Incorporation of Anisotropy through Azimuth for Machine Learning-Based Spatial Interpolation of Environmental Variables

Leonardo Locoselli Garcez[1], Vitor Vieira Vasconcelos[1] e Glauco Estácio Gonçalves[2]

[1] Universidade Federal do ABC (UFABC)

[2] Instituto de Tecnologia – Universidade Federal do Pará (UFPA)

leonardo.garcez@ufabc.edu.br

Abstract: *This study proposes a methodology for spatial interpolation using machine learning algorithms, with an emphasis on incorporating spatial anisotropy through azimuth classification. Using the Meuse dataset, we compare the performance of machine learning models while testing the hypotheses that decision tree-based algorithms are more efficient than in predicting regionalized variables and that incorporating anisotropy improves predictive results when an anisotropic component is present in the data. The results demonstrate that the inclusion of the classified azimuth variable as a predictor enhances the models' predictive capability, as evidenced by interpolated maps that capture the orientation of spatial patterns.*

1. Introduction

Spatial modeling is essential for predicting environmental variables at unsampled locations, supporting decision-making in areas such as environmental monitoring and contaminated site remediation. Traditional interpolation methods, such as Inverse Distance Weighting (IDW), and geostatistical techniques, such as kriging, have been successfully used, but they present limitations when the relationships between variables are nonlinear or when complex directional patterns are present (Li and Heap, 2014).

With the advancement of machine learning algorithms, techniques such as Random Forest, have shown promise due to their robustness, ability to capture nonlinear interactions, and tendency to avoid overfitting (Breiman, 2001). Recent studies have demonstrated that the integration of spatial attributes—such as distances—can further enhance predictions (Hengl et al., 2018). Kim et al. (2022) point out that machine learning methods (including Random Forests) proved to be more accurate in predicting housing prices compared to the spatial interpolation methods (IDW and kriging) used in the study.

Kopczewska (2022) states that Random Forest has frequently been the most accurate method in studies comparing machine learning models for spatial tasks, which has contributed to its growing popularity. According to this author, compared to geostatistical models, Random Forest requires fewer spatial assumptions and performs better with big data.

In this paper, we reproduce and expand existing approaches by applying three different decision tree-based machine learning algorithms to the Meuse dataset. An innovation of this study is the inclusion of variables that capture anisotropy through the

calculation and classification of azimuths between points, allowing the identification of directional patterns inherent to the distribution of contaminants or other natural phenomena.

According to Caers (2011), directions are important because spatial phenomena are often oriented according to a preferential direction. Anisotropy refers to the variation in the spatial structure of data as a function of direction (Goovaerts, 1997). Moreover, regionalized variables—i.e., variables with spatial dependence—are rarely truly isotropic (Leuangthong, 2008, as cited in Yamamoto, 2020).

Ordinary Kriging, a widely used geostatistical technique, already incorporates the concept of anisotropy through the use of the variogram (or covariance function), which is direction-specific (Caers, 2011). This means that when estimating a value at a given location, kriging assigns higher weights to points that lie in a direction of greater continuity (lower variability) and lower weights to points in directions of lesser continuity, even when the Euclidean distance is the same (Caers, 2011).

Thus, the hypotheses posed in this study are: (a) that decision tree-based machine learning algorithms are as efficient or more efficient in predicting regionalized variables, and (b) that incorporating anisotropy into machine learning models improves predictive outcomes in cases where an anisotropic component is associated with the data.

2. Materials and Methods

2.1 Dataset

The analysis relies on the Meuse dataset, which contains 155 irregularly spaced soil samples collected on the flood-plain of the River Meuse, the Netherlands. For each location the dataset reports planar coordinates (X, Y), concentrations of environmentally relevant metals expressed in ppm, and several ancillary covariates that may influence contaminant behaviour.

A distinctive feature of Meuse is its marked geometric anisotropy: spatial continuity is strongest along a northeast–southwest axis oriented at roughly 45 °. This directional structure makes the data a benchmark for interpolation studies and is explicitly modelled here through an azimuth-based predictor that links each sample to its nearest neighbours. The present work focuses on the zinc concentration field, using it as a test case to evaluate model performance under strongly anisotropic conditions.

