Machine learning reveals flowpath-structured distributions of pesticides and pharmaceuticals

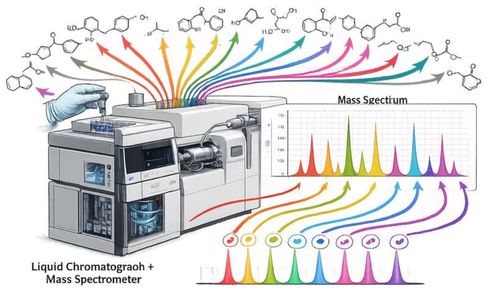
Shulamit Nussboim[1,2,3], Felicia Orah Rein[3]

1-Dept. of Environment, Planning and Sustainability, Bar ilan University; 2- Dept. of Geography and Environmental Studies University of Haifa; 3-Dept.of Environmental Resource Management, Soil Erosion Research Station, Ministry of Agriculture and Food Security;

1. Introduction and motivation

Organic pollutants in agricultural fields sourced from irrigation with treated effluents and pesticide application

Advanced in analytical techniques such as LC-MS, GS-MS

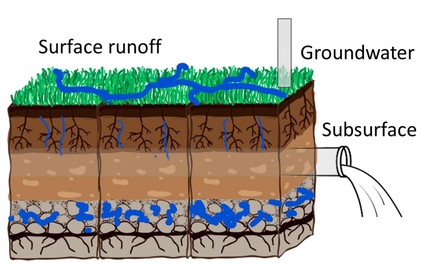


Tens of compounds per sample »

Thousands of data per a research field

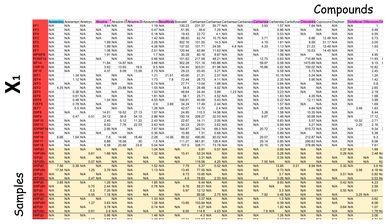
2. Methods

- 64 samples across flowpaths (OF, ditches, subsurface, GW)
- 81 compounds (LC-MS)
- Analysis:
- KCCA + Hellinger distance (HD) + Kruskal–Wallis (KW) with Dunn-Šidák.



Kernel CCA (KCCA) framework

Data (before centered, normalized)



Controls

Original configuration
Flowpaths
Field
Timing in the winter

Transposed:
Dominant flowpath
Mobility (Log Kow,Koc, solubility)
Half life

Mapping of original data into a latent space: $\Phi(x): R \rightarrow \mathcal{H}$ - - - → Mapping is unknown!

$\mathcal{H} - Hilbert\ (latent)\ space$

Kernel inner product (K) captures nonlinear relationships between variables.

- - - → This is the Kernel trick

Linear operations in latent space correspond to nonlinear functions of the original variables.

The inner product produces **canonical variables (Y_1, Y_2)** that maximize correlation in latent space.

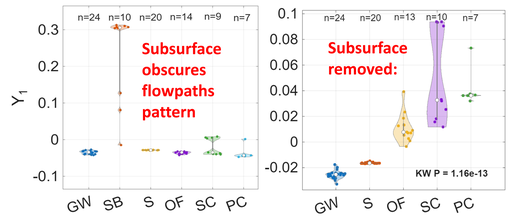
The **dominant flowpaths** are the highest concentration within the *90th percentile*. It was used as one of the controls (in transposed configuration).

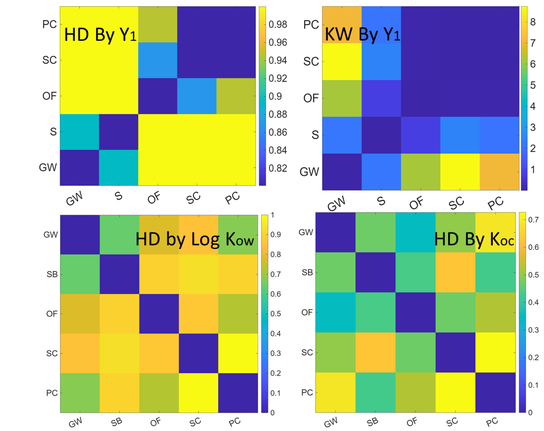
- Global significance was measured with **Kruskal-Wallis** test
- Significance within groups was estimated with **KW with Dunn-Šidák and Hellinger Distance, separately**

All significance analysis was taken over Y_1 (not original data)

3. Results

- 64 compounds were detected, including nine pharmaceuticals and 58 pesticides, totaling 2000 detected data.
- **Flowpath effects were highly significant (Kruskal–Wallis, $p \ll 0.001$).**
- Subsurface and stream variables obscured other flowpaths.
- After removing subsurface flowpaths, **differences among remaining flowpaths increased.**
- Hellinger distance showed more significant distinction of *flowpaths* (original config.), *dominant flowpath, and mobility measures* (transposed set).
- *Overland flow* is distinct from other sub-flowpaths: primary and secondary ditches





4. Discussion and conclusion

- **Flowpaths were shown to be a strong control on pollutant distribution within the field.**
- **Flowpaths structure pollutant composition more strongly (HD) than concentration levels (KW).**
- **Subsurface and stream contributions indicate integrative processes in time (subsurface- accumulation) and space (catchment integration).**
- Runoff sub-flowpaths distinction implies different processes involved in overland flow and ditches.
- Log Kow improved discrimination within flowpaths

Take home message:

KCCA reveals latent structure not observable in the original data, revealing **quantitatively the flowpath structure** that is not apparent in the original data.

shulamitnus@gmail.com
LinkedIn/ResearchGate: Shulamit Nussboim