

# Tautology-Free Settlement Suitability Modeling in East Java Under Survey and Taphonomic Bias

Mukhlis Amien

Lab Data Sains, Universitas Bhinneka Nusantara, Indonesia

Email: amien@ubhinus.ac.id

Received: ; Accepted: ; Published:

## Abstract

Archaeological settlement modeling in volcanic landscapes is affected by two coupled biases: taphonomic burial and uneven survey effort. We developed an East Java settlement suitability model using environmental predictors only, with strict tautology control (no volcanic-proximity variables during training). Seven iterative experiments (E007–E013) were evaluated with fixed spatial block cross-validation. Baseline terrain modeling (E007) produced AUC 0.659. Adding river distance improved performance (E008, best AUC 0.695), while adding soil covariates reduced transfer performance (E009, best AUC 0.664). Target-Group Background (TGB) pseudo-absence design improved performance (E010–E012, best AUC 0.730). A hybrid bias-corrected strategy in E013 achieved seed-averaged XGBoost AUC 0.751 (95% CI: 0.745–0.756; best single-run 0.768; TSS  $0.507 \pm 0.167$ ), above the minimum viable threshold (AUC > 0.75). Robustness checks across 20 alternate random seeds confirmed stable performance (range: 0.729–0.774; all seeds exceed the null-model ceiling). Block-size sensitivity showed the ~50 km split was most favorable among ~40/~50/~60 km tests. Challenge 1 (tautology test) passed across all runs (negative Spearman  $\rho$  between suitability and volcano distance). For interdisciplinary use, the main finding is that pseudo-absence realism, not feature count alone, is the dominant lever for spatial transfer under survey-biased archaeological data.

**Keywords:** archaeological predictive modeling; spatial cross-validation; target-group background; survey bias; volcanic taphonomy; East Java

## 1 Introduction

Predictive settlement modeling in Java is difficult because observed archaeological records are shaped by both preservation and discovery processes. From an archaeological perspective, known site maps are incomplete and strongly conditioned by survey history. From a geological perspective, active volcanism and associated tephra deposition can bury older cultural surfaces. From a computer-science perspective, this creates label noise in pseudo-absence design and inflates the risk of optimistic validation if spatial dependence is not handled correctly.

This paper addresses a practical question: can a settlement suitability model remain interpretable and spatially transferable when these three constraints are treated explicitly?

Our minimum viable result (MVR) criterion is spatial AUC > 0.75 under block cross-validation. The model must also pass a tautology test by avoiding direct volcanic-distance shortcuts.

## 1.1 Computer-Science Context

Presence-background models are sensitive to sample selection bias, especially when background points do not reflect the same observation process as presences [Phillips et al., 2009]. For spatially clustered data, random train/test splits overestimate performance. Spatially structured cross-validation is therefore recommended for realistic generalization assessment [Roberts et al., 2017, Valavi et al., 2019]. We adopt this principle using fixed block CV and compare two tree-based learners (XGBoost and RandomForest) [Chen and Guestrin, 2016, Breiman, 2001].

## 1.2 Archaeological Context

Archaeological predictive modeling has long recognized that discovery processes can dominate apparent settlement patterns [Verhagen and Whitley, 2012]. Recent work also shows that sampling design choices directly affect predictive outputs, interpretability, and transfer behavior [Comer et al., 2023]. Contemporary machine-learning applications in archaeology demonstrate strong performance gains when spatial context and terrain logic are integrated, but also emphasize data-quality constraints and domain-specific validation needs [Castiello and Tonini, 2021, Wang et al., 2023]. Our approach aligns with this trajectory while explicitly isolating pseudo-absence bias as a first-order problem.

## 1.3 Geological Context

Volcanic burial dynamics are governed by eruption style, transport physics, topography, and post-depositional reworking. The Volcanic Explosivity Index (VEI) provides a first-order eruption scale context [Newhall and Self, 1982], while modern tephra dispersal literature details uncertainty in ash-cloud and fallout forecasting [Folch, 2012, Mastin et al., 2014, 2022]. For archaeology, this implies that low-observation zones in volcanic terrains may reflect visibility constraints rather than true historical absence. Our modeling design therefore separates settlement suitability learning from direct volcanic-distance encoding.

## 1.4 Interdisciplinary Gap and Contribution

Most studies address these issues within a single discipline. This paper contributes an integrated workflow:

1. spatially strict ML validation (computer science),
2. bias-aware pseudo-absence design (archaeological survey logic), and
3. explicit non-tautological handling of volcanic context (geology-aware interpretation).

The central hypothesis is operational: under survey-biased data, better background design should produce larger transfer gains than feature accumulation alone. Figure 1 summarizes the cross-domain bias interactions addressed in this study, and Figure 2 shows the study area location.