2.2 Calculation of Geospatial Attributes

Distances between each pair of points were calculated using functions from the NumPy and SciPy libraries. These distances are used as predictor variables (features) to represent geographic proximity. The azimuth was calculated using the formula: azimuth = atan2(Δy, Δx), where Δy and Δx and are the coordinate differences.

Subsequently, the distance values were classified and weights were assigned to each class. For distances, higher weights were assigned to shorter distances, with weights decreasing as distance increased, based on the premise that geographically closer points are more similar. The distance classification was performed by dividing the maximum distance value by the number of intervals, generating equally sized bins. Then, each distance value was iteratively checked to determine its corresponding bin, receiving an

integer classification (from 1 to the number of intervals) that reflects the "distance class" to which it belongs.

Azimuth was computed for every pair formed by a target sample i and each of its k = 8 nearest neighbors j, using the standard convention $\theta = \text{atan2}(y_i - y_j, x_i - x_j)$ measured clockwise from geographic north. Directional variograms (Figure 1) revealed a major continuity axis at N45°, which we adopted as the reference. To encode this anisotropy, azimuths were discretised into three classes: (i) aligned: θ within ±22.5° of 45° (weight = +1); (ii) perpendicular: θ within ±22.5° of 135° (weight = –1); and (iii) other directions (weight = 0). This categorical azimuth feature was added to the predictor set, so that neighbours oriented along the principal trend exerted greater influence on the machine-learning models while still keeping the input dimensionality low.

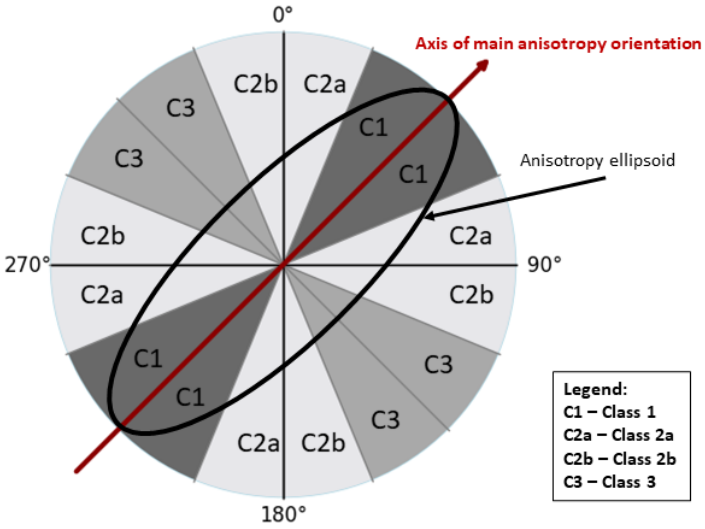

Figure 1 – Example of azimuth classification under 45° anisotropy

2.3 Algorithms and Implementation in scikit-learn

The following algorithms were implemented and compared:

(a) Random Forest (RF): introduced by Breiman (2001), this machine learning method builds multiple decision trees during training and combines their predictions to improve accuracy and control overfitting. It is effective for nonlinear data and can be used for both classification and regression tasks.

(b) ExtraTrees Regressor: a decision tree-based model similar to Random Forest, but it introduces more randomness in the selection of split points at each node. This improves generalization and reduces the risk of overfitting, making it efficient for nonlinear and high-dimensional problems (Geurts et al., 2006).

(c) XGBoost Regressor: a boosting algorithm that combines multiple weak decision trees to form a strong predictor. Introduced by Chen and Guestrin (2016), it incorporates optimizations such as regularization, parallelization, and overfitting control, making it efficient and popular in machine learning competitions.

These algorithms are widely used in machine learning tasks, all based on decision trees, but each with specific characteristics that make them suitable for different types of problems.

Each algorithm was trained using different combinations of predictor variables (features), and each combination was assigned an identification code. The initial letters refer to the algorithm—namely: RF for Random Forest, ET for ExtraTrees, and XGB for XGBoost—and the remaining letters indicate the features used, as follows: (1) XY: only coordinates; (2) XY Dist: coordinates and Euclidean distances; (3) Dist: only Euclidean distances; (4) XY Azim: coordinates and azimuths; (5) Azim: only azimuths; (6) XY Dist Azim: coordinates, distances, and azimuths; (7) Dist Azim: distances and azimuths; (8) ClassDist: classified distances; (9) ClassAzim: classified azimuths; (10) ClassDist ClassAzim: classified azimuths and classified distances; (11) XY ClassDist ClassAzim: coordinates, classified distances, and classified azimuth.

