Survey Protocol Cards for Crop Maps

Akram Zaytar Girmaw A. Tadesse Caleb Robinson Shabarinath S Nair Gerald Blasch
Jeroen Degerickx Mitelo Subakanya Juan Carlos Laso Bayas Gilles Q. Hacheme
Inbal Becker-Reshef Rahul Dodhia Juan Lavista Ferres

Postprint — author's accepted manuscript. This is the authors' peer-reviewed, accepted version of a manuscript published in *IEEE Geoscience and Remote Sensing Letters*, posted on EarthArXiv in accordance with IEEE's author-posting policy. It is *not* the publisher's final typeset Version of Record.

Published version (please cite): A. Zaytar *et al.*, "Survey Protocol Cards for Crop Maps," *IEEE Geoscience and Remote Sensing Letters*, 2026, doi: 10.1109/LGRS.2026.3708987.

EarthArXiv preprint DOI: 10.31223/x5wr1f.

Corresponding author ORCID: Akram Zaytar — https://orcid.org/0009-0003-7498-5260

SURVEY PROTOCOL CARDS FOR CROP MAPS

Akram Zaytar[1*] **Girmaw A. Tadesse**[1] **Caleb Robinson**[1] **Shabarinath S Nair**[2]
Gerald Blasch[3] Jeroen Degerickx[4] Mitelo Subakanya[3] Juan Carlos Laso Bayas[5]
Gilles Q. Hacheme[1] Inbal Becker-Reshef[1] Rahul Dodhia[1] Juan Lavista Ferres[1]
[1]Microsoft AI for Good Research Lab [2]NASA Harvest [3]CIMMYT [4]VITO [5]IIASA
[*]Corresponding author: akramzaytar@microsoft.com

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 the agricultural survey literature, and measure degradation on two 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 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. 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; 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 two datasets: a self-declared European benchmark (EuroCrops (Schneider et al., 2021)) and an African smallholder dataset (Zambia) 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 (Schneider et al., 2021), 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) whether duplicate/conflicting records exist. Table 3 in the Appendix 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 4), 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 dataset (baseline $F1 = 0.853$) and Zambia(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).

Representation Robustness To verify these tiers are not an artifact of a single feature representation (AlphaEarth (Brown et al., 2025)), we repeat the tolerance sweeps with two alternative

Dataset	Config	Noise Type	Intensity	F1	Importance
EuroCrops	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
Zambia	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.

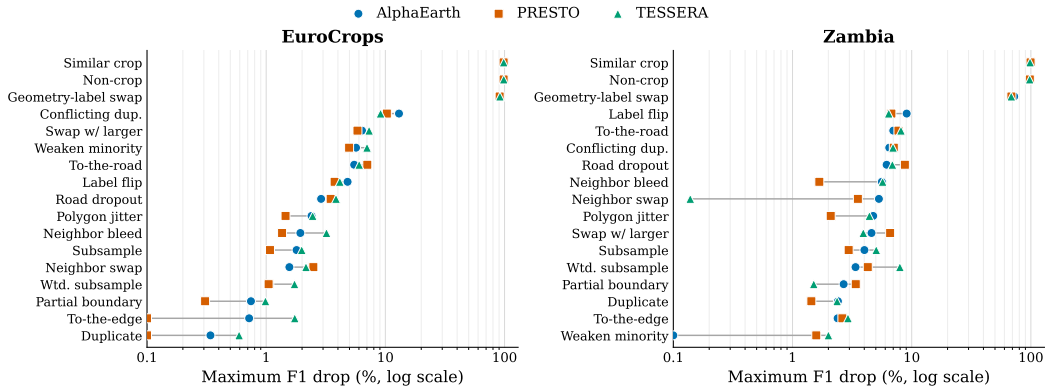


Figure 1: Per-operator noise sensitivity (maximum F1 drop) across three feature representations—AlphaEarth (AEF), PRESTO, and TESSERA—for EuroCrops (left) and Zambia (right), on a logarithmic axis. Each operator carries one marker per representation (circle/square/triangle); the grey connector spans the three values, so a short connector denotes strong cross-representation agreement. Operators are ordered by AEF sensitivity. The catastrophic top tier (similar-crop confusion, non-crop, geometry-label swap) and the low-impact bottom tier are preserved across all three representations; representation choice shifts absolute magnitudes but not the ranking.

