Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Classification Accuracy to Decision Value: Multimodal Sensor Fusion for Economic Prioritization of Landfill Mining Simulation-based Value-of-Information Analysis

Leon Sandler

sandler.leon@gmail.com

Independent Researcher

Research Article

Keywords: landfill mining, multimodal sensor fusion, spatial cross-validation, value of information, random forest, XGBoost, probability calibration, uncertainty quantification

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11323117/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Classification Accuracy to Decision Value: Multimodal Sensor Fusion for Economic Prioritization of Landfill Mining

Simulation-based Value-of-Information Analysis

Leon Sandler (corresponding author)

Independent Researcher, Northbrook, Illinois, USA

E-mail: sandler.leon@gmail.com · ORCID: 0009-0007-4584-808X

Abstract

Landfill mining projects are approved or shelved on incomplete information about what is underground, and a systematic assessment of European scenarios revealed that approximately 19% of them are profitable [1]. In this paper, by simulation, whether multimodal geophysical sensing combined with spatial machine learning and value-of-information (VOI) analysis can reduce that uncertainty before excavation is evaluated. On a synthetic but physically motivated 1,024 m² survey (five geophysical/gas channels: magnetometer, electromagnetic induction, ground-penetrating radar, thermal and gas, surveyed alongside surface topography for georeferencing), an XGBoost classifier reaches pooled F1 values of 0.99 (metal), 0.98 (organic) and 0.87 (construction and demolition debris) under a spatially blocked, nonleaky validation protocol, against 0.32 (plastic) and 0.24 (glass). Random per-cell splitting gave a small, statistically insignificant difference (95% CI including zero) relative to spatial blocking in this configuration, and held-out accuracy did not decrease with distance to the nearest training cell; we found no detectable performance inflation attributable to the tested spatial proximity mechanism in this specific setting, which is not the same as the general claim that spatial blocking is unnecessary. Sensor ablation shows that ground-penetrating radar contributes most to classification (macro-F1 decreases by 0.057 when removed), but an economic value-of-information analysis shows that only the magnetometer reliably increases the expected value of the excavation decision; classification improvements from the other channels did not translate into decision value in this model. Adding a near-infrared channel increases plastic F1 from 0.32 to 0.38 but leaves the VOI statistically unchanged because low-value materials contribute little to the economic layer regardless of classification accuracy. In a synthetic hazard-detection test whose ground truth and sensor reading share the same underlying generative model—an upper-bound test, not an independent validation—the gas-hazard classifier reaches a ROC-AUC of 0.94 and a 100% recall at a safety-oriented operating point, although at that operating point, only 52% of cells flagged as hazardous are true positives. Across the baseline synthetic scenarios and stated economic parameter distributions, pooling 40 independent survey realizations, the simulated survey has positive median VOI ($924, P(VOI>0) = 0.93) and a break-even survey cost of $1451; a two-dimensional

map over sensing-quality and economic-adversity severity is used to locate a model-specific VOI-zero boundary rather than a single degradation threshold, and a further test across deliberately extreme economic regimes reveals that the magnetometer's economic advantage is robust to price fluctuations but not to site composition: the magnetometer led at every tested metal/plastic price combination, yet a deliberately polymer-, debris-, or glass-rich site archetype shifted the top sensor to EMI or GPR. All the results are model predictions, and a field validation and falsification plan with acceptance criteria is given.

Keywords: landfill mining; multimodal sensor fusion; spatial cross-validation; value of information; random forest; XGBoost; probability calibration; uncertainty quantification

1 Introduction

Landfill mining projects convert a waste deposit into recovered material and avoid landfill taxes and reclaimed land, but their economics depend on what is actually underground: material composition, resource concentration, hazard locations, and the processing each fraction requires. A systematic assessment of 531,441 modeled European landfill-mining scenarios revealed only approximately 19% net profitability and approximately 80% return to a negative net present value [1], and follow-up work has investigated which factors most affect economic outcomes and how enhanced landfill mining performs on environmental and economic criteria [2]. The recurring finding is that uncertainty before excavation, not the difficulty of excavation itself, is what turns a potentially viable project into a cost center.

The geophysical characterization of landfills is well established. Electrical resistivity and induced-polarization surveys have long been used to investigate landfill structure and leachate [3], frequency-domain electromagnetic induction has been applied to landfill site characterization since at least the early 1990s [4], and, more recently, to imaging inside capped landfill sites [5], ground-penetrating radar has been used to estimate construction-waste landfill volume [6], and integrated electrical and electromagnetic prospecting has been proposed specifically for municipal waste landfills [3]. What this literature characterizes is subsurface structure and, at best, broad material classes; it does not, in general, convert that characterization into a per-zone economic excavation decision with a quantified value of doing the survey at all.

This paper evaluates one candidate architecture for closing that gap, previously described only as an industrial proposal, by treating it as a design hypothesis and testing it computationally: can multimodal geophysical sensing, fused by spatial machine learning with a spatially valid train/test protocol and propagated through an explicit economic layer, actually change a landfill-mining excavation decision, and by how much? The question is deliberately falsifiable. We do not assume that the answer is yes.

The contributions are (i) a physically motivated synthetic survey generator with five material–response sensor channels plus an optional near-infrared augmentation channel, material-dependent response, depth attenuation, moisture effects, spatially correlated noise and instrument dropout, released as open code so that every number in this paper can be regenerated; (ii) a systematic comparison of spatially blocked against randomly split validation, including a distance–to-nearest-training-observation diagnostic of the kind used in the spatial-statistics literature [7]; (iii) sensor ablation and near-infrared augmentation studies, evaluated on both classification F1 and economic value of information, which are shown to disagree; (iv) a Monte Carlo value-of-information and break-even survey-cost analysis with per-sensor value attribution; and (v) a synthetic-to-real degradation sweep that locates, rather than asserts, the point at which the survey stops being worth its cost.

2 Related work

2.1 Landfill geophysics

Electrical and electromagnetic methods have been applied to landfill investigations for several decades, from early frequency-domain electromagnetic induction soundings for site characterization [4] through resistivity surveys of leachate plumes [3] to recent electromagnetic induction imaging of a capped site [5] and integrated electrical/electromagnetic prospecting proposed specifically for municipal waste landfills [3]. Ground-penetrating radar has been used to estimate the volume of construction-waste landfills from density contrasts [6]. These studies establish that geophysical signals carry usable information about landfill structure and composition; however, no information is coupled to a quantified excavation decision.

2.2 Spatial machine learning and validation

Spatially structured data violate the independence assumption behind ordinary cross-validation, and ignoring that structure can seriously underestimate the predictive error or, conversely, produce optimistic validation statistics that do not hold outside the training footprint [7]. Spatial sorting bias in species distribution models has been shown to inflate cross-validated accuracy relative to spatially independent evaluation [8], and a large-scale demonstration using 11.8 million trees revealed that nonspatial validation of a random-forest biomass model substantially overstated its real predictive power relative to spatially blocked validation [9]. Random forests [10] and gradient-boosted trees such as XGBoost [11] are standard for this kind of tabular geophysical classification, and the spatial transferability of random-forest crop-type classification has itself been studied directly [12].

2.3 Probability calibration

A classifier that outputs a "72% probability of metal-rich material" is only useful to an economic decision layer if that 72% is empirically reliable; well-calibrated probabilities from standard classifiers should not be assumed, and modern flexible classifiers in particular can be poorly calibrated by default [13].

2.4 Landfill mining economics and value of information

Enhanced landfill mining has been reviewed for its resource-recovery potential across material, energy and land-use value streams [14], for its multiresource recovery prospects more broadly [14], for a coordinated European research programme spanning geophysics to material valorization [15], and for country-specific resource-recovery potential [16]. Value-of-information analysis, which involves comparing the expected outcome of a decision made with additional data against the best decision made without it, net of the data's cost, has a long history in mineral exploration under a Bayesian decision-theoretic framing [17] and has recently been reviewed for marine conservation management [18]; we are not aware of a prior application to landfill mining excavation decisions.

2.5 Explainability

Permutation importance is a model-agnostic way to rank feature contributions by the performance drop when a feature is shuffled [19]; we use the same logic at the level of whole sensor channels (ablation) and, in Section 5.4, at the level of economic value rather than classification accuracy, which is the more decision-relevant question for this application.

2.6 Research gap

To our knowledge, no prior study has evaluated an integrated multimodal-sensing-to-excavation-decision pipeline for landfill mining under a spatially valid protocol, with sensor values measured economically rather than only by classification accuracy and with an explicit, model-specific decision-viability boundary rather than a single robustness point estimate. This is what this paper contributes.