Hyperparameter optimization was performed using grid search (GridSearchCV) with 5-fold cross-validation, aiming to minimize the Mean Squared Error (MSE).

2.4 Data partitioning and validation

To ensure spatial representativeness and full reproducibility, the 155 samples were split using five-fold spatial block cross-validation. Blocks were defined via k-means clustering on standardized X/Y coordinates (random_state = 42), generating approximately 31 points per block with a minimum inter-centroid distance of ≈ 250 m. In each GridSearchCV iteration, four blocks (~80% of the data) were used for training while the remaining block was used for validation, ensuring that each block served once as test data. Because the target variable is continuous, no additional class balancing was required. The reproducibility of the entire procedure is ensured through a Python script available in the associated GitHub repository. [1].

2.5 Evaluation Metrics

To assess model performance, the following metrics were calculated: RMSE (Root Mean Squared Error) and R^2 (Coefficient of Determination). In addition, the spatial quality of each interpolated surface was evaluated both visually and quantitatively. We computed the global Structural Similarity Index (SSIM) between each model's prediction grid and the Ordinary Kriging reference surface—after normalizing the grids to [0,1] and filling missing values with the map-wise mean—to quantify overall structural resemblance. Furthermore, Moran's I was calculated on the model residuals at the original sampling locations (using an 8-nearest-neighbors weight matrix) to test for any remaining spatial autocorrelation.

3. Results

The results showed that including azimuth as a predictor variable—particularly in its classified form—contributed to improved predictive performance from the perspective of the adopted evaluation metrics.

Visual analysis of the predictions is highly important when evaluating the performance of spatial interpolation methods. The study by Li et al. (2011) demonstrates that methods with similar prediction errors can produce different spatial patterns, making visual inspection an essential step in assessing the performance of predictive methods. According to the authors, visual analysis helps detect artifacts or anomalies in the

[1] https://github.com/leogarcez75/ml_spatial_interpolation

predictions and allows verification of whether the predictions are realistic and consistent with knowledge of the study area.

From this perspective, the use of azimuths led to the appearance of block artifacts in the maps. The best results were obtained when azimuths (or classified azimuths) were used in combination with distance, rather than in isolation. In cases where the data exhibit a well-defined anisotropic component, as is the case with the Meuse dataset, azimuths appear to better capture this behavior compared to using distances alone.

As expected, the results from ExtraTrees were similar to those of Random Forest, since both are based on decision trees and use the Bagging (Bootstrap Aggregating) method—that is, training several decision trees independently and aggregating the results (using the mean in regression tasks). However, ExtraTrees introduces more randomness because the split points are selected randomly, rather than based on MSE as in Random Forest. This may explain its slightly superior performance in cross-validation, as it reduces overfitting.

The XGB model using classified distances and azimuths produced the best results in terms of evaluation metrics, even outperforming kriging. However, it produced predictions outside the range of the original data. To address this, the code was adjusted so that predictions remain within the data range: predictions above the maximum were replaced with the maximum value, and those below the minimum were replaced with the minimum value.

Figure 2 presents a comparison between kriging, IDW, and the best-performing model, taking into account both the metrics and visual assessment. Figures 3 illustrate selected interpolation maps obtained from the three algorithms, considering only a subset of the four tested feature combinations.

