SURVEY PROTOCOL CARDS FOR CROP MAPS

Akram Zaytar[1] **Girmaw A. Tadesse**[1] **Caleb Robinson**[1] **Shabarinath S Nair**[2]
Gilles Q. Hacheme[1] Inbal Becker-Reshef[1] Rahul Dodhia[1] Juan Lavista Ferres[1]
[1]Microsoft AI for Good [2]NASA Harvest

ABSTRACT

Crop type maps underpin food security decisions yet their accuracy depends on label quality, which in turn depends on survey design choices made under tight budgets. Survey planners must allocate limited resources across GPS devices, stratification strategies, sample size, worker training, and verification protocols, but lack quantitative guidance on which investments yield quality crop maps. We address this gap by modeling the full chain from survey design to downstream crop detection accuracy: survey choices map to costs, costs constrain achievable label noise levels, and noise levels affect crop mapping performance. We implement 17 noise functions grounded in documented errors from JECAM, World-Cereal, and LSMS-ISA, and measure degradation on 2 datasets: EuroCrops and Zambia. Our experiments reveal that label verification matters far more than GPS accuracy: crop misidentification causes up to 99% F1 loss while 30m GPS jitter causes only 4%. Dataset-specific noise-to-performance surrogate models achieve R^2=0.87, enabling millisecond what-if queries—but cross-dataset transfer shows mixed results: Spearman ρ=0.32–0.60 indicates rankings transfer asymmetrically, and negative R^2 reveals degradation predictions fail across contexts. We package these findings into a programmable protocol-card and web interface that optimizes survey design given budget constraints.

1 INTRODUCTION

Satellite-based crop type mapping has advanced rapidly with missions like Sentinel-2, yet persistent challenges around survey quality limit deployment. Consider a survey planner with budget B choosing between GPS devices, stratification strategies, sample sizing, and label verification protocols—without quantitative guidance, budgets are allocated by following common practices. Formally, we frame this problem as finding the set of survey protocol choices that minimizes the gap $\Delta = \mathcal{L}_{\text{realized}} - \mathcal{L}_{\text{ideal}}$ between classification loss under budget-constrained collection and loss achievable with perfect labels. Global initiatives have made significant progress toward operational systems for crop mapping: Van Tricht et al. (2023) demonstrated the feasibility of global-scale, seasonal crop mapping, though validation revealed substantially lower accuracies in Africa due to reference data gaps and agricultural landscape complexity. Meanwhile, emerging low-cost labeling pipelines—helmet-mounted cameras (Nakalembe et al., 2025), drive-by imagery (Paliyam et al., 2021), and smartphone crowdsourcing—greatly expand training data availability but introduce new noise sources: approximate geolocation, automated classification errors, and ambiguity from intercropping. Most directly relevant to our work, Azzari et al. (2021) systematically evaluated how survey choices affect crop classification accuracy in Malawi and Ethiopia, finding that georeferencing method quality causes 8–24% overestimation of maize area—seemingly small accuracy differences translate to 0.16–0.47 million hectares of error, showing that label collection methodology profoundly impacts downstream model performance.

To our knowledge, no existing work provides: (a) a systematic taxonomy of survey label noise encompassing GPS/centroid shifts, crop misidentification, duplicate labels, and selection bias; (b) tolerance curves quantifying performance degradation under controlled noise injection; (c) comparison of noise sensitivity patterns across agricultural contexts (European commercial vs. African smallholder); or (d) a compact surrogate model supporting constrained optimization queries. Our protocol-card fills this gap by linking survey design choices to expected performance loss, enabling practitioners to make informed tradeoffs between collection effort and downstream accuracy. We

ask two questions: (1) **How does mapping performance degrade under realistic survey-style label noise?** We measure tolerance curves under controlled noise injection and compare sensitivity across European commercial and African smallholder contexts. (2) **Can we provide actionable recommendations for survey designers?** We train a surrogate model that predicts degradation for unseen protocol settings and supports budget-constrained optimization.

