Data-driven environmental monitoring of soil potentially toxic elements using multisource remote sensing and Machine Learning

Maria Silvia Binetti [1] [2], Carmine Massarelli [2], Jonathan V. Solórzano[3], Jean Francois Mas[3], Emanuele Barca [2], and Vito Felice Uricchio [2]

[1] Department of Earth and Geoenvironmental Sciences, University of Bari Aldo Moro, 70125 Bari, Italy (maria.binetti@uniba.it)
[2] Environment and Territory Research Unit, Construction Technologies Institute, Italian National Research Council (ITC-CNR), 70124 Bari, Italy
[3] Centre for Research in Environmental Geography, National Autonomous University of Mexico (UNAM), Morelia, Mexico

This presentation participates in OSPP

Outstanding Student & PhD candidate Presentation contest

Introduction

PLS(Partial Least Squares Regression) + **LOOCV** (Leave-One-Out Cross-Validation) **6 heavy metals model comparison**

Predictive Mapping: Standard Random Forest vs. UMAP Integration

Conclusions

Introduction

Soil contamination monitoring in industrialized regions requires accurate, spatially continuous assessments.
We present an integrated remote sensing framework for predicting concentrations of soil Potentially Toxic Elements (Cd, Cr, V, Co, Be, As) near the industrial area of Taranto (southern Italy), a priority site for environmental risk assessment.
Preliminary exploratory analysis revealed significant relationships between PTEs and key soil properties, notably strong negative correlations with pH across all elements and positive associations with Organic Carbon (specifically for Cr, V, and Co).

Introduction

To isolate the unmixed mineralogical signal, initial chemometric modeling employed Partial Least Squares (PLS) regression with Leave-One-Out Cross-Validation (LOOCV), strictly applied to samples exhibiting pure bare soil PRISMA signatures to eliminate vegetation interference.

For continuous spatial predictive mapping, the framework integrated PRISMA hyperspectral imagery, Sentinel-2 multispectral indices, and DEM-derived topographic data. To systematically address hyperspectral multicollinearity, we evaluated an element-specific machine learning strategy, comparing a Standard Random Forest (RF) algorithm against an optimized variant utilizing UMAP (Uniform Manifold Approximation and Projection) for PRISM dimensionality reduction.

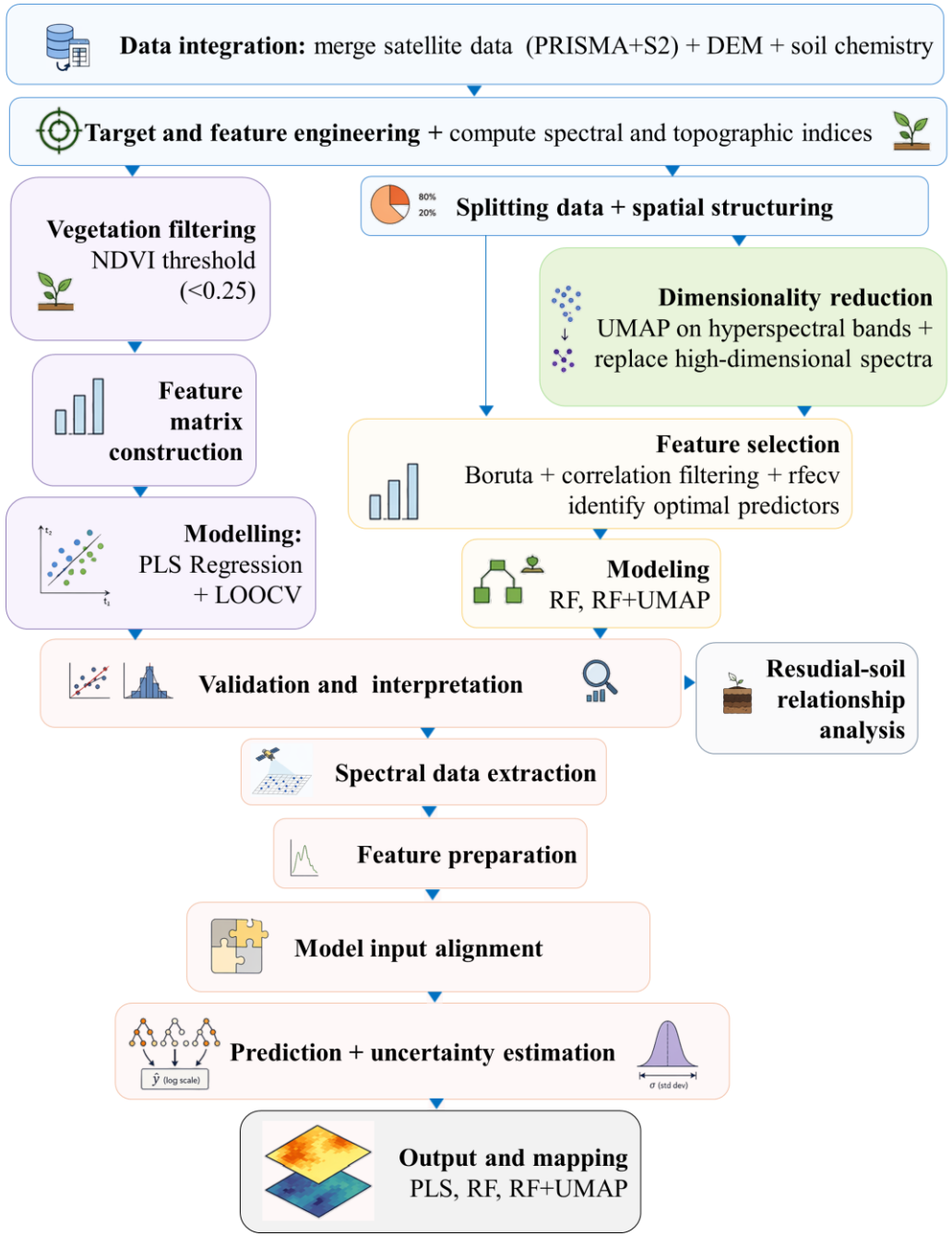
Results demonstrate that algorithm selection must be tailored to the target element: the RF-UMAP framework significantly enhanced predictive accuracy and successfully resolved residual spatial autocorrelation (Moran's Index) for Cd, V, Be, and As, achieving Test R^2 values up to 0.83 and relative errors (RRMSE) as low as 7.8%.