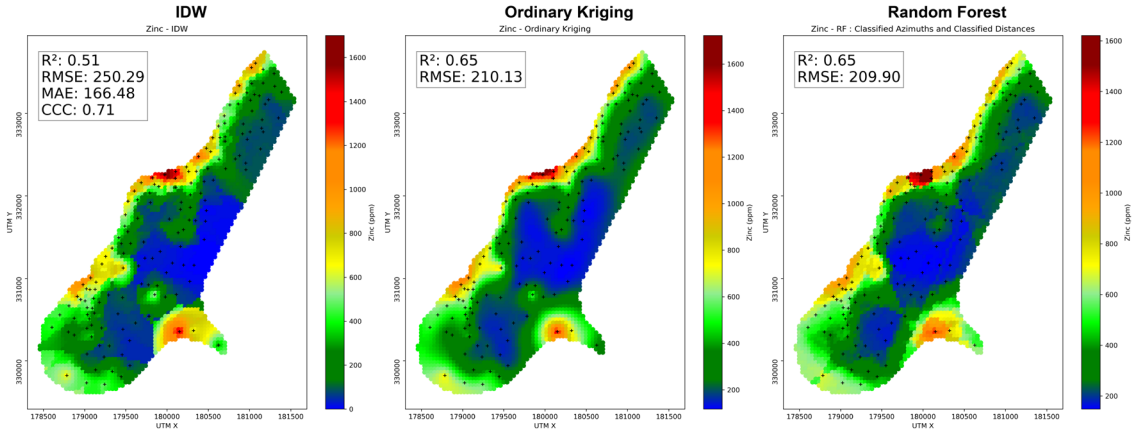

Figure 2. Comparison of Ordinary Kriging, Inverse Distance Weighting, and Random Forest using classified distances and azimuths (best result for random forest)

