NBER WORKING PAPER SERIES

ESTIMATING THE FOOTPRINT OF ARTISANAL MINING IN AFRICA

Darin Christensen
Tamma Carleton
Esther Rolf
Cullen Molitor
Shopnavo Biswas
Karena Yan
Graeme Blair

Working Paper 33646
http://www.nber.org/papers/w33646

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2025

We thank Jason Leland, Saloni Majmudar, Jarod Ngo, and Ophelia Sin for their invaluable assistance in hand-labeling satellite imagery for this research. We are also grateful to Jihae Hong, Juan Sebastian Leiva Molano, and Eki Ramadhan for overseeing the research assistants and providing essential guidance throughout the project. We thank Anna Boser for constructive feedback on the manuscript and Tracey Mangin for assistance with figure design. The labeling of satellite imagery was supported by a grant from the Hewlett Foundation. This work utilized high-performance computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Estimating the Footprint of Artisanal Mining in Africa
Darin Christensen, Tamma Carleton, Esther Rolf, Cullen Molitor, Shopnavo Biswas, Karena Yan, and Graeme Blair
NBER Working Paper No. 33646
April 2025
JEL No. Q32, Q49

ABSTRACT

Artisanal and small-scale mining (ASM) supplies livelihoods and critical minerals but has been linked to conflict and environmental degradation. We enable monitoring of this largely informal sector by creating high-resolution maps of ASM's footprint in Africa using machine learning models that integrate geographic features and satellite imagery. We find ASM is more extensive than documented: in five countries with on-the-ground surveys, we predict over 231,000 1-km2 grid cells [±2 standard errors: 170,153-297,710] contain ASM activity – over 40 times that recorded by surveyors. Adapting methods for spatial domain adaptation, we map ASM across 20 total countries, estimating that 4% [2-8%] of territory and 17% [10-30%] of the population are impacted by ASM, which encroaches on a larger share of settlements and ecosystems than previously understood.

Darin Christensen
UCLA Luskin School of Public Affairs
337 Charles Young Drive East
Los Angeles, CA 90095
darinc@luskin.ucla.edu

Tamma Carleton
Department of Agricultural and Resource Economics
University of California, Berkeley
Berkeley, CA 94720
and NBER
tcarleton@berkeley.edu

Esther Rolf
University of Colorado Boulder
Computer Science Department
430 UCB, 1111 Engineering Dr,
Boulder, CO, CO 80309
esther.rolf@colorado.edu

Cullen Molitor
University of California, Berkeley
cmolitor@berkeley.edu

Shopnavo Biswas
University of Chicago
Shop4navo@gmail.com

Karena Yan
Harvard University
kyan@g.harvard.edu

Graeme Blair
University of California at Los Angeles
graeme.blair@ucla.edu

A code repository is available at github.com/cullen-molitor/asm-paper

1 Introduction

Artisanal and small-scale mining (ASM) — low-technology, labor-intensive extraction of minerals — is estimated to directly employ 10 million people in sub-Saharan Africa and over 44 million globally [1]. From 1999 to 2019, employment in ASM increased five-fold, expanding from 24 to 40 countries in Africa [2]. With rising demand, particularly for minerals critical to the low-carbon energy transition, the sector will likely see continued growth [3]. While surveys enable estimates of the number of people employed in ASM, little is known about where this mining is happening or the extent of its spatial and environmental footprints. Government maps are notoriously incomplete, as many countries lack the capacity to maintain up-to-date registers of ASM, and even where they can, estimates suggest that 80-90% of the sector is unlicensed [1]. This knowledge gap hampers sustainable development across Africa: ASM is a major source of livelihoods [4], but simultaneously endangers workers and nearby communities due to unsafe working conditions and pollutants [5, 6], contributes to deforestation [7, 8], threatens biodiversity [9], and provokes conflict [10]. Scientific and policy efforts to measure and mitigate the sector's environmental and social harms or amplify its contribution to economic development are constrained by an inability to identify which communities and ecosystems are impacted by ASM.

Recent innovations in machine learning are emerging to help fill this data gap by predicting ASM activity using satellite imagery or other geographic information (e.g., rock lithology). Prior studies cover parts of the Amazon [7, 11], Ghana [12–16], Senegal [17], and the Philippines [18], showing feasibility within a single country or known mining area by demonstrating in-sample predictive capabilities. Some efforts have been larger scale: Couttenier *et al.* [19] use satellite imagery to predict ASM in 111 mining sites in West Africa and Rigterink *et al.* [20] use geologic information from three countries to predict ASM suitability across much of Africa. However, scarce training data and the use of standard in-sample evaluation techniques have limited the generation of reliable predictions outside of the small areas already being monitored. And yet, the large swaths of Africa where oversight is minimal are precisely where new measurement is needed and where environmental and social harms are likely to be most pronounced.

Here, we develop machine learning models to map the spatial footprint of ASM and assess its ecological footprint across sub-Saharan Africa. Our models synthesize data from both geographic information (e.g., geology, terrain) and daytime satellite imagery and are tailored for the challenging task of prediction across large regions where labeled data are sparse. To overcome the data limitations that have constrained past efforts, we design an application to facilitate manual annotation of high-resolution imagery, identifying mining activity across $\sim$23,000 0.01° grid cells in five countries with sizable ASM sectors: the Central African Republic (CAF), Democratic Republic of Congo (COD), Sierra Leone (SLE), Tanzania (TZA), and Zimbabwe (ZWE). Our sampling protocol enables us to assess performance in multiple policy-relevant subsets (e.g., by country or within and outside of areas with ground-based monitoring) and, importantly, to construct representative test sets that fairly adjudicate model performance for downstream prediction across entire countries, diverse landscapes, and built environments. We directly evaluate the value of automated ASM prediction for augmenting existing ground-based survey efforts, which are severely constrained by lack of resources and unsafe conditions [1].

We show, consistent with prior work in other settings [21], that developing machine learning models that successfully extrapolate to regions not represented in training data requires specialized approaches. We demonstrate that relatively simple solutions to this "domain adaptation" problem, such as modifying model feature sets, can substantially improve such spatial generalizability, leading to reliable large-scale mapping of ASM in regions yet to be surveyed. We construct what is, to our knowledge, the first estimate of the spatial extent of ASM activity across 20 countries and over 15 million square kilometers. We document the environmental and demographic footprint of the sector, identifying ASM intrusion into conservation areas, its overlap with biodiversity hotspots, and its proximity to population centers.

2 Approach

We compile coordinates of 11,518 suspected artisanal mining sites across five countries by combining publicly available data from the International Peace Information Service (IPIS) and Sierra Leone's National Minerals Agency (NMA). We augment these data with a stratified random sample of $0.01^\circ \times 0.01^\circ$ ($\sim$1km^2) grid cells that is structured to ensure that we can construct a representative sample of each country's area and that we have a sufficient number of no-mining observations (Methods Section 7.1.1). Using a custom application and the Google Maps Static API, research assistants manually label high resolution imagery in all sampled grid cells as containing mining or not. Where mining is present, they outline the boundaries of any mines. Our training sample comprises 23,061 manual labels (mapped in Fig. 1), where labels that were flagged as low confidence are removed for quality control (Methods; Supplementary Materials B).

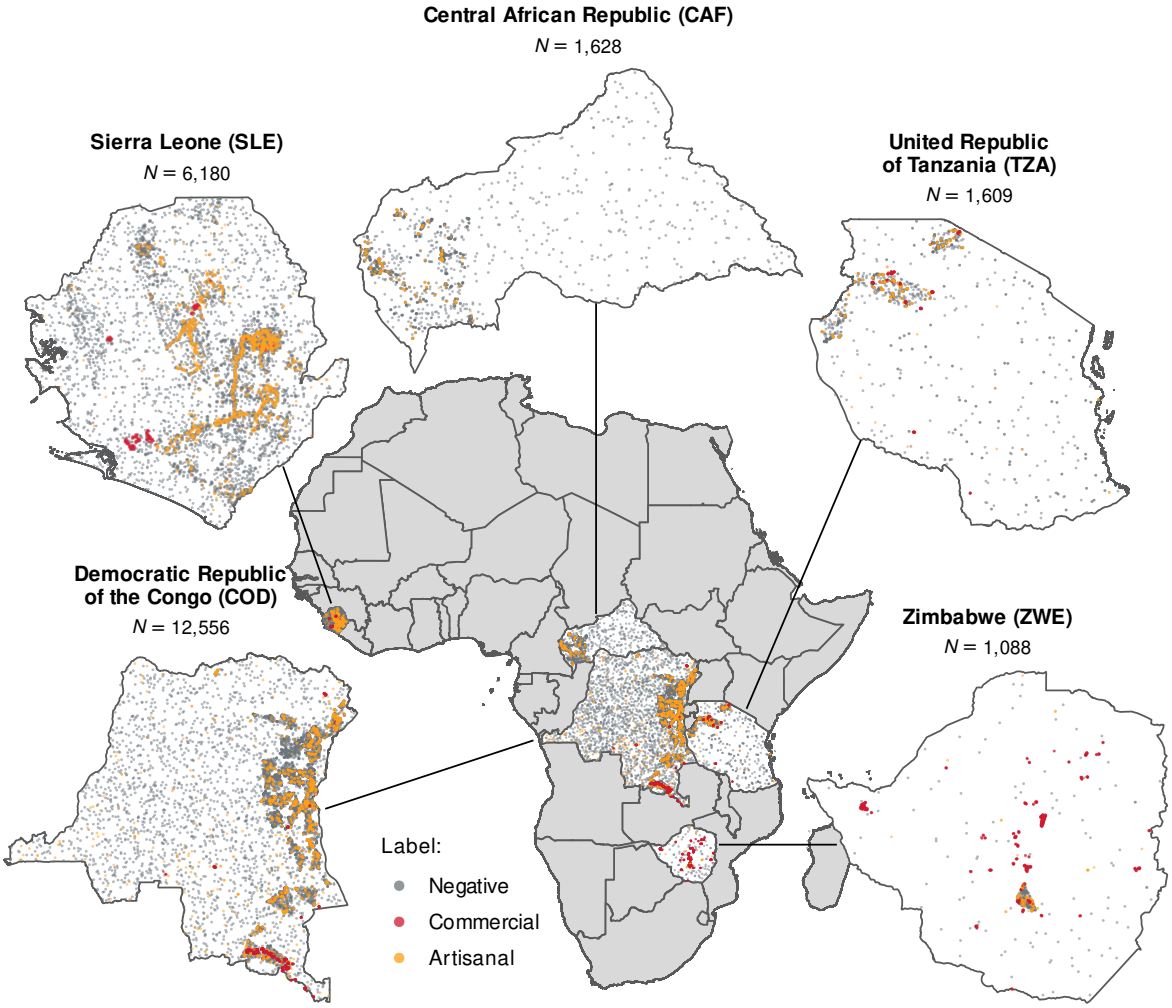

Figure 1: Sampling frame of artisanal and commercial mining across five nations. Map of 23,061 $0.01^\circ \times 0.01^\circ$ grid cells hand labeled using high-resolution satellite imagery as containing artisanal mines (orange) or no mining activity (grey) across five African nations. Auxiliary data from Maus *et al.* [22] are used to indicate grid cells with commercial mines (red). Total number of labels (N) available for each country is indicated (excluding cells with commercial mining). See Methods for details on sample construction.

We combine these labeled mining data with two sets of features. Our "geographic" features include 268 variables related to geology (lithology, presence of gold-suitable bedrock, distance to faults and deposits) [8, 23], topography (elevation, distance to rivers, surface water) [24–26], landcover (ecosys-

tems, landuse) [27–29], climate (rain, temperature) [30], and economic development (distance to roads, greenhouse gas emissions) [31, 32] (Supplementary Materials Table 3). The selection of these features is based on prior efforts to predict mining activity [20, 23, 33], though existing work only uses a subset of these variables. Our "imagery" features are 4,000 random convolutional features (RCF) extracted from composite four-channel 4.77 meter resolution satellite images from Planet Labs. We use RCF based on its demonstrated performance in a wide variety of remote sensing tasks and computational efficiency [34], which makes it feasible for us to conduct extensive experiments that assess the reliability and policy relevance of our predictions. We construct three predictive models relying on: geographic features alone; imagery features alone; and both feature sets in an ensemble model.

3 Automated ASM assessment augments on-the-ground survey efforts

Our predictive models reliably detect ASM activity across the five countries that comprise our training data (Fig. 2(a)). Models trained on geographic features achieve an average AUC – the area under the Receiver Operating Characteristic (ROC) curve – of 0.90 ± 0.004; imagery features, 0.85 ± 0.005; and ensemble models which combine both information sources, 0.91 ± 0.004 (where $\pm$ indicates two standard errors constructed by re-randomizing which grid cells are allocated to the training and test sets). We find performance gains, though small, from combining geographic and imagery-based feature sets. The ensemble model performance exceeds the performance reported in past work, which uses a subset of our features for similar tasks, including ASM prediction [20] or open-pit mining prediction [35], though direct performance comparisons are complicated by differences in geography, spatial resolution, and test set construction across studies. We show in Supplementary Materials Section C.1 that performance in our sample matches that from a segmentation model.

Our training data over-sample cells in areas surrounding suspected ASM activity relative to other parts of each country to ensure diverse types of mining are seen during training and evaluation. Although we follow the standard practice of reporting full-sample performance, we report the AUC, a metric robust to such label imbalance. In Fig. 2(b) we additionally show performance in representative test sets with cells drawn uniformly at random (UAR) from each of the five countries in our sample, which, to our knowledge, has never been evaluated in related prior work. These results indicate strong predictive power across the five countries in our sample, not only in the suspected mining areas that make up a disproportionate share of our labels.

3.1 Performance comparison in areas monitored with on-the-ground surveys

Machine learning augments existing ASM mapping only if it improves existing survey efforts by reducing error rates of on-the-ground surveys, detecting activity in areas that have not been surveyed, or both. Outside of Sierra Leone (SLE), current ASM mapping relies on on-the-ground surveys organized by IPIS (Methods Section 7.1.1). Within SLE, the National Minerals Agency (NMA) tracks artisanal mining through licensing. Both of these monitoring efforts can suffer from two types of errors: geolocating ASM sites where there is no mining (false positive) or failing to record an ASM mine where one is present (false negative). Here, we evaluate whether our predictive models can more accurately detect mining activity in the areas previously monitored by IPIS and/or the NMA, leveraging our labeled observations to reveal errors in existing surveys.

Fig. 2(c) reports model performance within areas that were monitored by previous ground-based efforts (see Methods Section 7.1.4 for definitions of monitored areas). Our models can reduce the false positive rate (FPR) by 0.47 while achieving the same true positive rate (TPR) as surveys, representing a significant reduction in false detection of mines. The AUC for these regions are very similar to that in the full sample (AUC = 0.87 ± 0.005 in the ensemble model), indicating that our predictions give a reliable signal of ASM activity within monitored regions.