3 LAND-MAP framework

3.1 Sensing architecture

The architecture is evaluated at a site with six coregistered channels—LiDAR/RGB surface topography, longwave infrared thermal imaging, a magnetometer, an electromagnetic-induction (EMI) sensor for bulk conductivity and moisture, ground-penetrating radar (GPR) for density contrast and depth, and a methane/VOC gas array—all of which are mounted on a georeferenced platform. Every sensing technology is individually mature and commercially available; the contribution evaluated here is the fusion, validation and economic-decision layer built above them, not new sensor hardware. The

classification model evaluated in Sections 4–5 draws on the five channels with a modeled material response (magnetometer, EMI, GPR, thermal, and gas); the LiDAR/RGB topography provides georeferencing and cell-depth context rather than an independent classification input, and a sixth near-infrared/optical channel is evaluated separately as a proposed augmentation (Section 4.4). The pipeline from the raw channels to the excavation decision is summarized in Figure 1.

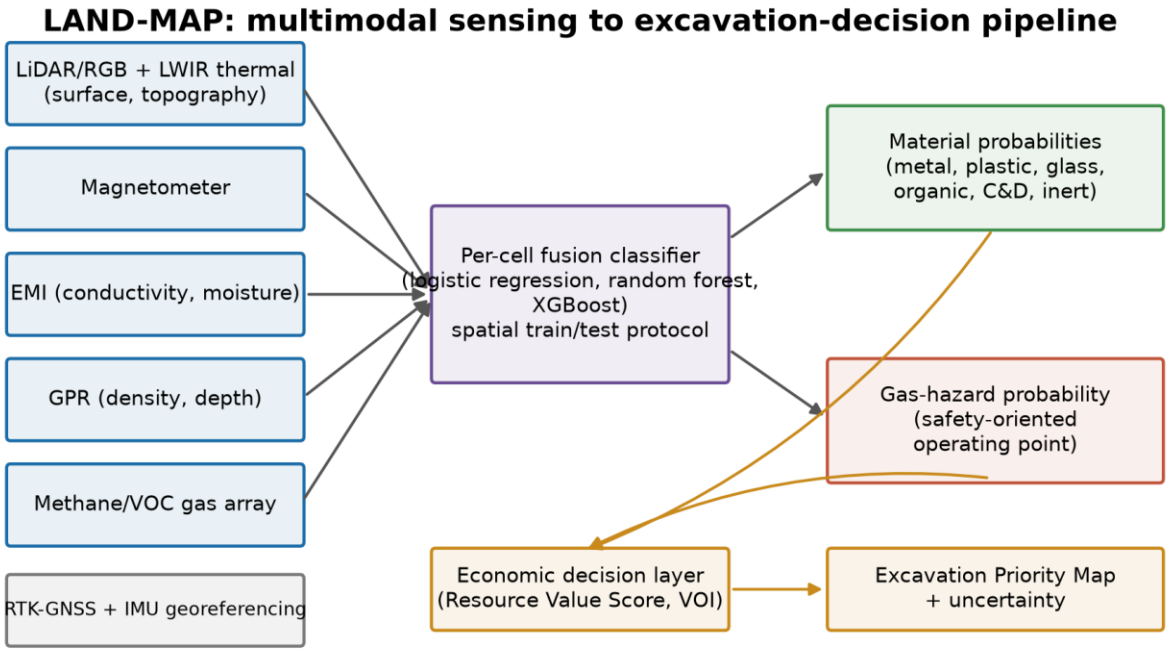


Fig. 1 LAND-MAP pipeline: coregistered sensing channels are fed a per-cell fusion classifier under a spatial train/test protocol, producing material and gas hazard probabilities that an economic decision layer converts into an excavation priority map.

3.2 Fusion and economic decision layers

Per-cell readings are combined by a trained classifier into probabilistic material classifications over six classes (metal, plastic, glass, organic, construction and demolition debris, and inert) and a separate gas hazard probability. Each cell's expected composition is converted into a resource value score by multiplying the predicted class probabilities by an assumed recoverable mass, recovery fraction and secondary material price and netting against excavation and processing cost; aggregating scores yields an excavation priority map. A value-of-information metric compares the expected outcome of survey-guided excavation against the best decision an operator could make without any survey, net of the survey's own cost.

4 Methods

4.1 Synthetic landfill generation

A 32 × 32 m (1,024 m²) survey area is discretised into 1 m×1 m cells over three coarse depth layers (0–1, 1–2, and 2–3 m). A latent dominant material is assigned to every cell and depth layer by thresholding a smooth two-dimensional Gaussian random field against the class prior, yielding spatially patchy (not salt-and-pepper) material zones consistent with how waste is dumped and compacted in lots, with a correlation length of 8.0 m. Cover depth to the dominant layer and moisture are independent smooth random fields.

$$x_{i,j} = f_j(M_i, z_i, m_i) + \varepsilon_{i,j} \quad (1)$$

where x_{i,j} is sensor j's reading at cell i, M_i is the material at each depth layer, z_i is the cover depth, m_i is the moisture, and f_j is a material-response function attenuated with depth as exp(−α_j z) and, for GPR and EMI, further modified by moisture. The response coefficients (Table 1) are literature-motivated orders of magnitude, not fitted values, and are stated once so, they can be perturbed systematically (Section 4.6). Sensor noise ε_{i,j} has both an independent per-cell component and a spatially correlated component (correlation length of 1.6 m, modeling instrument drift and footprint overlap), and each channel independently drops out at cell-specific rates reflecting real failure modes (highest for gas, drift and saturation, and GPR, from coupling and siting issues).

Channel	Metal	Plastic	Glass	Organic	C&D	Inert
MAG	3.00	0.05	0.05	0.05	0.55	0.05
EMI	2.60	0.15	0.10	0.90	0.75	0.20
GPR	2.20	0.70	0.90	0.40	2.90	0.30
THERM	0.10	0.10	0.10	1.60	0.15	0.10
GAS	0.05	0.05	0.05	2.40	0.05	0.05
NIR (optional)	0.30	1.80	1.20	0.90	0.60	0.40

Table 1. Assumed mean sensor response by material and channel, arbitrary standardized units, before depth attenuation, moisture effects and noise. Every value is an assumption motivated by the qualitative physics of each sensing modality (e.g., strong magnetometer/EMI response to metal, GPR response to density/dielectric contrast, thermal and gas response to organic decomposition); none is fitted to field data.

The gas hazard label is a noisy, logistic threshold of the prenoise gas-channel signal so that the hazard status is by construction what an ideal gas sensor would detect; the deployed sensor differs from that truth only by its own instrument noise and dropout, as a real hazard label derived from independent gas testing would relate to a field-deployed sensor. One representative synthetic survey is shown in Figure 2: the true near-surface material field, one raw noisy sensor channel, and the resulting gas hazard truth.

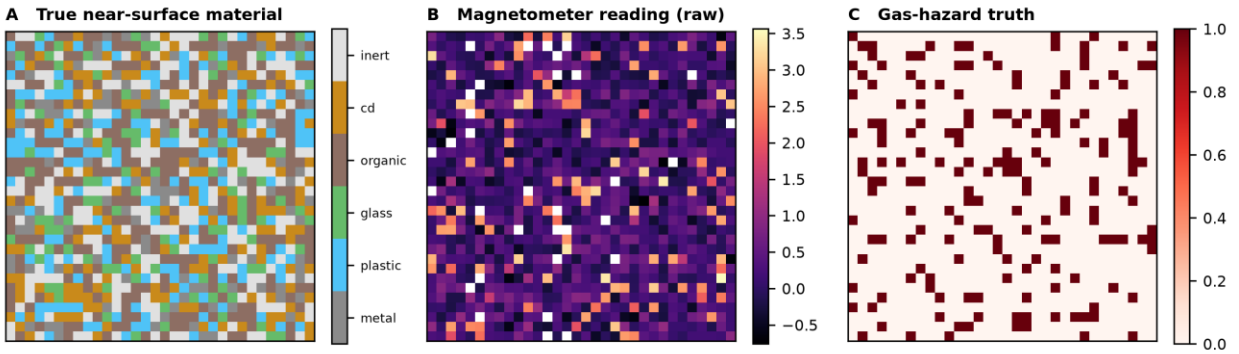


Fig. 2 One representative 32×32 m synthetic survey. (A) True near-surface dominant material; (B) raw magnetometer data (noisy, with dropout); (C) gas hazard truth derived from the prenoise gas signal.

4.2 Spatial validation protocol

Two held-out protocols are compared on the same synthetic surveys: a spatially blocked holdout, in which one contiguous 8×8 cell region is withheld entirely, and a random per-cell holdout of the same size scattered uniformly. To test the mechanism behind any difference directly rather than only at one block size, we also compute, pooling across 15 repeats and eight block sizes (4–20 cells), the distance from every held-out cell to its nearest training cell and report held-out accuracy as a function of that distance [7, 9].

4.3 Classifiers and calibration

Three classifiers are compared: logistic regression, random forest [10], and XGBoost [11], each with mean imputation for sensor dropout. Calibration is assessed by the one-vs-rest Brier score, macroaveraged over classes [13], because the economic layer consumes class probabilities directly and a poorly calibrated “72%” is not the same as an empirically reliable one.