2 METHODS

2.1 DATA

We study centroid-based crop classification on 2 datasets: a self-declared European benchmark and an African smallholder dataset representing target deployment contexts. Each sample is a field record with (i) crop label $y_i \in \{1, \ldots, C\}$, (ii) a geometry g_i (polygon), and (iii) a 64-dimensional embedding x_i from AlphaEarth Foundations (Brown et al., 2025), 10 m annual composites extracted at the field centroid. We split 80/20 train/test and retain classes with $\geq$100 samples (EuroCrops) or all classes (Zambia), excluding intercrops. Given a dataset $\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^{N}$, we evaluate a multi-class classifier h_θ trained on embeddings x_i to predict y_i.

EuroCrops. We use the EuroCrops dataset[1], which combines self-declared crop reporting data from various European Union countries. The dataset uses a hierarchical crop type taxonomy (HCAT) that harmonizes national classification schemes across member states. Farmers self-declare crop types annually for EU Common Agricultural Policy subsidies, with administrative verification processes—making these labels among the highest-fidelity publicly available. We focus on French parcels from 2018 to align with available satellite embeddings, using 6,916 training and 1,730 test samples across 15 classes.

Zambia In-Situ. We use a smallholder field dataset from Zambia comprising 621 field polygons with crop type labels collected via in-situ surveys. The dataset exhibits characteristics typical of African smallholder agriculture: small field sizes, high intercrop prevalence, and a long-tailed class distribution dominated by maize. The dataset contains 465 training and 156 test samples across 21 classes, creating a challenging sparse classification setting. This dataset tests whether noise sensitivity patterns generalize to data-scarce smallholder contexts with complex cropping systems.

2.2 MODELS

Pipeline We split each dataset randomly into train and test fields, keeping the test set clean throughout. We acquire satellite embedding rasters covering the spatial extent of all field polygons. During noise simulation, a protocol configuration p corrupts training labels and geometries via noise functions; we then collapse each (possibly corrupted) polygon to its centroid, extract the embedding, and train an XGBoost classifier $h_{\theta(p)}$, chosen for its fast training time which enables large-scale exploration of noise configurations. We apply inverse-frequency class weighting ($w_c = N/(C \cdot n_c)$) and standardized embeddings; XGBoost uses `multi:softprob` with log-loss evaluation. We define $\mathrm{F1}(p)$ as the weighted F1 on the clean test set, so changes isolate the impact of survey-style errors on downstream performance.

Protocol Mapping. A survey's cost is the sum of its labor (number of enumerators over the collection days), equipment, training, verification, and deduplication costs. Furthermore, survey design choices map to noise parameters through relationships: GPS device accuracy determines polygon jitter scale (phone 15m, handheld 5m, survey-grade 0.5m) and neighbor swap radius; training hours reduce the probability of making label mistakes exponentially; verification level multiplies all error rates. We created configurable heuristics for such mappings while protocol card users could change them to reflect different cost structures or error relationships in the interface.

Noise functions. A survey protocol maps directly to a set of noise functions. Let $\mathcal{D}_{\text{train}}$ be the clean training split. A protocol configuration p induces a transformation $\widetilde{\mathcal{D}}_{\text{train}} = \mathcal{T}_p(\mathcal{D}_{\text{train}})$, where $\mathcal{T}_p$ applies stochastic noise functions with explicit rates and severities. Noise functions may modify (i) which samples are collected, (ii) where they are located, (iii) what crop label is recorded, and (iv)

[1] `https://github.com/maja601/EuroCrops`

whether duplicate/conflicting records exist. Table 2 describes our 17 noise functions. Each operator is parameterized by a *rate* (i.e., ratio of the dataset impacted) and one or more *severity* parameters (e.g., meters of jitter). Like data augmentation pipelines, our noise operators are composable for simulating protocols where multiple error sources co-occur.

2.3 Directed & Random Search

Tolerance curves For each noise family, we run controlled one-dimensional sweeps: we vary a single rate/severity parameter over its range while holding other protocol dimensions fixed. Each configuration is repeated over 3 random seeds; we report the mean and standard error. This yields tolerance curves $\mathrm{F1}(\lambda)$ characterizing sensitivity to each noise type.

Search We explore the protocol space using two complementary strategies. Random search serves as a baseline: we sample configurations by selecting subsets of noise functions and drawing their parameters from predefined ranges (Table 3), producing a broad ledger of heterogeneous corruption regimes. To concentrate trials near high-impact regions, we augment this with Bayesian optimization (Optuna TPE, 100 trials/mode) over the same search space, looking for configurations that minimize or maximize cross-entropy loss and thereby identifying sharp degradation boundaries unlikely to appear under uniform sampling. Crucially, we constrain parameter ranges to exclude extreme values (e.g., $>$20% label flip, $>$80% subsampling) that would trivially degrade performance but rarely occur in practice. All trials are logged in a structured table that records protocol configurations and resulting performance degradation values. This ledger is the training data for the surrogate model below.