Standard RF preserved critical spectral variance for Cr and Co, yielding optimal performance without dimensionality reduction.

Feature importance analysis confirmed the dominant discriminative power of hyperspectral derivatives (UMAP components and specific narrow bands) alongside the Bare Soil Index (BSI).

This hybrid framework validates the efficacy of combining multisource satellite data for the rigorous spatial identification of soil contamination hotspots.

Methodology flowchart



PLS(Partial Least Squares Regression) + **LOOCV** (Leave-One-Out Cross-Validation) **6 heavy metals model comparison**

Here, we establish the baseline chemometric relationships. Partial Least Squares (PLS) regression, strictly validated via Leave-One-Out Cross-Validation (LOOCV), is applied exclusively to bare-soil PRISMA pixels. This targeted approach isolates the pure mineralogical signal from vegetation interference across all six target metals.

PLS(Partial Least Squares Regression) + LOOCV (Leave-One-Out Cross-Validation) 6 heavy metals model comparison

1. Data Loading & Integration
- Load datasets PRISMA S2 DEM chemical
- Standardize column names

2. Target Definition
- Select target element e.g. Be
- Log-transform: y = log1 + target

3. Feature Engineering
- NDVI Ferrous Index Color Index
- BSI Iron Oxide Ratio Clay Minerals

4. Vegetation Filtering
- NDVI less than 0.25
- Keep bare soil samples
- Remove NDVI

5. Feature Matrix Construction
- X = spectral bands + indices
- y = log target
- Standardization Z-score

6. Modeling: PLS + LOOCV
- Train PLS model
- Leave-One-Out CV
- Cross-validated predictions

7. Model Evaluation
- R^2 LOOCV
- RMSE
- MAE

8. Pedometric Metrics
- RPD = Std y / RMSE
- RPIQ = IQR / RMSE

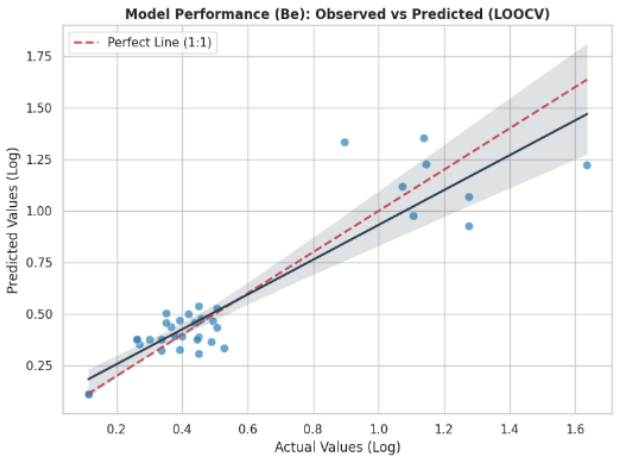
9. Residual Analysis
- Distribution normality
- Residuals vs predicted

