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

Mukhlis Amien

2026-02-24

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 the best single-run result (XGBoost AUC 0.768 ± 0.069; TSS 0.507 ± 0.167), above the minimum viable threshold (AUC > 0.75). Robustness checks across 20 alternate seeds yielded a seed-averaged XGBoost AUC of 0.751 (95% bootstrap CI: 0.745–0.756). 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 ρ 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 [1]. For spatially clustered data, random train/test splits overestimate performance. Spatially structured cross-validation is therefore recommended for realistic generalization assessment [2, 3]. We adopt this principle using fixed block CV and compare two tree-based learners (XGBoost and RandomForest) [6, 5].

1.2 Archaeological Context

Archaeological predictive modeling has long recognized that discovery processes can dominate apparent settlement patterns [7]. Recent work also shows that sampling design choices directly affect predictive outputs, interpretability, and transfer behavior [10]. 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 [8, 9]. 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 [11], while modern tephra dispersal literature details uncertainty in ash-cloud and fallout forecasting [12, 13, 14]. 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.

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 `data/processed/east_java_sites.geo`

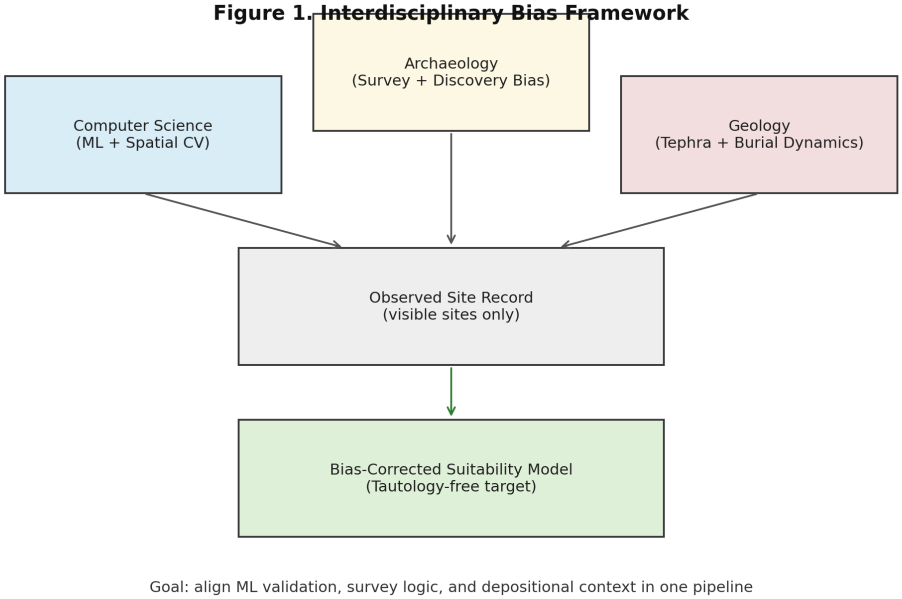

Figure 1: Interdisciplinary bias framework linking computer science validation, archaeological survey logic, and geological burial context.

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 `BLOCK_SIZE_DEG=0.45` (˜50 km). Primary metrics are AUC and TSS [4].

Decision rules:

- 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 ρ 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` $\in \{0.0, 0.15, 0.30\}$.

Figure 2 shows how these interventions were sequenced from E007 to E013.

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

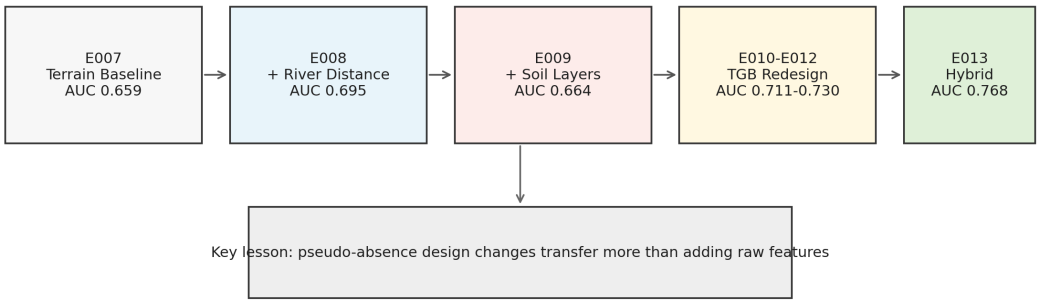

Figure 2: Experiment-to-decision pipeline from E007 baseline to E013 hybrid bias correction.

3 Results

3.1 Performance Progression

Feature-only expansion plateaued (E007–E009), while background redesign delivered monotonic gains (E010–E013). E013 exceeded MVR by +0.018. Performance trajectories are visualized in Figure 3.