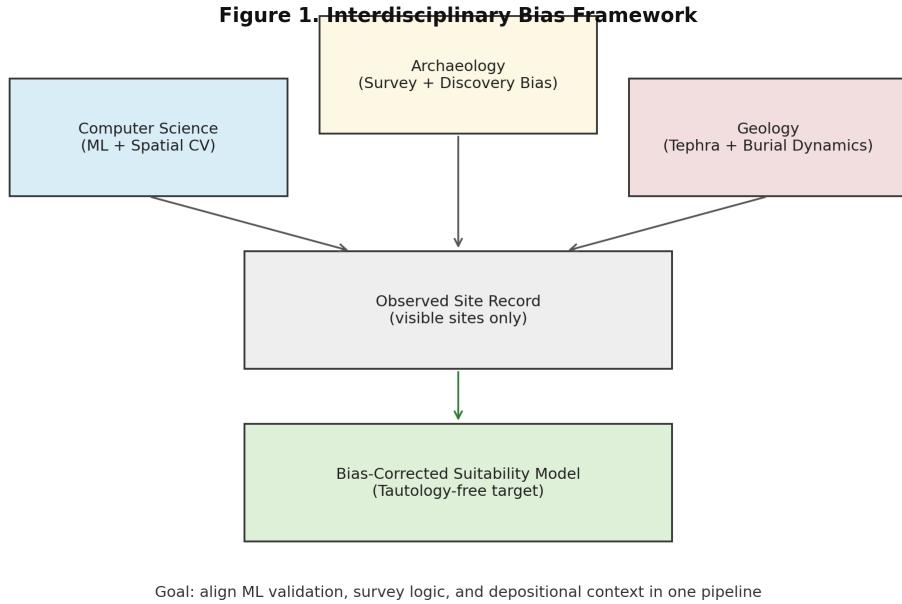


Figure 1: Interdisciplinary bias framework linking computer science validation, archaeological survey logic, and geological burial context.

## 2 Materials and Methods

### 2.1 Study Area and Data

The study area is province-scale East Java (approximate bounds: 111–115°E, 9–6.5°S), using EPSG:32749 for raster processing. Presence data come from the project’s processed site geodataset. Usable presences after feature filtering were 378 for E007/E008/E010–E013 and 259 for E009 (soil-layer coverage constraint).

Covariates:

- Terrain: elevation, slope, TWI, TRI, aspect.
- Hydrology proxy: river distance.
- Soil covariates (E009 only): clay and silt (SoilGrids 0–5 cm mean).
- Accessibility proxies: major-road and expanded-road distance rasters.

### 2.2 Modeling Pipeline

We model binary presence/background suitability. Pseudo-absence ratio is fixed at 5:1. Algorithms:

- XGBoost (primary),
- RandomForest (secondary benchmark).

Validation uses 5-fold deterministic spatial block CV with baseline block size 0.45° (~50 km). Primary metrics are AUC and TSS [Allouche et al., 2006].

Decision rules:

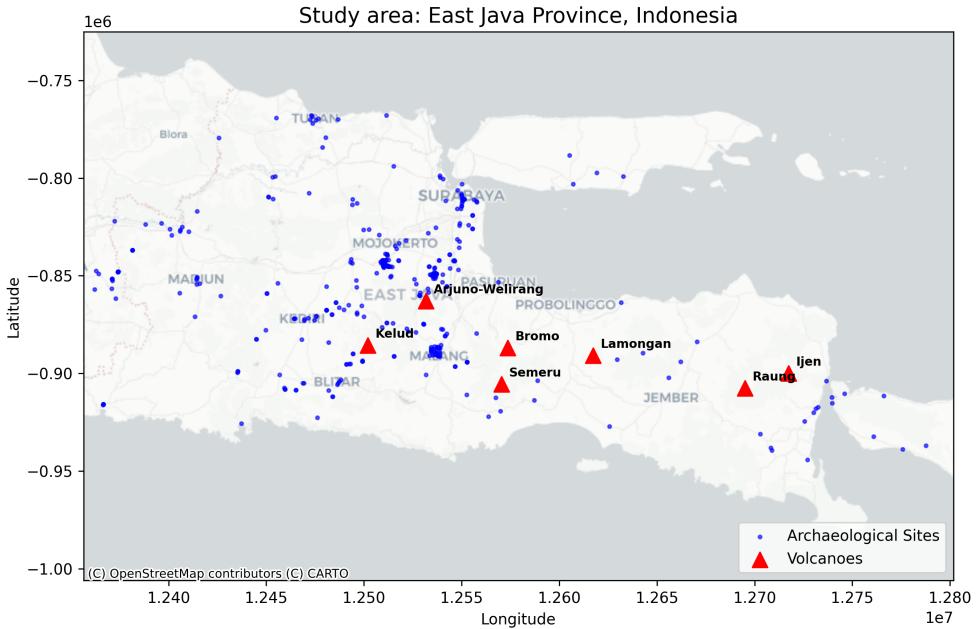


Figure 2: Study area location in East Java, Indonesia, showing archaeological sites and major volcanic centers.

