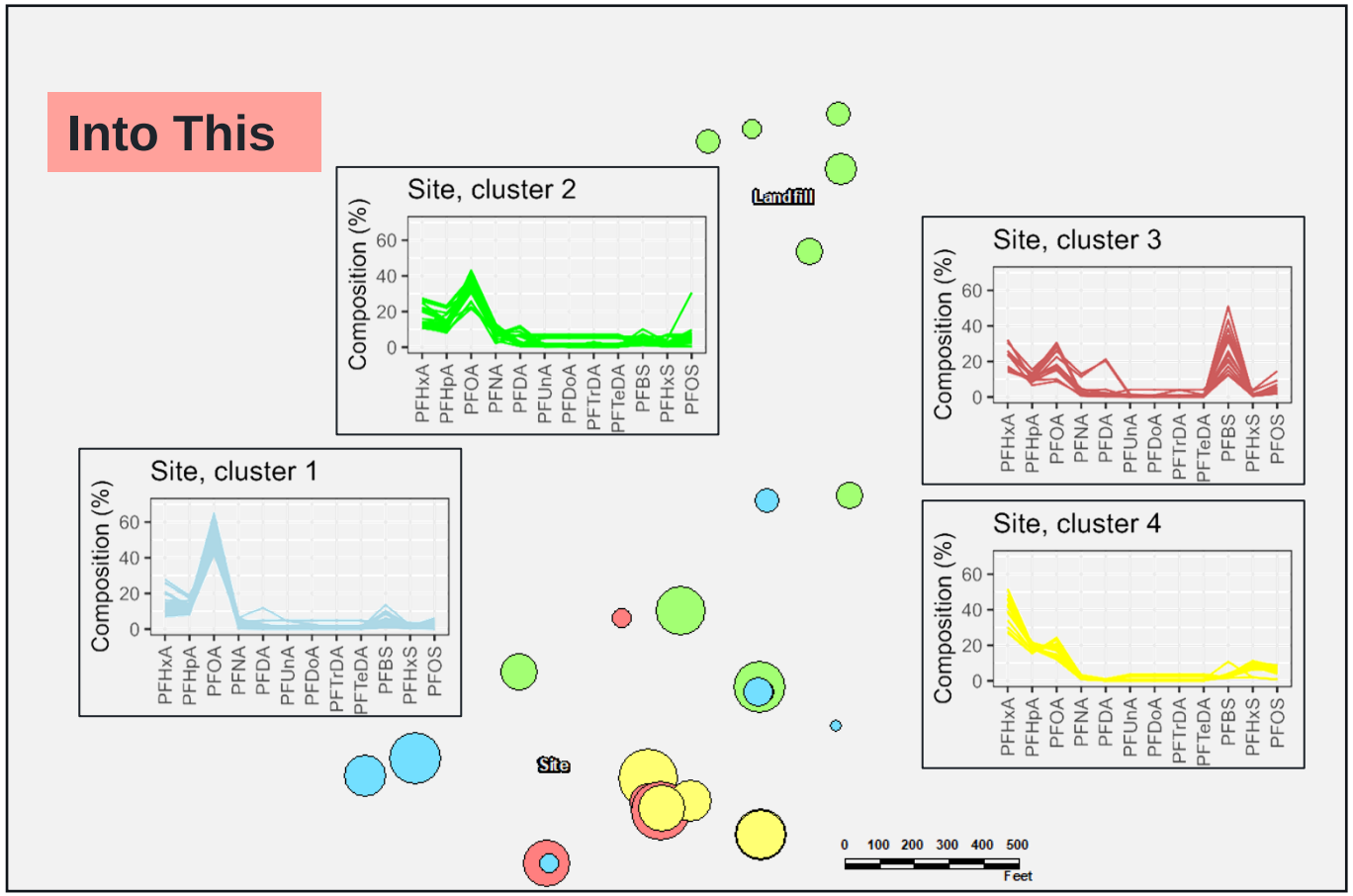
- PFHxS = key fingerprint component for signature # 7.
- **Question:** Can the TOP assay explain where a compound may have come from
- **Result:** PFHxS also key driver from TOP assay, identified by feature importance (3rd place). Potential link between PFHxS in the signature and precursor transformation.





Overall Takeaway - Decode Patterns, Gain Clear Insights





Key Takeaways and Lessons Learned

- ✓ Data matters – Garbage In = Garbage Out applies
- ✓ Data matters – Garbage In = Elevated Reporting Detection Limits Are Challenges
- ✓ Be Cautious of Over Analysis and Bias
- ✓ Improved Stakeholder Communication and Simplified Analysis of Complex Data
- ✓ Multiple Lines of Evidence
- ✓ Increased Confidence in Site Management
- ✓ Early insights promote better planning & more robust decisions.

More WSP PFAS Capabilities

Questions?

Contact Details:

Paul Hurst, M.Sc., P.Eng
WSP Canada

Paul.Hurst@wsp.com

Skyler Sorsby, M.Sc., P.G.
WSP USA

Skyler.Sorsby@wsp.com

Stefano Marconetto, M.Sc., P.Eng
WSP Canada

Stefano.Marconetto@wsp.com