4.4 Sensor ablation and NIR augmentation

Each of the five material-response channels is removed in turn, and the best-performing classifier is retrained, isolating each channel's contribution to the pooled macro-F1. A sixth, optical/near-infrared channel, motivated by the plastic/glass confusion in this paper's synthetic physics, is added to test whether it repairs that specific limitation under the assumed NIR response coefficients (Table 1); it is modeled as a surface-only channel (rapid depth attenuation) with distinct plastic and glass responses. Because those coefficients are assumptions rather than measured spectra, this tests a hypothetical NIR channel with a stated response structure, not a claim about any specific real instrument (Section 5.5).

4.5 Economic model and value of information

For material class c, the expected net value of excavating cell i is

$$v_i = \sum_c P(c|x_i) \cdot m_i \cdot \rho_c \cdot \pi_c - m_i(k_{ex} + k_{pr}) \quad (2)$$

where m_i is the assumed waste mass per cell, ρ_c is the recovery fraction, π_c is the secondary-material price, and k_{ex}, k_{pr} are the excavation and processing costs per tonne (Table 2; all assumptions are swept in Monte Carlo rather than fixed). The survey-guided decision excavates cell i if $v_i > 0$; the blind decision excavates every cell if the population-mean expected value is positive; otherwise, it is zero. The value of information is

$$VOI = EV_{survey} - EV_{blind} - C_{survey} \quad (3)$$

evaluated against the true (not expected) cell value, over 20000 Monte Carlo draws of the economic parameters. The break-even survey cost is $EV_{survey} - EV_{blind}$, with C_{survey} excluded, and the maximum rational operator should pay. The per-sensor value is $VOI_{all} - VOI_{without\ j}$, computed on the same economic draws for a paired comparison and averaged over 20 independent synthetic surveys because a single survey's VOI estimate is noisy.

Parameter	Range (uniform)
Price, $/t: metal	250–550
Price, $/t: plastic	60–180
Price, $/t: glass	10–40
Price, $/t: organic	0–15
Price, $/t: C&D	5–25
Price, $/t: inert	0–5
Recovery fraction: metal	0.55–0.85
Recovery fraction: plastic	0.15–0.40
Recovery fraction: glass	0.10–0.30
Mass per cell, t	0.9–1.6
Excavation cost, $/t	18–32
Processing cost, $/t	12–28
Survey cost, $ total	400–1200

Table 2. Economic assumptions (illustrative, as labeled throughout the source proposal), swept uniformly in the Monte Carlo analysis rather than fixed at a point estimate. The full ranges for every material class are in the released configuration file.

4.6 Synthetic-to-real robustness sweep

The greatest acknowledged risk in this design is that real, weathered, heterogeneous waste is less distinctive than any simulation's idealized material signatures. Rather than asserting a margin of safety, we degrade the simulation directly along five axes named in the source proposal's own risk assessment: overall sensor noise, dropout probability, GPR depth attenuation (modeling moisture/conductivity losses), moisture-driven GPR attenuation specifically, and material-signature overlap (each material's response decreases toward the six-class mean). Each axis is swept individually and then combined, and the combined sweep is searched to locate the point at which the mean VOI crosses zero; then, the same logic is extended to two dimensions with an independent economic-adversity axis (Section 5.11)—a model-specific zero-VOI boundary that this paper reports rather than assumes.

5 Results

5.1 Material classification

Class	Logistic Regression	Random Forest	XGBoost
Metal	0.975	0.987	0.991
Plastic	0.322	0.309	0.321
Glass	0.122	0.168	0.241
Organic	0.983	0.985	0.984
C&D	0.870	0.863	0.870
Inert	0.599	0.568	0.526
Macro-F1	0.645	0.646	0.655

Table 3. Pooled per-class F1 under the spatially blocked protocol, three classifiers, and 2560 pooled test cells across 40 repeats.

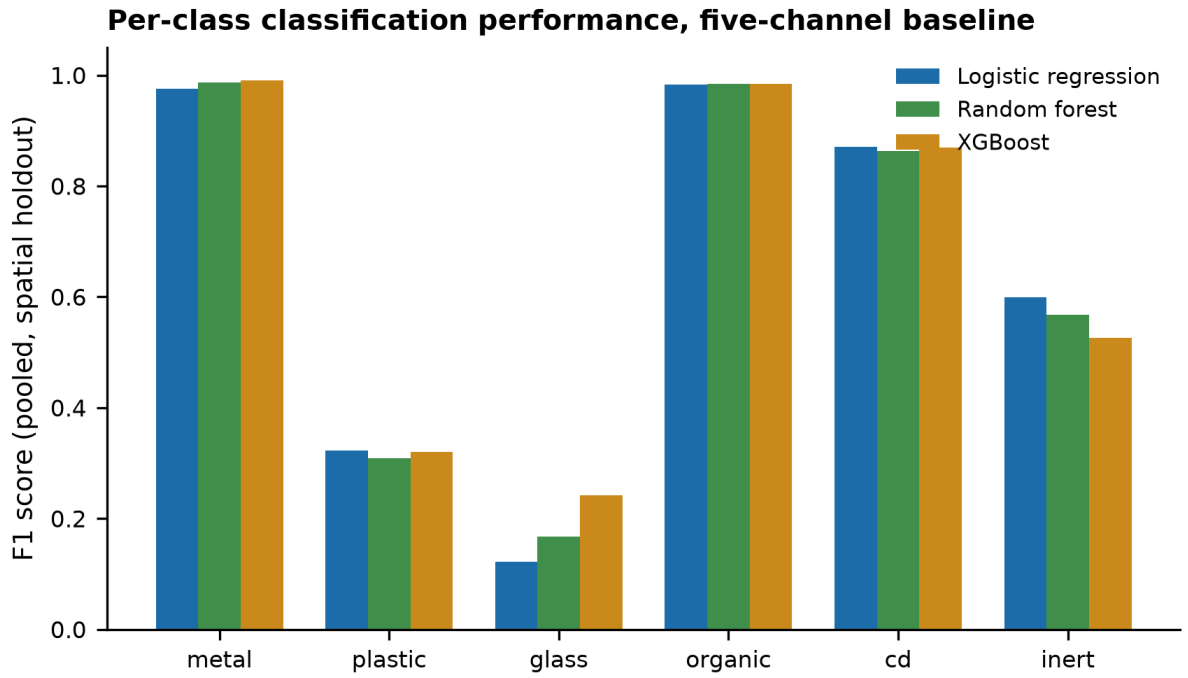


Fig. 3 Per-class F1 by classifier, five-channel material-response baseline, spatial holdout.

Metal (F1 0.98–0.99 across classifiers) and organic material (0.98–0.98) are classified reliably (Fig. 3), and construction and demolition debris reaches 0.86–0.87 once its magnetometer/EMI signature (motivated by embedded rebar and metal fragments) is included; these three values approach the reported range of 0.90–0.98. Plastic (0.31–0.32) and glass (0.12–0.24) remain poorly distinguished by this sensor set, which is consistent with the proposed diagnosis that both are weakly magnetic and lack a distinguishing signal in the magnetic/EMI/GPR/thermal/gas channel set. XGBoost gives the highest macro-F1 (0.655) and is used as the reference classifier for the remaining analyses unless stated otherwise.

5.2 Spatial versus random validation

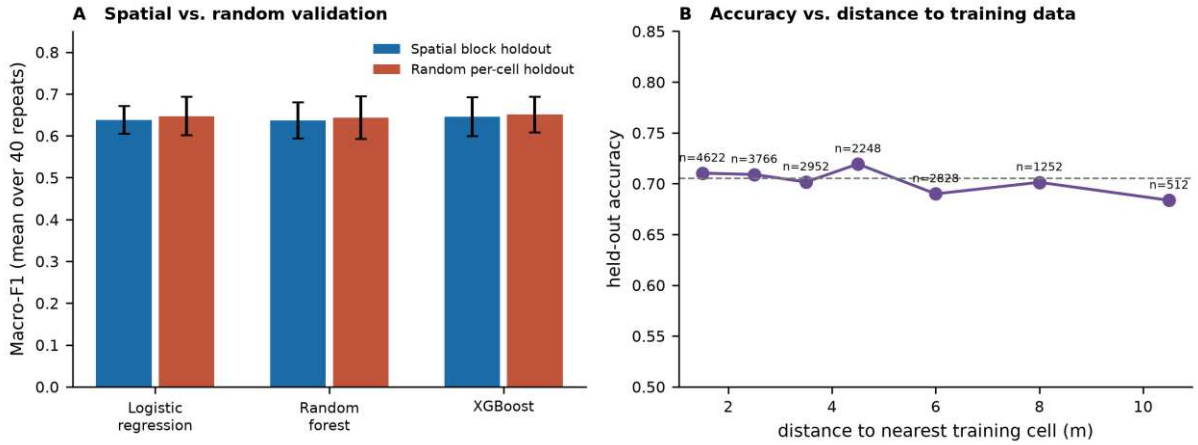


Fig. 4 (a) Macro-F1 under spatially blocked versus randomly split holdout, three classifiers, mean ± SD over 40 repeats. (b) Held-out accuracy against distance to the nearest training cell, pooled over 15 repeats and eight block sizes; the dashed line is the overall mean.