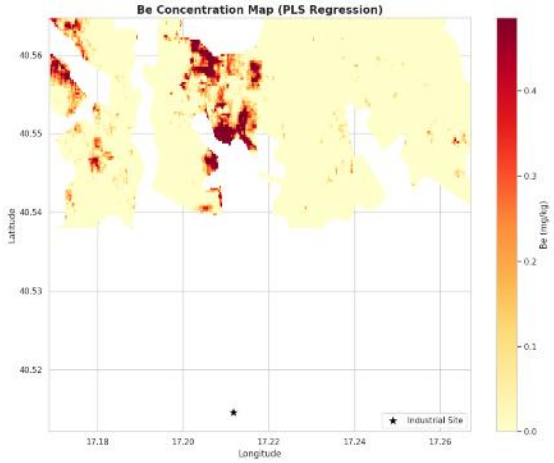
10. VIP Scores
- Compute VIP
- Rank variables
- Select VIP greater than 1

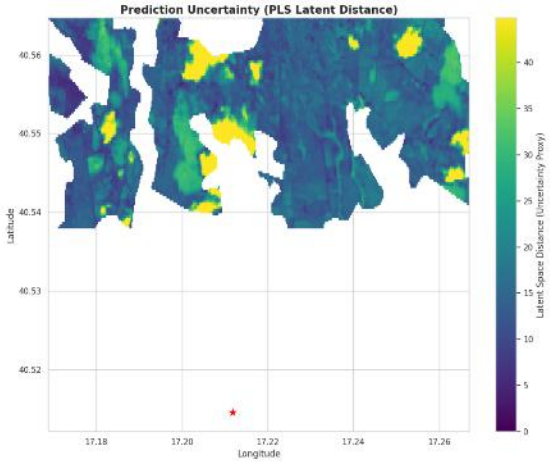
11. Output & Interpretation
- Model performance
- Key predictors
- Environmental insights

PLS(Partial Least Squares Regression) + LOOCV (Leave-One-Out Cross-Validation) 6 heavy metals model comparison

Metal	R^2	RPD	RPIQ	PRISMADelta	p-value	Overall Assessment
Cobalt (Co)	0.8895	3.0084	1.2353	**+0.0314**	0.0099	**Excellent — Quantitative Prediction**
Beryllium (Be)	0.8306	2.4298	1.1077	**+0.0147**	0.0099	**Good — Reliable Screening**
Vanadium (V)	0.7915	2.1901	1.5111	**-0.0100**	0.0099	**Good — Reliable Screening**
Cadmium (Cd)	0.7876	2.1698	2.4211	**-0.0022**	0.0099	**Good — Reliable Screening**
Arsenic (As)	0.7352	1.9431	2.8764	**+0.0309**	0.0099	**Moderate — Qualitative Indication**
Chromium (Cr)	0.7029	1.8346	1.4943	**-0.0308**	0.0099	**Moderate — Qualitative Indication**







Predictive Mapping: Standard Random Forest vs. UMAP Integration

This section presents the continuous spatial modelling framework. It systematically contrasts the predictive power, error metrics, and spatial validity (Moran's Index) of a Standard Random Forest algorithm against an optimized RF-UMAP pipeline, demonstrating the necessity of an element-specific modelling strategy.