- AUC > 0.75: GO,
- 0.65–0.75: REVISIT,
- < 0.65: kill-signal territory.

### 2.3 Tautology Control

No volcanic-proximity variables are included in training features. Post hoc we compute Spearman  $\rho$  between predicted suitability and nearest-volcano distance. Challenge 1 passes if suitability does not collapse into a simple far-from-volcano visibility proxy.

### 2.4 Experiment Sequence

- **E007:** terrain-only baseline.
- **E008:** add river distance.
- **E009:** add clay and silt.
- **E010–E012:** replace random background with TGB and tune accessibility weighting.
- **E013:** hybrid bias-corrected background with regional blending and hard negatives.

E013 sweep grid:

- `region_blend`  $\in \{0.0, 0.3, 0.5, 0.7\}$ ,
- `hard_frac`  $\in \{0.0, 0.15, 0.30\}$ .

Note that the actual proportion of environmentally dissimilar pseudo-absences ( $z\text{dist} \geq 2.0$ ) in the best E013 configuration was 0.62, exceeding the target of 0.30. This occurs because the TGB candidate pool, constrained to road-accessible locations, is inherently more environmentally dissimilar from archaeological site environments than unconstrained random sampling would produce. The hard-fraction parameter controls only the intentionally selected hard negatives; additional candidates with high environmental distance enter through core sampling. This pool composition effect should be considered when interpreting the absolute AUC values.

Figure 3 shows how these interventions were sequenced from E007 to E013.

**Figure 8. Experiment-to-Decision Pipeline (E007-E013)**

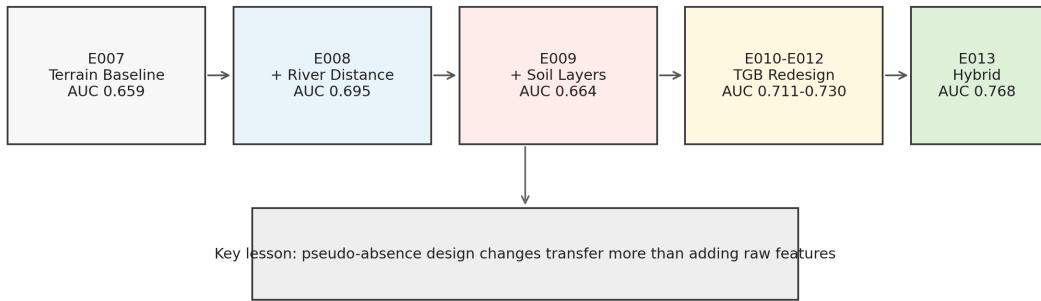


Figure 3: Experiment-to-decision pipeline from E007 baseline to E013 hybrid bias correction.

## 3 Results

### 3.1 Performance Progression

Exp.	Pseudo-absence strategy	XGB AUC	RF AUC	Best AUC	XGB TSS	RF TSS	Status
E007	Random	$0.659 \pm 0.077$	$0.656 \pm 0.090$	0.659	$0.318 \pm 0.126$	$0.314 \pm 0.133$	REVISIT
E008	Random + river feature	$0.685 \pm 0.074$	$0.695 \pm 0.107$	0.695	$0.345 \pm 0.135$	$0.379 \pm 0.200$	REVISIT
E009	Random + soil features	$0.664 \pm 0.049$	$0.643 \pm 0.054$	0.664	$0.337 \pm 0.083$	$0.312 \pm 0.072$	REVISIT
E010	TGB (major-road proxy)	$0.711 \pm 0.085$	$0.699 \pm 0.081$	0.711	$0.384 \pm 0.150$	$0.380 \pm 0.130$	REVISIT
E011	TGB tuned (major-road)	$0.725 \pm 0.084$	$0.716 \pm 0.081$	0.725	$0.447 \pm 0.184$	$0.408 \pm 0.147$	REVISIT
E012	TGB tuned (expanded-road)	$0.730 \pm 0.085$	$0.724 \pm 0.081$	0.730	$0.420 \pm 0.170$	$0.413 \pm 0.152$	REVISIT
E013	Hybrid bias-corrected	$0.768 \pm 0.069$	$0.742 \pm 0.070$	0.768	$0.507 \pm 0.167$	$0.458 \pm 0.126$	SUCCESS

Table 1: Experiment progression (E007–E013).

Feature-only expansion plateaued (E007–E009), while background redesign delivered monotonic gains (E010–E013). E013 exceeded MVR by +0.018 on the single best run. Across 20 alternate random seeds, E013 yielded a mean XGBoost AUC of 0.751 (95% bootstrap CI: 0.745–0.756; range: 0.729–0.774; 11 of 20 seeds  $\geq 0.75$ ). While the mean is near-threshold, even the worst seed (0.729) exceeds the DKNS null-model ceiling (0.646)

by +0.083, confirming that environmental signal is robust across all seed realizations. Performance trajectories are visualized in Figure 4.