Contrary to the hypothesis that random per-cell splitting inflates apparent performance through spatial autocorrelation leakage [7–9], the paired difference (random minus spatial) was small, and its 95% confidence interval included zero for every classifier: logistic regression: +0.009 (95% CI -0.008 to +0.027); random forest: +0.007 (95% CI -0.013 to +0.027); XGBoost: +0.005 (95% CI -0.012 to +0.022). Distance-to-nearest-training-cell analysis (Fig. 4b, pooled over 18180 observations) shows no accuracy decay to 10 m, the greatest distance sampled. We expected a gap, given the spatial correlation deliberately built into both the material field and the sensor noise; its absence indicates that, for a feature set built from single-cell sensor readings, per-cell classification difficulty in this system is dominated by genuine material overlap (plastic/glass/inert) rather than by a memorisable local spatial structure. We report this as a specific, tested finding, not a general claim that spatial blocking is unnecessary; the mechanism identified in the literature [7, 9] requires a model able to exploit local proximity, which single-cell features here evidently do not provide, and we recommend spatial blocking as standard practice regardless, because a less idealized real dataset need not share this property.

5.3 Sensor ablation

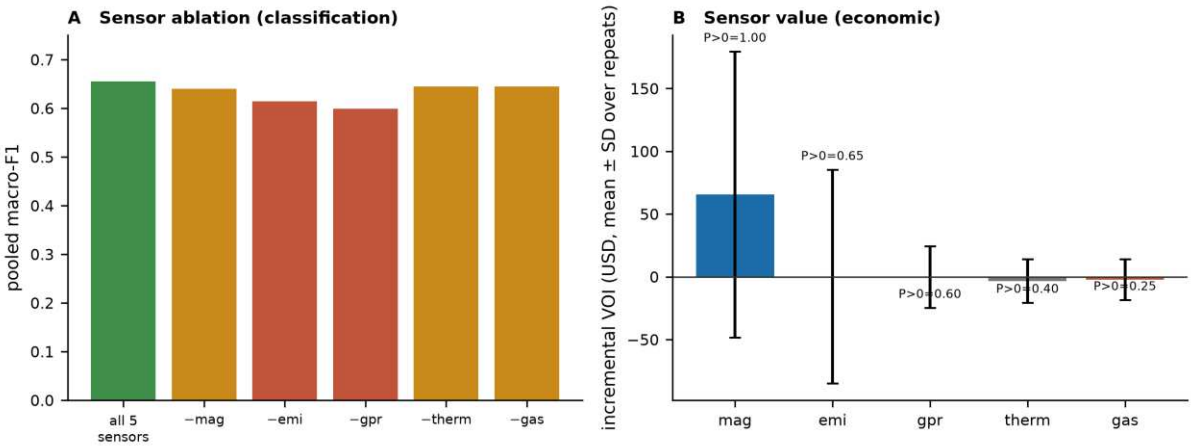


Fig. 5 (a) Pooled macro-F1 with each sensor removed in turn. (b) Incremental value of information (VOI with all sensors minus VOI without sensor j), mean ± SD over 20 independent surveys, with the fraction of surveys giving a positive incremental value.

Removing GPR costs the most classification performance (macro-F1 0.655 → 0.599), followed by EMI (→ 0.614); removing the magnetometer, thermal or gas channels costs little (→ 0.640, 0.645, 0.645, respectively). By classification accuracy alone, GPR and EMI are judged as the most valuable channels.

5.4 Sensor value: classification accuracy is not economic value

The economic picture is different, and this divergence is the paper's central empirical finding. Of the five channels, only the magnetometer shows a reliably positive incremental VOI (mean $65, positive in 100% of independent surveys); EMI, GPR, thermal and gas each have an incremental VOI whose sign is not consistent across surveys (positive in 65%, 60%, 40% and 25% of surveys, respectively). The reason is that the resource value score is dominated by metal, whose assumed price (Table 2) is several times that of any other class; thus, a channel's contribution to overall classification accuracy does not translate into decision value unless it specifically improves metal detection. GPR has the most unique subsurface information and is the strongest classification contributor, but that information is spread across several material classes with low individual economic weights.

5.5 Hypothetical optical/NIR augmentation

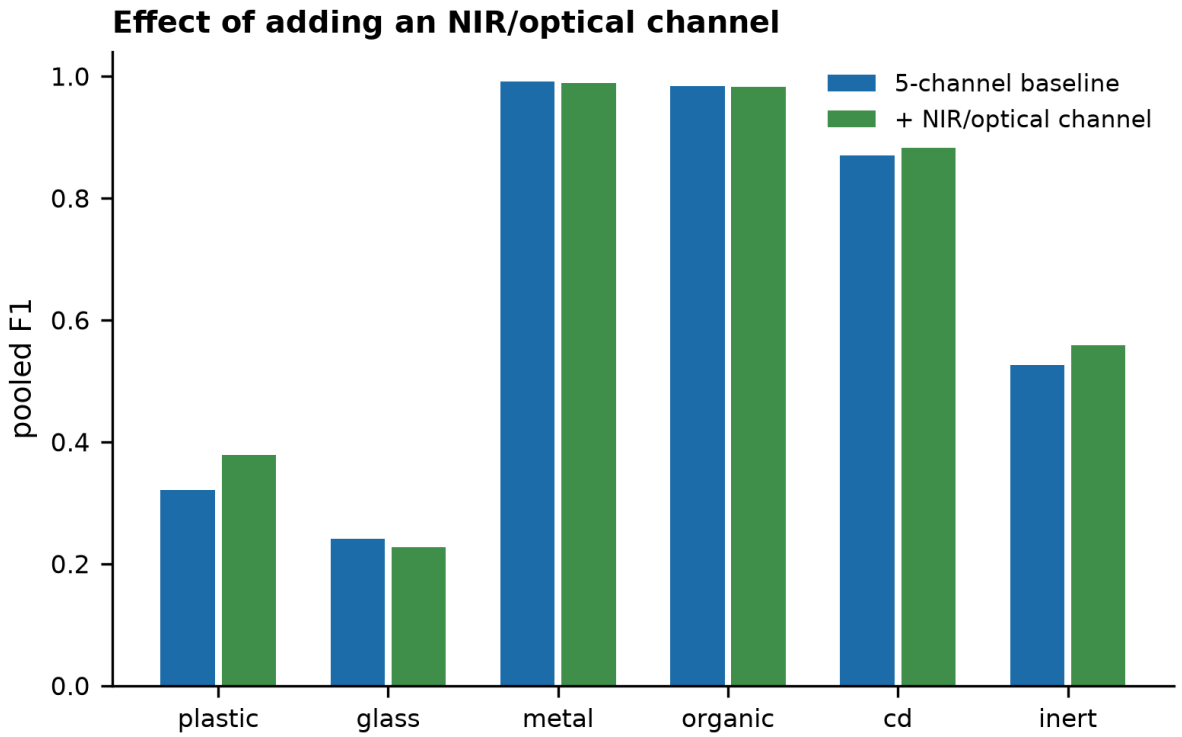


Fig. 6 Per-class pooled F1, five-channel material-response baseline versus baseline plus a hypothetical near-infrared/optical channel.

Under the assumed NIR response coefficients (Table 1) — a hypothetical surface optical channel with plastic- and glass-distinguishing responses, not spectra from any specific real instrument — adding it increases plastic F1 from 0.32 to 0.38 and macro-F1 from 0.655 to 0.669, but glass F1 moves from 0.24 to 0.23 (Fig. 6) — essentially unchanged, and in this simulation, because the rebalancing of the added channel and the classifier across the six classes do not uniformly benefit every class. The economic effect is still smaller: the mean incremental VOI from adding NIR is $-0.1 (SD $9.5), which is positive in only 50% of the surveys and statistically indistinguishable from no effect. Plastic and glass together carry a modest assumed price and a low assumed recovery fraction (Table 2), so even a genuine classification improvement for these classes barely moves the excavation decision. The proposal's own hypothesis — that an optical channel, not further classifier tuning, is needed to separate plastic and glass — is partially supported (plastic improves; glass does not), and in either case, the improvement is not what would justify the channel economically in this model.