2.4 Surrogate model

To support fast "what-if" queries for constrained optimization, we fit a surrogate model that predicts test loss degradation from protocol parameters. We encode a protocol configuration p as a feature vector $\phi(p)$ containing (i) which noise functions are active (binary indicators) and (ii) their numeric parameters (rates and severities). We then train an XGBoost regressor $\hat{f}: \phi(p) \mapsto \widehat{\Delta}(p)$ on the 29 features using 5-fold cross-validation for hyperparameter selection to predict test loss degradation $\Delta(p) = \mathcal{L}(p) - \mathcal{L}_{\text{clean}}$. We use the simulation ledger as supervised training data and evaluate generalization by holding out protocol configurations.

3 Results

Tolerance Curves We characterize model sensitivity through controlled one-dimensional sweeps on EuroCrops (European commercial agriculture, baseline F1=0.853) and Zambia (African smallholder fields, 465 training samples, 21 classes, baseline F1$\approx$0.68). Tolerance curves appear in Appendix D. We group noise types into three tiers: *catastrophic* operators (similar-crop confusion, geometry-label swap) cause 70–99% F1 loss universally at moderate intensities; *moderate-impact* operators show context-dependent sensitivity, with subsample and label flip causing $\sim$2$\times$ more degradation in sparse Zambia (5%, 10%) than data-rich EuroCrops (2%, 5%); *low-impact* operators (polygon jitter $<$5%, duplicates $<$1%, partial boundary $<$2%) remain robust across contexts. Road dropout shows moderate context dependency (3% in EuroCrops vs. 6% in Zambia).

Noise Search We use Bayesian optimization to both minimize and maximize test loss over the noise parameter space. Table 1 summarizes configurations and fANOVA importance rankings. Min-Deg identifies the most tolerable corruption ($<$1 point F1 loss in both datasets). For Max-Deg, both datasets share similar worst-case configurations: similar-crop confusion, subsampling, and partial boundary corruption cause maximum degradation—EuroCrops (baseline F1=0.853) loses 12.1 points while Zambia (baseline F1=0.684) loses 5.9 points. EC's larger degradation likely reflects its higher baseline (higher quality) and larger training set (6,916 vs. 465 samples); ZM's small noisy sample size introduces variance that limits directed search. Feature importance reveals that similar-crop confusion dominates max-degradation in both datasets (77% in EC, 27% in ZM), with context-specific secondary factors: EuroCrops is additionally sensitive to subsampling and polygon jitter, while Zambia shows sensitivity to partial boundary corruption and subsampling.

Dataset	Config	Noise Type	Intensity	F1	Importance
EC	Baseline	—	—	0.853	—
	Min	Partial Bound. Duplicate Similar Crop	0.35 (16%) 9% 0.33 (7%)	0.847	Neigh. Swap: 20.7 Label Flip: 19.1 Wtd. Subsamp.: 11.8
	Max	Similar Crop Subsample Poly. Jitter	0.31 (40%) 84% 3.2m (52%)	0.732	Similar Crop: 77.3 Subsample: 5.6 Poly. Jitter: 2.4
ZM	Baseline	—	—	0.684	—
	Min	Similar Crop Label Flip Poly. Jitter	0.52 (22%) 17% 9.1m (95%)	0.675	Wtd. Subsamp.: 28.0 Confl. Dup.: 21.5 Subsample: 8.3
	Max	Subsample Similar Crop Partial Bound.	51% 0.59 (27%) 0.40 (45%)	0.624	Similar Crop: 27.1 Partial Bound.: 12.4 Subsample: 12.1

Table 1: Noise search results. Min-Deg finds tolerable noise; Max-Deg finds worst-case configurations (top 3 noise types shown). Importance column shows fANOVA rankings for each search direction.