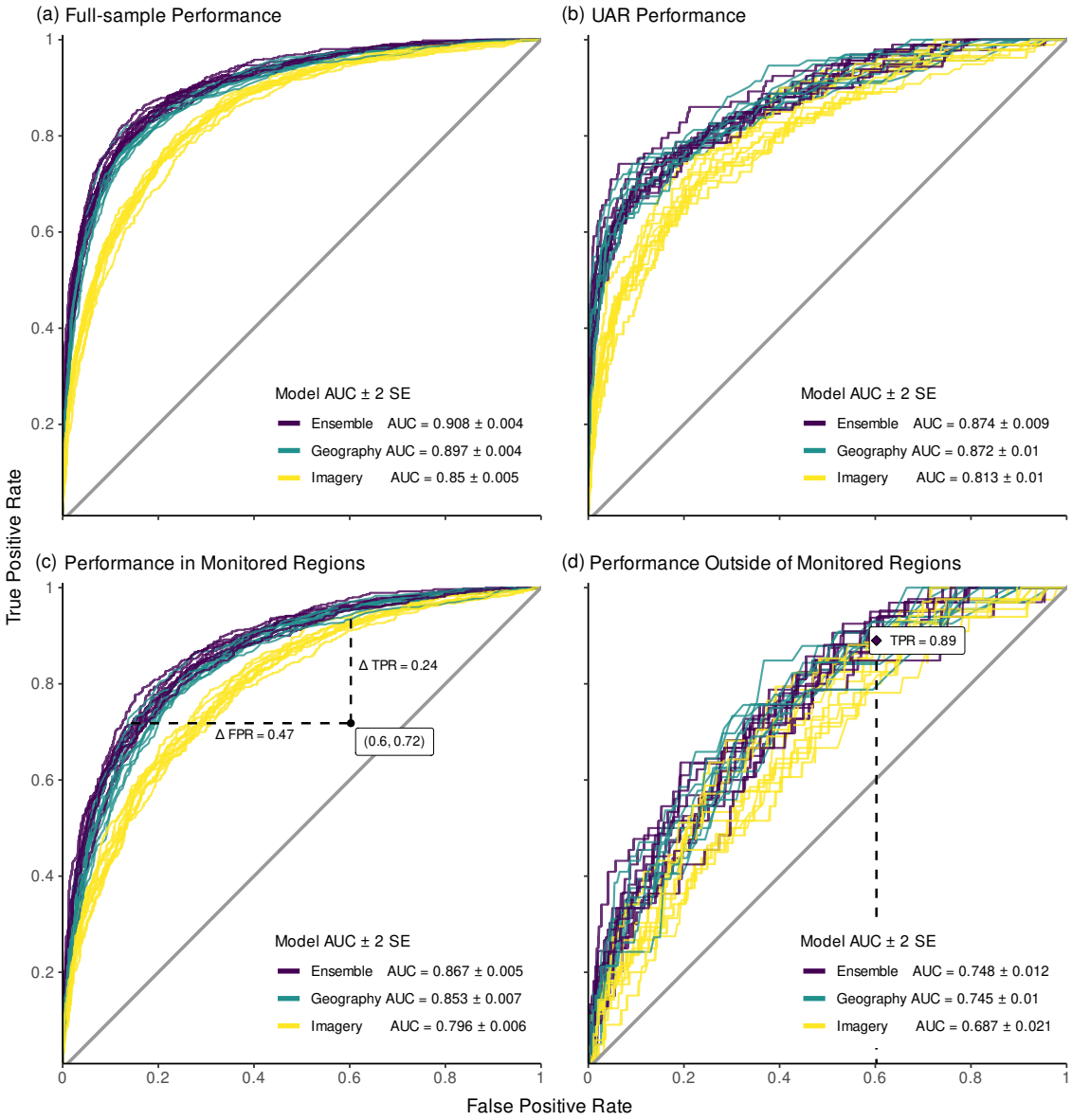

Figure 2: Predictive performance of automated ASM assessments. Receiver Operating Characteristic (ROC) curves and average area under the curve (AUC) shown for the performance of the geographic (green), imagery (yellow), and ensemble (purple) models across ten iterations that re-randomize labels into the training and test sets. $\pm$ values for AUC indicate two standard errors constructed from this re-randomization. Fig. 2(a) reports performance across the full sample. Fig. 2(b) reports performance for test sets drawn uniformly at random (UAR) in space and, thus, representative of the physical area of each country (Methods Section 7.1.5). Fig. 2(c) reports performance for test sets drawn only from areas monitored by on-the-ground surveyors (Methods Section 7.1.4). ΔFPR and ΔTPR report the difference in the false and true positive rates, respectively, between the on-the-ground efforts and the average ensemble predictive model. Fig. 2(d) reports performance on test sets drawn only from unmonitored areas. We report the TPR of our average ensemble model when the FPR is fixed to match the rate achieved by on-the-ground efforts in monitored areas (FPR=0.60).

3.2 Model performance outside areas monitored with on-the-ground surveys

Automated ASM mapping is particularly valuable outside of areas monitored by ground-based survey efforts. Unlike prior research [20, 35], our sampling protocol allows us to construct a random sample of such observations and therefore report representative estimates of performance (Methods Section 7.1.4).

Fig. 2(d) reports model performance in areas that have not been monitored by existing surveys. As expected, it is more difficult to predict ASM in these regions: the average AUC from our ensemble model

falls to 0.75 ± 0.012, declining by 39% of its margin above a baseline of random guessing, relative to the full sample (see Methods Section 7.3.2). However, this performance compares favorably to the error rates of on-the-ground efforts in monitored areas. The FPR achieved by on-the-ground efforts *within* monitored areas is 0.60 (vertical line in Fig. 2(d)); when evaluated at this FPR, our ensemble model generates a higher average TPR of 0.89 across the *unmonitored* areas than survey efforts do in monitored areas (average TPR of 0.72). (Note that the TPR of ground survey efforts in unmonitored areas is 0 by definition of them being unmonitored). Thus, our ensemble model generates predictions across large unmonitored regions in these five countries that are more accurate than on-the-ground efforts that survey much more geographically confined areas.

4 Generating reliable ASM predictions at scale

4.1 Evaluating performance in geographies not represented in training data

The policy relevance of geospatial machine learning models depends critically on the ability to deploy them to new locations [36]. In our setting, not only are certain subnational regions unmonitored, but entire countries known to contain ASM lack any ground-based data. These data-poor regions are precisely where the sector's scale and impacts may be largest, given lack of enforcement and licensing. Here, we simulate the deployment of automated ASM mapping to a new country with no ground-based data by training models using data from four countries and evaluating the performance of each model with data from the fifth held-out country (e.g., Fig. 3(a)). We repeat this out-of-domain procedure for all five countries in our data. This experiment is a far more difficult task than the more commonplace full-sample evaluation conducted above [36] and is made possible by both the spatial distribution of our training data and the computationally efficient approach we take to imagery-based prediction.

We find that the predictive performance of all three model types (solid lines, "main" in Fig. 3(c)) declines significantly when evaluated in a country not represented in the training data ("Out-of-domain"; triangles), versus when all countries are present in the training dataset ("Full-sample"; circles). To better contextualize these declines, we measure them relative to each model's AUC margin above random guessing (AUC = 0.50). Across countries, the (observation weighted) average performance of the ensemble model declines by 27% (Fig. 3(c)) of its margin above random relative to the test set performance shown in Fig. 2(a), with losses ranging from 20% (in SLE) to 45% (in ZWE). The size of the performance gaps is not consistently larger in countries with more observations and, thus, larger samples of training data, suggesting performance losses are due to the spatial distribution of training data rather than just the sample size. Performance gaps are much larger for models trained only using geographic information than those trained on satellite imagery — average declines are 48% and 13% for these two model classes, respectively.

Performance declines are similar, but less severe, when evaluating in subnational regions not present in training data (e.g., Fig. 3(b)). We evaluate this with all provinces in COD and districts in SLE, the two countries for which we have sufficient data for such an exercise. In this setting, mean performance (AUC) across the left-out regions ranges from 0.73 in the geography model to 0.81 in the ensemble model (Fig. 3(d)). We again find that performance declines are much larger for the geography models than for the imagery models when extrapolating to new subnational regions. Across all regions, the mean change in AUC between the test set shown in Fig. 2(a) and the leave-one-region-out experiment is -0.14 for the geography models, compared to just -0.02 for the imagery models.

4.2 Modifying models to improve spatial extrapolation

The challenge of out-of-domain generalization is common in many applications [37, 38]. Recent research shows that modest feature modifications can improve out-of-domain prediction in spatial and other settings [39–41]. We apply those insights and demonstrate the resultant gains, showing that feature modifications substantially improve our imagery-based models' full-sample and out-of-domain performance, while gains prove much harder to generate with geography-based models.

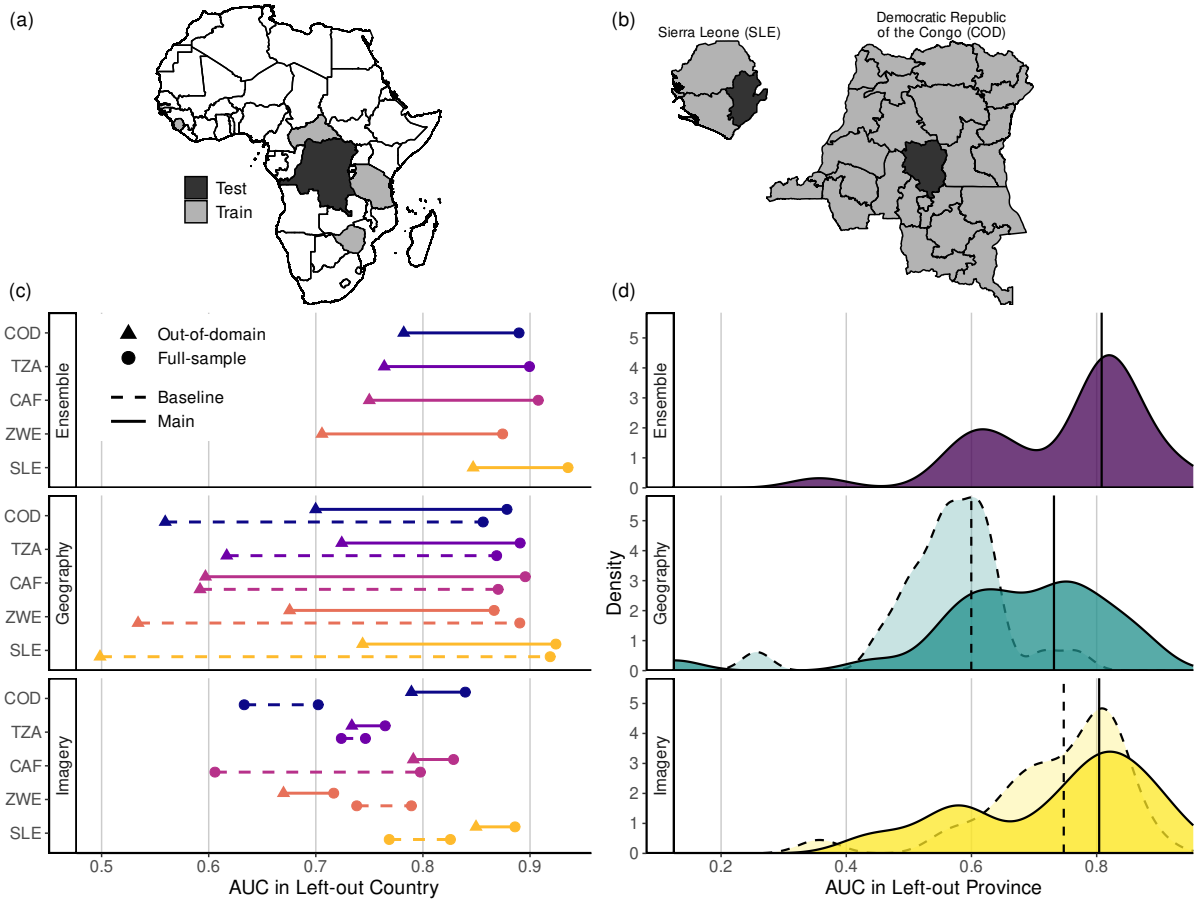

Figure 3: Model generalization across national and subnational borders. Results from spatial out-of-domain experiments, in which testing data from an entire target country or subnational region are removed from the training sample, as illustrated in Figures 3(a) and 3(b), respectively. Fig. 3(c) reports country-specific performance by model type when models are trained on the full dataset (circles) versus outside the target country (triangles). Fig. 3(d) shows the distribution of performance across held-out subnational regions in the Democratic Republic of Congo (COD) and Sierra Leone (SLE) by model type with mean AUC shown as vertical lines. Models used in (d) are country-specific. In both (c) and (d), solid lines indicate the main model trained and evaluated throughout the text, while dotted lines indicate a "baseline" benchmark. The "baseline" geography model is a spatial interpolation benchmark (Methods Section 7.4.2) and the "baseline" imagery model matches standard imagery featurization processes (Methods Section 7.4.1).

For the imagery-based models, we adapt image normalization techniques that have previously delivered gains for machine learning models built on satellite imagery [42], out-of-of-domain generalization in computer vision [41], and problems with imbalanced class labels [43]. Specifically, we conduct a grid search over a wide range of normalization parameters in which images are preprocessed differently before being input to our machine learning models (see Methods Section 7.4.1). The dashed lines in Fig. 3(c) report full-sample and out-of-domain performance using a "baseline" imagery model that standardizes imagery across the entire dataset, a common preprocessing approach in the literature (see Methods Section 7.4.1). By contrast, the model that performs best in the full sample ("main"; solid lines) uses local normalization, which maximizes contrast across pixels within a single band and image. This model dramatically outperforms the baseline model in the leave-one-country-out experiment: the (observation-weighted) average AUC across countries is 0.80 versus 0.68. Gains are smaller for the imagery model in the leave-one-region-out experiment, with mean AUC falling from 0.80 in our main model to 0.75 for the baseline model (Fig. 3(d), bottom), showing that normalization choices have less impact within smaller geographies. Importantly, our main model performs nearly as well in the more challenging out-of-domain experiment as the baseline model does at full-sample prediction (0.80 vs. 0.81; Fig. 3(c), bottom). We demonstrate more generally that using increasingly local statistics in im-

age normalization improves spatial extrapolation (Fig. D.2). While image normalization techniques that work well for out-of-domain prediction also tend to increase full-sample performance (Fig. 3(c) but for ZWE), gains in the latter are modest relative to the improvements in spatial extrapolation (Fig. D.2(a)). This can be explained by our finding that local image normalization minimizes the domain shift from the training sample to the held-out country (Fig. D.2(b) and (c)).