earth-observation embeddings—PRESTO (Tseng et al., 2023) (a temporal Sentinel-1/2 encoder) and TESSERA (Feng et al., 2025) (annual foundation embeddings)—holding the XGBoost classifier and all 17 noise operators fixed. Both embeddings are produced as dense rasters over the same areas of interest and years and sampled at the (perturbed) field centroid exactly as in the main pipeline.[1] The per-operator sensitivity ranking is strongly preserved across all three representations under two sensitivity definitions—the maximum F1 drop across the sweep, and the F1 drop at a fixed reference intensity: pairwise Spearman $\rho = 0.98$–1.00 (EuroCrops) and $\rho = 0.86$–0.88 (Zambia) at the fixed reference intensity, and $\rho \geq 0.97$ and $\rho \geq 0.78$ respectively by maximum F1 drop (all $p < 10^{-4}$). The catastrophic operators (non-crop, similar-crop confusion, geometry-label swap) rank highest for every representation in both datasets (Table 2, Fig. 1). Only absolute magnitudes and baseline F1 shift between representations; the noise-sensitivity tiering that drives our survey-design recommendations is representation-invariant.

[1]For Zambia, TESSERA uses its 2024 embedding (2025 was unavailable for this AOI); the 2024/25 season spans both calendar years, and TESSERA's clean baseline F1 (0.69) matches the season-aligned representations, so the offset does not affect the ranking.

Table 2: Cross-representation Spearman rank correlation (ρ) of per-operator noise sensitivity, under two sensitivity definitions (maximum F1 drop across the sweep; F1 drop at a fixed reference intensity). High ρ under both indicates the noise-sensitivity ranking is preserved across representations. AEF: AlphaEarth, PR: PRESTO, TE: TESSERA.

Dataset	Sensitivity	AEF–PR	AEF–TE	PR–TE
EuroCrops	Max F1 drop	0.978	0.985	0.973
	Fixed int.	0.983	0.998	0.980
Zambia	Max F1 drop	0.838	0.792	0.784
	Fixed int.	0.870	0.863	0.875

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.

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→ZM) and ρ=0.60 (ZM→EC), showing asymmetric transfer: ZM→EC rankings transfer moderately while EC→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. $\mathrm{cost}(p) \le B$, where $\mathrm{cost}(p)$ composes primitives into protocol-level cost. The card returns the recommended protocol configuration, the predicted degradation $\widehat{\Delta}$ with a 90% predictive interval, and tolerance curves for the relevant noise families. The interval comes from a bootstrap ensemble of 50 surrogates refit on resamples of the simulation ledger, so its width reflects the surrogate's confidence under the limited trial budget. Only relative degradation matters for protocol comparison. Appendix Figure 2 and Appendix C detail the optimizer, cost model, and survey-to-noise mapping.

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. Our robustness analysis varies the feature representation but holds the classifier and a standard supervised objective fixed, so noise-robust training remains a complementary direction we do not evaluate. 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. The full implementation—noise operators, simulation pipeline, surrogate, and protocol-card optimizer—will be released publicly.

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.

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, Siobhan Murray, and Alberto Zezza. Cheaper, faster, and more than good enough: Is GPS the new gold standard in land area measurement? *Survey Research Methods*, 11(3):235–265, 2017. doi: 10.18148/srm/2017.v11i3.6791.

Raphaël d'Andrimont, Momchil Yordanov, Laura Martinez-Sanchez, Javier Gallego, Guido Lemoine, Marijn van der Velde, Beatrice Eiselt, Alessandra Palmieri, Paolo Dominici, Hannes Isaak Reuter, and Christian Joebges. Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union. *Scientific Data*, 7:352, 2020. doi: 10.1038/s41597-020-00675-z.

Pierre Defourny, Ian Jarvis, and Xavier Blaes. JECAM guidelines for cropland and crop type definition and field data collection. Technical report, JECAM/GEOGLAM, 2014. URL https://jecam.org/wp-content/uploads/2018/10/JECAM_Guidelines_for_Field_Data_Collection_v1_0.pdf.

Arthur Elmes, Hamed Alemohammad, Ryan Avery, Kelly Caylor, J. Ronald Eastman, Lewis Fishgold, Mark A. Friedl, Meha Jain, Divyani Kohli, Juan Carlos Laso Bayas, Dalton Lunga, Jessica L. McCarty, Robert Gilmore Pontius Jr., Andrew B. Reinmann, John Rogan, Lei Song, Hristiana Stoynova, Su Ye, Zhuang-Fang Yi, Lyndon Estes, 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.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*, 2025.

