Environmental Pollution 390 (2026) 127447

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

# Beyond accuracy: Stabilizing feature importance in GWRF/RF for soil heavy metal mapping[☆]

## ARTICLE INFO

*Keywords:*
GWRF
Random forest
SHAP
Feature importance stability
Soil heavy metals

## ABSTRACT

Qin et al. (2025) integrate geographically weighted random forest with multi-source environmental covariates to map soil heavy metals, showing that optimal bandwidth selection and fusion of local GWRF with global RF markedly improve predictive accuracy. Using mean decrease in impurity and SHAP, they identify O3, topography, mining proximity, and rainfall as dominant predictors. However, feature importance lacks ground-truth validation, and rankings are model-dependent, risking misleading interpretation despite strong target prediction. Because SHAP explanations inherit supervised-model biases, interpretability cannot be inferred from accuracy. We recommend rank-stability auditing and label-agnostic validation: generate independent rankings (e.g., feature agglomeration, highly variable feature selection, Spearman's ρ), then perform a leave-one-out rank-stability test to assess order robustness. Transparent reporting of ranking instability should accompany predictive metrics.

Qin et al. (2025) conducted a comprehensive investigation using geographically weighted random forest (GWRF) to integrate multi-source environmental covariates for spatial prediction of soil heavy metals. Their research meticulously examined how varying bandwidth and weight parameters influence GWRF performance. The findings demonstrated that carefully selecting appropriate bandwidth parameters and strategically integrating local GWRF and global RF model results can significantly enhance prediction accuracy for soil heavy metal distribution mapping. They employed interpretable machine learning approaches, specifically Mean Decrease in Impurity (MDI) and Shapley Additive exPlanations (SHAP), to identify key environmental factors influencing soil heavy metal concentrations. These analyses revealed that air quality parameters (particularly O3 levels), topographic features, proximity to mining operations, and rainfall patterns emerged as the most significant predictors in their model framework (Qin et al., 2025).

It is essential to recognize that supervised models like GWRF and RF exhibit two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth measurements using labels, feature importance calculations lack corresponding validation benchmarks. This critical distinction was not adequately addressed by Qin et al. Due to this absence of ground truth for feature importance validation, they demonstrated that different models inevitably generate varying feature importance rankings, reflecting model-specific characteristics rather than objective reality, which can lead to potentially misleading interpretations. In other words, feature importances derived from supervised models are inherently skewed (Fisher et al., 2019; Steiner & Kim, 2016; Nalenz et al., 2024; Nazer et al., 2023; Ugirumurera et al., 2024; Alaimo Di Loro et al., 2023; Adler & Painsky, 2022; Dunne et al., 2023; Strobl et al., 2007; Wallace et al., 2023).

Furthermore, the Qin et al.'s implementation of SHAP for model explanation (explain = SHAP(model)) inherently bases interpretations solely on the underlying supervised model. This approach means that SHAP necessarily inherits—and may amplify—any biases present in the original model's feature importance calculations (Wu, 2025; Bilodeau et al., 2024; Huang & Marques-Silva, 2024; Kumar et al., 2021; Hooshyar & Yang, 2024; Lones, 2024; Molnar et al., 2022; Létoffé et al., 2025; Ponce-Bobadilla et al., 2024; Coupland et al., 2025). While SHAP itself offers robust explanation capabilities, its reliability ultimately depends on the quality of the feature importance signals from the supervised model. This dependency creates a critical limitation: even when a model achieves high prediction accuracy, its explanatory outputs may still lack validity, as predictive performance and interpretability represent fundamentally distinct dimensions of model evaluation. This distinction highlights the need for separate validation protocols for explanatory mechanisms beyond traditional predictive metrics.

Because feature-importance ranking orders, not merely the magnitude of individual scores, dictate which environmental covariates are labeled "most influential," Qin et al. must rigorously challenge the stability of those orders. We therefore recommend augmenting their GWRF/RF framework with label-agnostic strategies (e.g., feature agglomeration, highly variable gene selection) and non-target-driven metrics (such as Spearman's ρ) to generate independent ranking lists. They should then apply a simple leave-one-out rank-stability test: first, rank all covariates and select the top n among full features; next, remove the highest-ranked covariate and re-rank the remainder to identify the new top n-1; and finally, compare feature importance ranking orders between the original versus reduced lists. Large discrepancies in these ranking orders will expose the fragility of supervised feature rankings, warning against overinterpreting model-specific drivers and

[☆] This paper has been recommended for acceptance by: Sarah Michele Harmon.

https://doi.org/10.1016/j.envpol.2025.127447
Received 3 October 2025; Accepted 26 November 2025
Available online 18 December 2025

underscoring the imperative to validate and transparently report feature importance ordering.