Implications for Survey Design Label verification dominates: similar-crop confusion causes catastrophic failure (>95% F1 loss) universally. Protocols that separate visually similar crops (wheat/barley, maize/sorghum) deliver the highest returns. Geometry-label swap causes 70–92% degradation, making database integrity checks that verify parcel-to-label linkages critical. GPS accuracy requirements are modest—polygon jitter up to 30m causes only 2–4% degradation, confirming consumer-grade GPS suffices for centroid-based classification. Sample size matters more for sparse datasets: subsampling causes 2% degradation in EuroCrops but 5% in Zambia, so smallholder surveys should prioritize quality filtering alongside coverage. Duplicate detection has low priority (<1% harm) and can be deprioritized relative to label verification.

Surrogate Model To enable fast protocol queries without re-running simulations, we train dataset-specific regressors to predict Δloss from noise configurations. Each model takes 29 features: coverage rates and intensity parameters for active noise types. We use 5-fold cross-validation to tune hyperparameters and evaluate on held-out configurations. The EuroCrops surrogate achieves R^2=0.87, while the Zambia surrogate achieves R^2=0.57—reflecting higher variance in the sparse, multi-class setting. Cross-dataset evaluation yields Spearman ρ=0.32 (EC$\rightarrow$ZM) and ρ=0.60 (ZM$\rightarrow$EC), showing asymmetric transfer: ZM$\rightarrow$EC rankings transfer moderately while EC$\rightarrow$ZM transfer is weak. Negative R^2 in cross-dataset transfer reflects distribution shift in absolute degradation magnitude.

4 Protocol Card

We package the learned surrogate into a programmable *protocol card* that takes a budget cap B and cost primitives—per-sample cost, GPS cost model (linked to geolocation error), revisit multiplier, and polygon-vs-point collection ratio—and solves $p^\star(B) = \arg\min_p \widehat{\Delta}(p)$ s.t. $\mathrm{cost}(p) \leq B$, where $\mathrm{cost}(p)$ composes primitives into protocol-level cost. The card returns: recommended protocol configuration, predicted degradation $\widehat{\Delta}$ with uncertainty bounds, and tolerance curves for relevant noise families. Only relative degradation matters for protocol comparison. Figure 1 shows the interactive interface.

5 Conclusion

We presented a protocol-card framework linking survey design choices to crop mapping performance. While we quantify how noise types degrade performance, the mapping from survey design (training hours, salary, data cleaning effort) to noise levels remains heuristic—these human factors vary across contexts. Furthermore, surrogate models do not transfer across regions, limiting practical impact without a catalog of region-specific surrogates with flexible survey configuration. Next, we aim to address both gaps: gathering empirical survey data from field partners to calibrate survey-to-noise relationships, and expanding surrogates across geographies to disentangle universally important design factors from context-dependent ones.

REFERENCES

George Azzari, Shruti Jain, Graham Jeffries, Talip Kilic, and Siobhan Murray. Understanding the requirements for surveys to support satellite-based crop type mapping: Evidence from sub-Saharan Africa. *Remote Sensing*, 13(23):4749, 2021. doi: 10.3390/rs13234749. URL https://www.mdpi.com/2072-4292/13/23/4749.

Hendrik Boogaard, Arun Kumar Pratihast, Juan Carlos Laso Bayas, Santosh Karanam, Steffen Fritz, Kristof Van Tricht, Jeroen Degerickx, and Sven Gilliams. Building a community-based open harmonised reference data repository for global crop mapping. *PLOS ONE*, 18(7):e0287731, 2023. doi: 10.1371/journal.pone.0287731. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0287731.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Calogero Carletto, Sydney Gourlay, and Paul Winters. From guesstimates to GPStimates: Land area measurement and implications for agricultural analysis. *Journal of African Economies*, 24(5): 593–628, 2015. doi: 10.1093/jae/ejv011.

Arthur Elmes, Hamed Alemohammad, Ryan Avery, Kelly Caylor, J. Ronald Eastman, Lewis Fishgold, Mark A. Friedl, Meha Jain, Divyani Kohli, Juan Carlos Laso Bayas, David Luber, Jessica McCarty, Robert Gilmore Pontius Jr., and Curtis E. Woodcock. Accounting for training data error in machine learning applied to Earth observations. *Remote Sensing*, 12(6):1034, 2020. doi: 10.3390/rs12061034. URL https://www.mdpi.com/2072-4292/12/6/1034.