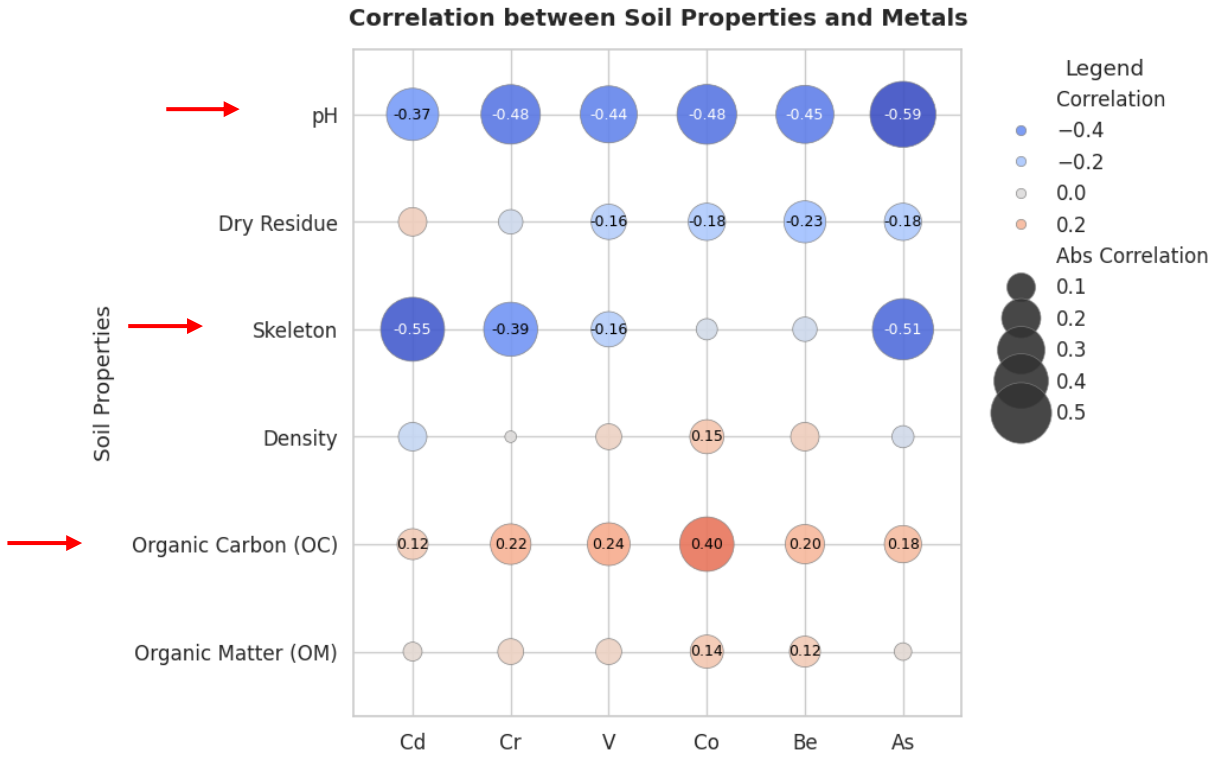
Predictive Mapping: Standard Random Forest vs. UMAP Integration

Configuration

	Cd		Cr		V		Co		Be		As	
	UMAP	NO UMAP	UMAP	NO UMAP	UMAP	NO UMAP	UMAP	NO UMAP	UMAP	NO UMAP	UMAP	NO UMAP
TEMPLATE CONFIGURATION												
n_estimators	500	500	500	500	500	500	500	500	500	500	500	500
max_depth	4	6	8	8	6	6	10	6	10	6	8	10
min_samples_split	2	8	4	4	2	4	2	2	2	4	4	2
min_samples_leaf	2	2	2	2	2	2	2	2	2	2	2	2
max_features	sqrt	log2	sqrt	log2	log2	log2	sqrt	sqrt	log2	log2	sqrt	sqrt

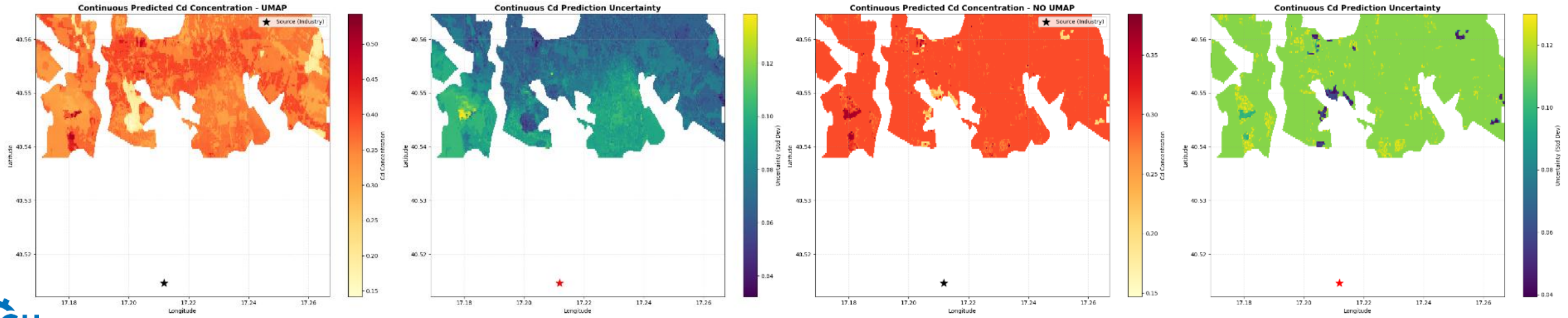
Predictive Mapping: Standard Random Forest vs. UMAP Integration