5.6 Calibration and hazard detection

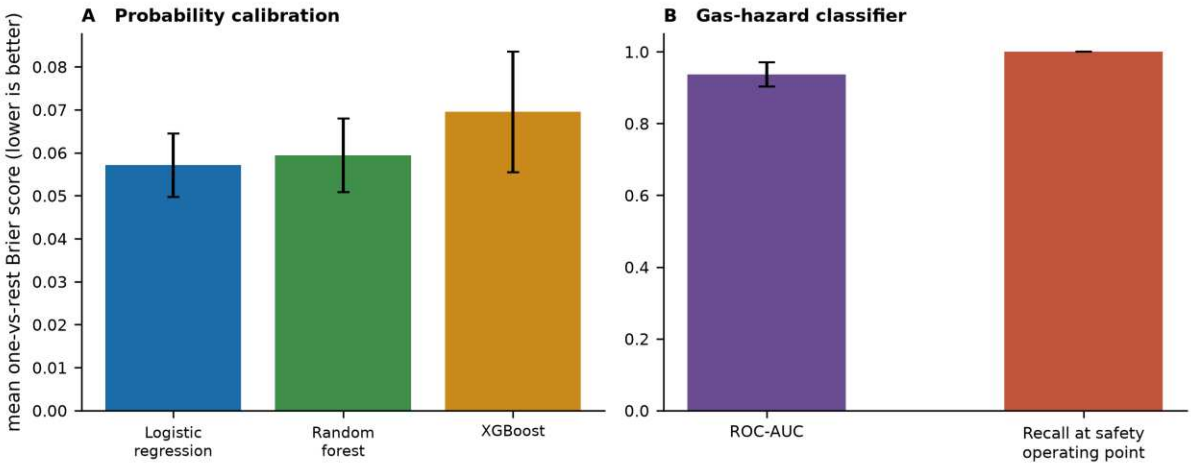


Fig. 7 (a) Mean one-vs-rest Brier score by classifier (lower is better calibrated). (b) ROC-AUC and recall of the gas-hazard classifier at a safety-oriented operating point (the lowest threshold achieving ≥95% recall), mean ± SD over 40 repeats.

Logistic regression is the best calibrated of the three classifiers (Brier 0.057; Fig. 7a), followed by random forest (0.059) and XGBoost (0.070); XGBoost's higher discriminative power (Table 3) therefore comes with somewhat less reliable probabilities, which matters because the economic layer consumes those probabilities directly rather than only the predicted class [13]. The gas hazard classifier reaches an ROC-AUC of 0.937 (0.034), and at an operating point chosen to guarantee at least 95% recall, it achieves 100.0% recall (0.000) in held-out testing (Fig. 7b), approaching 97%/0.97 figures, the source proposal reported from its own Stage 1 testing. Recall alone overstates this operating point's usefulness: a detector can reach 100% recall by flagging nearly everything. Pooled across the same 40 repeats, the precision at this operating point is only 52%, and the specificity is 82%, meaning that approximately 32% of cells are flagged as hazardous and that approximately two of the five flagged cells are false positives (PR-AUC 0.697); the operating point is deliberately chosen to guarantee recall for a safety application, and it does so, but at a real, quantified precision cost that the recall figure alone does not convey.

5.7 Economic prioritization and value of information

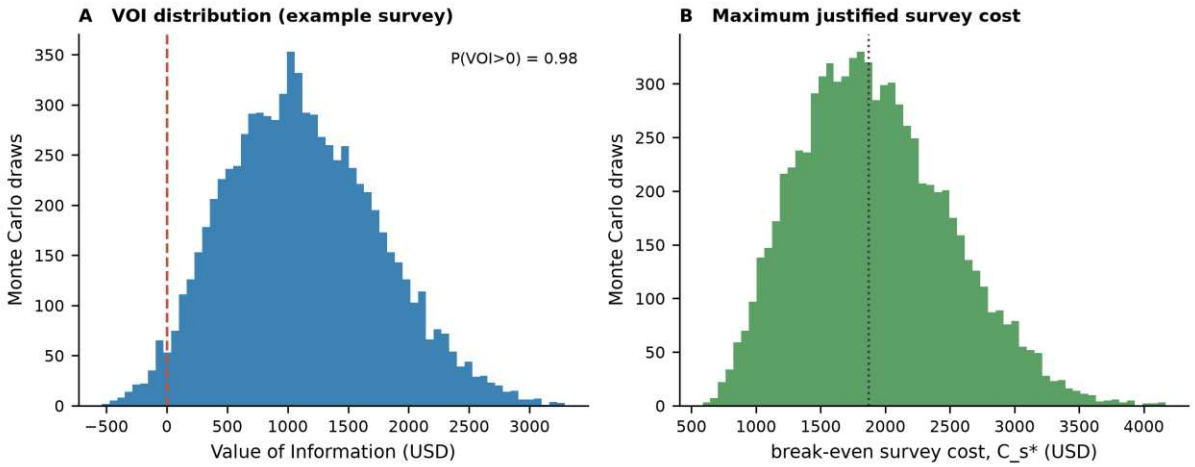


Fig. 8 (a) Monte Carlo distribution of the value of information for one representative survey (8000 draws). (b) Distribution of the break-even survey cost (the maximum a rational operator should pay), same draws.

For the representative survey used throughout Sections 5.3–5.6, the median VOI is $1080 (90% interval $154–$2243), and the survey is worth doing in 98% of the Monte Carlo draws (Fig. 8).

A single survey's mean VOI is itself a noisy estimate of the population quantity that this paper reports as its headline number; thus, it is averaged over many independent synthetic surveys rather than quoted from one. Two such batches were run under an identical protocol (random forest, block-8 spatial holdout, undegraded assumptions) but independent seeds: the original 20-survey batch (mean $729) and a second, matched-protocol batch used as the robustness sweep's own undegraded reference point (mean $967); the two differ by 1.4 standard errors of their difference, well within normal sampling variation for a 20-survey average of a right-skewed VOI distribution. Rather than reporting either batch alone, this paper's one canonical baseline VOI estimate pools both, 40 survey realizations in total: mean $848 (median $924, SE $86), which is positive in 92% of the surveys, and a median break-even survey cost of $1451—the amount up to which an operator should rationally be willing to pay for a survey of this kind under the stated economic assumptions. Section 5.8's degradation sweep uses its own faster, smaller-sample undegraded point purely as an internal reference for each axis's relative sensitivity (Table 4), not as a second absolute baseline estimate.

5.8 Synthetic-to-real robustness

Degradation axis	Baseline macro-F1	Most severe macro-F1	Baseline mean VOI ($)	Most severe mean VOI ($)
Sensor noise (× baseline SD)	0.637	0.389	1012	490
Dropout probability (× baseline)	0.637	0.508	1012	787
GPR depth attenuation (× baseline)	0.637	0.621	1012	1009
Moisture-driven GPR	0.637	0.598	1012	1023

attenuation (× baseline)				
Material-signature overlap (shrinkage fraction)	0.637	0.468	1012	810

Table 4. One-axis synthetic-to-real degradation sweep (Section 4.6): Each row varies only that axis, from baseline to the most severe level tested, holding all others fixed. The "Baseline" column is this sweep's own smaller-sample undegraded reference point (12 realizations, 1500 Monte Carlo draws), used only to measure each axis's relative degradation; it is not the paper's canonical VOI estimate, which Section 5.7 defines separately by pooling 40 realizations. The full multilevel sweeps are in the released results file.

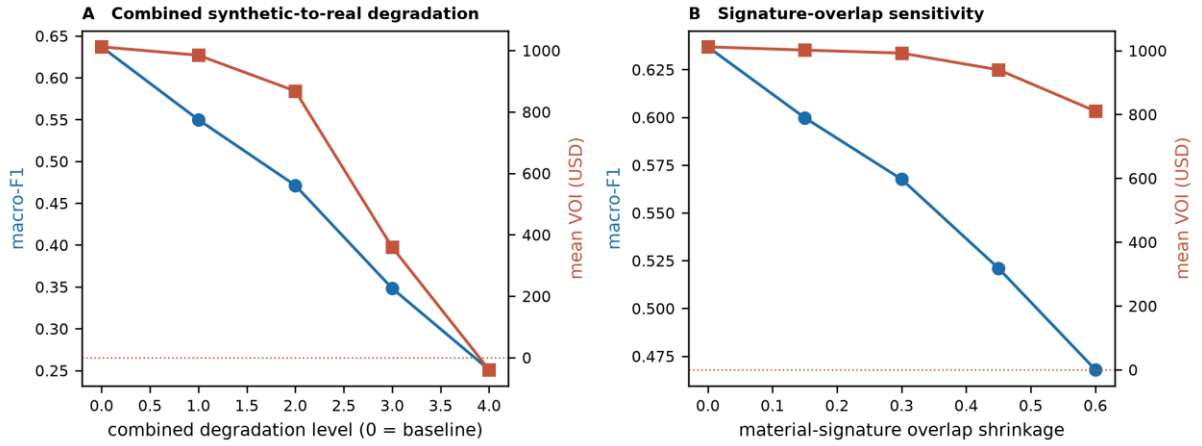


Fig. 9 (a) Macro-F1 (left axis) and mean VOI (right axis, dashed zero line) against combined synthetic-to-real degradation level moving noise, dropout, GPR attenuation, moisture attenuation and material signature overlap. (b) The same two quantities against material signature overlap alone.

