Contribution Analysis of Climate and Soil Feature Groups in Dominant Crop Classification Using Random Forest Algorithm

Fadhlan Ihsan Lubis, Muhammad Iqbal

Abstract

Crop recommendation systems commonly treat soil nutrients as the principal determinant of land suitability. This study tests that assumption and also asks how far the attribution methods used to support such claims can be trusted. The data cover 50,000 grid cells at 0.05 degree resolution, described by 16 features grouped into soil, climate, and topography, and labelled with the dominant crop across ten balanced classes. Group ablation, grouped permutation, and individual permutation were applied under four validation schemes spanning test-to-train distances of 5.6 to 209.9 km, repeated across three seeds, giving 270 Random Forest fits. Climate outweighed soil by a mean ratio of 3.72, range 3.40 to 4.23, and this held under every scheme. Conclusions about soil did not hold: its unique contribution was 0.0002 under random validation but 0.0293 under 5 degree spatial blocks (Wilcoxon $p = 1.2 \times 10^{-4}$; Friedman $p = 1.8 \times 10^{-4}$). Gini importance moved by only 0.0057 and missed the pattern.

Keywords: Feature Attribution, Spatial Autocorrelation, Crop Classification, Random Forest, Spatial Cross-Validation

Fadhlan Ihsan Lubis[1]
[1]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: fihsanlubis@gmail.com [1]

Muhammad Iqbal[2]
[2]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadiqbal@dosen.pancabudi.ac.id [2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Knowing which crops are grown where is a basic input to food planning and to monitoring progress towards the Sustainable Development Goals. The data available for such work has expanded sharply over the past five years. CROPGRIDS maps the cropping patterns of 173 crops at 0.05 degree resolution [1]. SoilGrids 2.0 provides global soil properties with quantified uncertainty [2], a global database supplies plant-available phosphorus [3], and WorldClim 2.1 supplies climate variables [4]. Random Forest [5] is widely applied to these data, and it remains a reference method for geospatial classification [6].

Crop recommendation systems usually treat nitrogen, phosphorus, potassium, and pH as their primary inputs. At field scale this is reasonable, since climate is effectively fixed and soil is the only factor a farmer can manage. At global scale the question changes, because the comparison runs between locations whose climates differ widely. Testing the assumption means measuring feature importance, and two obstacles appear at exactly that point.

The first is multicollinearity. Strobl et al. showed that the built-in importance measures of Random Forest favour correlated predictors through two mechanisms: a preference for correlated predictors while trees are grown, and an advantage conferred by the unconditional permutation scheme [7]. Nicodemus et al. confirmed this in a broader simulation study [8].

The second is spatial autocorrelation, the tendency of nearby locations to hold similar values [9]. Ploton et al. found that mapping models which look accurate may have near-zero predictive power once they are validated spatially [10]. Kattenborn et al. reported performance inflation of up to 28 percent in deep learning models [11], and Meyer and Pebesma argued that random cross-validation on global maps gives biased assessments [12].

This study measures the contribution of the climate and soil feature groups with three attribution methods, each applied under four validation schemes. What is new here is the question of whether the ranking of importance between feature groups itself shifts when the validation scheme changes. Existing work has not examined that. Three objectives follow: to quantify the relative weight of the two groups and how stable it is, to test whether conclusions depend on the validation scheme, and to judge how far built-in attribution measures can be trusted on geospatial data.

Literature Review

The debate over validation schemes for spatial data is unsettled. Wadoux et al. objected that spatial cross-validation can produce assessments that are too pessimistic [13]. Several middle paths have since been proposed. These include fold construction that adapts to the intended prediction target, namely NNDM [14] and its k-fold variant [15], restriction of the prediction area through the area of applicability [16], and assessment of how well global models transfer [17]. Nowosad et al. treat the matter as one of the unresolved challenges of spatial machine learning [18].

In the attribution literature, Meyer et al. showed that conventional variable selection tends to favour predictors that reproduce the training data over those that predict into new space, and proposed forward feature selection guided by spatial validation [19]. This study moves one step along the same line. Instead of selecting features, it asks whether the ranking between groups shifts. Both bodies of work trace back to the same problem, dependence between observations, yet they have developed apart.

Research Methodology

The data comprise 50,000 grid cells of 0.05 degree distributed globally, with no missing values and a perfectly balanced class distribution of 5,000 cells for each of ten crop classes. The random-guess baseline is therefore 0.100. The 16 features were grouped by measurement origin, and the grouping was fixed before any experiment ran.