Residuals vs. Soil Properties Correlation



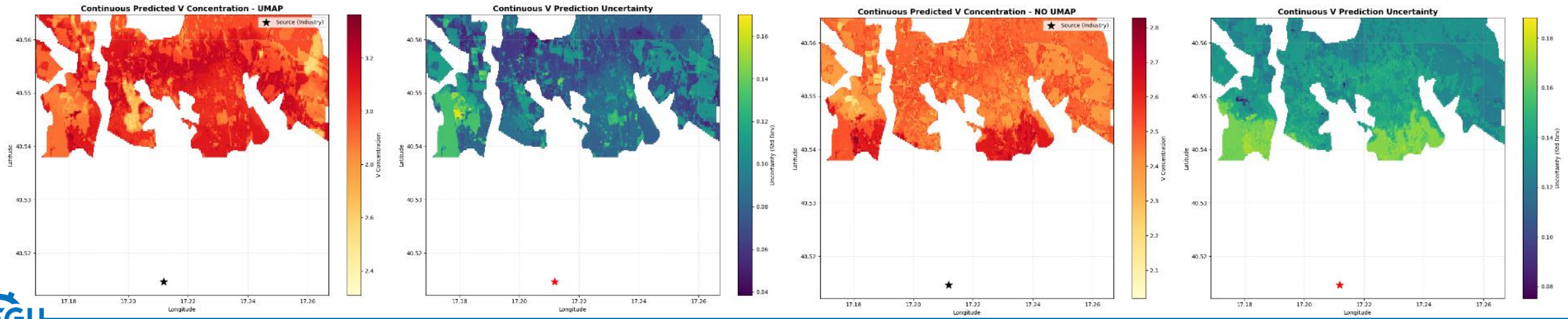
Cd

	UMAP	NO UMAP
Residual Moran's Index	0.068 (p=0.13)	0.220 (p=0.029)
Spatial autocorrelation	No	Yes (p<0.05)
Top feature	UMAP_1 (38.7%)	2409.98nm (55.3%)



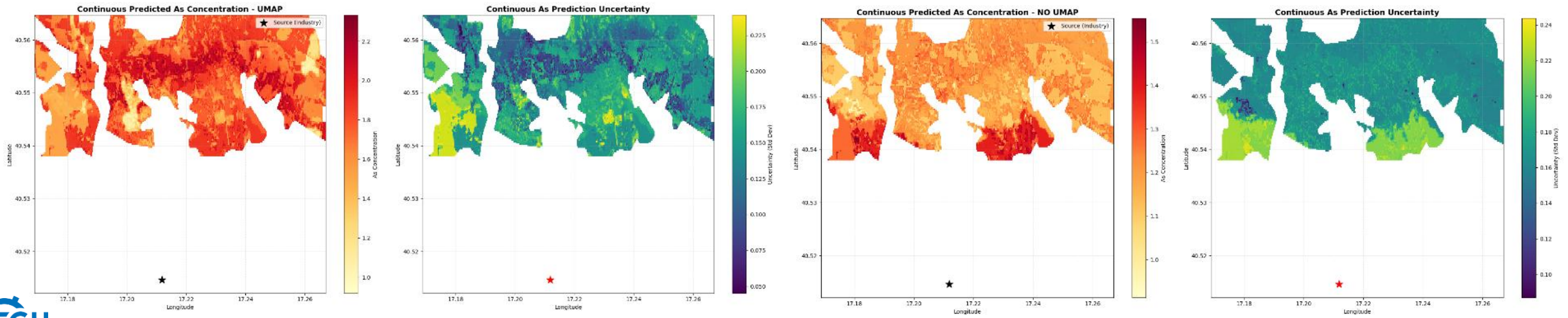
V

Key Aspect	**UMAP**	**NO UMAP**
Residual Moran's Index	0.016 (p=0.230)	0.253 (p=0.017)
Spatial autocorrelation	No	Yes (p<0.05)
Top feature	UMAP_1 (28.0%)	1962.17nm (24.5%)