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

Table 1: Experiment progression (E007–E013).

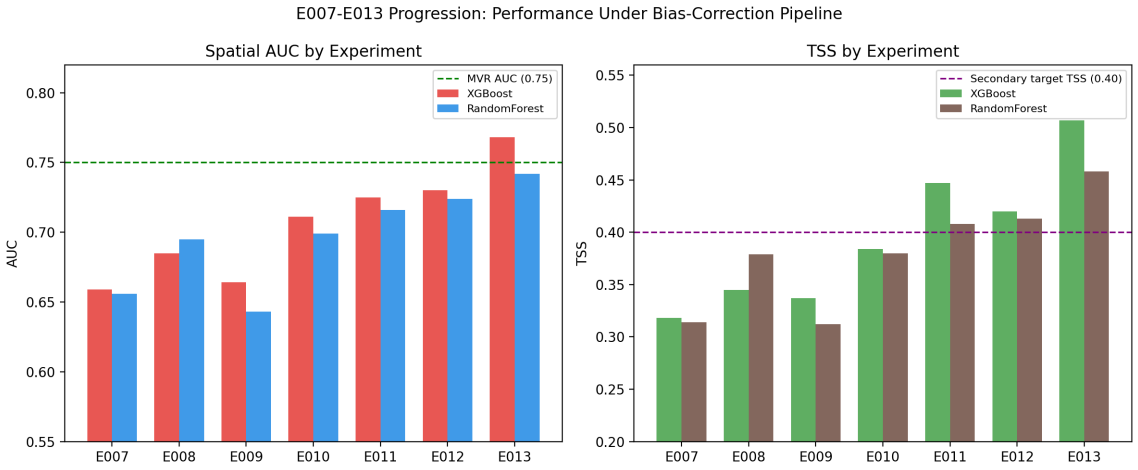

Figure 3: AUC and TSS progression across E007–E013 (XGBoost and RandomForest).

3.2 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).

3.3 Tautology Test

Challenge 1 remained negative across all experiments: E007 -0.095; E008 -0.153; E009 -0.266; E010 -0.142; E011 -0.169; E012 -0.160; E013 -0.229. High-suitability cells within 50 km volcano radius in E013: 57.9%.

3.4 Robustness and Block-Size Sensitivity

For fixed E013 hybrid parameters and 20 alternate seeds:

- XGBoost mean AUC 0.751 (95% CI 0.745–0.756), mean TSS 0.465.
- RandomForest mean AUC 0.744 (95% CI 0.740–0.749), mean TSS 0.458.

Block-size sensitivity:

- ˜40 km: XGB AUC 0.725 (0.718–0.733), RF AUC 0.742 (0.738–0.746).
- ˜50 km: XGB AUC 0.751 (0.746–0.757), RF AUC 0.744 (0.740–0.749).
- ˜60 km: XGB AUC 0.742 (0.737–0.747), RF AUC 0.732 (0.729–0.736).

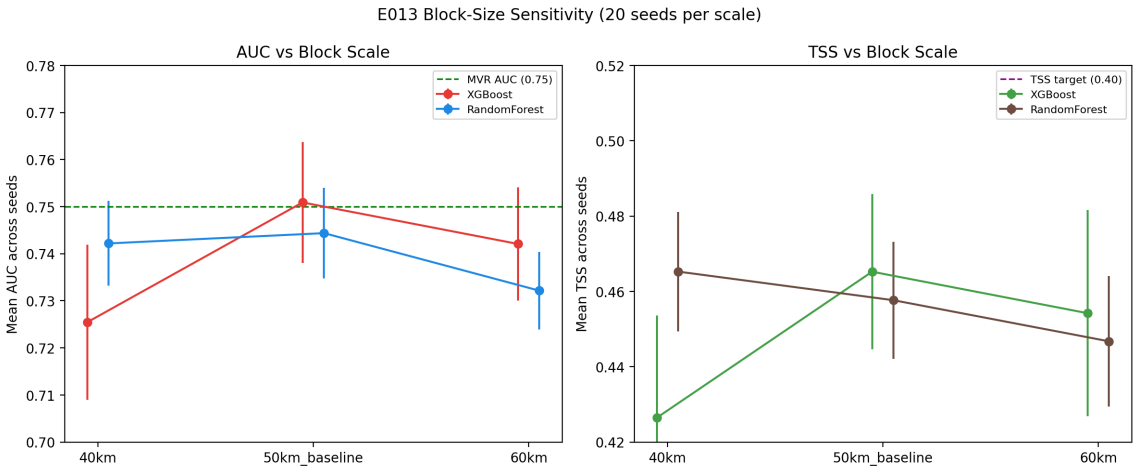

Figure 4: Block-size sensitivity of E013 across ˜40/˜50/˜60 km CV scales.