Table 1. Feature grouping

Group	n	Features	Source
Soil	8	N, P, exchangeable bases, pH, organic carbon, clay, sand, CEC	[2], [3]
Climate	7	Mean temperature, coldest-month temperature, warmest-month temperature, humidity, precipitation, rainfall seasonality, solar radiation	[4]
Topography	1	Elevation	[4]

Two properties of the data shaped the choice of method. Multicollinearity is strong, with the highest correlation reaching 0.951 between mean temperature and coldest-month temperature. Spatial autocorrelation also differs between groups. Mean Moran's I is 0.944 for climate, 0.824 for topography, and 0.765 for soil, measured on a random sample of 5,000 cells using eight nearest neighbours.

Table 2. The four validation schemes compared

Scheme	Fold construction	Blocks	Test-to-train distance
Random	StratifiedKFold with shuffling	–	5.6 km
2 degree blocks	Whole blocks kept within one fold	1,747	54.4 km
5 degree blocks	Whole blocks kept within one fold	479	108.7 km
10 degree blocks	Whole blocks kept within one fold	171	209.9 km

Test-to-train distance is the median haversine distance from each test cell to its nearest training cell. Blocks were spread across five folds by a greedy rule: blocks were sorted from largest to smallest, and each was assigned to whichever fold held the fewest cells at that moment. Block order was shuffled first so that the partition varied across seeds.

Three attribution methods were used. Group ablation trains three models on different feature sets, namely soil only, climate only, and all features. The difference between the full model and the model without a given group gives that group's unique contribution, while the accuracy of a single-group model gives its total information content. Grouped permutation shuffles all columns of a group jointly in the test set, which preserves correlation inside the group while severing the link to the target. Individual permutation and Gini importance are reported as comparators rather than as evidence, since both are vulnerable to multicollinearity [7], [8]. The first two methods are immune to it, because features that substitute for one another are always treated together.

The model was a Random Forest with 100 trees, the Gini criterion, and unrestricted depth, identical across every feature set and scheme. Hyperparameters were left untuned on purpose, so that comparisons between schemes stay on equal footing. Each experiment was repeated across three random seeds, which gives 15 estimates per experimental cell and 270 fits in total. Differences between schemes were tested with the paired Wilcoxon signed-rank test and the Friedman test at a significance level of 0.05. The full run took 25.5 minutes on a dual-core machine using Python 3 and scikit-learn.

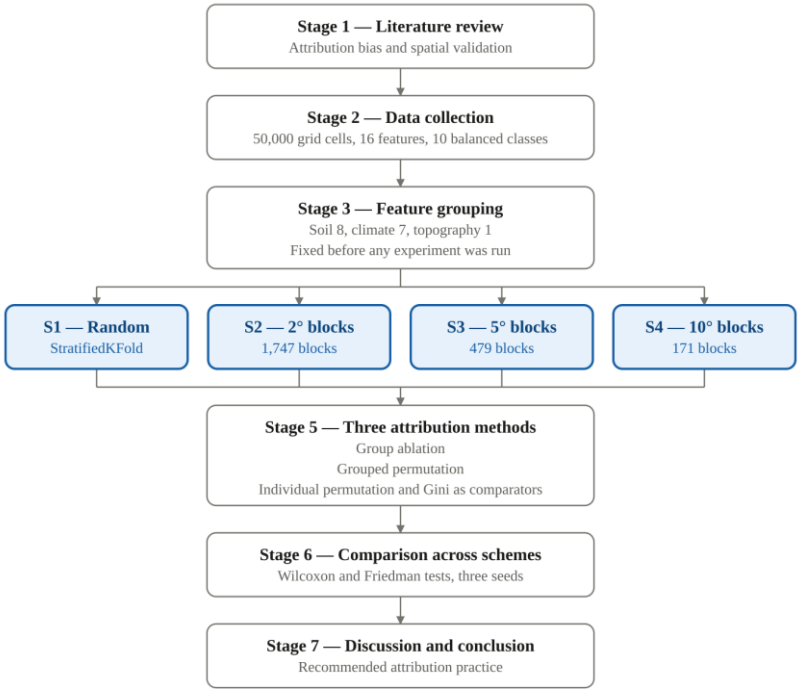

Figure 1. Research workflow

Results

Table 3 reports the accuracy of models trained on different feature sets, as the mean and standard deviation of 15 estimates.

Table 3. Accuracy by feature set and validation scheme (mean ± SD, n = 15)

Feature set	Random	2 deg. blocks	5 deg. blocks	10 deg. blocks
Soil only	0.790 ± 0.005	0.610 ± 0.025	0.505 ± 0.043	0.388 ± 0.055
Climate only	0.915 ± 0.003	0.728 ± 0.024	0.630 ± 0.028	0.502 ± 0.057
All features	0.915 ± 0.003	0.753 ± 0.029	0.660 ± 0.034	0.521 ± 0.080

Both groups carry substantial information. A model using only the eight soil features reaches 0.790 under random validation, close to eight times the baseline, and still attains 0.388 under 10 degree blocks. The climate model is higher under all four schemes despite using one feature fewer, with a steady margin of between 11.4 and 12.5 points.