Prior work also suggests that feature selection could improve spatial extrapolation with our geography-based model [21, 44, 45]. We find that removing subsets of features (e.g., distances to specific fault lines or mineral deposits) improves performance in the out-of-domain experiment (Fig. D.1(c)). However, removing these features generates a nearly offsetting decline in full-sample performance (Fig. D.1(a)). This exemplifies a more general tradeoff: subsets of our geographic features generate predictions more highly correlated with pure spatial interpolation (i.e., a random forest based only on latitude and longitude); these subsets tend to generate better full-sample performance but poorer spatial extrapolations (Fig. D.1(b,d)). This tradeoff is emphasized in Fig. 3(c)-(d), where a pure spatial interpolation "baseline" (dashed lines) performs competitively with our 268-variable random forest when evaluated on the full sample, but is little better than random guessing when evaluated out-of-domain.

To generate predictions across the five countries where we have labels, we use the ensemble model, given its high performance in the full sample. To generate predictions beyond these five countries, we use our main imagery model, based on its high performance in new domains.

5 Estimating the scope and environmental footprint of ASM in Africa

We estimate the geographic extent of ASM across 20 African countries: the five in-sample countries described above, as well as 15 additional countries that are estimated to have sizable ASM sectors [1], but lack publicly-available training data. Fig. 2 maps these predictions across 15 million km^2, aggregating estimates to 0.05° (∼5 km) resolution to avoid disclosing the precise location of potentially unlicensed ASM. Continuous predicted probabilities from each model are transformed into binary predictions using a custom thresholding approach designed to best match labeled data on ASM prevalence (Methods Section 7.5).

We find that ASM is widespread, extending beyond existing monitored areas and near to population centers. For the five countries in our training sample, on-the-ground mapping efforts report ASM in more than 5,000 cells, but we estimate a full 231,000 cells [95% confidence interval: 170,153–297,710], representing 6.5% [4.8-8.3%] of all grid cells, intersect with ASM activity (Table 1). In some cases, new detection of ASM is even more extreme; for example, in ZWE we estimate that just 0.4% [0.26-0.86%] of predicted ASM activity is recorded in existing ground-based surveys. Across the five countries, more than 90% of our predicted ASM activity falls outside areas currently monitored by ground-based efforts (as defined in Methods Section 7.1.1). A hotspot analysis detects over 1,900 clusters of ASM (each containing at least 10 cells with predicted ASM within 3 km of each other) across the sample, with 88% of these hotspots located more than 3 km away from mines reported by on-the-ground survey efforts. We predict that ASM often operates in close proximity to human settlements: across our five countries, 44% [37-49%] of people are estimated to reside within 1km of ASM activity (Table D.1).

We also predict substantial ASM activity in 15 out-of-sample countries shown in Fig. 4. Although these results should be regarded more cautiously, given necessary spatial extrapolation, we estimate that 3% [1-9%] of cells contain ASM activity, clustered in up to 2,871 hotspots. In these countries, 10% [3-25%] of the population is estimated to reside within 1km of ASM activity (Table D.1). Although this represents a more limited population exposure to ASM than in the five in-sample countries where ASM is regularly reported on, it demonstrates the wide socioeconomic reach of the sector, even in countries where no public ASM data exist.

ASM's large spatial footprint has critical environmental implications. Using boundaries of protected areas from UNEP-WCMD & IUCN [46] and of biodiversity hotspots from Hoffman *et al.* [47], we predict that in our five in-sample countries, 2% [sensitivity range: 2-3%] of cells within protected areas and 18% [15-20%] of the cells within biodiversity hotspots contain ASM (Table 1; columns 3-4). This inter-

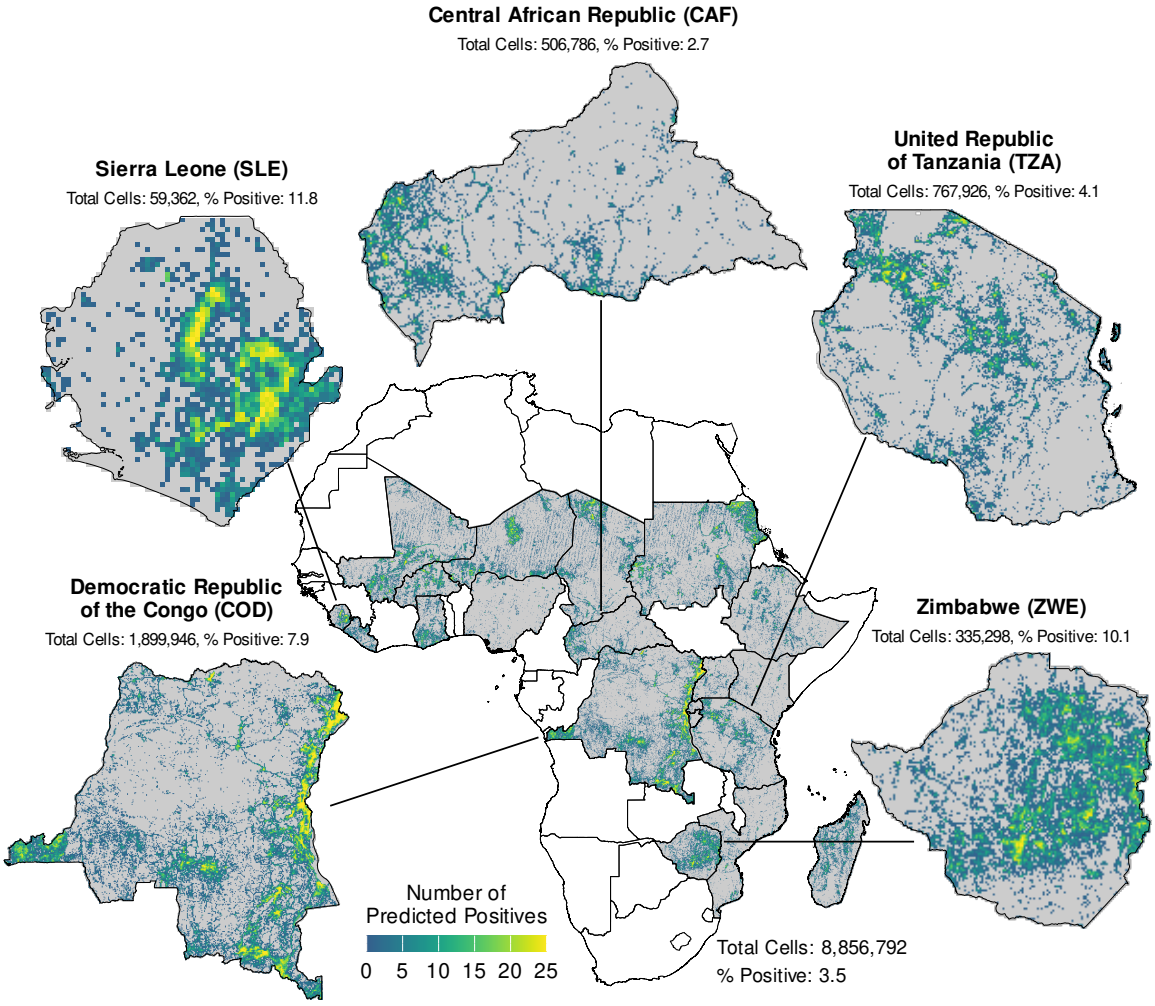

Figure 4: Predicted footprint of artisanal mining across 20 African nations. Predicted number of $0.05^\circ \times 0.05^\circ$ cells with artisanal and small-scale mining (ASM) across 20 countries known to have sizable ASM sectors [1]. Grey areas indicate no predicted artisanal mining and white areas indicate countries not evaluated for ASM. Blow-out maps of the five countries in the training sample use an ensemble model combining information from satellite imagery and geographic variables. The continent-scale map uses a model trained only on imagery and tuned for out-of-country extrapolation. Cells labeled as commercial mining in Fig. 1 are coded as negatives and used only for out-of-country extrapolation (Methods Section 7.1.2 and Appendix B).

section is much larger than suggested by on-the-ground mapping efforts: in Sierra Leone, for example, government data suggests that 1% of cells within protected areas contain ASM, whereas our estimate is 9% [8-10%]. In COD and ZWE, 0-1% of cells within biodiveristy hotspots contain ASM according to ground-based surveys, but we estimate that a full 30% [28-33%] (COD) and 40% [24-52%] (ZWE) of these cells have ASM activity. Although this environmental impact is less severe outside our training data, we predict that across all 20 countries, over 3% [1-5%] of protected areas and 5% [3-10%] of biodiversity hotspots contain ASM. Despite efforts to set aside land for conservation, our findings suggest that the sector infringes on ecologically important areas to an extent not previously documented.

6 Discussion

We show that low-cost machine learning methods can reliably map ASM activity across sub-Saharan Africa. We find that these methods valuably augment on-the-ground efforts to monitor ASM, both by uncovering ASM activity that surveyors missed and by detecting activity in areas that were not surveyed.

		Monitored Areas		Protected Areas	Biodiversity Hotspots
		Inside	Outside		
CAF	*Area* (km^2)	5,415	501,371	88,659	0
	Est. % Area With ASM				
	Administrative Labels	7%	-	0%	0%
	Ensemble Predictions	25%	2%	1%	0%
		[14-32%]	[1-4%]	[<1-1%]	[0-0%]
COD	*Area* (km^2)	31,886	1,868,060	266,967	143,172
	Est. % Area With ASM				
	Administrative Labels	8%	-	< 1%	1%
	Ensemble Predictions	41%	7%	4%	30%
		[39-43%]	[6-8%]	[4-5%]	[28-33%]
SLE	*Area* (km^2)	9,825	49,537	8,100	38,268
	Est. % Area With ASM				
	Administrative Labels	16%	-	1%	3%
	Ensemble Predictions	43%	6%	9%	15%
		[40-46%]	[4-7%]	[8-10%]	[13-16%]
TZA	*Area* (km^2)	3,638	764,288	306,678	129,762
	Est. % Area With ASM				
	Administrative Labels	9%	-	< 1%	< 1%
	Ensemble Predictions	42%	4%	1%	4%
		[34-48%]	[2-6%]	[<1-2%]	[2-5%]
ZWE	*Area* (km^2)	760	334,538	93,826	5,849
	Est. % Area With ASM				
	Administrative Labels	18%	-	0%	0%
	Ensemble Predictions	41%	10%	2%	40%
		[35-48%]	[5-15%]	[1-4%]	[24-52%]
Other	*Area* (km^2)	-	8,847,890	1,159,916	1,955,737
	Est. % Area With ASM				
	Imagery Predictions	-	3%	3%	3%
		-	[1-8%]	[1-7%]	[1-9%]

Table 1: Estimated extent of new ASM detection. Table reports the number of cells (equivalent to km^2) within and outside of areas monitored by ground-based surveys (Methods Section 7.1.1), as well as those within protected areas and biodiversity hotspots (Methods Section 7.5). The share of these cells containing ASM activity is computed based on administrative labels (second row) and our machine learning predictions (third row). Predictions and clipping thresholds are constructed identically to those mapped in Figure 4. For the first five in-sample countries listed, we construct bounds (in square brackets) by setting the clipping thresholds to $\widehat{q} \pm 1.96 \times \sqrt{\widehat{q}(1-\widehat{q})/N}$, where $\widehat{q}$ is the proportion of cells in each country that contain ASM, estimated using a uniform-at-random sample of labeled training data. For the remaining countries, clipping threshold bounds are a 95% prediction interval around $\widehat{q}$, which we estimate using a linear model that relates the share of cells with ASM to per capita employment in ASM (see Methods 7.5 for details).

We show that even in countries with ongoing efforts to map ASM activity, the sector's footprint is likely much larger than previously documented, with over 90% of predicted ASM locations falling outside of areas covered by on-the-ground mapping endeavors. We predict that ASM encroaches on ecologically important areas and human settlements, raising important management concerns related to deforestation, water and soil pollution, and social stability.

Automated ASM mapping, however, has important limitations, which we systematically evaluate using a novel sampling strategy and experimental design. Specifically, our results reveal the challenge of spatial extrapolation across large, heterogeneous regions. While spatial out-of-domain prediction is known to be a difficult machine learning problem [21], past work on ASM has not diagnosed or ad-

dressed this challenge. Training and evaluation data typically over-represent artisanal mining areas from a small number of provinces or countries where ASM sites have been geolocated. Based on our results, we anticipate sharp declines in performance when models based on geographic features extrapolate to areas beyond the training data.

We address this challenge by identifying low-cost opportunities to improve performance in new spatial settings, both in geography-based and imagery-based predictive models. Our labeling and sampling protocols enable us to rigorously evaluate these solutions to the spatial extrapolation problem without dispatching surveyors. Using feature-tuning techniques that permit more accurate spatial extrapolation, we map the footprint of ASM at scale. While past surveys estimate that large shares of people are employed in ASM across much of Africa, our analysis reveals *where* this activity takes place and, thus, what settlements or ecosystems could be targeted by policymakers looking to support or regulate the sector.

We expect that future work will enhance the granularity and scope of our assessment of ASM in Africa. First, while our low-cost models have comparable performance to more complex segmentation models when evaluated at grid cell level (Appendix C.1), the latter approach could provide pixel-level predictions that pinpoint, for example, precisely where mines infringe on surface water or transgress the boundaries of licensed or protected areas. While our 1km resolution analysis represents a finer scale mapping than many prior efforts (e.g., Rigterink *et al.* [20]), further increasing the precision of spatial assessments may enable policy interventions or scientific analyses that are not possible using grid-scale predictions.

Second, our estimates are static, providing just one cross-sectional assessment of ASM. Because both the visual signal and the environmental impacts of ASM persist over time, our assessment captures current and prior ASM activity relevant to current management of the sector. However, we expect this sector will continue to evolve due, in part, to the growing demand for minerals critical to the clean energy transition, and much can be learned from mapping this sector over time. To do so, new predictive algorithms must be designed and evaluated explicitly on temporal variation – prior work has shown that predicting variation over time using satellite imagery is often more difficult than variation over space (e.g., Barenblitt *et al.* [12] and Khachiyan *et al.* [48]).

More broadly, our findings suggest that satellite-based automated mapping may be able to fill critical data gaps in other unregulated sectors where training data remain sparse and predictions are needed at scale. These related settings, such as identifying illicit drug cultivation [49] or monitoring refugee settlements [50], similarly require the development of predictive models that can spatially generalize beyond limited training datasets. Our low-cost methods for evaluating and improving domain adaptation suggest a feasible path towards large-scale mapping in such settings.

7 Methods

7.1 Label data

7.1.1 Sample selection

Our training, evaluation, and test sample is composed of a set of locations across five countries that we manually label as containing ASM or not. To construct this sample, we begin by compiling coordinates of suspected artisanal mining sites from two sources: in Sierra Leone (SLE), the National Minerals Agency (NMA) provides 7,762 polygons areas licensed for artisanal mining. In CAF, COD, TZA, and ZWE, the International Peace Information Service (IPIS) geolocates 3,756 artisanal mines. The NMA data are from 2017–2019 and include all artisanal mining licenses in Sierra Leone. We drop 0.01% of NMA polygons, which were improperly constructed. The IPIS data were collected in waves from 2009–2020. IPIS includes both licensed and informal artisanal mines, but it only covers a subset of known mining areas in each country (e.g., in COD, IPIS only surveys eastern provinces). We refer to these ASM sites as suspected, because the data contain two types of false positives: (1) in SLE a license holder may not (yet) be mining a site; (2) in CAF, COD, TZA, and ZWE, IPIS enumerators do not visit every mine

they locate (e.g., due to safety concerns) and, in those instances, they record the coordinates of a nearby landmark or town. To reduce overlap in the sample locations provided to us by NMA and IPIS, we use the `hclust` algorithm in `R` to cluster locations within 500 meters of another sample location. This reduces the total number of suspected ASM locations from 11,518 to 4,441.