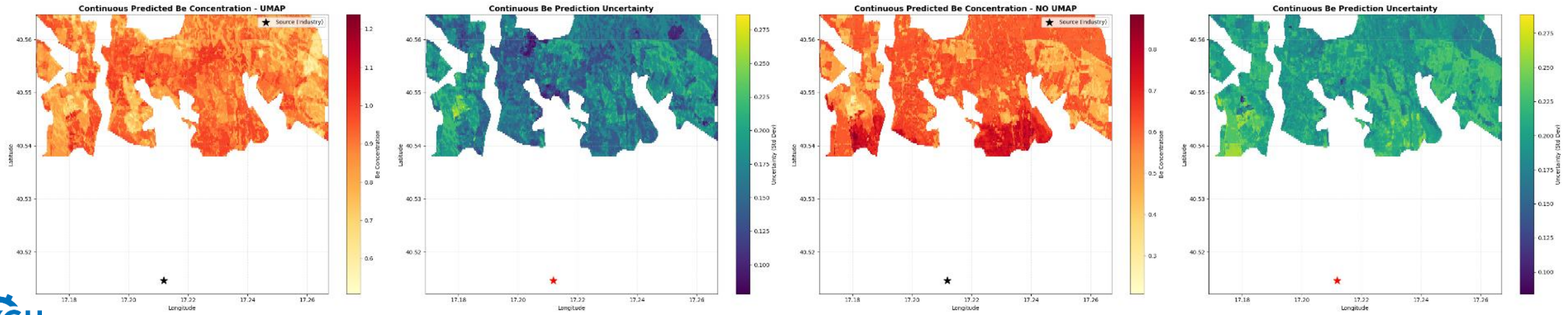
As

Key Aspect	UMAP	NO UMAP
Residual Moran's Index	–0.030 (p=0.356)	0.352 (p=0.004)
Spatial autocorrelation	No	Yes (p=0.004)
Top feature	UMAP_1 (38.9%)	1953.43nm (27.1%)



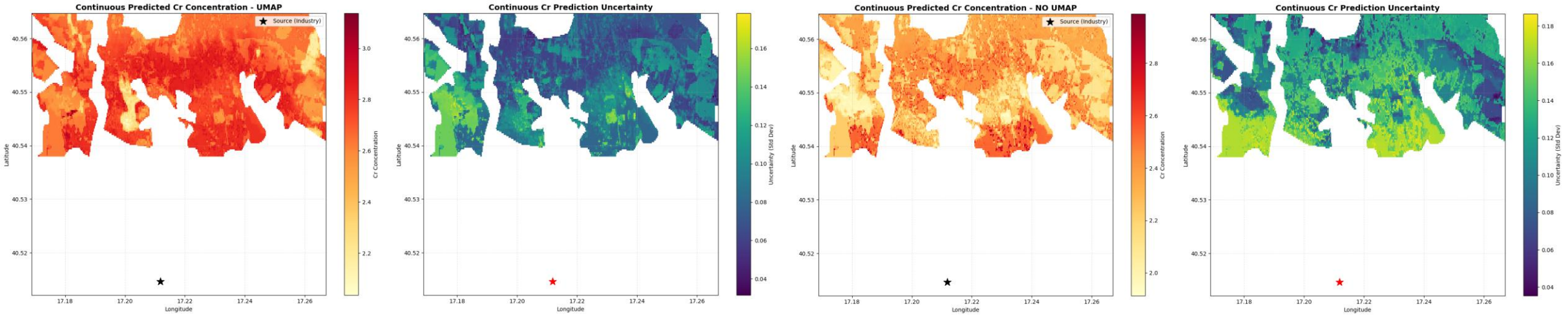
Be

Key Aspect	UMAP	NO UMAP
Residual Moran's Index	0.171 (p=0.044)	0.294 (p=0.003)
Spatial autocorrelation	Borderline (p=0.044)	Yes (p=0.003)
Top feature	UMAP_1 (22.4%)	BSI (24.1%)