Table 4. Unique contribution of each feature group

Group	Random	2 deg. blocks	5 deg. blocks	10 deg. blocks
Climate	+0.1252 ± 0.0035	+0.1430 ± 0.0170	+0.1545 ± 0.0428	+0.1332 ± 0.0454
Soil	+0.0002 ± 0.0016	+0.0247 ± 0.0103	+0.0293 ± 0.0159	+0.0194 ± 0.0376

Under random validation, adding the eight soil features to the climate model raises accuracy by 0.0002, which is smaller than its own standard deviation. All soil information is already represented by the climate features. This is where the familiar conclusion comes from, that soil features are of no use. Under spatial block validation the conclusion does not survive. The unique contribution of soil features becomes 0.0247 under 2 degree blocks and 0.0293 under 5 degree blocks. A paired Wilcoxon test between random validation and 5 degree blocks

gives W = 1.0 with p = 1.2×10^{-4}, and a Friedman test across all four schemes gives chi-squared = 19.88 with p = 1.8×10^{-4}.

The explanation that fits the data best is the difference in spatial autocorrelation between the two groups. Climate features are far more spatially clustered, with mean Moran's I of 0.944 against 0.765. Under random validation every test cell neighbours a training cell, so the climate features alone are enough to identify where the test cell sits, and soil features have no room left to contribute. Under spatial block validation that identification is no longer available, and the model has to fall back on genuine biophysical relationships.

The relationship is not monotonic. Under 10 degree blocks the unique contribution of soil features falls back to 0.0194, and its standard deviation inflates to 0.0376. What emerges is an inverted-U curve peaking near 5 degree blocks. The Pearson correlation between the logarithm of test-to-train distance and the unique contribution of soil features is r = 0.831, but with only four observations the associated p-value is 0.169 and is not significant. At 10 degree blocks the accuracy of the full model drops to 0.521 and the unique contribution of climate declines as well, so tightening the validation scheme without limit does not make attribution more accurate.

Table 5. Grouped permutation importance and Gini importance

Measure and group	Random	2 deg.	5 deg.	10 deg.	Range
Permutation, climate	0.7749	0.6061	0.5086	0.3773	0.3976
Permutation, soil	0.2140	0.1782	0.1402	0.0892	0.1248
Ratio climate : soil	3.62	3.40	3.63	4.23	0.83
Gini, climate	0.6279	0.6309	0.6332	0.6334	0.0055
Gini, soil	0.3008	0.2977	0.2960	0.2951	0.0057

The ratio of climate to soil permutation importance runs from 3.40 to 4.23, with a mean of 3.72. The model's reliance on climate features more than halves across the four schemes, while its ratio to soil features shifts by less than one unit. The dominance of climate is therefore robust to the choice of validation scheme.

Gini importance for the soil group moves by only 0.0057. Grouped permutation for the same group moves by 0.1248, twenty-two times as much. This is what one would expect, since Gini is computed from tree structure on the training data and never touches the test set. A researcher reporting Gini alone would therefore not discover the finding in Table 4. The direction of its bias differs too. Under random validation Gini assigns the soil group 0.3008 while permutation assigns 0.2140, which matches the finding of Strobl et al. that Gini favours continuous, mutually correlated features [7].

The effect of multicollinearity on per-feature attribution is measurable as well. Mean annual temperature receives an individual permutation importance of only 0.0129 under random validation, which places it almost on a par with soil features. The reason is its correlation of 0.951 with coldest-month temperature. Summing the individual permutation importances of all soil features yields 0.0465, whereas their grouped permutation importance is 0.2140, a factor of 4.6 larger.

A sensitivity check with 500 trees on two schemes gave a unique soil contribution of 0.0013 under random validation and 0.0340 under 5 degree blocks. Both the direction and the magnitude of the finding survive a fivefold increase in forest size. Four limitations should be stated. All measures are predictive rather than causal. The source datasets differ in resolution, so part of the weakness of soil features may come from there. The results have not been tested on algorithms other than Random Forest. Finally, the relationship between distance and unique contribution rests on four points only.

Conclusion

Climate features outweigh soil features by a mean ratio of 3.72, with a range of 3.40 to 4.23, and this holds under every validation scheme tested. The common assumption that soil

nutrients are the principal determinant of crop suitability is therefore not supported at global scale. It remains reasonable at field scale, where climate is effectively fixed.

Conclusions about soil features are less stable. Their unique contribution is 0.0002 under random validation but 0.0293 under 5 degree spatial blocks, a difference significant under both the paired Wilcoxon test and the Friedman test. The relationship is not monotonic and falls back at 10 degree blocks, so tightening the validation scheme without limit is not a remedy. Gini importance captures none of this, and individual permutation understates the soil group by a factor of 4.6 because of multicollinearity.