To ensure both a sufficient number of negatives (i.e., locations with no ASM activity) and a representative sample of locations from each country, we sample additional locations. We use a geospatial grid with 0.01 degree resolution and whose borders intersect Null Island (0°N 0°E) such that each cell is centered along an axis of 0.005°. Specifically, we cluster the suspected ASM sites described above using K-Means [51] and draw a convex hull around each cluster. We randomly sample from all 1×1 km grid cells outside of these convex hulls. Within the convex hulls, we randomly sample from grid cells that are not adjacent to a cell with a suspected site. We draw an equal number of cells from outside and within the convex hulls until we reach double the number of suspected ASM sites in each country. This leads us to a final sample of 13,952 locations for manual labeling, 4,441 of which were original suspected mining sites from IPIS and NMA. Our final training sample includes 23,061 grid cells of size 1km^2, as the suspected mining sites can overlap up to 4 cells on the standardized grid (see Appendix B.2 for more details).

7.1.2 Labeling

The sampling procedure described above generates a set of 13,952 geolocated points. To label each location as containing ASM activity or not, we design a manual labeling procedure in which research assistants review high-resolution satellite imagery and identify ASM activity by hand. Specifically, we develop a custom application that tasks research assistants with reviewing high-resolution satellite imagery (from Google Maps Static API) for each sampled location. For each location, we superimpose on the image a circle that is one kilometer in diameter, centered on each sampled point. A researcher is asked to visually scan the imagery and draw polygons around any signs of artisanal or commercial mining activity, which often can be seen as open pits or pools. While the app allows the researcher to pan and zoom, they only flag activity that intersects this circle. If a researcher locates one or more mines within the circle, they: (1) draw a polygon that traces the boundaries of each mine; (2) indicate whether the mines are artisanal, commercial, or both; and (3) rate their confidence in the label on a five-point scale. Researchers were asked to explain low-confidence ratings: 70% were due to blurry imagery, and the remainder were due to ambiguous activity. For a random sample of 10% of locations, we assign two research assistants to label the same point, providing a higher quality subsample in which we can cross-check hand labeling outcomes. For example, in locations where both researchers express moderate to very high confidence, in 83% both individuals agree about whether an artisanal mine is present. The details of this labeling process, including the specific instructions provided to research assistants, are provided in Appendix A.

We augment our sample of labels with a recent dataset that manually traces the boundaries of commercial mining activity [22]. These data include 388 commercial mining polygons across our five countries, which intersect with 1,773 1km^2 grid cells. If a grid cell intersects any commercial mine, we label it as a commercial mining site. We label a grid cell as containing artisanal mining if it intersects one or more of the polygons that the researchers drew around artisanal mines and contains no commercial mines. All other manually inspected grid cells are labeled negative (i.e., not containing mining activity). Fig. 1 maps the 23,061 labeled grid cells across our five countries.

At the end of this process, each 1km^2 grid cell has an associated binary label of 1 ("mine") or 0 ("no mine"). Cells with a label of 1 additionally have the geo-referenced polygons of mining activity. While past work has cast ASM detection from satellite imagery as a semantic segmentation task [52] (essentially predicting whether each pixel is a mine), we instead output predictions at the grid cell level for two reasons. First, this structure of predictions is suitable for policy-relevant outputs, where monitoring and enforcement efforts rely on deployment of on-the-ground inspection based on knowledge of where mining activity is taking place (rather than outlines of individual mines). Second, predictions at our grid resolution are more suitable for comparing and combining models built on imagery with

those built with geological covariates, which are common in previous literature but cannot be expected to resolve mining up to the ∼5m pixel resolution of the satellite imagery.

7.1.3 Data selection

After labeling, we are left with a dataset that contains 23,061 locations. Throughout the main text, we use a subset of these locations (N=14,638) after filtering out observations that do not meet a set of quality criteria. This filtering primarily addresses three issues. (1) *Confidence thresholds in manual labeling*: labels are excluded if the research assistant expressed low certainty (confidence scores <3 on a 1-5 scale). (2) *Cell inspection levels*: for some locations, the research assistant's view window only partially overlapped a target grid cell. When that overlap was below 20%, we exclude the cell to ensure sufficiently thorough inspection (except when the researcher identified a mine, in which case we retained the cell). (3) *Commercial mines*: commercial mines are excluded from training and testing because their large and well-defined footprint could confound models intended to detect smaller-scale artisanal operations. Appendix B provides more details on data selection criteria and additional results under alternative filtering approaches. Table **??** summarizes all data selection decisions and final sample sizes.

7.1.4 Defining areas monitored by existing ground-based efforts

IPIS's enumeration teams do not canvas entire countries: their missions focus on specific provinces or smaller sub-provincial areas — often known mining hotspots. Unfortunately, IPIS only reports suspected mining locations and does not record the area where their enumerators search for mines. To approximate the spatial extent of these monitoring efforts, we define a monitored area to include any grid cell within three kilometers of an ASM site in the IPIS or NMA data. To implement this on our standardized 1 km grid, we begin by estimating the average size of an ASM site using our hand drawn polygon labels. We then buffer each suspected ASM point (from IPIS or NMA) by the equivalent circular radius of the average mine size. Intersecting these buffered circles with our 1 km grid yields a set of grid cells that have ground based monitoring labels of ASM. We then expand each of these grid cells by three additional grid cells (i.e., 3 km) in all directions, and take the unary union of all resulting squares to define the "monitored area" – i.e., grid cells likely monitored by enumerators searching for mines.

For the purposes of comparing our predicted ASM outputs with the quality of existing ground-based survey efforts (e.g., in Fig. 2(c)-(d)), we consider all grid cells that intersect ASM sites present in the IPIS or NMA data to be positive as evaluated by ground-based efforts. All remaining grid cells within the monitored area are considered negative. Crucially, this is a presence-only definition: IPIS and NMA identify only those mines they suspect to exist, without documenting the regions they surveyed but deemed non-mining. Hence, areas enumerators may have visited and found to be negative do not appear in our data. Moreover, we note that these suspected sites can include false positives, such as sites not yet mined or with spatial inaccuracies. Despite these limitations, defining a monitored area in this manner allows us to compute indicative measures of true and false positive rates within the ground-based monitoring data, providing a reference point against which we can compare the performance of our machine-learning-based predictions.

7.1.5 Defining a uniform at random sample

Our protocol allows us to reconstruct a near-random sample of grid cells from each country, which is used as a "uniform-at-random" (UAR) sample in performance evaluation (e.g., see Fig. 2(b)). To do this, we retain all cells outside of the convex hulls (defined in Methods Section 7.1.1) and resample a smaller number of grid cells within the convex hulls, such that all areas have an equal probability of inclusion. Using this representative sample, we can unbiasedly estimate the share of cells containing ASM activity for each of the five countries in our training sample.

7.2 Feature sets

7.2.1 Imagery-derived random convolutional features

Our "imagery" models use 4,000 random convolutional features (RCF) [53] extracted from satellite images. Recent research shows that RCF performs well in a variety of remote sensing tasks, while being highly computationally efficient [34, 54]. The input imagery is from Norway's International Climate and Forests Initiative (NICFI) Satellite Data Program with source imagery from Planet Labs, Inc. [55, 56]. The images are composite biannual surface reflectance mosaics from the first half of 2020. The surface reflectance mosaics have four channels (red, green, blue, and near-infrared) at 4.77 m spatial resolution. This data product is preprocessed to remove cloud cover, other obfuscations, and distortions. Importantly, some image artifacts may persist. We further process the satellite imagery by normalizing pixel values prior to feature extraction. Several normalization strategies are considered; these are detailed in Section 7.4.1.

Following normalization, we employ the `RCF` class in the TorchGeo Python module [57] to generate our random convolutional features. Specifically, the `RCF` class constructs a single-layer convolutional network with randomly initialized filters that remain fixed throughout training. We use 4,000 output features, specifying the kernel size (4), bias (-0.1), and mode (empirical); all of which we tune in model selection (Table 2; Section 7.4.1). A nonlinear activation (e.g., ReLU) is applied to the convolved outputs, and the resulting feature maps are flattened to produce a feature vector for each image. Because the filters are not updated via backpropagation, this process is computationally efficient while preserving relevant color, texture, spatial patterns in the data.

7.2.2 Geographic features

Our "geographic" models use 268 unique features related to geology (lithology, presence of gold-suitable bedrock, distance to faults and deposits) [8, 23], topography (elevation, distance to rivers, surface water) [24–26], landcover (ecosystems, landuse) [27–29], climate (rain, temperature) [30], and economic development (distance to roads, greenhouse gas emissions) [31, 32]. We experiment with including electromagnetism following [33], but find that it does not boost performance, as these data are missing for large parts of COD. We identify geographic model features based on those that have been used in prior efforts to predict mining activity, though existing work only uses a subset of the variables we collect. We group these 268 features into 6 categories (e.g., "topography", "landcover"), as shown in Table 3. This grouping enables the feature selection experiments described in Section 7.4.2.

7.3 Model training and evaluation

All code for model training and evaluation is available at `github.com/cullen-molitor/asm-paper`. We use the scikit-learn [58] implementation of ridge classifier, random forest, and isotonic regression models.

7.3.1 Model overview and data separation practices

We train three types of models for our main experiments. First, an imagery-based model uses RCF features in a penalized ridge regression. Second, a geography model uses geographic features in a random forest algorithm. Finally, an ensemble model is a weighted average of predictions from the other two models, with weights determined endogenously, as described below.

We divide our labeled data into training and testing sets at the ratio of 4:1, stratifying the sample by country. Within the training set, we use five-fold cross-validation to pick hyperparameters that optimize out-of-fold performance using the area under the ROC curve (AUC) for each model: in the random forest, the number and depth of trees; in the ridge regression, the penalization strength; in the ensemble model, the weight placed on predictions from each of the two prior models. These optimal hyperparameters are then used to train the final models on all training data. Then, we generate predictions for the test

set and report corresponding performance. To characterize uncertainty in model predictions, we repeat this procedure ten times; in each of these iterations, we re-randomize which observations are placed into the training and testing sets.

7.3.2 Predictive models

For the implementation of our imagery-based model, we closely follow the methods of Rolf *et al.* [34]. Specifically, we train a ridge classifier with a custom 5-fold cross-validation to pick the optimal penalization parameter in an expanding grid search to ensure the chosen optimal parameter is not the minimum or maximum of all supplied. We apply isotonic calibration to the model outputs to convert the linear predictions into estimated probabilities.

For the geography model, we implement a random forest with a hyperparameter grid search over the number of trees (`n_estimators`) and maximum tree depth (`max_depth`). Specifically, we search over `n_estimators` in $\{50, 100, 200\}$ and `max_depth` in $\{4, 8, \text{None}\}$.

We combine the predictions from our geographic (random forest) and imagery (ridge) models via a simple weighted average:

$$\widehat{p}_{i,ensemble} = \omega\, \widehat{p}_{i,geography} \;+\; (1-\omega)\, \widehat{p}_{i,imagery},$$

where $\hat{p}_{i,geography}$ and $\hat{p}_{i,imagery}$ are the estimated probabilities that grid cell i contains ASM from the geography and imagery models, respectively. The scalar $0 \leq \omega \leq 1$ is the ensemble weight placed on the geographic model's predictions. To select ω, we search over a grid of possible weights to maximize the area under the ROC curve (AUC) on out-of-fold predictions (i.e., the validation set) across all candidate hyperparameter configurations for the two base models. Once the best weight and base-model hyperparameters have been identified, we retrain these models on the full training set and calculate their predictions on the test set. Finally, we combine the two sets of predictions using the optimal ensemble weight, yielding the ensemble predictions reported.

Throughout the main text, we report proportional changes in performance across model modifications relative to a baseline of random guessing (AUC = 0.5). To do so, we compute:

$$\frac{|\text{AUC(Original)} - .50| - |\text{AUC(Modified)} - 0.50|}{|\text{AUC(Original)} - .50|} \times 100$$

.

7.4 Modifying features for improved spatial extrapolation

7.4.1 Imagery normalization and RCF tuning

As described above, we extract random convolutional features (RCFs) from Planet imagery using a single-layer convolutional network with randomly initialized filters (Methods Section 7.2). Before generating these RCFs, we first normalize the raw imagery and select which bands to include. We additionally tune several parameters related to the RCF extraction process itself (i.e., the kernel size, the bias term, and the method for sampling the random convolution filters). To make these decisions optimally, we conduct a grid search over all possible combinations of our selected normalization and tuning choices, leading to 470 distinct imagery models. Table 2 details the relevant parameters, their possible values, and brief descriptions.

Before computing random convolutional features, we choose which imagery bands to include (`RGB` or `RGB&NIR`) and how to normalize pixel values prior to feature extraction (`max`, `min-max`, or `z-score`). We also vary the reference group for normalization (`image-band`, `image-all`, `quad-band`, `country-band`, or `dataset-band`). At one extreme, *dataset-band* uses the full dataset as the reference group for normalization. For example, for z-score normalization with a *dataset-band* reference group, we compute the mean and standard deviation across all images for each band and translate each pixel value into a corresponding z-score. At the other extreme, *image-band* limits each normalization reference group to

Parameter	Values	Description
`patches`	{empirical, gaussian}	Distribution for sampling random convolution filters. *empirical*: draws filter weights by resampling from pixel intensity distributions. *gaussian*: draws filter weights from a standard normal distribution.
`kernel`	{3, 4, 6}	Size of the convolution kernel, i.e., the spatial footprint each filter covers. Larger kernels capture a broader spatial context.
`bias`	{-1.0, -0.1, -0.01}	Constant added to feature maps before activation. More negative values can lead to sparser activations.
`bands`	{RGB, RGB&NIR}	Spectral channels used for generating RCFs. *RGB*: red, green, blue. *RGB&NIR*: includes near-infrared as well.
`norm`	{max, min-max, z-score}	Type of normalization applied to pixel intensities. *max*: divides pixel values by the (local or global) maximum. *min-max*: linearly scales to the [0,1] range. *z-score*: standardizes values to zero mean, unit variance.
`reference`	{image-band, image-all, quad-band, country-band, dataset-band}	Reference group for computing normalization parameters (e.g., min, max, mean, std). *image-band*: compute stats for each band *within each image*. *image-all*: compute stats jointly over *all bands* within the same image. *quad-band*: compute stats for each band across an entire *quad* ($\sim$380 km^2). *country-band*: compute stats for each band across an entire country. *dataset-band*: compute stats for each band using the full dataset.