Steffen Fritz, Linda See, Christoph Perger, Ian McCallum, Christian Schill, Dmitry Schepaschenko, Martina Duerauer, Mathias Karner, Christopher Dresel, Juan-Carlos Laso-Bayas, Myroslava Lesiv, Inian Moorthy, Carl F. Salk, Olha Danylo, Tobias Sturn, Franziska Albrecht, Liangzhi You, Florian Kraxner, and Michael Obersteiner. A global dataset of crowdsourced land cover and land use reference data. *Scientific Data*, 4:170075, 2017. doi: 10.1038/sdata.2017.75.

Moti Jaleta, Kindie Tesfaye, Andrzej Kilian, Chilot Yirga, Endeshaw Habte, Habekiristos Beyene, Bekele Abeyo, Ayele Badebo, and Olaf Erenstein. Misidentification by farmers of the crop varieties they grow: Lessons from DNA fingerprinting of wheat in Ethiopia. *PLOS ONE*, 15(7): e0235484, 2020. doi: 10.1371/journal.pone.0235484.

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*, 12(1):1496, 2025. doi: 10.1038/s41597-025-05762-7.

Madhava Paliyam, Catherine Nakalembe, Kevin Liu, Richard Nyiawung, and Hannah 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.

Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, Claire Marais Sicre, and Gérard Dedieu. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing*, 9(2):173, 2017. doi: 10.3390/rs9020173.

Julien Radoux, Céline Lamarche, Eric Van Bogaert, Sophie Bontemps, Carsten Brockmann, and Pierre Defourny. Automated training sample extraction for global land cover mapping. *Remote Sensing*, 6(5):3965–3987, 2014. doi: 10.3390/rs6053965.

Maja Schneider, Amelie Broszeit, and Marco Körner. EuroCrops: A pan-European dataset for time series crop type classification. In *Proceedings of the Conference on Big Data from Space (BiDS)*, pp. 125–128, 2021. doi: 10.2760/125905.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

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, Bertran 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/`.

François Waldner, Nicolas Bellemans, Zvi Hochman, Terence Newby, Diego de Abelleyra, Santiago R. Veron, Sergey Bartalev, Mykola Lavreniuk, Nataliia Kussul, Guerric Le Maire, Margareth Simoes, Sergii Skakun, and Pierre Defourny. Roadside collection of training data for cropland mapping is viable when environmental and management gradients are surveyed. *International Journal of Applied Earth Observation and Geoinformation*, 80:82–93, 2019a. doi: 10.1016/j.jag.2019.01.002.

François Waldner, Yang Chen, Roger Lawes, and Zvi Hochman. Needle in a haystack: Mapping rare and infrequent crops using satellite imagery and data balancing methods. *Remote Sensing of Environment*, 233:111375, 2019b. doi: 10.1016/j.rse.2019.111375.

A NOISE TAXONOMY

Operator	Description	Evidence
Label Errors		
Label flip	Random label reassignment due to misidentification at visit	Pelletier et al. (2017); Defourny et al. (2014)
Similar-crop	Confusion between similar crops (wheat/barley, corn/sunflower)	Pelletier et al. (2017)
Non-crop	Fallow or infrastructure incorrectly labeled as crop	Van Tricht et al. (2023); Pelletier et al. (2017)
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	Lobell et al. (2020); Carletto et al. (2017)
Partial bound.	Drop polygon vertices from incomplete boundary capture	Carletto et al. (2017)
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	Defourny et al. (2014); Waldner et al. (2019a)
Road dropout	Remove samples far from roads due to remote undersampling	Azzari et al. (2021); Waldner et al. (2019a)
Neighbor bleed	Expand polygon toward neighbor via adjacent label propagation	Radoux et al. (2014)
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	Waldner et al. (2019b); Azzari et al. (2021)
Swap w/ larger	Replace minority with dominant class label	Jaleta et al. (2020)
Subsample	Uniform size reduction from reduced sample collection	Azzari et al. (2021)
Wtd. subsamp.	Class-weighted retention for non-uniform sampling	Waldner et al. (2019b); Lobell et al. (2020)
Data Quality		
Duplicate	Exact row copies from same field recorded twice	d'Andrimont et al. (2020)
Conflict. dup.	Same location with different labels from disagreement	Fritz et al. (2017); d'Andrimont et al. (2020)

Table 3: Noise taxonomy: 17 operators with descriptions and literature evidence.

B SEARCH SPACE

Table 4 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 4: Noise parameter search space. Coverage is the fraction of training samples affected; intensity controls error severity.