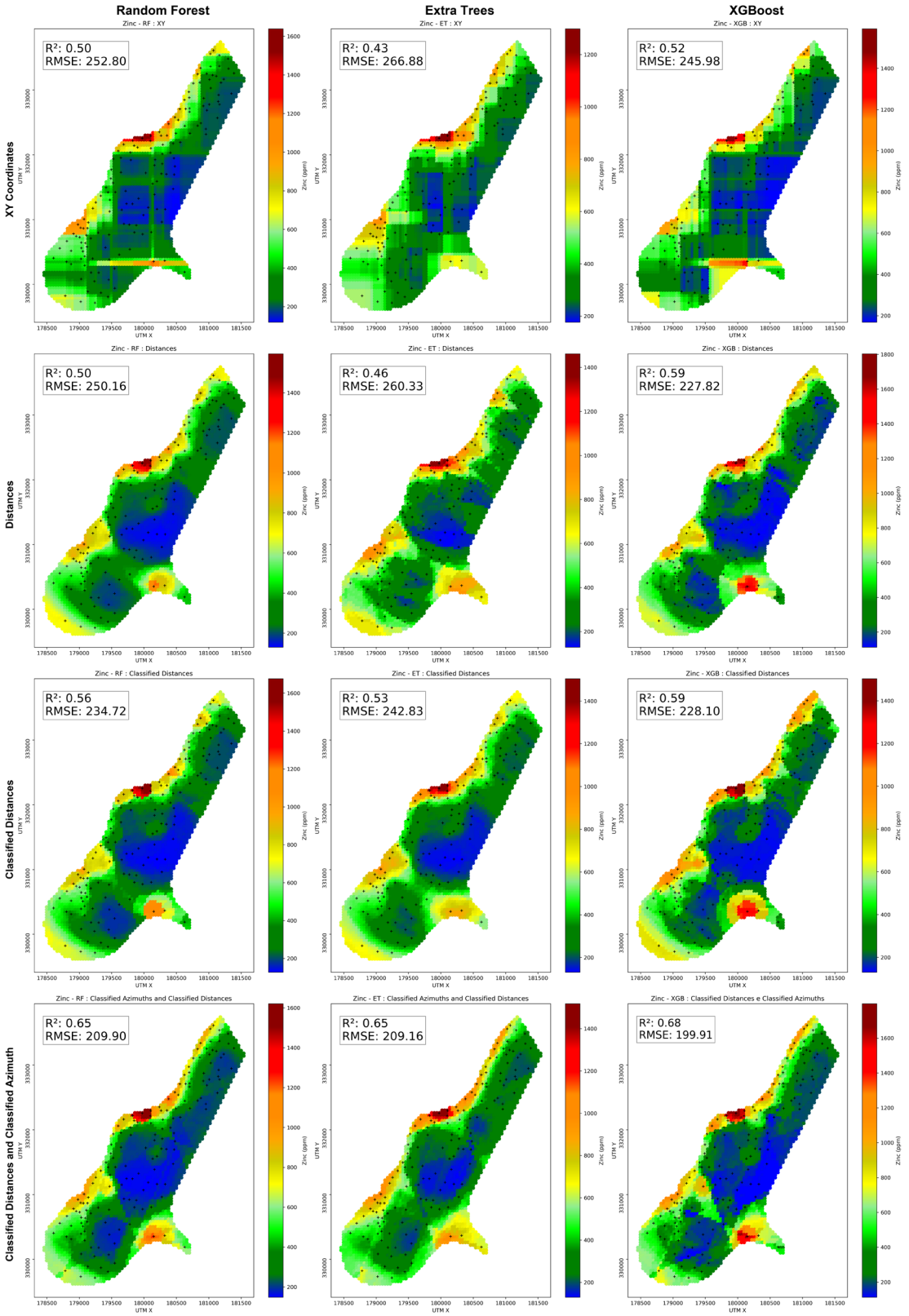

Figure 3. Interpolated zinc maps (subset of feature combinations). Columns: algorithms (RF, ET, XGB); rows: feature sets. Colors show predicted concentrations (µg/kg); R² and RMSE on each map.

Figure 4 presents a forest plot showing the R^2 results of all models at their best fit. Models shown in blue are those whose R^2 values exceeded that of kriging, which was adopted as the reference value.

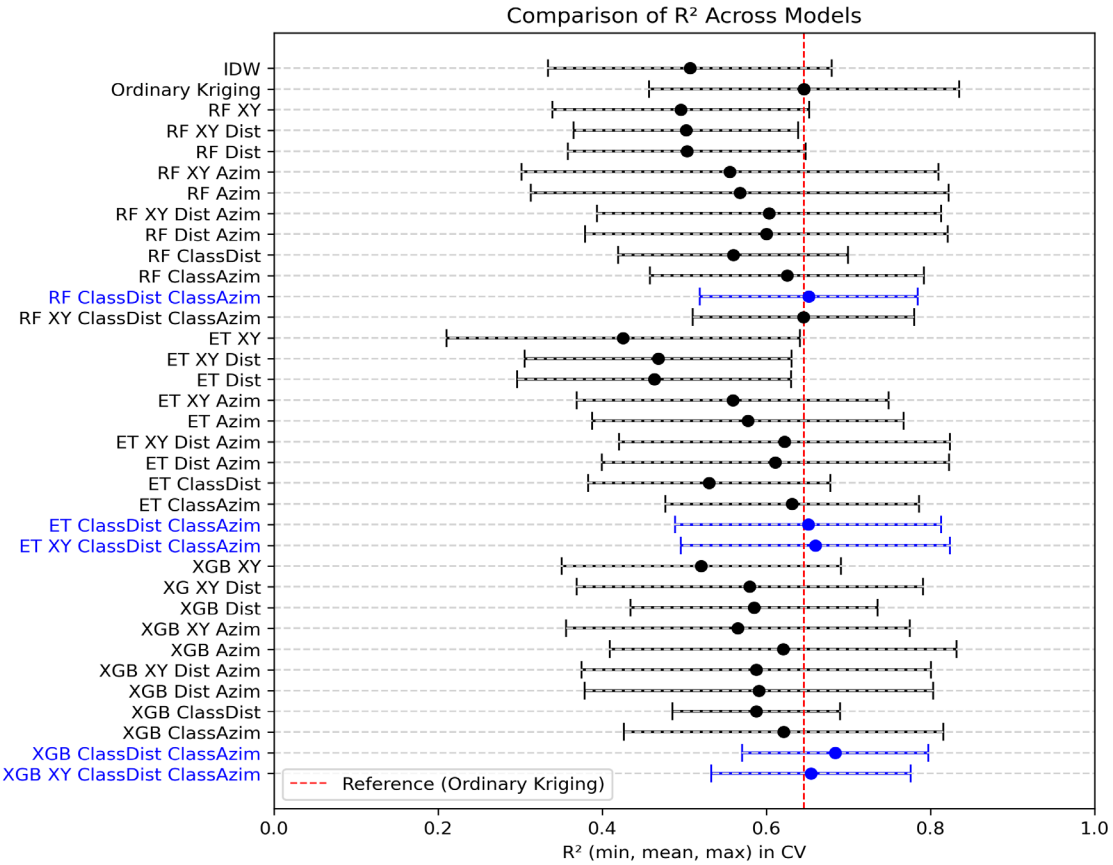

Figure 4. Model comparison across different combinations of predictor features. Blue values outperform ordinary kriging.