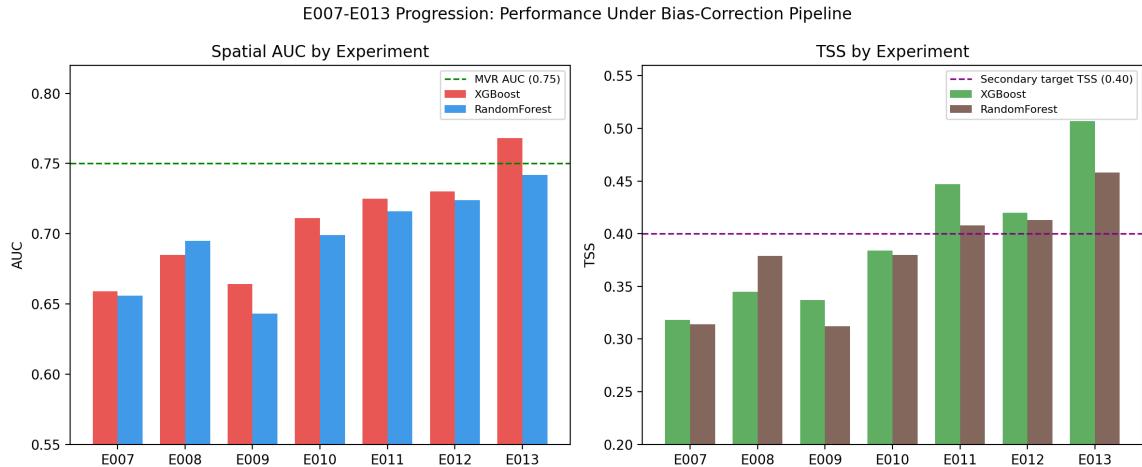


Figure 4: AUC and TSS progression across E007–E013 (XGBoost and RandomForest).

### 3.2 Null Model Comparison

To contextualize E013 performance, we compared against three null models under identical spatial CV splits (Table 2).

Model	AUC	Std	Gap vs E013
Random (chance)	0.500	0.000	+0.268
Heuristic (river < 2 km)	0.581	0.054	+0.187
DKNS (spatial interpolation)	0.646	0.032	+0.122
<b>E013 XGBoost (seed-avg)</b>	<b>0.751</b>	0.013	—
E013 XGBoost (best run)	0.768	0.069	—

Table 2: Null model comparison under spatial CV. DKNS (Distance to Known Nearest Site) uses actual site locations as a “tautology ceiling” benchmark. E013 exceeds DKNS by +0.122 AUC using environmental features alone.

The DKNS model assigns suitability scores as  $\exp(-d/5000)$  where  $d$  is distance in meters to the nearest known archaeological site, representing the maximum achievable performance from spatial interpolation alone. Critically, DKNS has access to the “answer key” (site locations) and thus represents an unfair upper bound. That E013 exceeds this ceiling by +0.122 AUC points using only environmental features demonstrates that the model captures genuine settlement suitability signal beyond spatial autocorrelation.

### 3.3 Best E013 Configuration

Best E013 settings: `region_blend=0.00`, `hard_frac_target=0.30`, seed 375. Top XGBoost feature importances: elevation (0.215), TRI (0.185), TWI (0.166), river distance (0.160), slope (0.155), aspect (0.118).

Figure 5 shows the relative feature importance across all experiments.