Joint Experiment for Crop Assessment and Monitoring. JECAM guidelines for cropland and crop type definition and field data collection. Technical report, GEOGLAM, 2018. URL https://www.jecam.org/.

David B. Lobell, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. Eyes in the sky, boots on the ground: Assessing satellite- and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, 102(1):202–219, 2020. doi: 10.1093/ajae/aaz051.

Catherine Nakalembe, Ivan Zvonkov, Hannah Kerner, et al. Helmets labeling crops: Kenya crop type dataset created via helmet-mounted cameras and deep learning. *Scientific Data*, 2025.

Aigerim Orynbaikyzy, Ursula Gessner, and Christopher Conrad. Crop type classification using fusion of Sentinel-1 and Sentinel-2 data: Assessing the impact of feature selection, object size, parcel boundary, and class label noise. *Remote Sensing*, 12(17):2779, 2020. doi: 10.3390/rs12172779.

Megha Paliyam, Catherine L. Nakalembe, Kaiyu Liu, Ritvik Nyiawung, and Hannah R. Kerner. Street2sat: A machine learning pipeline for generating ground-truth geo-referenced labeled datasets from street-level images. In *ICML Workshop on Tackling Climate Change with Machine Learning*, 2021.

Maja Schneider, Amelie Broszeit, and Marco Körner. EuroCrops: A pan-European dataset for time series crop type classification. *arXiv preprint arXiv:2106.08151*, 2021.

Kristof Van Tricht, Jeroen Degerickx, Sven Gilliams, Daniele Zanaga, Marjorie Battude, Alex Grosu, Joost Brombacher, Myroslava Lesiv, Juan Carlos Laso Bayas, Santosh Karanam, Steffen Fritz, Inbal Becker-Reshef, Belen Franch, Belén Mollà-Bononad, Hendrik Boogaard, Arun Kumar Pratihast, Benjamin Koetz, and Zoltan Szantoi. WorldCereal: A dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, 15:5491–5515, 2023. doi: 10.5194/essd-15-5491-2023. URL https://essd.copernicus.org/articles/15/5491/2023/.

A NOISE TAXONOMY

Operator	Description	Evidence
Label Errors		
Label flip	Random label reassignment due to misidentification at visit	Joint Experiment for Crop Assessment and Monitoring (2018); Van Tricht et al. (2023)
Similar-crop	Confusion between similar crops (wheat/barley, maize/sorghum)	Orynbaikyzy et al. (2020)
Non-crop	Fallow or infrastructure incorrectly labeled as crop	Van Tricht et al. (2023)
Geometry Errors		
Polygon jitter	Gaussian noise on coordinates from GPS positioning error	Azzari et al. (2021)
To-the-edge	Shift toward polygon boundary from boundary-walking bias	Orynbaikyzy et al. (2020); Lobell et al. (2020)
Partial bound.	Drop polygon vertices from incomplete boundary capture	Lobell et al. (2020); Schneider et al. (2021)
Geom-label swap	Geometry linked to wrong label via adjacent record linkage	Elmes et al. (2020)
Spatial Bias		
To-the-road	Shift toward road from oversampling near roads	Van Tricht et al. (2023); Joint Experiment for Crop Assessment and Monitoring (2018)
Road dropout	Remove samples far from roads due to remote undersampling	Azzari et al. (2021); Carletto et al. (2015)
Neighbor bleed	Expand polygon toward neighbor via adjacent label propagation	Elmes et al. (2020)
Neighbor swap	Swap labels between nearby fields due to spatial confusion	Elmes et al. (2020)
Class Imbalance		
Weaken minor.	Drop minority samples; rare crops underrepresented	Schneider et al. (2021); Azzari et al. (2021)
Swap w/ larger	Replace minority with dominant class label	Joint Experiment for Crop Assessment and Monitoring (2018); Van Tricht et al. (2023)
Subsample	Uniform size reduction from reduced sample collection	Azzari et al. (2021); Carletto et al. (2015)
Wtd. subsamp.	Class-weighted retention for non-uniform sampling	Lobell et al. (2020)
Data Quality		
Duplicate	Exact row copies from same field recorded twice	Joint Experiment for Crop Assessment and Monitoring (2018); Orynbaikyzy et al. (2020)
Conflict. dup.	Same location with different labels from disagreement	Boogaard et al. (2023)

Table 2: Noise taxonomy: 17 operators with descriptions and literature evidence.