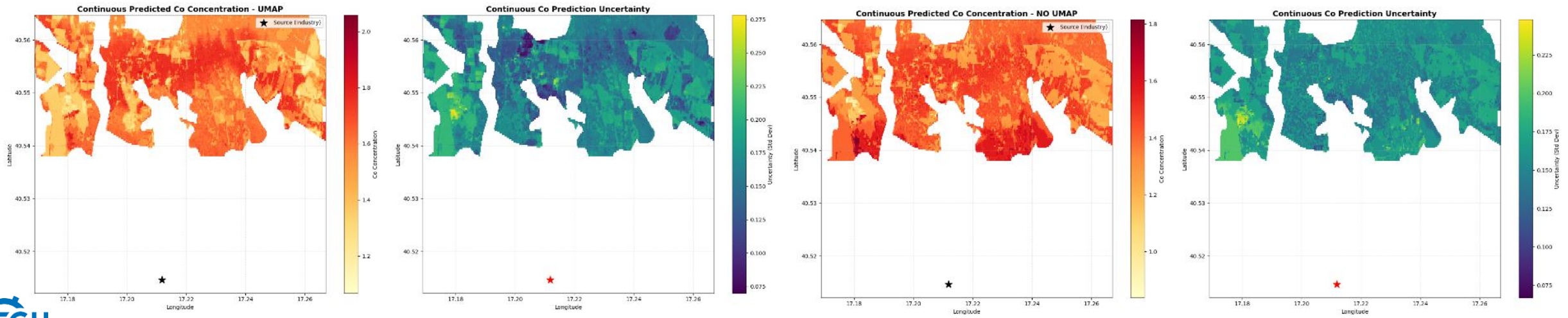


Cr

Key Aspect	**UMAP**	**NO UMAP**
Residual Moran's Index	–0.109 (p=0.308)	0.093 (p=0.099)
Spatial autocorrelation	No	No
Top feature	UMAP_1 (34.1%)	1962.17nm (36.2%)



Key Aspect	**UMAP**	**NO UMAP**
Residual Moran's Index	0.143 (p=0.060)	0.091 (p=0.117)
Spatial autocorrelation	No (borderline)	No
Top feature	UMAP_1 (23.2%)	BSI (30.4%)



Continuous spatial prediction and uncertainty maps were generated using an **element-specific framework**: the PRISMA-UMAP dimensionality reduction pipeline was applied exclusively to elements where it demonstrably resolved spatial autocorrelation and improved accuracy (Cd, V, Be, As), while standard high-dimensional Random Forest was retained for Cr and Co to prevent signal loss.

Conclusions

The final section synthesizes the key findings, confirming the efficacy of integrating hyperspectral data with ML. It highlights how tailored dimensionality reduction and multisensory data fusion provide a robust framework for identifying environmental contamination hotspots.

Conclusions

The results of this study demonstrate the high potential of integrating **PRISMA hyperspectral data** with multispectral indices and topographic features for the high-resolution mapping of **soil Potentially Toxic Elements (PTEs)** in complex industrial landscapes.
The implementation of an **element-specific modelling framework** proved to be the most robust approach:

- **Hybrid modelling efficiency -** dimensionality reduction via UMAP was essential to resolve residual spatial autocorrelation and maximize predictive accuracy for Cd, V, Be, and As (achieving Test R^2 values up to 0.83). Conversely, a Standard RF approach was superior for Cr and Co, demonstrating that retaining high-dimensional spectral variance is preferable for specific mineralogical signatures.

- **Proxy validity -** the consistent negative **correlation between PTEs and pH**, alongside **positive associations with Organic Carbon**, confirms the scientific validity of using these soil properties as reliable indirect proxies for trace metal detection via remote sensing.

- **Spatial reliability -** systematic validation through the **Moran's Index** ensured that the generated continuous prediction maps are spatially sound and free from bias, providing a **high-fidelity representation of contamination patterns**.

- **Early-warning and regulatory compliance** - The continuous prediction maps offer high-fidelity spatial reliability for environmental risk assessment. Evaluated against strict Italian thresholds (D.Lgs. 152/2006 & D.M. 46/2019), all six modelled PTEs currently remain below the legal agricultural safety limits. Models successfully identified the industrial dispersion footprint highlighting Be and Cd as the elements with the narrowest

In conclusion, this **multi-sensor and hybrid machine learning framework provides a scalable and cost-effective tool for environmental risk assessment, enabling the precise identification of soil contamination hotspots in priority industrial sites** like **Taranto**.