Individually, the five degradation axes are not equally damaging (Table 4). At the most severe level tested on each axis alone, overall sensor noise does the most harm (macro-F1 0.64 → 0.39, mean VOI $1012 → $490), followed by material-signature overlap (F1 → 0.47, VOI → $810) and dropout probability (F1 → 0.51, VOI → $787); GPR-specific depth attenuation and moisture attenuation, tested over the same relative range, show little standalone effect on either metric (VOI → $1009 and $1023, respectively), which is notable given the importance of GPR to raw classification (Section 5.3) and suggests that the economic result is more sensitive to generic sensor noise than to the physics of this specific channel. Compared with any individual axis, the combined scenario, in which all five axes are moved together, is more damaging: the mean VOI decreases from $1012 at baseline to $985, $868, $359 and $-40 across increasing combined severity, and it becomes negative before the most severe combined scenario is tested. A more detailed search along the same combined axis reveals the point at which the VOI changes sign 98% of the way from baseline to the most severe scenario evaluated—a model-specific zero-VOI boundary—not a general claim about when surveying stops being worthwhile: beyond it, this simulation's economic prediction changes sign. In Section 5.11, this one-dimensional severity fraction is replaced with an explicit two-dimensional decision-viability map that gives the boundary in terms of the underlying degradation parameters.

Degradation axis	Baseline	Value at VOI = 0	Most severe scenario
Sensor noise	1.0	2.47	2.5
Dropout probability	1.0	3.93	4.0
GPR depth attenuation	1.0	2.47	2.5
Moisture-driven GPR attenuation	1.0	2.47	2.5
Material-signature overlap	0.0	0.59	0.6

Table 5. The combined-degradation VOI-zero boundary (97.7% of the way from the baseline to the most severe scenario; Fig. 9a) is expressed in the underlying degradation parameters used by the combined-severity sweep rather than as a bare percentage. Sensor noise and dropout are expressed as multiples of their baseline standard deviation/probability; GPR and moisture attenuation are expressed as multiples of their baseline attenuation coefficient and overlap as the fraction of each material's response decreases toward the six-class mean.

5.9 Classifier choice and probability calibration for the economic layer

The economic layer (Sections 5.4 and 5.7) uses random forest rather than XGBoost, the classifier with the highest raw macro-F1 (Section 5.1), because XGBoost's probabilities are the least well calibrated of the three (Brier 0.070 vs. 0.059 for random forest; Section 5.6), and the VOI consumes the predicted probabilities directly rather than only the predicted class. This choice is tested, not merely asserted: averaged over 20 independent surveys, the mean VOI is $729 with random forest, $720 with raw XGBoost (difference $-9) and $731 with Platt-calibrated XGBoost (sigmoid calibration, 3-fold; Brier improves to 0.066). All three are within the normal sampling variation of one another, and all give $P(\text{VOI}>0) \geq 95\%$; the classifier choice for the economic layer does not materially change the economic conclusion in this simulation, which is itself informative—it means that the classifier-calibration concern raised by Section 5.6's Brier score gap, while real, is not what is driving this paper's headline VOI estimate. This comparison uses one fixed batch of 20 survey seeds so that the three classifiers are compared on identical draws (a paired design); Section 5.7's canonical baseline VOI pools this same random-forest batch with an independent second batch for a lower-variance point estimate, which is why its value differs slightly from the $729 quoted here for random forest alone.

5.10 Sensor selection depends on site composition, not price alone

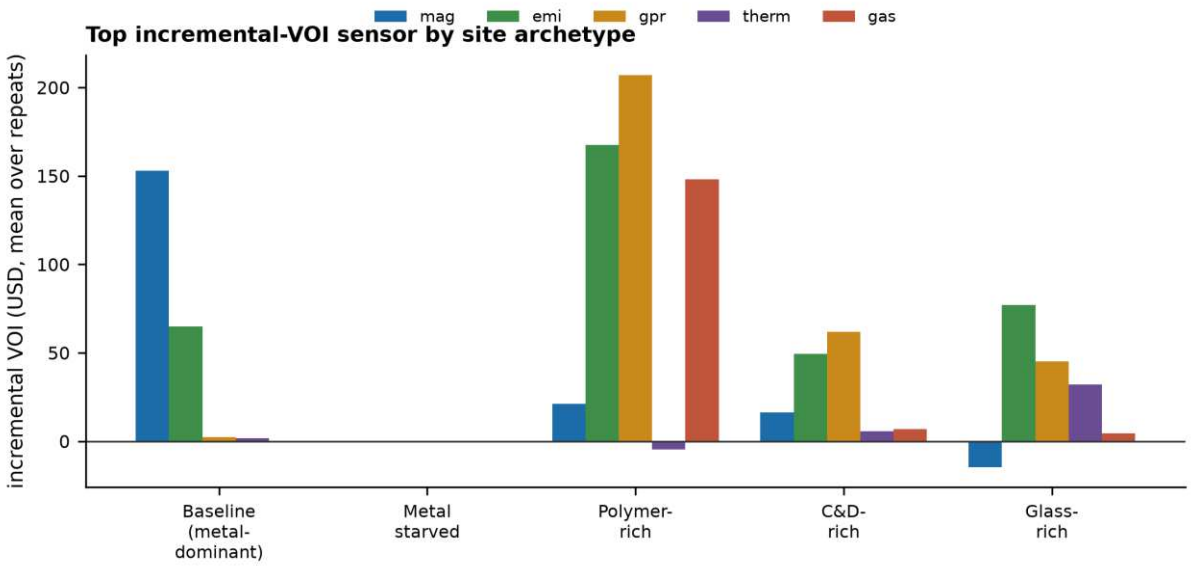


Fig. 10 Incremental VOI (mean over 15 repeats) for each of the five channels at baseline economics and four deliberately extreme site archetypes: metal devalued to 30% of baseline with no other change; the same metal devaluation plus plastic price and recovery fraction raised 5.0x/4.0x (polymer-rich); plus C&D price/recovery raised 6.0x/2.5x (C&D-rich); plus glass price/recovery raised 8.0x/3.5x (glass-rich).

A grid over metal price (0.5–1.6x baseline) crossed with plastic price (0.4–3.0x baseline), 5×5 points, 15 repeats per point, never changed the winner away from the magnetometer (Fig. 5b's single-point result therefore generalizes across ordinary price fluctuations, not just the one-price table in Table 2). However, deliberately extreme, still economically nameable site archetypes do flip the ranking (Fig. 10): starving metal's value alone collapses every sensor's incremental VOI toward zero (no single channel matters when no material is worth much); combined with a hypothetical high-value, high-recovery plastics market, GPR becomes most valuable ($207), ahead of EMI ($168) and gas ($148); with a high-value C&D market, GPR again leads ($62); with a high-value glass market, EMI leads ($77). The classification-optimal sensor (GPR, Section 5.3) and the economic-optimal sensor are therefore not fixed properties of the sensing architecture: they are conditional on the site's resource-value structure, and the general claim this paper supports is that sensor suites should be selected against site-specific decision value, not against a generic classification benchmark or a single assumed price table.