## Consent to participate

Not applicable.

## Ethics approval

Not applicable.

## Consent for publication

Not applicable.

## Availability of data and material

Not applicable.

## Code availability

Not applicable.

## Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

**According to ScholarGPS,** Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

## AI use

Not applicable.

## Funding

This research has no fund.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## References

Adler, A.I., Painsky, A., 2022. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy 24 (5), 687. https://doi.org/10.3390/e24050687.

Alaimo Di Loro, P., Scacciatelli, D., Tagliaferri, G., 2023. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat. Methods Appl. 32, 237–270. https://doi.org/10.1007/s10260-022-00643-4.

Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. Proc. Natl. Acad. Sci. 121 (2), e2304406120. https://doi.org/10.1073/pnas.2304406120.

Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Hulvej Rod, N., Mishra, S., Bhatt, S., Unwin, H.J.T., 2025. Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health 25 (1), 1520. https://doi.org/10.1186/s12889-025-22705-4.

Dunne, R., Reguant, R., Ramarao-Milne, P., Szul, P., Sng, L.M.F., Lundberg, M., Twine, N. A., Bauer, D.C., 2023. Thresholding gini variable importance with a single-trained random forest: an empirical bayes approach. Comput. Struct. Biotechnol. J. 21, 4354–4360. https://doi.org/10.1016/j.csbj.2023.08.033.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 177.

Hooshyar, D., Yang, Y., 2024. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access 12, 137472–137490. https://doi.org/10.1109/ACCESS.2024.3463948.

Huang, X., Marques-Silva, J., 2024. On the failings of shapley values for explainability. Int. J. Approx. Reason. 171, 109112. https://doi.org/10.1016/j.ijar.2023.109112.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S., 2021. Shapley residuals: quantifying the limits of the shapley value for explanations. Adv. Neural Inf. Process. Syst. 34, 26598–26608.

Létoffé, O., Huang, X., Marques-Silva, J., 2025. Towards trustable SHAP scores. Proc. AAAI Conf. Artif. Intell. 39 (17), 18198–18208. https://doi.org/10.1609/aaai.v39i17.34002.

Lones, M.A., 2024. Avoiding common machine learning pitfalls. Patterns 5 (10), 101046. https://doi.org/10.1016/j.patter.2024.101046.

Molnar, C., et al., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., Samek, W. (Eds.), Xxai - Beyond Explainable AI, 13200. Springer, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.

Nalenz, M., Rodemann, J., Augustin, T., 2024. Learning de-biased regression trees and forests from complex samples. Mach. Learn. 113, 3379–3398. https://doi.org/10.1007/s10994-023-06439-1.

Nazer, L.H., Zatarah, R., Waldrip, S., et al., 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health 2 (6), e0000278. https://doi.org/10.1371/journal.pdig.0000278.

Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S., 2024. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin. Transl. Sci. 17 (11), e70056. https://doi.org/10.1111/cts.70056.

Qin, Z., Peng, Q., Jin, C., Xu, J., Xing, S., Zhu, P., Yang, G., 2025. Geographically weighted random forest fusing multi-source environmental covariates for spatial prediction of soil heavy metals. Environmental Pollution 385, 127135. https://doi.org/10.1016/j.envpol.2025.127135.

Steiner, P.M., Kim, Y., 2016. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. J. Causal Inference 4 (2), 20160009. https://doi.org/10.1515/jci-2016-0009.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 8, 25. https://doi.org/10.1186/1471-2105-8-25.

Ugirumurera, J., Bensen, E.A., Severino, J., Sanyal, J., 2024. Addressing bias in bagging and boosting regression models. Sci. Rep. 14 (1), 18452. https://doi.org/10.1038/s41598-024-68907-5.

Wallace, M.L., Mentch, L., Wheeler, B.J., Lyons, M., Reichmann, W.M., 2023. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. BMC Med. Res. Methodol. 23 (1), 144. https://doi.org/10.1186/s12874-023-01965-x.

Wu, L., 2025. A review of the transition from shapley values and SHAP values to RGE. Statistics 1–23. https://doi.org/10.1080/02331888.2025.2487853.

Yoshiyasu Takefuji
*Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan*
*E-mail address:* takefuji@keio.jp.