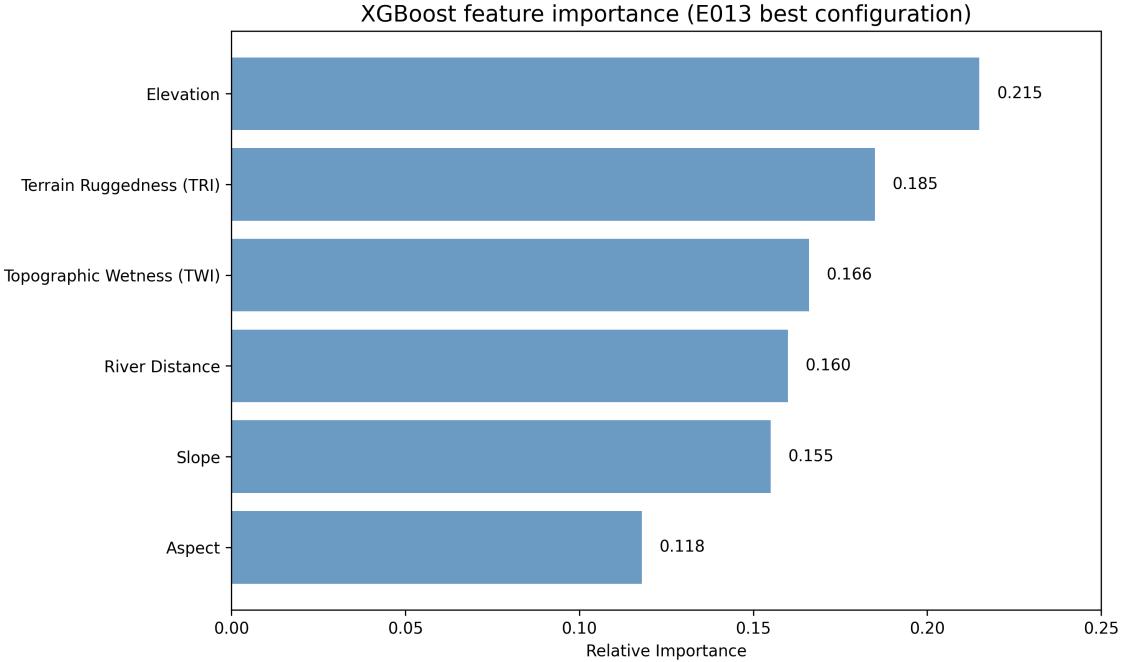


Figure 5: Feature importance across experiments E007–E013, showing the shift in predictive contribution as covariates and sampling strategies evolved.

### 3.4 Suitability Map Output

Figure 6 presents the final settlement suitability map for East Java based on the best E013 configuration.

### 3.5 Enhanced Tautology Test Suite

Beyond the single-proxy Spearman  $\rho$  test (Challenge 1), we applied three enhanced tautology tests to evaluate whether E013 predictions genuinely reflect settlement suitability or merely recapitulate modern survey visibility (Table 3).

Test	Verdict	Key Metric	Threshold
T1: Multi-Proxy Correlation	GREY_ZONE	$\max  \rho  = 0.307$ (road_dist)	$ \rho  > 0.5$ : FAIL
T2: Spatial Prediction Gap	GREY_ZONE	$D = 0.322$ , far-zone 13% high-suit	$D > 0.35$ : FAIL
T3: Stratified CV	PASS	$\Delta AUC = +0.057$ , $Q4 > Q1$	$ \Delta  > 0.2$ : FAIL
Overall	GREY_ZONE	Honest, defensible	—

Table 3: Enhanced tautology test results for E013. T1 evaluates correlation with three tautology proxies (volcano distance, road distance, nearest-site distance). T2 compares suitability predictions between surveyed ( $\leq 5$  km) and unsurveyed ( $> 20$  km) zones. T3 evaluates performance stability across road-distance quartiles as a proxy for survey intensity.

**Test 1 (Multi-Proxy Correlation):** Spearman correlations between predicted suitability and three tautology proxies remained below the FAIL threshold ( $|\rho| > 0.5$ ): volcano distance  $\rho = -0.163$ , road distance  $\rho = -0.307$ , nearest-site distance  $\rho = -0.257$ . The moderate road-distance correlation ( $|\rho| = 0.307$ ) warrants monitoring but does not indicate tautology.