Bibliography

•**Adamo, P., Iavazzo, P., Albanese, S., Agrelli, D., De Vivo, B., & Lima, A.** (2024). Leveraging machine learning for sustainable cultivation of Zn-enriched crops in Cd-contaminated karst regions. *Science of The Total Environment*, *954*, 176650. https://doi.org/10.1016/j.scitotenv.2014.08.085

•**Bai, L., Ding, S., Huang, X., Chen, X., Chen, Y., Cao, X., Wang, X., Yu, X., & Dai, J.** (2024). Predicting Cd accumulation in crops and identifying nonlinear effects of multiple environmental factors based on machine learning models. *Science of The Total Environment*, *951*, 175787. https://doi.org/10.1016/j.jclepro.2023.138081

•**Cheng, S., Zhang, G., Yang, X., & Lei, B.** (2023). A multiscale geographically weighted regression kriging method for spatial downscaling of satellite-based ozone datasets. *Frontiers in Environmental Science*, *11*, 1267752. https://doi.org/10.3389/fenvs.2023.1267752

•**Han, H., & Suh, J.** (2024). Spatial Prediction of Soil Contaminants Using a Hybrid Random Forest–Ordinary Kriging Model. *Applied Sciences*, *14*(4), Articolo 1666. https://doi.org/10.3390/app14041666

•**Karamalidis, A. K., Torres, S. G., Hakala, J. A., Shao, H., Cantrell, K. J., & Carroll, S.** (2012). Trace Metal Source Terms in Carbon Sequestration Environments. *Environmental Science and Technology*, *47*(1), 322–329. https://doi.org/10.1021/es304832m

•**Keskin, H., & Grunwald, S.** (2018). Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma*, *326*, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

•**Lafuente, A. L., González, C., Quintana, J. R., Vázquez, A., & Romero, A.** (2008). Mobility of heavy metals in poorly developed carbonate soils in the Mediterranean region. *Geoderma*, *145*(3-4), 238–244. https://doi.org/10.1016/j.geoderma.2008.03.012

•**Li, Y., Xiang, B., Wang, T., He, Y., Liu, X., Li, Y., Ren, S., Wang, E., & Guo, G.** (2025). Applications of machine learning in potentially toxic elemental contamination in soils: A review. *Ecotoxicology and Environmental Safety*, *295*(9), 118110. https://doi.org/10.1016/j.ecoenv.2025.118110

•**Manaf, M., Ali, Z., & Scholz, M.** (2026). Integrating random forest-based regression kriging for analyzing spatial variability of rainfall in arid and semi-arid regions. *Scientific Reports*, *16*(1), 5298. https://doi.org/10.1038/s41598-026-36074-4

•**Petronio, B. M., Cardellicchio, N., Calace, N., Pietroletti, M., Pietrantonio, M., & Caliandro, L.** (2011). Spatial and Temporal Heavy Metal Concentration (Cu, Pb, Zn, Hg, Fe, Mn, Hg) in Sediments of the Mar Piccolo in Taranto (Ionian Sea, Italy). *Water, Air, & Soil Pollution*, *223*(2), 863–875. https://doi.org/10.1007/s11270-011-0908-4

•**Trifuoggi, M., Pagano, G., Oral, R., Gravina, M., Toscanesi, M., Mozzillo, M., Siciliano, A., Burić, P., Lyons, D. M., Palumbo, A., Thomas, P. J., D'Ambra, L., Crisci, A., Guida, M., & Tommasi, F.** (2018). Topsoil and urban dust pollution and toxicity in Taranto (southern Italy) industrial area and in a residential district. *Environmental Monitoring and Assessment*, *191*(1), 43. https://doi.org/10.1007/s10661-018-7164-7

•**Xie, K., Ou, J., He, M., Peng, W., & Yuan, Y.** (2024). Predicting the Bioaccessibility of Soil Cd, Pb, and As with Advanced Machine Learning for Continental-Scale Soil Environmental Criteria Determination in China. *Environment & Health*, *2*(9), 631–641. https://doi.org/10.1021/envhealth.4c00035

The end

Thank you for your time 🧡

This presentation participates in OSPP

Outstanding Student & PhD candidate Presentation contest