Three practical recommendations follow. For questions about feature groups, use group ablation and grouped permutation rather than built-in importance measures. Report attribution results under at least two validation schemes, since disagreement between them is a finding rather than a nuisance. Report the test-to-train distance and not only the name of the scheme, so that results stay comparable between studies. Further work should add intermediate block sizes to map the inverted-U curve, apply SHAP [20] to examine the direction of each feature's influence, test spatially aware Random Forest variants [21], [22], bring in local training-data density measures [23], and compare against other tree-based ensembles used on tabular data [24].

References

[1] F. H. M. Tang, T. H. Nguyen, G. Conchedda, L. Casse, F. N. Tubiello, and F. Maggi, "CROPGRIDS: A global geo-referenced dataset of 173 crops," Sci. Data, vol. 11, no. 1, art. 413, 2024.

[2] L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter, "SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty," SOIL, vol. 7, no. 1, pp. 217–240, 2021.

[3] R. W. McDowell, A. Noble, P. Pletnyakov, and P. M. Haygarth, "A global database of soil plant available phosphorus," Sci. Data, vol. 10, no. 1, 2023.

[4] S. E. Fick and R. J. Hijmans, "WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas," Int. J. Climatol., vol. 37, no. 12, pp. 4302–4315, 2017.

[5] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[6] I. Colkesen and M. Y. Ozturk, "A comparative evaluation of state-of-the-art ensemble learning algorithms for land cover classification using WorldView-2, Sentinel-2 and ROSIS imagery," Arab. J. Geosci., vol. 15, no. 10, art. 942, 2022.

[7] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," BMC Bioinformatics, vol. 9, art. 307, 2008.

[8] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," BMC Bioinformatics, vol. 11, art. 110, 2010.

[9] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," Econ. Geogr., vol. 46, pp. 234–240, 1970.

[10] P. Ploton et al., "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," Nat. Commun., vol. 11, no. 1, 2020.

[11] T. Kattenborn, F. Schiefer, J. Frey, H. Feilhauer, M. D. Mahecha, and C. F. Dormann, "Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks," ISPRS Open J. Photogramm. Remote Sens., vol. 5, art. 100018, 2022.

[12] H. Meyer and E. Pebesma, "Machine learning-based global maps of ecological variables and the challenge of assessing them," Nat. Commun., vol. 13, art. 2208, 2022.

[13] A. M. J.-C. Wadoux, G. B. M. Heuvelink, S. de Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," Ecol. Modell., vol. 457, art. 109692, 2021.

[14] C. Milà, J. Mateu, E. Pebesma, and H. Meyer, "Nearest neighbour distance matching leave-one-out cross-validation for map validation," Methods Ecol. Evol., vol. 13, no. 6, pp. 1304–1316, 2022.

[15] J. Linnenbrink, C. Milà, M. Ludwig, and H. Meyer, "kNNDM CV: k-fold nearest-neighbour distance matching cross-validation for map accuracy estimation," Geosci. Model Dev., vol. 17, no. 15, pp. 5897–5912, 2024.

[16] H. Meyer and E. Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," Methods Ecol. Evol., vol. 12, no. 9, pp. 1620–1633, 2021.

[17] M. Ludwig, A. Moreno-Martínez, N. Hölzel, E. Pebesma, and H. Meyer, "Assessing and improving the transferability of current global spatial prediction models," Glob. Ecol. Biogeogr., vol. 32, no. 3, pp. 356–368, 2023.

[18] J. Nowosad et al., "Navigating challenges in spatial machine learning: Validation, uncertainty, algorithms, and reproducibility," Erdkunde, 2026.

[19] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications: Moving from data reproduction to spatial prediction," Ecol. Modell., vol. 411, art. 108815, 2019.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] S. Georganos et al., "Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," Geocarto Int., vol. 36, no. 2, pp. 121–136, 2021.

[22] S. Georganos and S. Kalogirou, "A forest of forests: A spatially weighted and computationally efficient formulation of geographical random forests," ISPRS Int. J. Geo-Inf., vol. 11, no. 9, art. 471, 2022.

[23] H. Meyer, C. Milà, M. Ludwig, J. Linnenbrink, and F. Schumacher, "Estimation of local training data point densities to support the assessment of spatial prediction uncertainty," Geosci. Model Dev., vol. 18, art. 10185, 2025.

[24] "Analisis perbandingan model machine learning tree-based dan non-tree-based untuk tugas klasifikasi," J. Teknol. Inf. dan Ilmu Komput., vol. 12, no. 4, pp. 875–884, 2025.