B PROTOCOL CARD INTERFACE

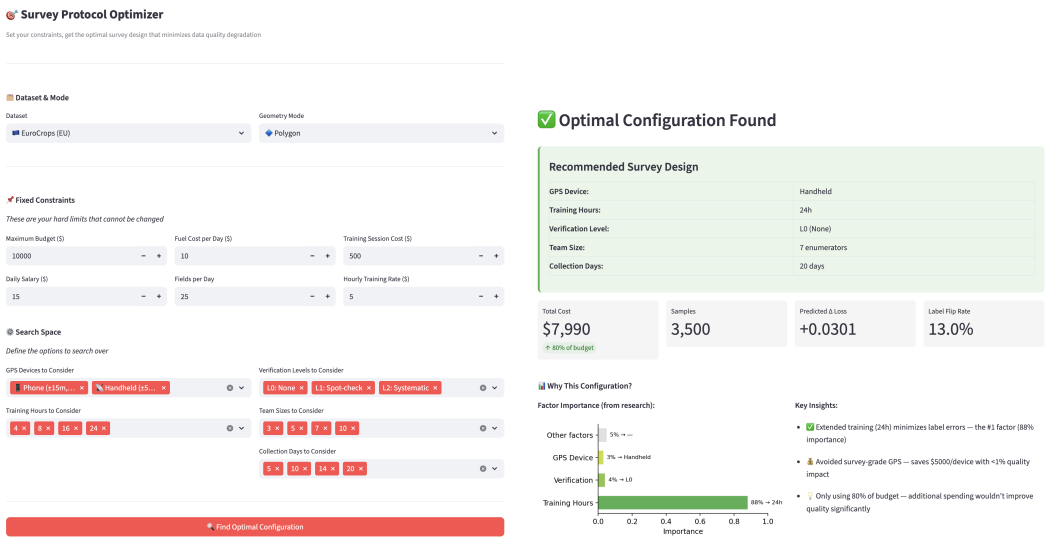


Figure 1: Interactive protocol-card interface. Users configure survey design choices and budget constraints (left); the tool maps them through the cost model and surrogate to produce an optimized protocol recommendation with predicted performance impact (right).