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 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.

4.3 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.

4.4 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 5 illustrates this interpretation bridge.

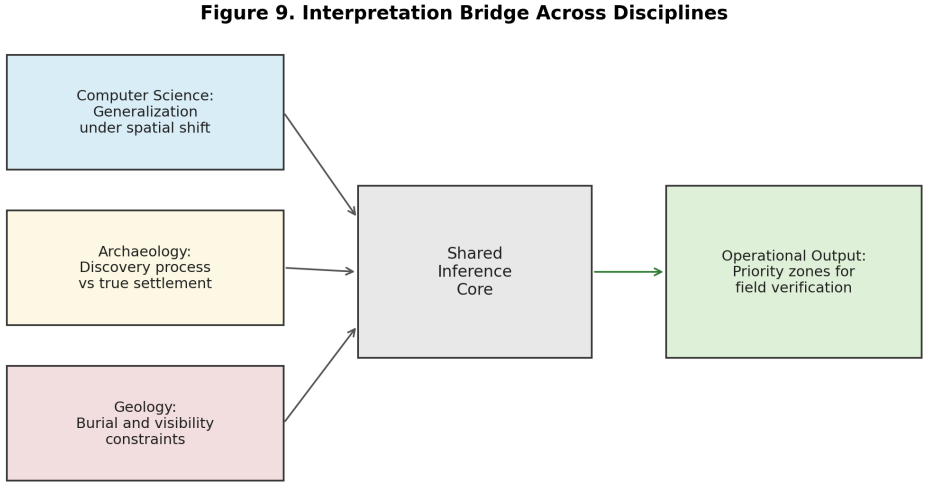

Figure 5: Interpretation bridge showing how CS, archaeology, and geology converge on operational survey prioritization.

4.5 Limitations

1. Pseudo-absences remain inferred labels.
2. Road accessibility is an imperfect survey-effort proxy.
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 (single-run AUC 0.768), while robustness estimates remain near-threshold but stable (seed-averaged AUC 0.751). 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.

Data Availability Statement

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

Code Availability Statement

Model and manuscript scripts are available in experiments/E007_...E013_... and papers/P2_settlement_model/.

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.

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

[1] Phillips, S.J.; Dudik, M.; Elith, J.; Graham, C.H.; Lehmann, A.; Leathwick, J.; Ferrier, S. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications* **2009**, *19*, 181–197. https://doi.org/10.1890/07-2153.1.

[2] Roberts, D.R.; Bahn, V.; Ciuti, S.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. https://doi.org/10.1111/ecog.02881.

[3] Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Arroita, G. 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* **2019**, *10*, 225–232. https://doi.org/10.1111/2041-210X.13107.

[4] Allouche, O.; Tsoar, A.; Kadmon, R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **2006**, *43*, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

[5] Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

[6] Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of KDD '16*; 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[7] Verhagen, P.; Whitley, T.G. Integrating archaeological theory and predictive modeling: a live report from the scene. *Journal of Archaeological Method and Theory* **2012**, *19*, 49–100. https://doi.org/10.1007/s10816-011-9102-7.

[8] Castiello, M.; Tonini, M. Machine learning applications in archaeological predictive modelling. *Journal of Computer Applications in Archaeology* **2021**, *4*, 110–125. https://doi.org/10.5334/jcaa.71.

[9] Wang, C.; et al. A machine-learning approach to geospatial predictive modelling in archaeology. *ISPRS International Journal of Geo-Information* **2023**, *12*, 238. https://doi.org/10.3390/ijgi12060238.

[10] Comer, D.C.; et al. Sampling methods for archaeological predictive modeling. *Journal of Archaeological Science: Reports* **2023**, *47*, 103824. https://doi.org/10.1016/j.jasrep.2022.103824.

[11] Newhall, C.G.; Self, S. The Volcanic Explosivity Index (VEI): An estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research: Oceans* **1982**, *87*, 1231–1238. https://doi.org/10.1029/JC087iC02p01231.

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

[13] Mastin, L.G.; Van Eaton, A.R.; Lowenstern, J.B.; Wessels, R.L. Modeling ash fall distribution from a Yellowstone supereruption. *Atmospheric Chemistry and Physics* **2016**, *16*, 9399–9415. https://doi.org/10.5194/acp-16-9399-2016.

[14] Mastin, L.G.; et al. How best to forecast eruptions? Lessons from the 18 May 1980 eruption of Mount St. Helens. *Bulletin of Volcanology* **2022**. https://doi.org/10.1007/s00445-022-01613-0.