Table 2: Imagery normalization and RCF tuning parameters. Summary of parameters explored in our grid search for extracting random convolutional features (RCFs). Each parameter is combined with each of the others, producing 470 unique models.

be within a single image for each band. For example, for z-score normalization with an *image-band* reference group, the pixel values in the near-infrared (NIR) band would be normalized using a mean and standard deviation computed only over the NIR pixel values in the specific image being processed.

Once the imagery is pre-processed, the next step is to determine the random convolutional feature extraction parameters. We focus on three parameters; 1) `bias`, 2) `kernel size`, and 3) `patches`. These parameters and the values we search over in our tuning process are described in Table 2. We consider bias values of {-1.0, -0.1, -0.01}. More negative biases can zero out large portions of the feature map, potentially increasing sparsity. We test kernel sizes of {3, 4, 6}. Intuitively, larger kernels capture more contextual information. We compare two distributions for sampling patches (filter weights). `empirical` draws filter weights by resampling from the observed distribution of pixel intensities in our image dataset, while `gaussian` draws from a standard normal distribution.

After conducting a full grid search, we select a set of imagery parameterizations to maximize full-sample average validation AUC, while utilizing all 4 spectral bands. We find that a model using `RGB&NIR` bands, normalizing values by the `max` in each `image-band`, and specifying a `bias` of -0.1, `kernel` size of 4, and `empirical` weight sampling delivers the highest full-sample AUC while leveraging all available spectral bands. The imagery models in all experiments adopt these choices for pre-processing and RCF extraction.

We compare imagery models based on these specifications to a "baseline" imagery model that uses the RCF pipeline from the MOSAIKS API (available at https://www.mosaiks.org/). This comparison is made to demonstrate the effects of imagery normalization and feature tuning on the ability of the

model to spatially generalize (see Figures 3 and D.2). This "baseline" employs the `RGB` bands, divides all pixel values by 255 (equivalent to `max` normalization at the `dataset` level), and uses `empirical` sampling of weights with a `bias` of -1.0 and `kernel` sizes of 4 and 6 (each producing 2,000 features, for a total of 4,000). These normalization choices are common in the broader literature [42] and therefore provide a valuable comparison specification for both the full-sample and out-of-domain experiments.

7.4.2 Geographic feature selection

In Figures 3 and D.1, we show how the geography model performs at predicting ASM in new geographies. To investigate the possibility that feature selection could improve out-of-domain performance, we group our 268 geographic features into seven categories and evaluate performance on models that iteratively omit each category of features. Table 3 shows these groupings: *Coordinates*, *Infrastructure*, *Climate*, *Topography*, *Geologic Distance*, *Geologic Classes*, and *Landcover*. As noted below, the *Coordinates* features (i.e., latitude and longitude) serve as a spatial interpolation baseline and are not considered when developing our final geographic models.

Category	Features
Coordinates	Longitude, Latitude
Infrastructure	Carbon Dioxide Emissions (2019), Min. Dist. to Road, # Intersecting Roads
Climate	Maximum Monthly Temperature, Minimum Monthly Temperature, Total Precipitation
Topography	Elevation, Min. Dist. to River, # Intersecting Rivers, Surface Water
Geologic Distance	Min. Dist. to {Astrobleme, Carbonatite, Kimberlite, Volcano}, Min. Dist. to {Fault, Inferred Fault, Inferred Normal Fault, Inferred Thrust Fault, Normal Fault, Thrust Fault}
Geologic Classes	Gold Suitability, 28 Lithological Classes (GLiM), 40 Stratigraphic Ages, 61 Geologic Notations, 14 Lithologies, Presence of {Astrobleme, Carbonatite, Kimberlite, Volcano}
Landcover	60 Ecosystem Types, 21 Land Cover Types (GlobCover), 16 Land Cover Types (Copernicus)

Table 3: Geographic feature category groupings. The coordinate category is used as a benchmark for model performance and is not considered when evaluating models.

To systematically evaluate which of the six categories contribute most to predictive performance in spatially held out evaluation sets, we conduct an exhaustive search over all non-empty subsets of these six categories. Since there are $2^6 = 64$ such subsets, and we additionally allow inclusion or exclusion of the *Coordinates* category for completeness, we evaluate a total of $2^7 - 1 = 127$ unique category combinations. For each combination, we train a random forest (Methods Section 7.3.2) on data from four of our five countries, then evaluate performance in the held-out country. We repeat this out-of-domain process five times, each time leaving out a different country. To summarize each model's performance across all five test folds in a single statistic, we compute a sample-weighted average of the area under the ROC curve (AUC). Comparing the weighted AUCs across all 127 feature-category combinations reveals which geographic features provide the greatest gains in out-of-country predictive power. Results can be seen in Fig. D.1.

In Fig. D.1, we compare the performance from various versions of this geography-based model to a benchmark of spatial interpolation between labeled data points (i.e., the *Coordinates*-only model). To construct this benchmark, we train a random forest model using only the location's latitude and longitude as predictors. We apply the same parameter grid search described above to tune this model. To measure how similar each feature group's predictions are to a spatial interpolation baseline, we store the test set predictions for each category combination and compute the coefficient of determination (R^2) when regressing those predictions on *Coordinates*-only predictions trained and evaluated on the same data splits. An R^2 close to 1 indicates that the final predictions from the two models are nearly identical. We

collect and plot these R^2 values against the performance in the full-sample experiment (Fig. D.1(b)) and against the sample weighted average in the out-of-country experiment (Fig. D.1(d)).

7.5 Generating predictions of ASM activity at scale

7.5.1 Estimating ASM probability and generating binary predictions

Each of our trained models produces a continuous predicted probability that a given grid cell contains ASM activity. To generate binary classifications (i.e., ASM or no ASM) and map the sector's spatial footprint, we convert these probabilities to 0/1 predictions by applying country-specific thresholds. Below, we describe how we set these thresholds, how we aggregate predictions for display, and how predictions are used in subsequent analyses.

We start by applying each of the 30 trained models (10 for each model type based on 10 random train/test splits of the sample, as described above), to each grid cell in all 20 countries (5 in-sample countries and 15 out-of-sample countries). We then find the median predicted probability for each model type in each location. Often, practitioners convert such continuous predictions into binary classifications using a single global threshold (e.g., 0.5 [59]) or by choosing the cutoff that optimizes a metric like Youden's index (e.g., Youden [60]). Because of the diversity of ASM activity across our countries of interest, we instead set country-specific thresholds.

Specifically, for the 5 countries in our training sample (SLE, CAF, COD, TZA, ZWE), we rely on our manually collected labels in our unbiased representative sample of locations (UAR; see Section 7.1.5) to estimate the share of grid cells that contain ASM activity in each country. Specifically, for the predictions of a given model, we select the threshold that ensures the predicted proportion of positive cells matches the proportion estimated in our labeled data from the UAR sample. To capture uncertainty in this threshold, we repeat the sampling process outlined above in Section 7.1.5 that generates the UAR sample 1,000 times. For each resample, we compute the fraction of cells that contain ASM in the labeled data. The mean of these fractions across all 1,000 samples is used to generate a threshold value and resulting point estimate for the prevalence of ASM. The 2.5^{th} and 97.5^{th} percentiles of these fractions are used to form a threshold range, resulting in a 95% confidence interval of ASM prevalence.

For the 15 countries not included in our training set, we lack the representative labeled data that we use to calibrate thresholds for in-sample countries. Instead, we estimate the share of mining-affected cells in each out-of-sample country using independent auxiliary data. Specifically, we collect per capita ASM employment data from the World Bank [1] and use our five in-sample countries to regress the fraction of positive ASM cells observed in our labeled training set on total ASM employment at country level (slope coefficient $\hat{\beta}$=0.002 (SE=0.001); R^2=0.66). We then predict the fraction of mined cells in each new country using country-level data on ASM employment. This prediction is then used to set a probability threshold in the same manner as described above. Using the residual variability from this regression, we form a lower (2.5%) and upper (97.5%) bound for the predicted fraction of mined cells.

Once we obtain a point estimate (and bounds) for each country's fraction of mined cells, we convert these to country-specific classification thresholds by taking percentiles of our grid cell level predicted probabilities. For example, if a country's expected prevalence is $\hat{p}$, we set the threshold at the $100\% - \hat{p}$ percentile of predicted probabilities. The lower and upper bounds analogously determine low and high thresholds. Grid cells whose median probabilities exceed the threshold are classified as ASM-positive, while all others are classified as negative.

7.5.2 Assessing the spatial and environmental footprint of ASM

To visualize predicted ASM activity (e.g., in Fig. 4) while maintaining a level of privacy, we aggregate predicted ASM labels in each country's $0.01^\circ \times 0.01^\circ$ grid onto a coarser $0.05^\circ \times 0.05^\circ$ grid. To do so, we sum the total number of predicted positive 0.01° cells within each coarse 0.05° grid cell, thereby preserving the spatial distribution of ASM activity without disclosing exact coordinates of any potentially unauthorized mining sites.

In Table 1, we measure the extent to which ASM activity intersects with: protected areas, using polygons from UNEP-WCMD & IUCN [46]; biodiversity hotspots, using polygons from Hoffman *et al.* [47]; and monitored areas, using regions defined in Section 7.1.4. Each $0.01^\circ \times 0.01^\circ$ grid cell is classified as overlapping one of these three regions of interest if its centroid lies within any relevant polygon. We then identify whether that cell was also classified as positive for ASM, providing an estimate of the fraction of these areas that is affected by artisanal mining.

In Table D.1, we quantify how many people live near ASM activity. To do so, we use population rasters from Schiavina *et al.* [61], assigning each $0.01^\circ \times 0.01^\circ$ cell its corresponding population. Summing the population over cells classified as containing ASM provides a direct measure of the total population living in mined cells. Additionally, we label each grid cell as within an urban center or not using the Florczyk *et al.* [62] polygons.

We detect clusters of ASM activity by computing adjacency graphs in which each $0.01^\circ \times 0.01^\circ$ cell classified as containing ASM is linked to its neighbors (within 3km). We define a cluster as any connected component containing at least 10 positive cells. Specifically, we use the `DBSCAN` algorithm in `Python` to detect clusters. For each country, we first select all $0.01^\circ \times 0.01^\circ$ cells classified as mining and extract their centroids, treating positive cells closer than 0.03° as neighbors. Any group of at least ten points is labeled a valid cluster. To measure how far each cluster is from officially documented ASM data (e.g., IPIS/NMA), we compute, for each predicted cluster, the minimum distance to any administratively labeled ASM site. This metric aids in distinguishing newly detected hotspots from those proximate to sites in monitored areas, guiding analyses of whether our predictions identify previously undocumented mining areas or extend known mining regions.

Acknowledgments

We thank Jason Leland, Saloni Majmudar, Jarod Ngo, and Ophelia Sin for their invaluable assistance in hand-labeling satellite imagery for this research. We are also grateful to Jihae Hong, Juan Sebastián Leiva Molano, and Eki Ramadhan for overseeing the research assistants and providing essential guidance throughout the project. We thank Anna Boser for constructive feedback on the manuscript and Tracey Mangin for figure design assistance.

This work utilized high-performance computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara.

Data availability

All data needed to evaluate the conclusions of this study will be made publicly available at the time of publication. Because of the sensitive nature of the training and prediction data, low-resolution (5 km) aggregations of the data will be accessible through an online repository, while high-resolution labeled training data and model predictions will be provided upon request.

Code availability

All code used in this study will be made publicly available at the time of publication. Code will be available at (`github.com/cullen-molitor/asm-paper`).

Funding statement

The labeling of satellite imagery was supported by a grant from the Hewlett Foundation.

Competing interests statement

The authors declare no competing interests.

References

1. Bank, W. *2020 State of the Artisanal and Small-Scale Mining Sector* (World Bank, Washington, D.C., 2020). https://www.delvedatabase.org/resources/2020-state-of-the-artisanal-and-small-scale-mining-sector.
2. Bank, W. *2019 State of the Artisanal and Small-Scale Mining Sector* (World Bank, Washington, D.C., 2019). https://www.delvedatabase.org/resources/state-of-the-artisanal-and-small-scale-mining-sector.
3. Laing, T. & Pinto, A. N. Artisanal and small-scale mining and the low-carbon transition: Challenges and opportunities. *Environmental Science & Policy* **149,** 103563. ISSN: 1462-9011. https://www.sciencedirect.com/science/article/pii/S1462901123002125 (2023).
4. Hilson, G. Farming, small-scale mining and rural livelihoods in Sub-Saharan Africa: A critical overview. *The Extractive Industries and Society* **3,** 547–563. ISSN: 2214-790X. https://www.sciencedirect.com/science/article/pii/S2214790X16300132 (2016).
5. Gibb, H. & O'Leary, K. G. Mercury Exposure and Health Impacts among Individuals in the Artisanal and Small-Scale Gold Mining Community: A Comprehensive Review. *Environmental Health Perspectives* **122,** 667–672. https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.1307864 (2014).
6. Landrigan, P. *et al.* Reducing disease and death from Artisanal and Small-Scale Mining (ASM) - the urgent need for responsible mining in the context of growing global demand for minerals and metals for climate change mitigation. *Environmental Health* **21.** https://doi.org/10.1186/s12940-022-00877-5 (2022).
7. Caballero Espejo, J. *et al.* Deforestation and Forest Degradation Due to Gold Mining in the Peruvian Amazon: A 34-Year Perspective. *Remote Sensing* **10.** ISSN: 2072-4292. https://www.mdpi.com/2072-4292/10/12/1903 (2018).
8. Victoire Girard, T. M.-M. & Vic, G. *Artisanal Mining in Africa* Working Paper Series, ISSN 2183-0843, Working Paper No. 2021. Mar. 2022.
9. Gandiwa, E. & Gandiwa, P. Biodiversity conservation versus artisanal gold mining: a case study of Chimanimani National Park, Zimbabwe. *Journal of Sustainable Development in Africa* **14,** 29–37 (2012).
10. Blair, G., Christensen, D. & Rudkin, A. Do commodity price shocks cause armed conflict? A meta-analysis of natural experiments. *American Political Science Review* **115,** 709–716 (2021).
11. Saavedra, S. Technology and State Capacity: Experimental Evidence from Illegal Mining in Colombia. *SSRN*. https://ssrn.com/abstract=3933128 (2024).
12. Barenblitt, A. *et al.* The large footprint of small-scale artisanal gold mining in Ghana. *Science of the Total Environment* **781,** 146644 (2021).
13. Mahboob, M. A. & Genc, B. *Evaluation of ISODATA Clustering Algorithm for Surface Gold Mining Using Satellite Data* in *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (2019), 1–6.
14. Forkuor, G., Ullmann, T. & Griesbeck, M. Mapping and Monitoring Small-Scale Mining Activities in Ghana using Sentinel-1 Time Series (2015–2019). *Remote Sensing* **12.** ISSN: 2072-4292. https://www.mdpi.com/2072-4292/12/6/911 (2020).
15. Gallwey, J., Robiati, C., Coggan, J., Vogt, D. & Eyre, M. A Sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in Ghana: Applying deep learning to shallow mining. *Remote Sensing of Environment* **248,** 111970. ISSN: 0034-4257. https://www.sciencedirect.com/science/article/pii/S0034425720303400 (2020).