To better understand the spatial behavior of the predictors used by the XGBoost model, SHAP (SHapley Additive exPlanations) values were computed and aggregated by feature group. Figure 5 shows aggregated SHAP maps for distance-based and azimuth-based predictors. Positive (red) values denote locations where the corresponding feature group increases the zinc estimate, whereas negative (blue) values lower it. The azimuth map clearly reproduces the NE–SW anisotropic trend, confirming that directional information is effectively captured. Distance classes exert stronger influence in the central flood-plain, whereas azimuth dominates along the 45° structural axis, illustrating the complementary roles of both feature groups.

For the best-performing XGBoost model (using classified distances and azimuths), residuals showed no significant spatial autocorrelation (Moran's $I = -0.033$, $p = 0.235$), indicating that the model adequately captured the spatial pattern in the data.

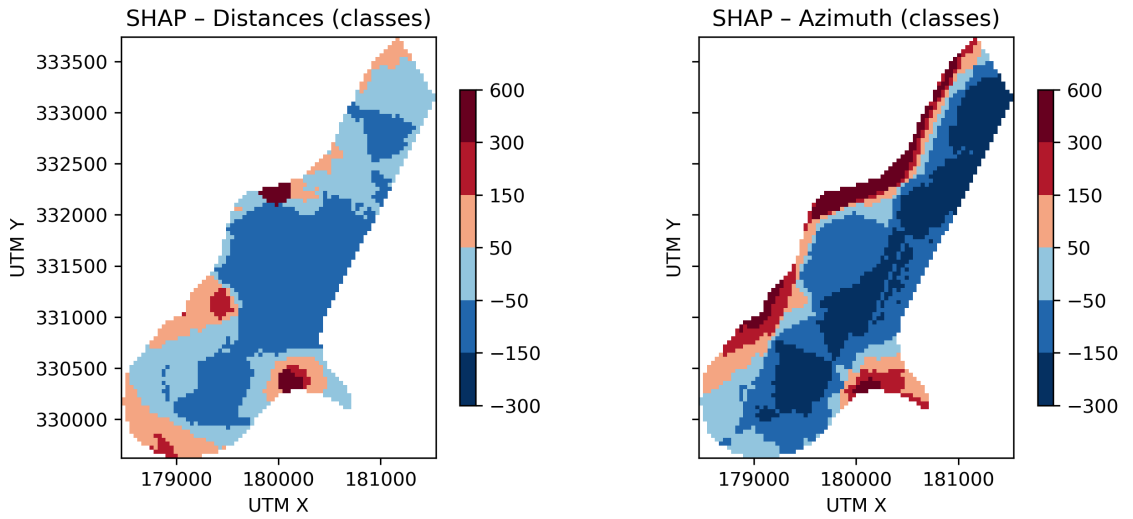

Figure 5. SHAP contribution maps for predictor groups: classified distances (left) and classified azimuths (right).

4. Discussion

Several studies have demonstrated that machine learning-based methods can outperform traditional interpolation approaches, especially in scenarios with complex variability and nonlinear relationships. Breiman (2001) showed that Random Forest combines multiple decision trees to reduce overfitting and capture nonlinear interactions. In our results, Random Forest outperformed ordinary kriging by 0.9% in R^2 and 0.1% in RMSE, validating its effectiveness in modeling complex spatial patterns.

Li and Heap (2014) highlighted that machine learning methods often surpass both IDW and kriging in nonlinear contexts. Our findings confirm this trend: XGBoost showed the best performance, with a 5.9% increase in R^2 and 4.9% reduction in RMSE over kriging. Random Forest and Extra Trees also achieved gains, though more modest. By contrast, IDW had the weakest performance, with R^2 21.5% lower and RMSE 19.1% higher than kriging. These results reinforce the value of ML-based interpolators in environmental modeling.

Hengl et al. (2018) demonstrated the value of spatial features derived from coordinates. Our SHAP maps support this, revealing that distance-based classes strongly influence predictions in the central floodplain, underscoring the benefit of including spatial features.

Nwaila et al. (2024) emphasize the practicality of ML over classical kriging. Our spatial validation supports this: even without variograms, the RF model (cDist + cAzim) produced patterns very similar to kriging (Moran's $I \approx 0$).