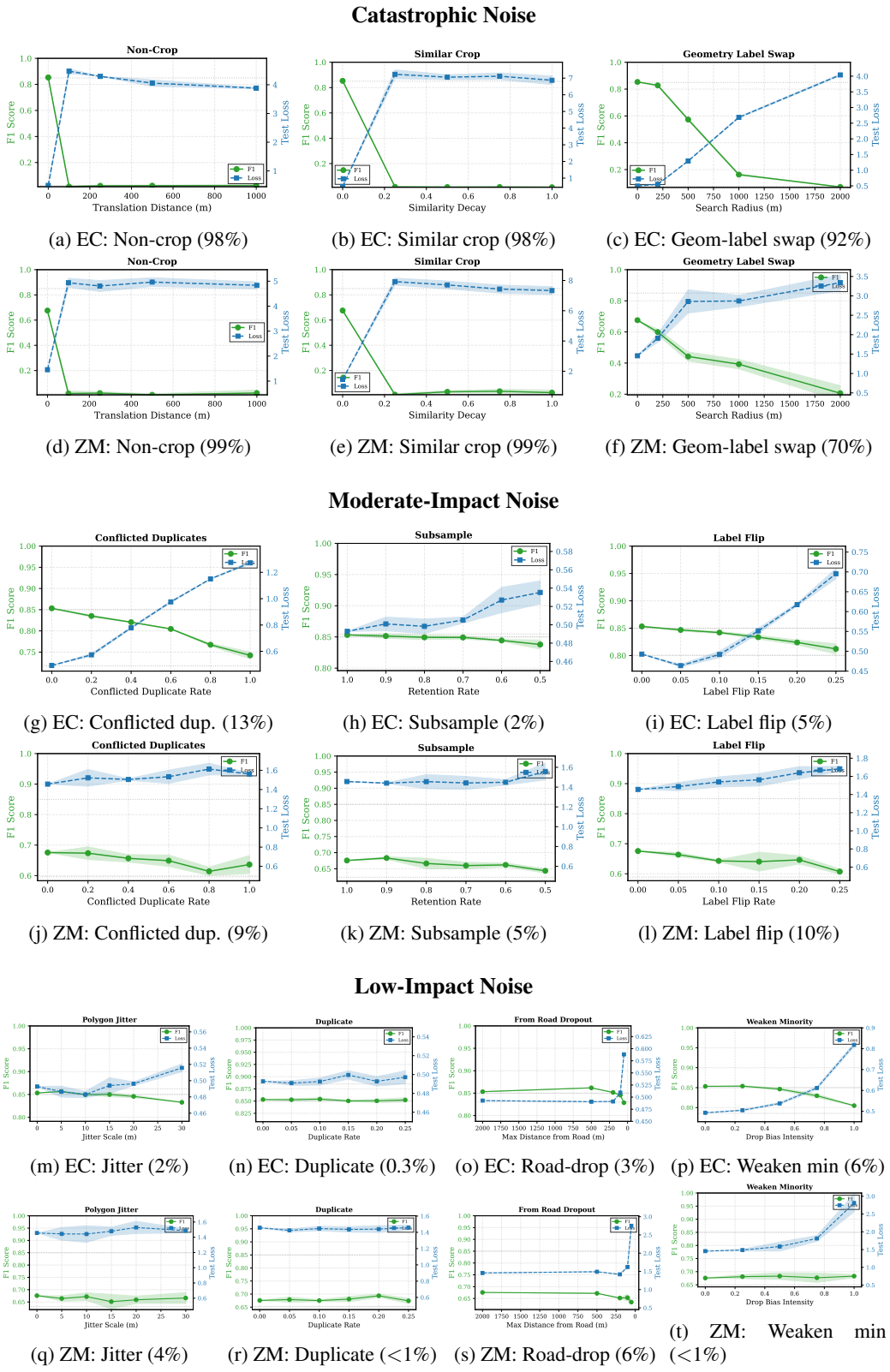
C SEARCH SPACE

Table 3 lists the parameter bounds used in both random and Bayesian search. Each trial activates a random subset of noise types and samples parameters uniformly within these bounds.

Noise Type	Coverage Range	Intensity Range	Unit
Label flip	0.05–0.25	—	—
Similar-crop	0.05–0.40	0.3–0.8	confusion prob.
Polygon jitter	0.10–1.00	3–15	meters
To-the-edge	0.10–0.50	0.1–0.5	fraction
To-the-road	0.10–0.50	50–500	meters
Partial boundary	0.10–0.50	0.1–0.4	fraction
Neighbor bleed	0.10–0.50	5–30	meters
Non-crop	0.01–0.10	10–100	meters
Subsample	0.50–1.00	—	retention rate
Wtd. subsample	0.10–0.50	—	—
Road dropout	0.10–0.50	100–1000	meters
Swap w/ larger	0.05–0.20	100–1000	meters
Weaken minority	0.10–0.50	0.2–0.6	drop fraction
Duplicate	0.01–0.15	—	—
Neighbor swap	0.01–0.10	10–100	meters
Confl. duplicates	0.01–0.10	—	—
Geom-label swap	0.01–0.10	50–200	meters

Table 3: Noise parameter search space. Coverage is the fraction of training samples affected; intensity controls error severity.

D TOLERANCE CURVES



(a) EC: Non-crop (98%) (b) EC: Similar crop (98%) (c) EC: Geom-label swap (92%)

(d) ZM: Non-crop (99%) (e) ZM: Similar crop (99%) (f) ZM: Geom-label swap (70%)

(g) EC: Conflicted dup. (13%) (h) EC: Subsample (2%) (i) EC: Label flip (5%)

(j) ZM: Conflicted dup. (9%) (k) ZM: Subsample (5%) (l) ZM: Label flip (10%)

(m) EC: Jitter (2%) (n) EC: Duplicate (0.3%) (o) EC: Road-drop (3%) (p) EC: Weaken min (6%)

(q) ZM: Jitter (4%) (r) ZM: Duplicate (<1%) (s) ZM: Road-drop (6%) (t) ZM: Weaken min (<1%)

Figure 2: **Tolerance curves across noise types.** Top: Catastrophic noise causes near-complete failure (>95% F1 drop). Middle: Moderate-impact noise (9–13% for conflicted duplicates). Bottom: Low-impact noise (<5% degradation). Percentages indicate maximum F1 degradation at full noise injection.