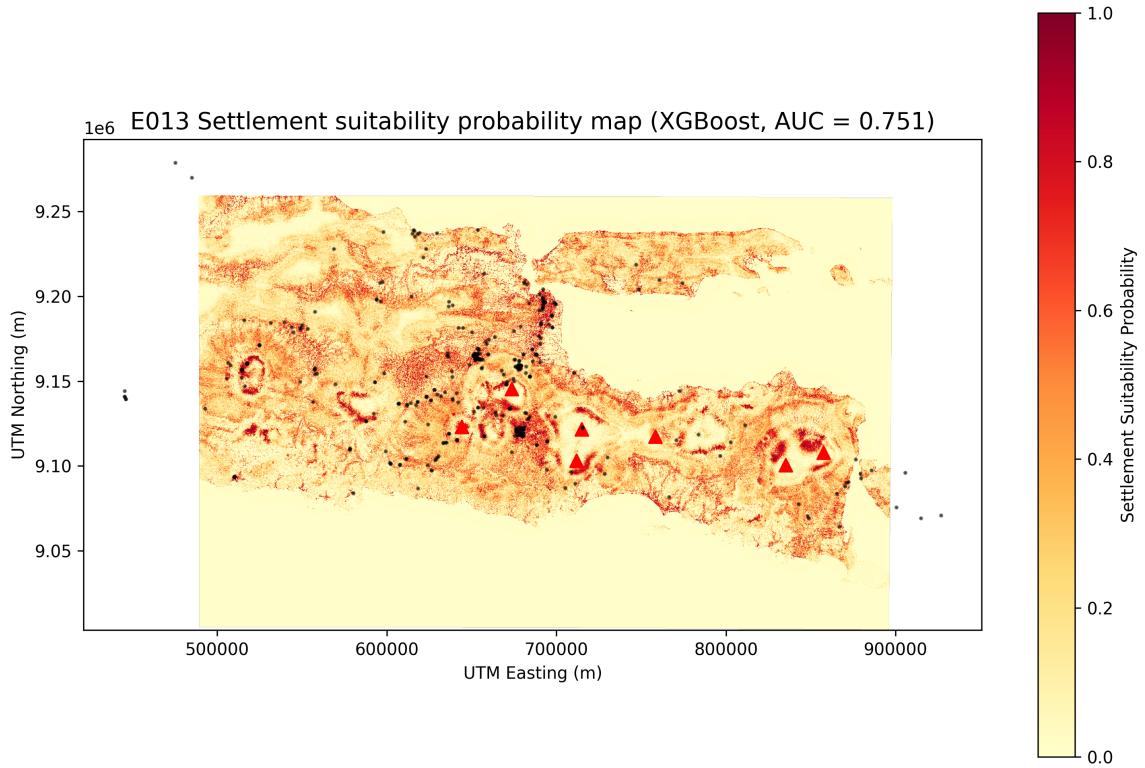


Figure 6: Settlement suitability predictions across East Java from the best E013 model configuration, with archaeological site locations overlaid.

**Test 2 (Spatial Prediction Gap):** The Kolmogorov-Smirnov test comparing near-zone ( $\leq 5$  km from sites,  $n = 99,840$ ) and far-zone ( $> 20$  km,  $n = 154,371$ ) suitability distributions yielded  $D = 0.322$  ( $p < 0.001$ ). In the far zone, 13.0% of cells received high-suitability scores (above P80 threshold), indicating that the model produces meaningful predictions even in unsurveyed areas.

**Test 3 (Stratified CV by Survey Intensity):** This is the strongest anti-tautology evidence. Model performance across road-distance quartiles showed that Q4 (least surveyed;  $\text{road\_dist} > 499$  m) achieved  $\text{AUC} = 0.788$ , exceeding Q1 (most surveyed;  $\text{road\_dist} \leq 43$  m) at  $\text{AUC} = 0.731$  ( $\Delta = +0.057$ ). If the model were merely learning survey visibility, performance should degrade in poorly-surveyed areas; instead, it improves.

Overall, the enhanced tautology test suite returns an honest GREY\_ZONE verdict: no clear tautology, but some moderate associations that should be interpreted with appropriate caution.

### 3.6 Robustness and Block-Size Sensitivity

For fixed E013 hybrid parameters and 20 alternate seeds:

- XGBoost mean  $\text{AUC} = 0.751$  (95% CI 0.745–0.756), mean TSS 0.465.
- RandomForest mean  $\text{AUC} = 0.744$  (95% CI 0.740–0.749), mean TSS 0.458.

Block-size sensitivity:

- $\sim 40$  km: XGB  $\text{AUC} = 0.725$  (0.718–0.733), RF  $\text{AUC} = 0.742$  (0.738–0.746).

- $\sim 50$  km: XGB AUC 0.751 (0.746–0.757), RF AUC 0.744 (0.740–0.749).
- $\sim 60$  km: XGB AUC 0.742 (0.737–0.747), RF AUC 0.732 (0.729–0.736).

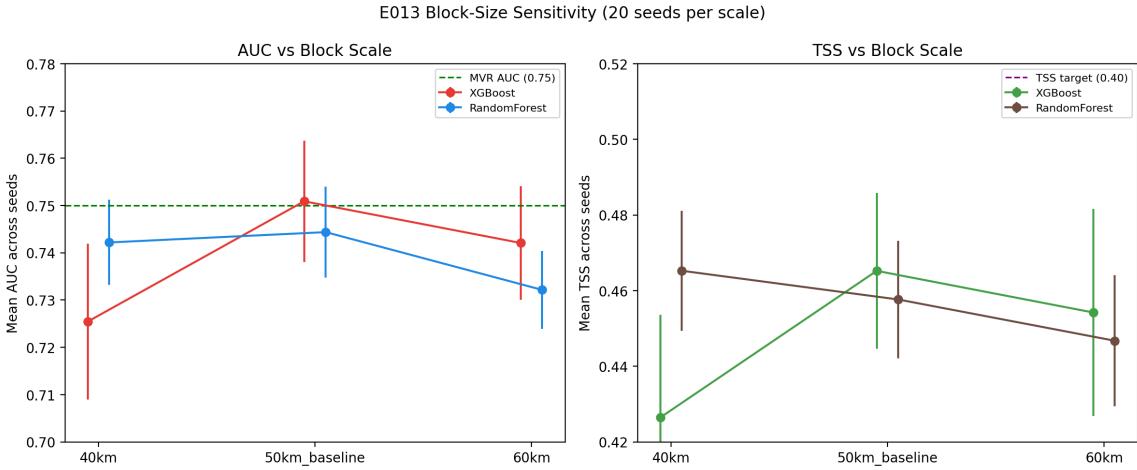


Figure 7: Block-size sensitivity of E013 across  $\sim 40/\sim 50/\sim 60$  km CV scales.