16. Nyamekye, C., Ghansah, B., Agyapong, E. & Kwofie, S. Mapping changes in artisanal and small-scale mining (ASM) landscape using machine and deep learning algorithms.-a proxy evaluation of the 2017 ban on ASM in Ghana. *Environmental Challenges* **3,** 100053 (2021).

17. Ngom, N. M. *et al.* Mapping artisanal and small-scale gold mining in Senegal using Sentinel 2 data. *GeoHealth* **4,** e2020GH000310 (2020).

18. Chaussard, E. & Kerosky, S. Characterization of Black Sand Mining Activities and Their Environmental Impacts in the Philippines Using Remote Sensing. *Remote Sensing* **8.** ISSN: 2072-4292. https://www.mdpi.com/2072-4292/8/2/100 (2016).

19. Couttenier, M., Di Rollo, S., Inguere, L., Mohand, M. & Schmidt, L. Mapping artisanal and small-scale mines at large scale from space with deep learning. *PLOS ONE* **17,** 1–13. https://doi.org/10.1371/journal.pone.0267963 (Sept. 2022).

20. Rigterink, A. S., Ghani, T., Lozano, J. S. & Shapiro, J. N. Mining Competition and Violent Conflict in Africa: Pitting Against Each Other. *The Journal of Politics.* Publisher: The University of Chicago Press, 000–000. ISSN: 0022-3816. https://www.journals.uchicago.edu/doi/full/10.1086/730743 (2025) (Apr. 3, 2024).

21. Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E. & Meyer, H. Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography* **32,** 356–368 (2023).

22. Maus, V. *et al.* A global-scale data set of mining areas. *Scientific Data* **7,** 179 (2020).

23. (ggis@un-igrac.org), I. A. *AFR CGMW-Brgm 1:10m geological units* International Groundwater Resources Assessment Centre (IGRAC), July 2020.

24. NASA JPL. *NASADEM Merged DEM Global 1 arc second V001* NASA EOSDIS Land Processes DAAC, 2020.

25. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540,** 418–422 (2016).

26. Lehner, B. & Grill, G. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes* **27,** 2171–2186 (2013).

27. ESA. *GlobCover: Global Land Cover Map* 2009.

28. Buchhorn, M. *et al.* Copernicus Global Land Cover Layers - Collection 2. *Remote Sensing* **12,** 1044 (2020).

29. Sayre, R. G. *et al.* A new map of standardized terrestrial ecosystems of Africa. *U.S. Geological Survey* (Mar. 2013).

30. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a High-Resolution Global Dataset of Monthly Climate and Climatic Water Balance from 1958-2015. *Scientific Data* **5,** 170191 (2018).

31. Center for International Earth Science Information Network - CIESIN - Columbia University & Information Technology Outreach Services - ITOS - University of Georgia. *Global Roads Open Access Data Set, Version 1 (gROADSv1)* https://doi.org/10.7927/H4VD6WCT. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC), 2013.

32. Earth Science Data Systems, N. *Greenhouse Gases Data Pathfinder* Sept. 2021.

33. Rigterink, A. S. Diamonds, Rebel's and Farmer's Best Friend: Impact of Variation in the Price of a Lootable, Labor-intensive Natural Resource on the Intensity of Violent Conflict. *Journal of Conflict Resolution* **64.** Publisher: SAGE Publications Inc. ISSN: 0022-0027. https://doi.org/10.1177/0022002719849623 (Jan. 1, 2020).

34. Rolf, E. *et al.* A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications* **12,** 4392 (2021).

35. Saavedra, S. & Romero, M. Local incentives and national tax evasion: The response of illegal mining to a tax reform in Colombia. *European Economic Review* **138,** 103843 (2021).

36. Rolf, E. Evaluation challenges for geospatial ML. *arXiv preprint arXiv:2303.18087* (2023).

37. Benson, V. & Ecker, A. Assessing out-of-domain generalization for robust building damage detection. *arXiv preprint arXiv:2011.10328* (2020).

38. Kerner, H., Sundar, S. & Satish, M. *Multi-Region Transfer Learning for Segmentation of Crop Field Boundaries in Satellite Images with Limited Labels* in *AAAI Conference on Artificial Intelligence (AAAI) Workshops* (2023). https://ai-2-ase.github.io/papers/14%5CSubmission%5CField_boundary_delineation___AAAI_2023_AI2SE-camera-ready.pdf.

39. Ortiz, A. *et al. Local context normalization: Revisiting local normalization* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 11276–11285.

40. Kellenberger, B., Tasar, O., Bhushan Damodaran, B., Courty, N. & Tuia, D. Deep domain adaptation in earth observation. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences,* 90–104 (2021).

41. Huang, X. & Belongie, S. *Arbitrary style transfer in real-time with adaptive instance normalization* in *Proceedings of the IEEE international conference on computer vision* (2017), 1501–1510.

42. Corley, I., Robinson, C., Dodhia, R., Ferres, J. M. L. & Najafirad, P. *Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), 3162–3172.

43. Rußwurm, M., Wang, S., Kellenberger, B., Roscher, R. & Tuia, D. Meta-learning to address diverse Earth observation problems across resolutions. *Communications Earth & Environment* **5,** 37 (2024).

44. Meyer, H., Reudenbach, C., Wöllauer, S. & Nauss, T. Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecological Modelling* **411,** 108815 (2019).

45. Rolf, E., Jordan, M. I. & Recht, B. *Post-estimation smoothing: A simple baseline for learning with side information* in *International Conference on Artificial Intelligence and Statistics* (2020), 1759–1769.

46. UNEP-WCMD & IUCN. *Protected Planet: The World Database of Protected Areas (WDPA)* version 1.6 (UNEP-WCMC and IUCN, Aug. 2023). https://www.protectedplanet.net/en/thematic-areas/wdpa?tab=WDPA.

47. Hoffman, M., Koenig, K., Bunting, G., Costanza, J. & Williams, K. J. *Biodiversity Hotspots (version 2016.1)* version 2016.1 (Zenodo, Sept. 2019). https://doi.org/10.5281/zenodo.3261807.

48. Khachiyan, A. *et al.* Using neural networks to predict microspatial economic growth. *American Economic Review: Insights* **4,** 491–506 (2022).

49. Murillo-Sandoval, P. J. *et al.* The post-conflict expansion of coca farming and illicit cattle ranching in Colombia. *Scientific Reports* **13,** 1965 (2023).

50. Van Den Hoek, J. & Friedrich, H. K. Satellite-based human settlement datasets inadequately detect refugee settlements: a critical assessment at thirty refugee settlements in Uganda. *Remote Sensing* **13,** 3574 (2021).

51. Likas, A., Vlassis, N. & J. Verbeek, J. The global *k*-means clustering algorithm. *Pattern Recognition. Biometrics* **36,** 451–461. ISSN: 0031-3203. https://www.sciencedirect.com/science/article/pii/S0031320302000602 (2025) (Feb. 1, 2003).

52. Couttenier, M., Di Rollo, S., Inguere, L., Mohand, M. & Schmidt, L. Mapping artisanal and small-scale mines at large scale from space with deep learning. *PLOS ONE* **17,** 1–13. https://doi.org/10.1371/journal.pone.0267963 (Sept. 2022).

53. Rahimi, A. & Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems* **21** (2008).

54. Sherman, L., Proctor, J., Druckenmiller, H., Tapia, H. & Hsiang, S. M. *Global high-resolution estimates of the United Nations Human Development Index using satellite imagery and machine-learning* tech. rep. (National Bureau of Economic Research, 2023).

55. Climate, N. I. & (NICFI), F. I. *Satellite Data Program* Kongens gate 20, Oslo, 2020. https://www.nicfi.no/.

56. PBC, P. L. *Planet Application Program Interface: In Space for Life on Earth* Planet, 2020. https://api.planet.com.

57. Stewart, A. J. *et al. TorchGeo: Deep Learning With Geospatial Data* in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22)* (Association for Computing Machinery, Seattle, Washington, Nov. 2022), 1–12. ISBN: 9781450395298. https://dl.acm.org/doi/10.1145/3557915.3560953.

58. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

59. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning* ISBN: 978-1-4899-0519-2 978-0-387-21606-5. http://link.springer.com/10.1007/978-0-387-21606-5 (2025) (Springer, New York, NY, 2001).

60. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3,** 32–35. ISSN: 0008-543X (Jan. 1950).

61. Schiavina, M., Freire, S., Carioli, A. & MacManus, K. *GHS-POP R2023A - GHS population grid multitemporal (1975-2030)* (European Commission, Joint Research Centre, May 8, 2023). http://data.europa.eu/89h/2ff68a52-5b5b-4a22-8f40-c41da8332cfe.

62. Florczyk, A. *et al. GHS Urban Centre Database 2015, multitemporal and multidimensional attributes, R2019A* Dataset. Publisher: European Commission, Joint Research Centre (JRC). Jan. 28, 2019. http://data.europa.eu/89h/53473144-b88c-44bc-b4a3-4583ed1f547e.

Supplementary Materials

Appendix

Table of Contents

A Hand Labeling of Satellite Imagery

After compiling a sample of candidate locations, we use a custom application to manually label high-resolution optical satellite imagery with polygons drawn around ASM mining operations. Additional metadata such as the research assistant's confidence (scale of 1-5) in their label and the suspected mine type (none, artisanal, or commercial) are recorded. This process is standardized among multiple research assistants by providing a detailed training with corresponding instructional material, as outlined in this Appendix. The document copied below in Appendix A.1 was presented to each research assistant and was used as guidance for labeling each provided location. In total, 4 undergraduate student assistants participated in labeling with some locations presented to more than one assistant as a means of checking for consistency.

A.1 Instructions for manual labeling

1. Open the remote sensing task application on the following website: `https://YY.shinyapps.io/geocodemines/?coder=XX`

2. Replace your name in the "XX" in the above URL to access your assigned geopoints.

    - research assistant 1
    - research assistant 2
    - research assistant 3
    - research assistant 4

3. Once you have entered your name, you should see a landing site similar to Figure 1. You are now ready to begin the task of identifying mining areas.

Figure A.1: Remote sensing app landing page

4. Do you see a mining area on screen?

   If No: Go to (13)

   If Yes: Go to (5)

   Please use the following identification process to confirm whether you are seeing a mining area:

    a. Zoom out until the full circle is visible on screen.

    b. Scan (within the circle) for areas that contrast in color to their immediate surroundings (lighter colored areas should attract your attention the most)

        - Cleared (more precision) vs not-cleared/forest (higher confidence)

c. Zoom in to area of interest and assess:
   - Looking at the land "texture" are there any sure signs of ASM?
     (i) Pits
     (ii) Scratched rock/exposed rocks/man-made clearing
     (iii) Disrupted soil (lighter dirt color and may have textured surface or holes)
     (iv) Discolored pools (do not match the surrounding environment and are generally light brown or green)
   - If not, is there evidence of human activity nearby?
     (i) Shacks
     (ii) Roads or trails (to settlements)
     (iii) Settlements (usually white boxes)
     (iv) Unnatural pools of water
   - If not, is the area next to a key landmark? (increases chances of ASM)
     (i) River
     (ii) Delta
     (iii) Deforestation
   - If not, is this area visibly different from the areas immediately surrounding it? Zoom out and pan around if necessary.
   - If not, is the content of the circle as a whole different from the area surrounding it? Really zoom out and pan around to get an assessment of the surrounding landscape.

d. If any of the above questions are YES, the likelihood of the area being ASM increases (with higher weight given to the first few questions)

e. Repeat the process with all other areas of interest within the circle.

Keep in mind:

- You can use the "+" and "-" on the left side of the application to zoom in and out (or scroll using your mouse) to confirm whether you can see a mining area.
- You can also use the point's coordinates (lat and long) shown on the app to review the point on Google Earth Pro
- Human activity, such as settlements, or key landmarks, such as deforestation, by itself do not automatically mean there is ASM activity. These have to be accompanied by other features in 4.C.a.i.
- Tree shadows may look like pits. If there are surrounding forests, the holes are likely trees

5. Is the mining area within the 1km circle?

   If No: Go to (13)

   If Yes: Go to (6)

   Keep in mind:

   - If a mining area is completely within the circle, the answer is YES
   - If any part of the mining area is within the circle, the answer is YES (even if part of the mining area extends outside of the circle)
   - If a mining area is completely outside of the circle, the answer is NO

   If you DO see a mining area within the 1 km circle:

6. How many different mining areas do you see within the 1 km circle?

Figure A.2: Research assistant view window with multiple polygons drawn.

- Note: It is possible to see multiple mining areas within the 1 km circle.

7. Use the drawing tool on the application to draw a polygon over the identified mining area(s) as seen in Figure 2. Go to (8) when you have completed drawing the polygons.

   Keep in mind:

   - When drawing, try to follow the boundaries of the mining area as closely as possible; zooming in will help increase precision.
   - Try to include the least amount of "non-mining area" inside the polygon as possible; this means drawing multiple polygons if needed to break up big spaces of greenery in between mining areas, individual holes, and separate plots will require multiple polygons.
   - If multiple ASM pits (>10) are clustered together into a single area, draw a single polygon for this specific area.

8. Answer the following question for the identified mining area(s): On a scale from 1 to 5, where 1 is "No confidence" and 5 is "Very confident", how confident are you that this IS a mining area?

   - Please use the following rating scale for guidance:
     - **Very high confidence (5):** "Ticks all the boxes", i.e. it includes all features from 4.A.C.a
       - Definitely a mining area, no doubt about it, textbook ASM. There is evidence of: lightly colored areas, definitive pits, clearly exposed rock, shacks, and/or (discolored) water pools nearby.
     - **High confidence (4):** "Very close to 5, but one feature is off"
       - A very strong candidate to be a mining area, but one feature within the image may suggest otherwise.
       - There is exposed rock and/or discolored pools near cleared areas, but it is hard to tell whether there are definitive pits (they may be trees); or
       - It is unclear whether the boundaries drawn are accurate
     - **Moderate confidence (3):** "It ticks one or two of the boxes"

        - It is likely a mining area and there are vague signs of ASM (areas that resemble pits, cleared land, lightly colored water), yet...
        - The boundaries of the mine are unclear, and/or
        - There are some possible sites that were not drawn due to ambiguity, and/or
        - The disturbed soil is not farmland or deforestation but does not resemble typical ASM, and/or
        - There is human activity, but it is unknown if for the purpose of ASM.
    - **Low confidence (2):** "It ticks only one box"
        - It may be a mining area, there is a sign of ASM activity. Yet, the feature that suggests ASM may be ambiguous.
    - **No confidence (1):** "No boxes are ticked"
        - There is poor/blurry imagery (e.g., cloud cover, insufficient satellite imagery).
        - Ambiguous activity.
- Select your answer from the drop-down menu.