C PROTOCOL CARD IMPLEMENTATION

Constrained optimizer. The card solves $p^\star(B) = \arg\min_p \widehat{\Delta}(p)$ s.t. $\text{cost}(p) \le B$ by exhaustive grid enumeration over seven design axes: GPS device, training hours, verification level, team size, collection days, a multi-source flag, and a minority-bias weight. Each configuration is scored by the cost model and the surrogate; configurations exceeding the budget B are discarded. Among the feasible set the optimizer takes the minimum predicted degradation $\widehat{\Delta}_{\min}$, keeps all configurations within a tolerance band $\widehat{\Delta} \le \widehat{\Delta}_{\min} + \tau$ (default $\tau = 0.005$), and breaks ties by maximizing the achievable sample count. Enumeration is exact rather than heuristic; the feasible space is small (hundreds to a few thousand configurations) so the search runs in well under a second with a cached surrogate.

Cost model. Total cost is the sum of five components. Labor is $n_e\, d\, (s + f)$ for n_e enumerators over d days at daily salary s and daily fuel f. Training is $\lceil n_e/20 \rceil\, c_{\text{sess}} + n_e\, h\, r$, a fixed per-session cost c_{sess} plus h training hours at hourly rate r. Equipment is $n_e\, g_{\text{dev}}$ with per-device GPS cost $g \in \{0, 200, 5000\}$ for phone / handheld / survey-grade. Verification is $S\, v_\ell$ with per-sample cost $v \in \{0, 2, 5, 15\}$ for levels L0–L3, and deduplication adds $S\,(0.5)$ per sample when enabled, where the achievable sample count is $S = n_e\, d \cdot (\text{fields per day})$. Defaults: $s = \$15$, $f = \$10$, $c_{\text{sess}} = \$500$, $r = \$5/\text{h}$, fields per day $= 25$. All primitives are editable in the interface.

Survey-to-noise map. Survey choices set noise parameters through two latent factors: a knowledge factor $\kappa = 0.3 + 0.7\, \min(1, s/30)\, \big(0.4 + 0.6(1 - e^{-0.1h})\big)$ that rises with salary and training hours, and a motivation factor $\mu = \min(1,\ 0.3{+}0.7\, s/25)$ that rises with salary. The GPS device sets the polygon-jitter scale and neighbor-swap radius (phone 15 m, handheld 5 m, survey-grade 0.5 m). The uniform label-flip probability is $\big(0.03 + 0.27(1 - \kappa)\big)\, m_\ell$, and every error rate is multiplied by a verification factor $m_\ell \in \{1.0, 0.7, 0.5, 0.3\}$ for levels L0–L3, a 0–70% reduction. Crop-confusion, boundary, accessibility, and duplicate errors scale analogously with $1 - \kappa$ or $1 - \mu$.

Surrogate and uncertainty. The surrogate is an XGBoost regressor over 29 features (the coverage of 17 noise operators plus 12 intensity parameters), with hyperparameters chosen by 5-fold cross-validation on the simulation ledger. To quantify uncertainty we fit a bootstrap ensemble of 50 regressors, each trained on a resample of the ledger using the tuned hyperparameters. The reported point estimate is the ensemble mean and the 90% predictive interval is the 5th–95th percentile of the ensemble predictions; intervals widen where the ledger offers little support, flagging low-confidence recommendations.

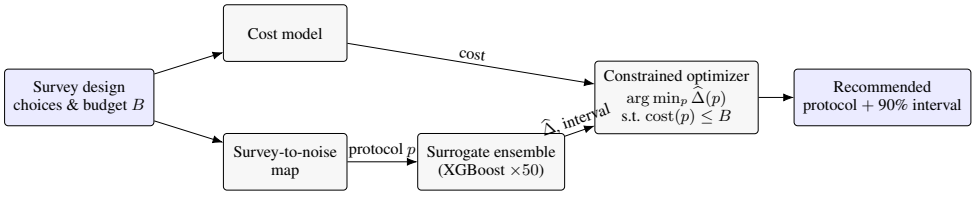
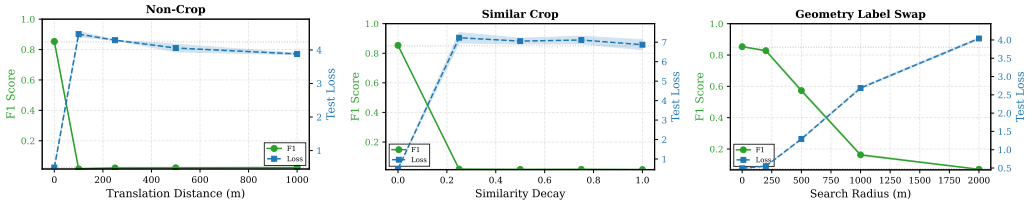