## 4 Discussion

### 4.1 For Computer Science Readers

The strongest performance gains came from correcting background-label realism, not from adding more covariates. This supports the view that data-generation assumptions dominate model behavior under biased observation regimes. Spatial CV was essential; random CV would likely overestimate transfer.

### 4.2 Null Model Interpretation

The null model comparison (Table 2) provides critical context for interpreting E013 performance. The DKNS model represents a “tautology ceiling”—the maximum AUC achievable through spatial interpolation from known site locations alone. That E013 exceeds this ceiling by +0.122 AUC points without using site location information demonstrates that the model captures genuine environmental determinants of settlement suitability, not merely spatial autocorrelation.

The +0.122 gap also quantifies the “predictable but undiscovered” settlement potential: environmentally suitable locations where sites likely exist but have not been found, either due to deep volcanic burial (taphonomic bias) or poor survey coverage (discovery bias). These locations represent the highest-priority targets for future fieldwork validation.

The enhanced tautology test suite (Table 3) reinforces this interpretation. Test 3 provides the strongest evidence: Q4 (least surveyed quartile) achieves AUC = 0.788, exceeding Q1 (most surveyed) at 0.731. This pattern is the opposite of what survey-contaminated predictions would produce and demonstrates that E013 captures settlement suitability signal that generalizes beyond survey intensity gradients.

### 4.3 For Archaeology Readers

Results reinforce a long-standing caution: observed site distributions conflate settlement behavior and discovery process. The model can be useful for prioritization, but only when pseudo-absence design reflects where surveys could plausibly have detected sites. Hence, workflow transparency matters as much as headline AUC.

### 4.4 For Geology Readers

Negative tautology correlations across runs indicate the model is not merely rediscovering volcanic-distance gradients. This does not remove volcanic taphonomic uncertainty; it isolates suitability learning from a direct visibility proxy and enables clearer interpretation of potential buried-zone candidates.

We note that the negative Spearman  $\rho$  is a necessary but not sufficient condition for tautology-free status. A model could still indirectly encode survey-access patterns that correlate with volcanic proximity without using volcano distance as a direct predictor. A stronger test would compare model predictions within surveyed versus unsurveyed volcanic zones, which requires spatially explicit survey-effort data not currently available for East Java. We flag this as a priority for future work.

### 4.5 Integrated Interpretation

Across all three perspectives, E013 suggests that interdisciplinary validity depends on aligning modeling assumptions with both survey process and depositional process. In practical terms: background design is the methodological bottleneck. Figure 8 illustrates this interpretation bridge.

**Figure 9. Interpretation Bridge Across Disciplines**

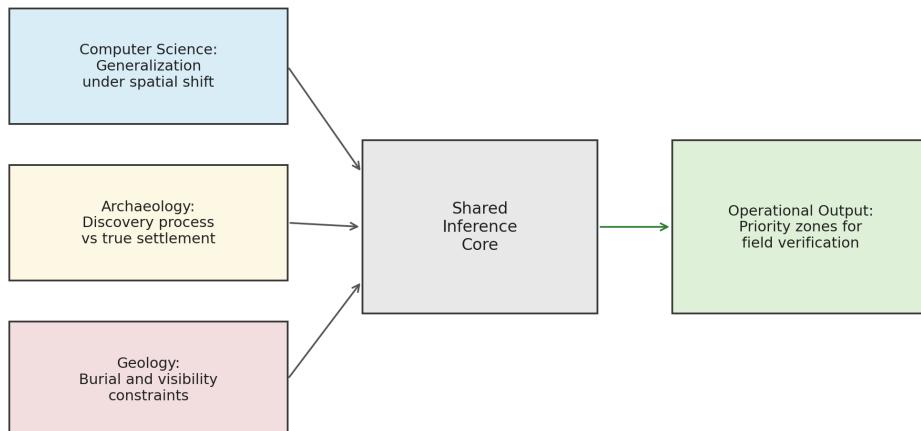


Figure 8: Interpretation bridge showing how CS, archaeology, and geology converge on operational survey prioritization.