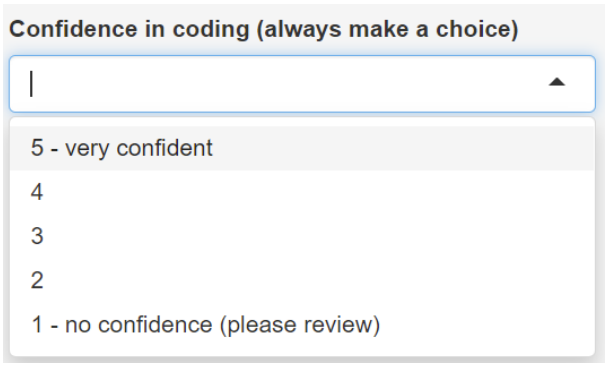

Figure A.3: Confidence score menu

- Note: All mining areas must be accompanied by a confidence rating.

- If you answered "1 – no confidence": Go to (9).
- If you answered otherwise: Go to (10).

9. Answer the following question for the identified mining area(s):

    a. Select your answer from the drop-down menu.

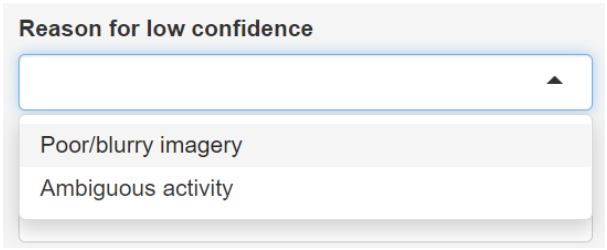

Figure A.4: Low confidence reason menu

10. Answer the following question for the identified mining area(s):

    - What type of mining area did you draw a polygon over?

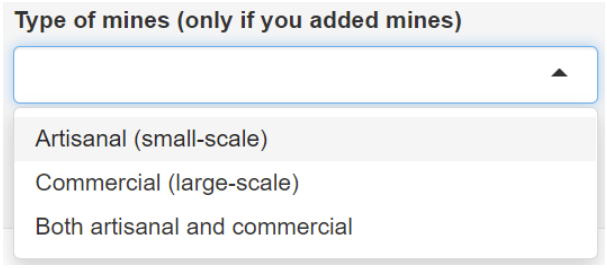

Figure A.5: Mine-type menu

- Select your answer from the drop-down menu.
  Note: All mining areas must be labeled with a mining type.
  Keep in mind:
  * Most sites are artisanal, so unless you suspect otherwise, choose "Artisanal".
  * However, if the site contains deep and large pits that are clearly delineated and/or large, straight roads and/or large structures/trucks/machinery, choose "Commercial".
  * If the "Commercial" site also includes smaller pits in a random pattern in a separate area within the circle, choose "Both Artisan and Commercial".

11. Once you have completed capturing the data for the identified mining area(s), please click "Finish task and save".

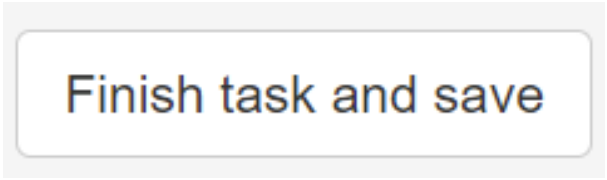

Figure A.6: Finish and save button

12. A new geopoint to inspect will appear once you click on "Finish task and save". To review the instructions to restart the task, go to (4). If you are finished identifying mining areas, please close the app.

    If you do NOT see a mining area on screen:

13. Answer the following question for the identified non-mining area:

    On a scale from 1 to 5, where 1 is "No confidence" and 5 is "Very confident", how confident are you that this is NOT a mining area?

    Please use the following rating scale for guidance:

    - **Very high confidence (5):**
      - Definitely not a mining area: no doubt about it, not a single suspicious area in sight (e.g., complete green forest, middle of the water, large human settlement within radius).
    - **High confidence (4):**
      - Most likely not a mining area: strongly indicative of the natural environment, perhaps a tiny bit of uncertainty in some places (features that remotely look like ASM), but not much.

  - Non-natural.
- **Moderate confidence (3):**
  - Likely not a mining area: probably a result of natural forces, although somewhat resembles human activity. Usually characterized by potential areas that differ in color from their immediate surroundings and have deceiving properties of ASM (e.g., close to settlement or river).
  - Cleared land with no pits/objects that are most likely trees.
- **Low confidence (2):** (ambiguous activity - has now become 2)
  - Borderline not a mining area: difficult to tell but I would say it is more likely to be a non-mining area than a mining area. Usually characterized by areas that closely resemble pits and exposed rock, but not for certain.
  - Cleared land but no sign of pits/scratched surface.
- **No confidence (1):**
  - Poor/blurry imagery (e.g., cloud cover, insufficient satellite imagery): I usually only mark a confidence level of 1 for data points where it is impossible to see the Earth's surface (poor/blurry imagery). If the imagery is sufficient, I usually try to make a definitive decision on whether there is mining activity or not and mark it with a confidence level of 2-3.
  - If the image is blurry and is brown or contains brown; If the circle contains human activity that can't be classified.
  - No idea — very blurry or ambiguous (but lighter area somewhat visible).

- Select your answer from the drop-down menu.

Note: All non-mining areas must be accompanied by a confidence rating

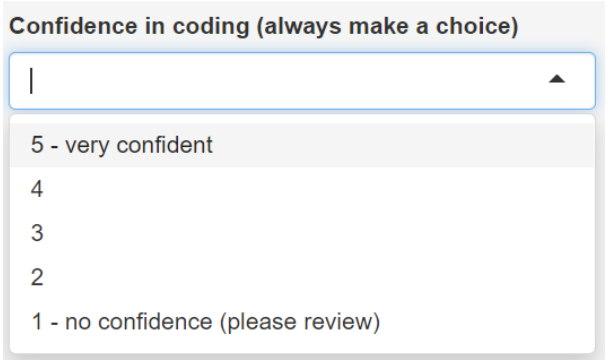

Figure A.7: Confidence menu

- If you answered "1 – no confidence": Go to (14).
- If you answered otherwise: Go to (15).

14. Answer the following question for the identified mining area(s):
    - Select your answer from the drop-down menu
15. Once you have completed capturing the data for the identified non-mining area, please click "Finish task and save".
16. A new geopoint to inspect will appear once you click on "Finish task and save". To review the instruction to restart the task, go to (4). If you are finished identifying mining areas, please close the app.

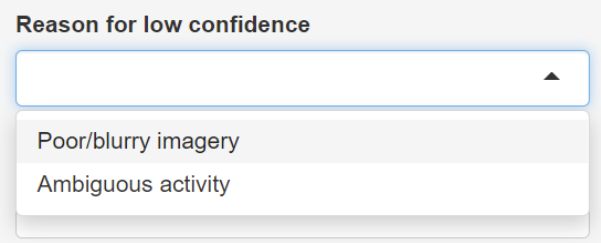

Figure A.8: Low confidence reason menu

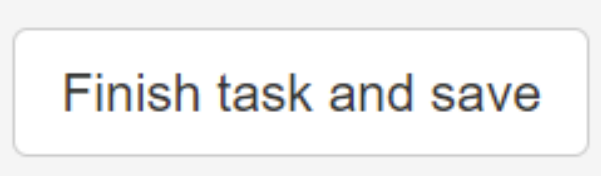

Figure A.9: Finish and save button

A.2 Manual labeling FAQs

A set of frequently asked questions emerged during the manual labeling procedure. After an initial labeling pilot, we created the following FAQs and included these answers in the instructions provided to research assistants.

A.2.1 Identifying mining areas

- How do I identify a mining area?
  - To understand how to identify mining and non-mining areas, please review the slide deck prepared for training. This document has images that showcase different types of mining and non-mining areas that can be used as guidance for visually identifying these areas.

A.2.2 Labeling the data

- How do I choose a confidence level?
  - 5: if you have very high confidence that you have identified a (non-) mining area
    - "Definitely a mine, no question about it, boundaries are clear"
  - 4: if you have high confidence that you have identified a (non-) mining area
  - 3: if you have moderate confidence that you have identified a (non-) mining area
    - "Pretty sure that this is a mine, boundaries unclear"
  - 2: if you have low confidence that you have identified a (non-) mining area
  - 1: if you have no confidence that you have identified a (non-) mining area
    - Use this option if you are completely unsure on how to label the data, we will review all entries with confidence levels equal to 1.
      - You are unsure because: the image is blurry or has poor resolution
      - You are unsure because: there is ambiguous activity on screen (i.e. it can be mining, but it also can be agriculture or a human settlement).
- How do I select a mining type?

  - Artisanal (small-scale):
    * "ASM is a collective term embracing both small scale and artisanal mining. It covers formal or informal mining which is characterized by low capital intensity and high labor intensity and relatively simple methods for exploration, extraction and processing." - World Gold Council
    * Visual characteristics for this type of mining include:
      · The mining areas are small, usually smaller than 500m
      · It is common to see multiple small pits in a mining area, when compared to larger commercial mines, it is evident that these have been dug using rudimentary tools and machinery (less organized, no clear patterns).
  - Commercial (large-scale):
    * "Large-scale mining is highly mechanized, and has industrial and capital-intensive operations that are usually run by multinational companies." – Extractives Hub
    * Visual characteristics for this type of mining include:
      · The mining areas are very large
      · Areas around commercial mines usually lack vegetation, large machinery has deforested these areas or large roads are built to transport the minerals on trucks.
      · The pits being dug are clean-cut. It is common to see large pits delineated by straight lines or with ridges, a result from using heavy machinery.

- How do I label data when I have identified multiple mining areas on the same screen?
  - You must answer the questions for all mining areas on screen, not for each individual mining area identified
    * 1 mining area: 1 response on confidence and type
    * 2 mining areas: 1 response on confidence and type
    * N mining areas: 1 response on confidence and type
  - If you are unsure of one mining area (confidence level of 2), and sure of the other (confidence level of 4), please answer with the minimum confidence level

A.2.3 Using the application

- How do I draw a polygon?
  - Select the "Draw a polygon" on the left-hand side of the interface

Figure A.10: Polygon icon

  - Use your mouse to click the first point of the polygon. With each subsequent click you add an additional point on your polygon. To complete the drawing, you must either click back to the first point or select "Finish" on the toolbar.

- What is the correct way to draw a polygon?
  - The polygon should contain the least amount of non-mining area possible. That is, the area should stick close to / closely delineate the mining area, even if it means drawing an irregular shape to avoid capturing non-mining areas.

Figure A.11: Polygon menu

- Should I always draw a polygon?
  - If possible, you should always draw a polygon. However, in instances where the mining area is too small i.e., a single pit, you should add a marker.
- How do I get to the original zoom level?
  - Use the following tool reset zoom / center image tool to arrive at the predetermined zoom level.

Figure A.12: Zoom icon

B Sample selection

We restrict the full dataset of 23,061 labeled 0.01° grid cells to a smaller sample used for training and evaluation based on three quality criteria:

1. Inspection certainty: We restrict data based on the confidence score provided by research assistants during their investigation of high-resolution imagery.
2. Inspection coverage: We restrict data based on the fraction of a grid cell that was directly inspected by a research assistant.
3. Mining category: We exclude grid cells containing commercial mines.

In the following subsections, we detail our procedure for sample selection for each of these criteria.

B.1 Inspection certainty

When labeling satellite imagery, research assistants were asked to outline mining activity and attempt to identify the type (none, commercial, or artisanal). Additionally, they were asked to rank their confidence in the chosen designation on a scale from 1-5 (as detailed in Appendix A). Confidence levels 1 and 2 represent ambiguous activity, and compose just 3% of our sample (Table B.1). Therefore, throughout our analysis we set a minimum confidence level of 3, lowering the estimating sample from the original 23,061 observations to 21,348.

Table B.1 reports the data loss under this sample selection criteria, as well as hypothetical data losses under more strict confidence cutoffs. It additionally reports cross-validated AUC results under any confidence threshold ranging from keeping all the data (confidence threshold $\geq$1) to only retaining the highest certainty observations (confidence threshold ≥ 5). Results show that predictive performance improves when higher confidence labels are retained.

	Data Loss		Model AUC		
Confidence Threshold	Observations Dropped	Percentage of Data Dropped	Imagery	Geography	Ensemble
≥ 1	0	0.00%	0.828	0.872	0.884
≥ 2	644	3.04%	0.834	0.879	0.89
≥ 3	1,700	8.02%	0.849	0.897	0.907
≥ 4	4,986	23.51%	0.881	0.93	0.938
$= 5$	11,542	54.43%	0.917	0.957	0.967

Table B.1: Data loss and model performance metrics by confidence threshold. In each row, confidence thresholds indicates the certainty value above which data is retained (certainty values are reported by research assistants as they manually label high-resolution imagery, as outlined in Appendix A). Commercial mines are excluded from this table. Corresponding model performance is shown evaluated on the held out test set. The decision to to use a confidence threshold of $\geq$3 was decided upon before construction of this table, which is provided for transparency and diagnostic purposes only.

B.2 Inspection coverage

During labeling, each latitude-longitude sampling point was presented to a research assistant with a 0.005° circle super-imposed over very high resolution satellite imagery (see Figure A.1). For the locations not provided by IPIS and NMA, the centroid of the circle was the exact center of a 0.01° grid cell, as our sampling regime uses this standardized grid (see Methods Section 7.1.1). Thus, for these observations, the research assistant inspected a minimum of 78.5% of the grid cell (the percentage of total that area a circle occupies inside a square). There is a potential that a research assistant inspected past

the boundary of the circle (and thus coverage was higher than 78.5%) for two reasons: 1) the boundary of a mine extended through the circle, in which case they were instructed to continue drawing outside of the circle; or 2) a research assistant's attention was naturally drawn to the areas outside of the circle if it looked like potential mining activity.

In contrast, the locations identified by IPIS and NMA as suspected ASM activity were not centered on our standardized grid and thus required spatial merging with the grid (Methods Section 7.1.1). In these cases, a single latitude-longitude point and its corresponding circular view window can intersect with as many as 4 grid cells. We therefore calculate the overlap between each viewing window and our standardized grid to compute a cell inspection coverage value indicating the fraction of the grid cell covered by the research assistant's circular viewing window.[1]

Figure B.1 shows the distribution of inspection coverage values for all 23,061 grid cells in our full dataset. The vast majority of cells have coverage of 78.5%. Inspection coverage values above 78.5% occur due to the close proximity of sampling points in mining clusters, which causes overlap in view windows. In addition, the commercial mining dataset from Maus *et al.* [1] contains polygons outlining large mining operations, some of which fully cover one or more grid cells.