5.11 A two-dimensional decision-viability map

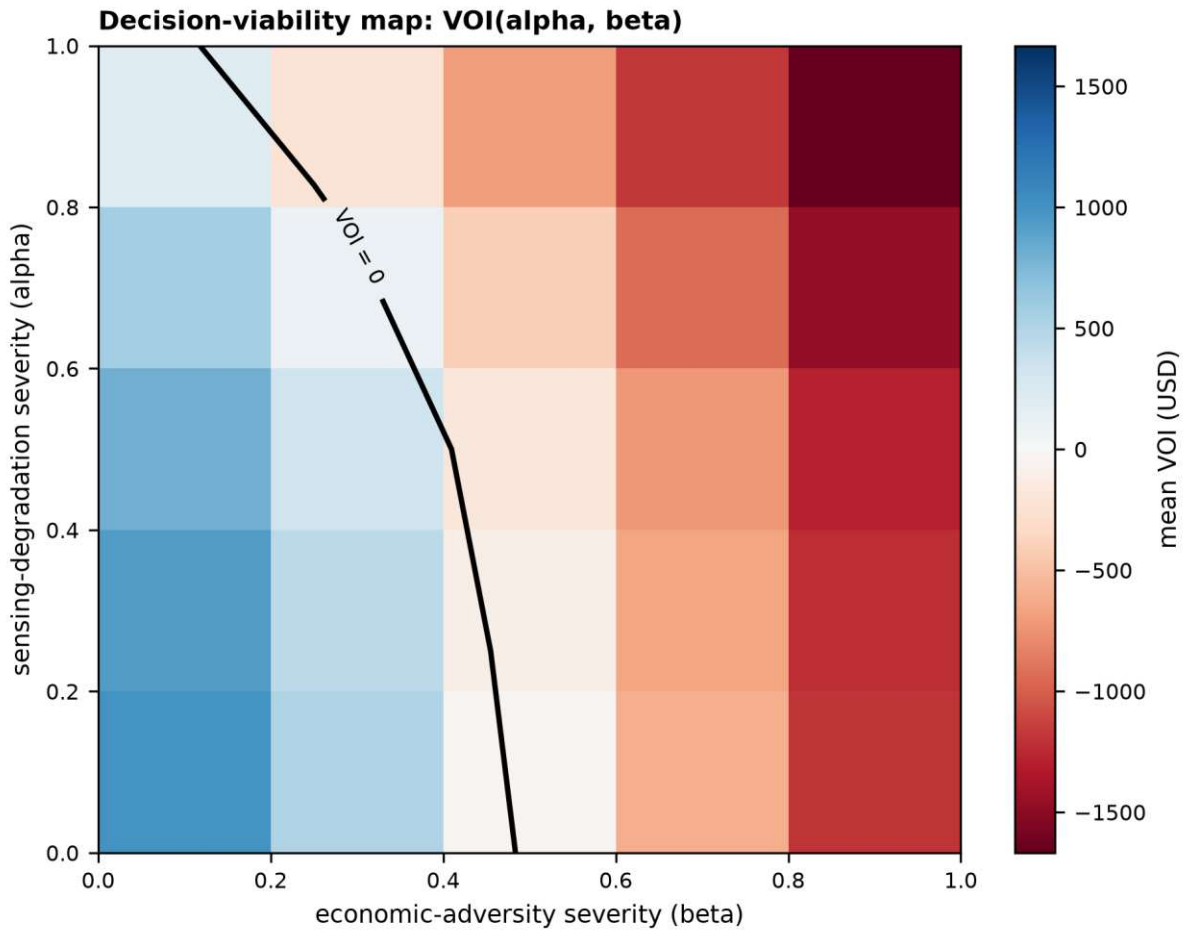


Fig. 11 Mean VOI as a function of sensing-degradation severity (alpha, the same combined parameterization as in Fig. 9a) and economic-adversity severity (beta: secondary-material prices decreased toward zero and survey cost increased), 5×5 grid, 10 repeats per point, with the VOI = 0 contour drawn in black.

The one-dimensional combined-degradation sweep (Fig. 9a, Table 5) is the beta = 0 slice of this surface (Fig. 11). Moving along that slice alone, the VOI-zero boundary sits at 98% combined severity, but the boundary is far more sensitive to economic adversity than to sensing degradation: at alpha = 0 (no added sensing degradation at all), the mean VOI already crosses zero at beta ≈ 0.48, and at alpha = 1 (the most severe sensing degradation tested), it crosses at beta ≈ 0.12—a relatively small shift. The decision-viability region $\mathscr{L}_+ = \{(\alpha,\beta): \mathrm{VOI}(\alpha,\beta) > 0\}$ in this simulation is therefore bounded mainly by whether secondary material markets and survey costs stay favorable, not by how well the sensors themselves perform; a site operator worried about the economics of this approach should stress-test market assumptions before stress-testing sensor specifications.

6 Discussion

6.1 Classification accuracy is not the right sensor value metric, and the right sensor is not fixed

Section 5.4's central finding—that GPR is the strongest classification contributor but that only the magnetometer shows reliable economic value at baseline economics—argues against sizing a multimodal survey by F1 or macro-F1 alone. Section 5.10 goes further: the magnetometer's advantage survives ordinary price fluctuations but not deliberate changes in what the site actually contains, and GPR or EMI become the economically optimal channel under polymer-rich, debris-rich or glass-rich archetypes. The general, falsifiable claim this paper supports is therefore that predictive utility is not decision utility and that the economically optimal sensor is a function of site composition, market economics and sensor uncertainty jointly — not "magnetometers are economically most useful," which is only the baseline-economics special case. A sensor package for this application should be chosen and priced against its contribution to the value-of-information calculation for the specific material-price structure and composition of the site, not against a generic classification benchmark or a single assumed price table. This is a design implication that the original proposal's Stage 1 testing, reporting only F1 scores under one assumed economic table, could not have surfaced.

6.2 What the spatial-validation null result does and does not show

We tested, rather than assumed, whether random splitting overestimates performance for this system and found no detectable effect (Section 5.2). This should not be read as evidence that spatial blocking is unnecessary in general; the mechanism in the literature [7, 9] depends on a model exploiting local spatial proximity, and our feature set (per-cell sensor readings plus raw coordinates) evidently gives XGBoost, random forest and logistic regression little opportunity to do so. A design that adds explicit neighborhood-averaged features, or a real dataset whose noise correlation structure differs from our assumptions, could behave differently, and spatial blocking costs nothing to keep as standard practice.

6.3 Plastic/glass limitation arises from assumed sensor-response physics, not from classifier tuning

The proposal's diagnosis—that plastic and glass are weakly magnetic and poorly distinguished by this channel set and that no amount of classifier tuning fixes this—is supported by our results: three different classifier families fail on these two classes by a similar margin (Table 3), and adding a channel with distinct plastic/glass responses (NIR) improves plastic specifically without improving glass (Section 5.5). This finding points to plastic and glass needing materially different remedies (for example, glass may need a different spectral band or a density-based separation step) rather than one "optical fix" for both.

6.4 Limits of the economic model

The VOI conclusion in Section 5.7 depends on the price and recovery fraction ranges of Table 2, explicitly labeled illustrative, and on treating the excavate/do-not-excavate decision as a single per-cell

threshold rather than a portfolio decision across the whole site. Real secondary-material markets are more volatile than the ranges swept here, and the practical strength of the framework, as the source proposal itself notes, is that the VOI can be recomputed under site- and market-specific inputs rather than relying on any single value estimate.

6.5 Relation to established landfill mining economics

The finding that survey-guided decisions can convert an unprofitable blind-excavation baseline into a profitable baseline is consistent with the broader literature's finding that landfill-mining profitability is highly sensitive to a small number of factors that are, in principle, knowable before excavation [1, 2]. This paper's contribution is to make "how much would it be worth to know" a computable, falsifiable quantity for a specific sensing architecture, rather than a general observation about uncertainty.

6.6 Limitations

Every quantitative result in this paper is a prediction from a synthetic, physically motivated but unvalidated model. Sensor-response coefficients (Table 1), noise levels, dropout rates, spatial correlation lengths, and every economic parameter (Table 2) are stated assumptions, not measurements; the robustness sweep (Section 5.8) shows that the central economic conclusion depends on assumptions held within roughly the first 98% of the degradation range explored. The gas-hazard truth label is constructed from the same underlying model as the gas sensor reading, so the reported hazard performance is an upper bound on what independent ground-truth gas testing would show. The three-depth-layer material field is a simplification of real, continuously heterogeneous stratigraphy. No field, laboratory, or test-bed data were used in this study.

6.7 Load-bearing assumptions

Four assumptions carry most of the conclusions in this paper, and different conclusions lean on different subsets of them; thus, a failure in one needs not undermine the others.

A1 — synthetic sensor signatures approximate real landfill sensor separability. Load-bearing for every classification result (Sections 5.1–5.3, 5.5): the metal/organic/C&D-high, plastic/glass-low F1 pattern and the GPR-dominant ablation ranking all depend on it. If real materials are less separable than modeled, classification F1 (and everything downstream of it) will be lower than reported.

A2 — the material–price/recovery–fraction distributions represent the intended decision environment. Load-bearing for every economic result (Sections 5.4, 5.7–5.11): the baseline VOI advantage of the magnetometer, the break-even survey cost and the beta axis of the decision-viability map depend on it directly. Section 5.10 explicitly describes this dependence by varying it.

A3—the spatial structure of the simulator approximates real landfill heterogeneity. Load-bearing mainly for the spatial-vs-random null result (Section 5.2) and the distance-decay diagnostic; a real site with different spatial correlations in its material field or sensor noise could show a leakage gap that this simulation did not.

A4 — The predicted probabilities are sufficiently well calibrated for expected-value arithmetic. Load-bearing for the economic layer's use of predicted probabilities rather than only predicted classes; Section 5.9 tests this directly and finds that the conclusion of the VOI is not sensitive to it in this simulation, which weakens, but does not remove, its load-bearing status.

7 Field validation and falsification plan

The predictions above can be confirmed or overturned by a staged test program. Table 6 lists the measurements, the test bed, and the acceptance criterion for each.

Prediction	Test	Failure criterion
Metal, organic, C&D F1 remain high on real waste	Controlled test bed with known buried materials of each class	F1 below the values in Table 3 by more than the robustness sweep's noise-equivalent margin (Fig. 9)
Plastic/glass remain the hard classes without an optical channel	Same test bed, five-channel configuration	Unexpectedly high plastic/glass F1 would indicate the synthetic physics undermodeled real discriminability
NIR improves plastic but not necessarily glass	Same test bed, six-channel configuration with an added optical/NIR unit	Glass F1 improves as much as plastic (would contradict Section 5.5); or neither improves (would contradict the source proposal's diagnosis)
Sensor economic ranking follows site composition (Section 5.10), not a fixed classification-accuracy ranking	Ablation repeated on field data with the same economic model applied to real prices and the site's actual measured composition	Field outcome inconsistent with simulation: GPR (or another nonmagnetometer channel) produces a reliably positive incremental VOI exceeding the magnetometer's on a site the simulation's regime map would have classified as metal-dominant
Spatial vs. random validation gap remains small	Repeat the spatial-block/random-split comparison on field data	A gap appears whose 95% CI excludes zero, indicating the null result of Section 5.2 does not transfer
Gas-hazard recall remains high at a safety operating point	Independent, ground-truthed gas testing (not derived from the same sensor)	Recall below 90% at the chosen operating point
VOI remains positive at the site's actual survey cost	Full field pilot with measured excavation and processing costs and actual secondary-material prices	Measured VOI is negative, or the break-even survey cost is below the field survey's actual cost

Table 6. Field validation and falsification criteria. A pilot at the full site scale is proposed only after the controlled test-bed stage has passed.