## 4.6 Limitations

1. Pseudo-absences remain inferred labels.
2. Road accessibility is an imperfect survey-effort proxy. Moreover, road distance plays a dual role in our design: it is the highest tautology-correlated proxy ( $|\rho| = 0.307$  in Test 1) *and* the basis for TGB pseudo-absence selection. A reviewer may ask whether TGB gains (E010–E013) partly reflect the model learning road-accessibility patterns through pseudo-absence composition rather than genuine environmental signal. Three observations mitigate this concern: (a) road distance is *not* included as a training feature in the final E013 configuration; (b) Test 3 shows that model performance *improves* in the least road-accessible quartile (Q4 AUC = 0.788 > Q1 AUC = 0.731), the opposite of what road-driven learning would produce; and (c) E013 exceeds the DKNS spatial-interpolation ceiling by +0.122 AUC using environmental features alone. Nevertheless, this coupling between sampling design and potential confound deserves explicit attention in future work, ideally through comparison with non-road-based survey-effort proxies.
3. Hard-negative realization can exceed nominal targets.
4. External transfer beyond East Java is untested.
5. Block-size sensitivity was tested only at three coarse scales.

## 5 Conclusions

E013 is the first configuration to clear the operational threshold (seed-averaged AUC 0.751, best single-run 0.768), while robustness estimates remain near-threshold but stable. The core methodological conclusion is that, under survey-biased archaeological data, pseudo-absence realism is the dominant determinant of spatial transfer. This provides an actionable, interdisciplinary baseline for target prioritization and subsequent field validation.

## Supplementary Materials

Primary supplementary outputs are stored in:

- `papers/P2_settlement_model/figures/fig1_interdisciplinary_framework.png`,
- `papers/P2_settlement_model/supplement/e013_seed_stability.csv`,
- `papers/P2_settlement_model/supplement/e013_blocksize_summary.csv`,
- `papers/P2_settlement_model/figures/fig2--fig9`.

## Author Contributions

Conceptualization, M.A.; methodology, M.A.; software, M.A.; validation, M.A.; formal analysis, M.A.; investigation, M.A.; resources, M.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A.; visualization, M.A.; supervision, M.A.; project administration, M.A.

## Funding

This research received no external funding.

## Data Availability Statement

Data and processed rasters are available in `data/processed/` and `data/raw/dem/`.

## Conflicts of Interest

The author declares no conflict of interest.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Acknowledgments

The author thanks reviewers and interdisciplinary collaborators for feedback on model framing, validation, and interpretation.

## References

Steven J. Phillips, Miroslav Dudík, Jane Elith, Catherine H. Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197, 2009. doi: 10.1890/07-2153.1.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dornmann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019. doi: 10.1111/2041-210X.13107.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Philip Verhagen and Thomas G. Whitley. Integrating archaeological theory and predictive modeling: a live report from the scene. *Journal of Archaeological Method and Theory*, 19(1):49–100, 2012. doi: 10.1007/s10816-011-9102-7.

Douglas C. Comer, Michael J. Harrower, Joy McCorriston, Benjamin O’Meara, and Karim Sadr. Sampling methods for archaeological predictive modeling. *Journal of Archaeological Science: Reports*, 47:103824, 2023. doi: 10.1016/j.jasrep.2022.103824.

Marco Castiello and Marj Tonini. Machine learning applications in archaeological predictive modelling. *Journal of Computer Applications in Archaeology*, 4(1):110–125, 2021. doi: 10.5334/jcaa.71.

Cheng Wang, Ying Li, Yijian Liu, Huan Liu, Jian Liu, Yijian Su, and Huadong Guo. A machine-learning approach to geospatial predictive modelling in archaeology. *ISPRS International Journal of Geo-Information*, 12(6):238, 2023. doi: 10.3390/ijgi12060238.

Christopher G. Newhall and Stephen Self. The Volcanic Explosivity Index (VEI): An estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research: Oceans*, 87(C2):1231–1238, 1982. doi: 10.1029/JC087iC02p01231.

Arnau Folch. A review of tephra transport and dispersal models: evolution, current status, and future perspectives. *Journal of Volcanology and Geothermal Research*, 235–236:96–115, 2012. doi: 10.1016/j.jvolgeores.2012.05.020.

Larry G. Mastin, Alexa R. Van Eaton, and Jacob B. Lowenstern. Modeling ash fall distribution from a Yellowstone supereruption. *Geochemistry, Geophysics, Geosystems*, 15(8):3459–3475, 2014. doi: 10.1002/2014GC005469.

Larry G. Mastin, Alexa R. Van Eaton, David J. Schneider, Kristi L. Wallace, Donald A. Swanson, Charles R. Bacon, Manuel Nathenson, Richard B. Waittt, Cynthia A. Gardner, Alison B. Till, et al. Lessons for forecasting eruptions at Yellowstone and other large volcanoes from the 1980 eruption of Mount St. Helens. *Bulletin of Volcanology*, 84(7):78, 2022. doi: 10.1007/s00445-022-01613-0.

Omri Allouche, Asaf Tsoar, and Ronen Kadmon. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6):1223–1232, 2006. doi: 10.1111/j.1365-2664.2006.01214.x.