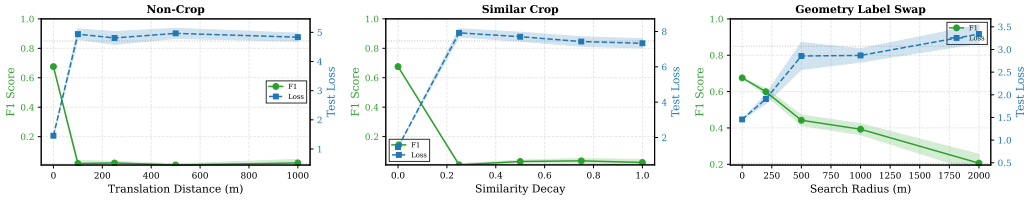
Figure 2: Protocol-card system architecture. Survey design choices and a budget cap B feed both a cost model and a heuristic survey-to-noise map; the resulting protocol p is scored by a bootstrap ensemble of surrogates that returns the predicted degradation $\widehat{\Delta}$ with a 90% predictive interval. A constrained optimizer enumerates feasible configurations, selects the minimum-$\widehat{\Delta}$ protocol within budget, and returns it with its uncertainty.

D TOLERANCE CURVES

Catastrophic Noise

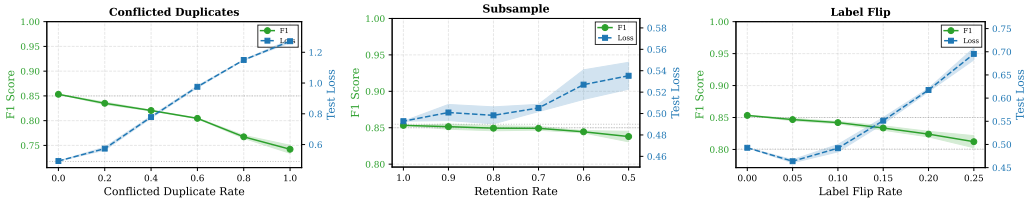


(a) EC: Non-crop (98%) (b) EC: Similar crop (98%) (c) EC: Geom-label swap (92%)

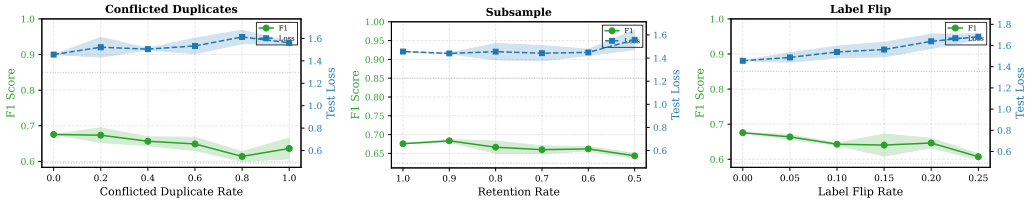


(d) ZM: Non-crop (99%) (e) ZM: Similar crop (99%) (f) ZM: Geom-label swap (70%)

Moderate-Impact Noise

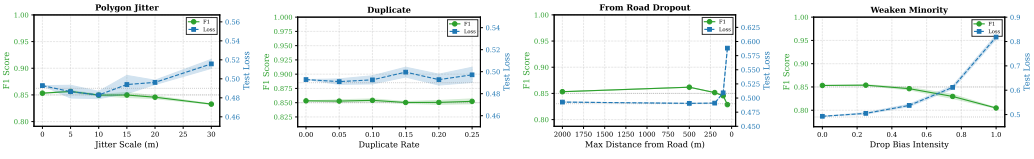


(g) EC: Conflicted dup. (13%) (h) EC: Subsample (2%) (i) EC: Label flip (5%)

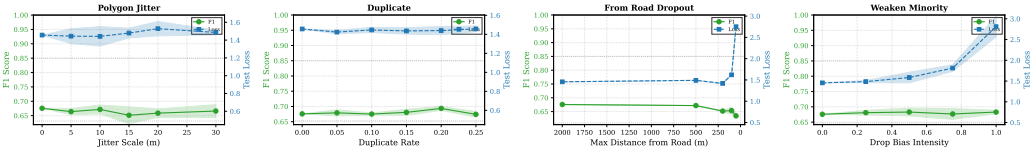


(j) ZM: Conflicted dup. (9%) (k) ZM: Subsample (5%) (l) ZM: Label flip (10%)

Low-Impact Noise



(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 3: **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.