7.1 Status of claims and assumptions

Every claim in this paper falls into one of four evidentiary categories; conflating them is the most common way a simulation study overstates itself, so Table 7 explicitly states each claim's status rather than leaving it implicit in the prose.

Claim	Status	Evidence/basis
Random splitting inflates classification	Not supported in simulation	95% CI includes zero for LR/RF/XGB

performance relative to spatial blocking		(Section 5.2)
GPR contributes most to macro-F1 among the five channels	Simulation result	Leave-one-sensor-out ablation (Section 5.3)
Magnetometer contributes the most reliable incremental VOI at baseline economics	Simulation result, economics-conditional	20-survey incremental-VOI analysis (Section 5.4); shown conditional in Section 5.10
NIR improves plastic classification	Conditional simulation result	Assumed synthetic NIR response coefficients (Table 1), not measured spectra
NIR increases VOI	Not supported	Incremental VOI statistically indistinguishable from zero (Section 5.5)
Baseline median VOI is positive	Model-dependent numerical result	Monte Carlo economic model, stated price/recovery/cost ranges (Section 5.7)
VOI becomes negative under severe combined degradation	Model-dependent boundary	Robustness sweep and 2-D decision-viability map (Sections 5.8, 5.11)
The economically optimal sensor depends on site composition, not just price	Supported within tested simulation	Economic-regime grid and extreme-archetype test (Section 5.10)
Classifier choice for the economic layer does not change the VOI conclusion	Supported within tested simulation	RF vs. raw/calibrated XGBoost comparison (Section 5.9)
Field performance will match these simulated values	Untested prediction	Requires the prospective validation in Table 6

Table 7. Status of claims and assumptions. "Simulation result" means the finding is reproducible from the released code under the stated assumptions; it is not thereby a claim about real landfills, which is what Table 6 exists to test.

8 Conclusions

A multimodal sensor-fusion architecture for preexcavation landfill characterization was evaluated by simulation against the questions it raises rather than the claims originally made for it. Metal, organic material and construction and demolition debris are classified with high simulated F1 (up to 0.99) under a spatially valid protocol; plastic and glass are not (F1 0.12–0.32) because of the assumed sensor–response physics rather than classifier choice. Random per-cell splitting did not detectably inflate performance relative to spatial blocking in this configuration, a tested and reported null result rather than an assumption. The paper's central finding is that predictive utility is not decision utility and is not fixed: GPR maximizes classification accuracy, but the economically optimal sensor at baseline economics is the magnetometer, and it changes to GPR or EMI under deliberately different, still economically nameable site compositions (Section 5.10) — optimal sensor selection is a function of site composition, market economics and sensor uncertainty, not a fixed property of the sensing architecture. Pooling 40 independent survey realizations under baseline economics, the simulated survey has a positive median VOI ($924) and a break-even cost of $1451, and this classifier-layer choice was itself tested and found not to matter (Section 5.9). A two-dimensional decision-viability map (Section 5.11) replaces the single combined-degradation boundary with a full VOI(α,β) surface, showing that the economic conclusion is considerably more sensitive to adverse market conditions than to sensor degradation itself. All the results are predictions, explicitly labeled by evidentiary status in Table 7, and the field validation plan (Table 6) specifies the measurements that would confirm or overturn them.

Declarations

Funding. This work received no external funding.

Competing interests. The author declares that there are no competing interests.

Ethics approval. Not applicable. No human participants or animals were involved.

Consent to participate/for publication. Not applicable.

Prior dissemination. An earlier, nonquantitative description of this architecture was submitted by the author in response to an industrial open-innovation challenge focused on landfill mining. That submission was not peer reviewed or published and contained none of the models, validation protocol, or results reported here.

Data availability. No field data were generated. All the model inputs are stated in the paper and the released configuration file.

Code availability. The synthetic survey generator, classification and economic models, robustness sweep, decision-viability and sensor-selection maps, and figure generators are openly available at https://github.com/sandlerleon/landmap-sensor-fusion and are archived on Zenodo (DOI 10.5281/zenodo.22984668). Run landmap_model.py, landmap_econ.py, landmap_robustness.py, landmap_hazard_extra.py, landmap_calib_voi.py, landmap_sensor_map.py, landmap_sensor_map_extreme.py, landmap_voi_map.py and make_figures.py reproduces every number and figure.

Author contributions. L.S. conceived the study, developed the models, performed the analysis and wrote the manuscript.

Use of generative AI. During the preparation of this work, the author used Claude (Anthropic) to assist with the literature search, implementation of the models and drafting of the text. The author reviewed and edited all the content, verified every cited reference against its source, and takes full responsibility for the content of the article.

References

1. Laner D, Esguerra JL, Krook J, Horttanainen M, Kriipsalu M, Rosendal RM, et al. (2019) Systematic assessment of critical factors for the economic performance of landfill mining in Europe: What drives the economy of landfill mining?. Waste Management 95:674-686. https://doi.org/10.1016/j.wasman.2019.07.007

2. Danthurebandara M, Van Passel S, Vanderreydt I, Van Acker K (2015) Assessment of environmental and economic feasibility of Enhanced Landfill Mining. Waste Management 45:434-447. https://doi.org/10.1016/j.wasman.2015.01.041

3. De Donno G, Melegari D, Paoletti V, Piegari E (2024) Electrical and Electromagnetic Prospecting for the Characterization of Municipal Waste Landfills: A Review. In: Springer Water. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-52633-6_1

4. Jansen J, Haddad B, Fassbender W, Jurcek P (1992) Frequency Domain Electromagnetic Induction Sounding Surveys for Landfill Site Characterization Studies. Groundwater Monitoring & Remediation 12(4):103-109. https://doi.org/10.1111/j.1745-6592.1992.tb00068.x

5. Piero Deidda G, De Carlo L, Clementina Caputo M, Cassiani G (2022) Frequency domain electromagnetic induction imaging: An effective method to see inside a capped landfill. Waste Management 144:29-40. https://doi.org/10.1016/j.wasman.2022.03.007

6. Zhang T, Zhang D, Zheng D, Guo X, Zhao W (2022) Construction waste landfill volume estimation using ground penetrating radar. Waste Management & Research: The Journal for a Sustainable Circular Economy 40(8):1167-1175. https://doi.org/10.1177/0734242x221074114

7. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913-929. https://doi.org/10.1111/ecog.02881

8. Hijmans RJ (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93(3):679-688. https://doi.org/10.1890/11-0826.1

9. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, et al. (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11(1):4540. https://doi.org/10.1038/s41467-020-18321-y

10. Breiman L (2001) Random Forests. Machine Learning 45(1):5-32. https://doi.org/10.1023/a:1010933404324

11. Chen T, Guestrin C (2016) XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. ACM. https://doi.org/10.1145/2939672.2939785

12. Orynbaikyzy A, Gessner U, Conrad C (2022) Spatial Transferability of Random Forest Models for Crop Type Classification Using Sentinel-1 and Sentinel-2. Remote Sensing 14(6):1493. https://doi.org/10.3390/rs14061493

13. Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05, 625-632. ACM Press. https://doi.org/10.1145/1102351.1102430

14. Jones PT, Geysen D, Tielemans Y, Van Passel S, Pontikes Y, Blanpain B, et al. (2013) Enhanced Landfill Mining in view of multiple resource recovery: a critical review. Journal of Cleaner Production 55:45-55. https://doi.org/10.1016/j.jclepro.2012.05.021

15. Vollprecht D, Machiels L, Jones PT (2021) The EU Training Network for Resource Recovery through Enhanced Landfill Mining—A Review. Processes 9(2):394. https://doi.org/10.3390/pr9020394

16. Yi S (2019) Resource recovery potentials by landfill mining and reclamation in South Korea. Journal of Environmental Management 242:178-185. https://doi.org/10.1016/j.jenvman.2019.01.101

17. Rendu JM (1976) Bayesian Decision Theory Applied to Mineral Exploration and Mine Valuation. In: Advanced Geostatistics in the Mining Industry. Springer Netherlands. https://doi.org/10.1007/978-94-010-1470-0_28

18. Luhede A, Verma P, Upmann T (2025) Value of Information analysis for marine conservation decision making. Environmental Science & Policy 171:104164. https://doi.org/10.1016/j.envsci.2025.104164

19. Huang N, Lu G, Xu D (2016) A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest. Energies 9(10):767. https://doi.org/10.3390/en9100767