To further assess spatial reliability, we computed Moran's I on model residuals. Kriging showed high and significant spatial autocorrelation ($I = 0.403$, $p = 0.001$), indicating incomplete capture of spatial structure. In contrast, RF ($I = –0.037$) and XGB ($I = –0.033$) had non-significant results, suggesting better spatial coverage. Extra Trees showed significant but negative autocorrelation ($I = –0.074$, $p = 0.021$), indicating residual spatial bias.

Júnior et al. (2019) warn about artifacts when using coordinates alone. We observed similar effects: models using only X/Y produced blocky maps, which improved with the addition of classified distances and azimuths, or mild post-processing.

Block artifacts observed in some predictions stem from discretizing continuous features into distance and azimuth bins, leading to piecewise-constant surfaces. This can be smoothed using Gaussian kernels or finer binning to improve readability without losing anisotropic structure.

Yamamoto (2020) defined anisotropy as directional variation in spatial behavior, and Journel and Huijbregts (1978) emphasized its importance in improving predictions. Our study confirmed that adding directional features (classified azimuths along N45°) significantly improves results. For instance, adding azimuths to Random Forest increased R^2 from 0.560 to 0.652 (+9.2%) and reduced RMSE from 234.7 to 209.9 (–10.6%). Similar gains occurred for Extra Trees and XGBoost, confirming that encoding directionality strengthens model performance.

The literature supports this integrative approach: combining spatially informative features like azimuths enhances model fidelity and overcomes limitations of traditional methods.

Nonetheless, some challenges remain. Direction selection (N45° here) may not generalize. The method was tested only on an anisotropic dataset. XGBoost required post-hoc clipping to the observed range. Results remain sensitive to hyperparameter choices and feature discretization. Future work should explore hybrid strategies like co-kriging or RF-kriging to leverage both ML and geostatistical strengths.

5. Conclusion

This study demonstrated that tree-based machine learning models, particularly Random Forest, Extra Trees, and XGBoost, can surpass ordinary kriging in both accuracy and spatial consistency when applied to environmental data interpolation. A key driver of this improvement was the incorporation of directional features—specifically, classified azimuths—which allowed the models to better capture anisotropic patterns typical of real-world geospatial processes.

While the models achieved higher performance metrics overall, their use also introduced sharper transitions in the interpolated surfaces, reflecting the influence of discretized spatial features. Nevertheless, the ability to model nonlinear relationships and reduce spatial autocorrelation in the residuals highlights the potential of ML-based approaches as robust alternatives to traditional geostatistical methods.

Future research could explore broader applications across different spatial contexts, test hybrid techniques (e.g., RF-kriging), and refine feature engineering strategies to balance predictive accuracy with cartographic smoothness.

6 Acknowledgments

This work was partially supported by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) – Funding Code 001.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Caers, J. (2011) Modeling Uncertainty in the Earth Sciences, John Wiley & Sons, Ltd., Chichester, UK.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42.

Goovaerts, P. (1997) Geostatistics for Natural Resource Evaluation, Oxford University Press, New York.

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. Ecological Modelling, 394, 1–12.

Journel, A. G., & Huijbregts, C. J. (1978). Mining Geostatistics. Academic Press.

Anonymous. Information omitted for anonymous submission. (2019)

Kim, J.; Lee, Y.; Lee, M.-H.; Hong, S.-Y. (2022) "A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices", In: Sustainability, 14, 9056.

Kopczewska, K. (2022) "Spatial machine learning: new opportunities for regional science". In: The Annals of Regional Science, v. 68, p. 713–755, Springer, Cham.

Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software, 53, 173–189.

Li, J., Heap, A.D., Potter, A., Daniell, J.J. (2011) "Application of machine learning methods to spatial interpolation of environmental variables", Environmental Modelling & Software, 26, 1647-1659

Nwaila, K.; de Schutte, J.; van der Westhuizen, W. (2024) "Spatial Interpolation Using Machine Learning: From Patterns and Regularities to Block Models". In: Natural Resources Research, 33, 105–132, Springer, Cham.

Yamamoto, J. K. (2020) Estatística, análise e interpolação de dados geoespaciais, Gráfica Paulos, São Paulo.