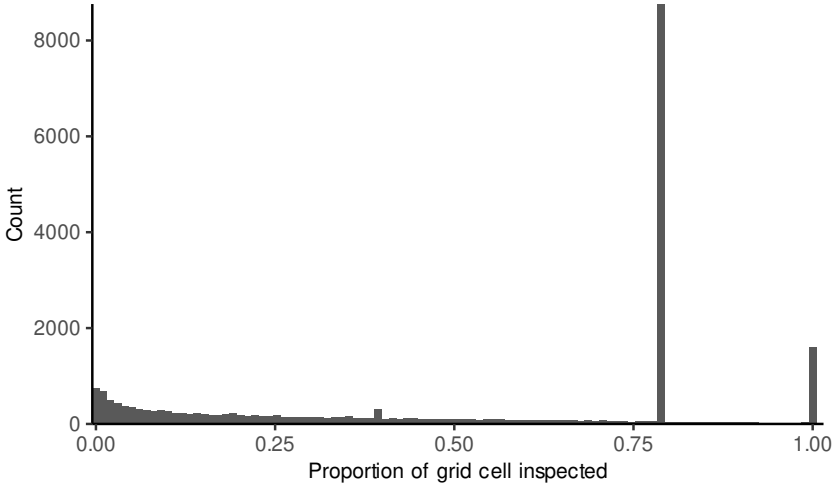

Figure B.1: Distribution of grid cell inspection coverage. Figure displays the distribution of grid cell inspection coverage values, indicating the proportion of the grid cell overlapping with manual labeler's circular viewing windows.

To ensure high quality of our training data, we exclude locations with a low inspection coverage level, unless the grid cell has been labeled as a positive (i.e., containing ASM activity). We retain positive cells as our aim is to predict presence or absence of ASM and any presence of ASM (even if detected within a small area of a grid cell) renders the grid cell a positive. To determine a threshold for inclusion based on inspection coverage, we balance data quality against data quantity and label imbalance across positive (i.e., ASM detected) and negative (i.e., no ASM detected) classes. Table B.2 shows, for inspection coverage thresholds increasing from 0 to 0.35, the quantity of data lost as well as the effects on class imbalance in the observations located within the convex hulls of suspected mining activity defined in Methods Section 7.1.1.

[1] As noted above, it is likely that research assistants looked at areas outside of their circular viewing window. Therefore, these inspection coverage values serve as lower bounds on the areas visually inspected in practice.

Inspection Threshold	Data Loss			Model AUC		
	Observations Dropped	Percentage of Data Dropped	Ratio of Positives to Negatives (near suspected ASM)	Imagery	Geography	Ensemble
≥ 0	0	0.00%	0.3368	0.828	0.849	0.869
≥ 0.05	2621	11.37%	0.4242	0.836	0.870	0.883
≥ 0.10	3852	16.70%	0.4829	0.839	0.876	0.889
≥ 0.15	4728	20.50%	0.5357	0.840	0.886	0.896
≥ 0.20	5437	23.58%	0.5878	0.849	0.897	0.907
≥ 0.25	6008	26.05%	0.6378	0.852	0.901	0.910
≥ 0.30	6471	28.06%	0.6851	0.856	0.905	0.914
≥ 0.35	6867	29.78%	0.7314	0.860	0.907	0.916

Table B.2: Data loss and model performance metrics by inspection coverage threshold. In each row, inspection coverage thresholds indicate the overlap between the 0.01° grid cell and the circular view window shown to research assistants. Commercial mines are excluded from the table. The ratio of positives to negatives is computed excluding commercial mines and sampling points outside convex hulls surrounding suspected mining clusters (see Methods 7.1.1) as inspection thresholds are not relevant in those cases. Model performance is measured on a held out test set and is used for diagnostic purposes only; the selection of a threshold of $\geq$0.20 was made before AUC values were calculated.

Based on the first panel of Table B.2 indicating data loss and class imbalance, we chose an inspection threshold of 20% as it provides a balanced tradeoff between the total number of observations lost and the quality of the data, while maintaining an appropriate ratio of positives to negatives. The second panel of Table B.2 indicates that our results are not very sensitive to this choice, although they improve when more restrictive inspection coverage thresholds are used.

B.3 Commercial mine labels

Our aim is to detect artisanal and small-scale mining. However, it is possible that labeled data on commercial mining activity could improve model performance when evaluated on ASM, as the visual signature of commercial mines may be similar to ASM. In this section, we evaluate whether including commercial mines in our training dataset improves our ability to detect ASM in our held-out test dataset. To do so, we follow the prior sample selection criteria and set a confidence level of ≥ 3 and an inspection coverage threshold of $\geq$ 20%. We then use the same modeling workflow as our main results, and iterate through the inclusion and exclusion of commercial mines in our training data. When we include commercial mines in training, we exclude them from our validation data to better match our test set. This is only to pick hyperparameters, after which, the models are trained on the full training data which includes commercial mines. We repeat this experiment for the in-sample experiments and the out-of-domain experiments in which an entire country is left out of training.

Table B.3 shows the results of this exercise. Including commercial mines training in the full sample marginally reduces model performance when evaluated on the randomly held out test set (first two rows). However, for many countries, including commercial mines improves predictive performance in the out-of-domain experiment in which that country's data is left out of the training sample (remaining rows of the table; including commercial mines improves out-of-domain performance for all countries except Sierra Leone). We therefore include commercial mines in the training data when evaluating out-of-domain performance and when predicting ASM in 15 out-of-sample countries.

		Model AUC		
Test Location	Commercial Mines	Imagery	Geography	Ensemble
Full sample	exclude	0.849	0.897	0.907
	include	0.845	0.89	0.899
CAF	exclude	0.788	0.579	0.733
	include	0.791	0.576	0.744
COD	exclude	0.766	0.64	0.753
	include	0.768	0.713	0.797
SLE	exclude	0.854	0.654	0.845
	include	0.85	0.725	0.843
TZA	exclude	0.695	0.66	0.715
	include	0.733	0.654	0.743
ZWE	exclude	0.63	0.653	0.66
	include	0.669	0.667	0.713

Table B.3: Comparison of model performance with versus without inclusion of commercial mines in the training sample. "Full sample" indicates a random train and test split of the full data sample across 5 countries, where the train and test sets are stratified by country (see Methods Section 7.3). Reported results for these two rows represent the average AUC score over 10 random train and test splits, as in Methods Section 7.3. Reported results for each country represent the AUC score when that country is held out of training as the evaluation set.

C Image Model Benchmark

C.1 Image segmentation

Much of the past work using satellite imagery and machine learning to detect ASM or other mining operations has cast the problem as a semantic segmentation problem [2–5], in which the task is to mask (or "segment") out which pixels in an image correspond to a mine and which do not. For the segmentation task, past work has found that (different variants of) U-Net architectures [6] outperform per-pixel random forests for this task, which can themselves be a strong baseline in machine learning with remote sensing data. In contrast to this prior literature, in this study we make image-level predictions at the 1km$\times$1km grid cell resolution, as described in Section 7.5. This is motivated by the policy relevance of identifying mining areas, as opposed to outlining individual mines.

To validate our choice of the random convolutional features (RCF) model for image-level predictions, we compare our approach to an alternative computer vision approach that casts our problem first as a semantic segmentation problem before then converting per-pixel predictions into an image-level prediction. Based on the findings of past work, we train a U-Net model to optimize a segmentation (pixel-level) loss. We then convert the per-pixel predictions output by the U-Net into image-level predictions by one of two methods: (a) taking a simple average of per-pixel predictions, and (b) training a small convolutional network to take in segmentation predictions and output image-level predictions.

C.2 Training Details

We use a U-Net architecture [6] with a ResNet18 backbone and pre-trained ImageNet weights, which we find gives comparable-to-better performance than using Sentinel-2 pre-trained weights or training from scratch. We adjust the standard U-Net model to have four input channels (RGB+NIR) and two output channels (no mine/mine). We also include batch normalization and dropout layers after each nonlinear activation layer to address overfitting, consistent with past work [2]. Without dropout, we observed a large disparity in pixel-level performance (IoU) between train and validation. After experimenting with different dropout rates (0.05, 0.1, 0.3) and alternate methods of regularization (increasing weight decay, reducing encoder depth), we find a dropout rate of 0.1 to be the most effective in both improving test performance and closing the gap between train and validation metrics. We supplement this with image augmentation by introducing random flips and rotations during training.

We normalize the images per channel with min-max normalization, such that the values in each individual image channel range from 0 to 1, as described in Section 7.4.1. We also try using dataset-level normalization, in which we divide all images by the maximum value of each channel across the entire dataset. This produces comparable global metrics but lower image-level performance for out-of-domain experiments (Supplementary Table C.2). We resize all images and target masks after normalization to be 256$\times$256 pixels, employing bilinear interpolation with antialiasing for the images and nearest-neighbor interpolation for the binary label segmentation masks.

We perform a 64%-20%-16% train-test-validation split stratified by country on the imagery data. Training is performed with a batch size of 64 and a weighted cross-entropy function to account for class imbalance. In light of results of past work [2, 3], we also experimented with a focal loss objective function, which adjusts standard cross-entropy loss to focus on misclassified examples [7], but found that this led to unstable training at higher learning rates and had comparable performance to weighted cross-entropy at lower learning rates.

We utilize the AdamW optimizer with a learning rate scheduler that drops the global learning rate by a factor of ten after ten epochs without a drop in validation loss. The maximum number of training epochs is 100, with the potential for early stopping after ten epochs without a decrease in validation loss. We perform a hyperparameter sweep over a grid of learning rate $(10^{-3}, 10^{-4}, 10^{-5})$ and class weights ([0.5,0.5], [0.3,0.7], [0.1,0.9]). We take the best hyperparameter configuration to be the one that achieves the highest positive class IoU (Intersection Over Union) at a prediction threshold of 0.5 during training.

The selected parameters were a learning rate of 10^{-4} and class weights of $[0.3, 0.7]$. The resulting

model was trained for 66 epochs (truncated due to early stopping) on dataset of 9,369 images took roughly 90 minutes on a NVIDIA A100-SXM4-40GB GPU. Test-time inference on 2,929 images took roughly 80 seconds.

We consider two methods to translate the per-pixel predicted probabilities from the U-Net into image-level output. We perform this translation both to allow for direct performance comparison with RCF, as well as to simulate downstream tasks that require image-level predictions.

1. Take the average of pixelwise probabilities to be the image-level score. We consider alternatives, such as taking the maximum pixel probability or the average of pixelwise probabilities exceeding a given threshold, but find a simple average to produce the best performance.

2. Train a small CNN to take in a heat map of predicted probabilities as input and classify the image as either containing a mine or not. This model consists of three convolutional layers with 16, 32, and 64 filters separated by 2×2 max pooling and ReLU activation layers, followed by a final fully connected layer that transforms the flattened convolutional output into a two-channel output.

C.3 Results

Table C.1 compares the performance of our U-Net implementation to RCF on the basis of image-level prediction, the key task in this paper. Full sample performance of the two U-Net models and the RCF model are comparable, and taking the mean of pixelwise probabilities performs on par with using a small CNN to translate predictions from the pixel level to the image level. The U-Nets have slightly better out-of-domain performance than RCF in CAF, COD, and TZA, reflected by less or no drop in performance going from country-specific global inference to country-specific out-of-domain inference. However, the U-Net incurs a greater performance drop than RCF in SLE (-0.07 versus -0.04) and ZWE (-0.09 versus -0.04), which we note are geographically further from the remaining three countries, likely making spatial extrapolation more difficult. Figure C.1) shows that the U-Net qualitatively performs well at finding difficult mines after visual inspection (Figure C.1.

Table C.2 shows U-Net performance under different image normalization approaches, as detailed in Methods Section 7.4.1 for the RCF approach. Image normalization has very minimal impact on both full sample and out-of-domain performance for the U-Net. However, consistent with RCF, performance is slightly higher in the out-of-domain experiment when image-level normalization is used.

Full sample performance						
	CAF	COD	SLE	TZA	ZWE	All
RCF	0.815	0.843	0.891	0.782	0.740	0.851
U-Net + pixel average	0.815	0.847	0.882	0.747	0.809	0.852
U-Net + small CNN	0.822	0.847	0.886	0.742	0.812	0.854
Out-of-domain performance						
	CAF	COD	SLE	TZA	ZWE	Aggregated
RCF	0.794	0.775	0.854	0.758	0.698	0.798
U-Net + pixel average	0.824	0.811	0.816	0.798	0.724	0.810
U-Net + small CNN	0.831	0.807	0.817	0.798	0.719	0.809

Table C.1: Comparing image-level predictive performance of our U-Net segmentation model with RCF. Aggregated AUC for out-of-domain experiments is calculated as the average of individual country-specific out-of-domain AUCs, weighted by the number of samples in each country.

Full sample performance						
	CAF	COD	SLE	TZA	ZWE	All
Image-level normalization	0.815	0.847	0.882	0.747	0.809	0.852
Dataset-level normalization	0.807	0.839	0.908	0.765	0.891	0.857
Out-of-domain performance						
	CAF	COD	SLE	TZA	ZWE	Aggregated
Image-level normalization	0.824	0.811	0.816	0.798	0.724	0.810
Dataset-level normalization	0.794	0.817	0.799	0.746	0.685	0.802

Table C.2: Comparing image-level predictive performance of a U-Net model trained with image-level versus dataset-level normalization. For image-level normalization we perform min-max normalization on each channel of an individual image, while for dataset-level normalization we divide channels across all images by the maximum channel values in the dataset.

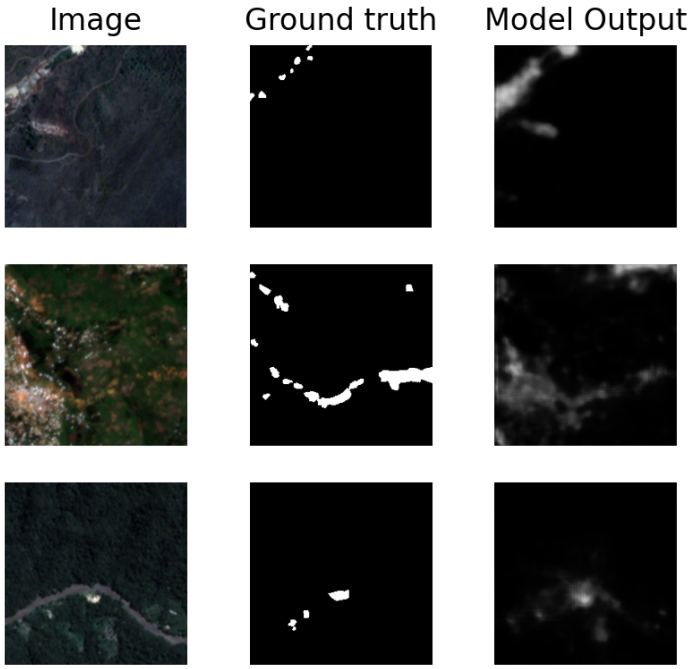

Figure C.1: Normalized input image, ground truth target, and U-Net output for three examples. The model output is a heat map of predicted ASM probabilities and takes values from 0 (black) to 1 (white).

These results show that the RCF model used in our main experiments performs comparably with our U-Net implementation for the task of image-level prediction, with slightly lower performance for three of five countries in the out-of-domain setting in which predictions are made in a country held out of training. These findings, taken together with the relative ease of training and deploying RCF models, motivate our choice of using an RCF vision model for the main experiments. Of course, it is possible that extensions to a U-Net or other architecture could boost performance.

D Supplementary Results

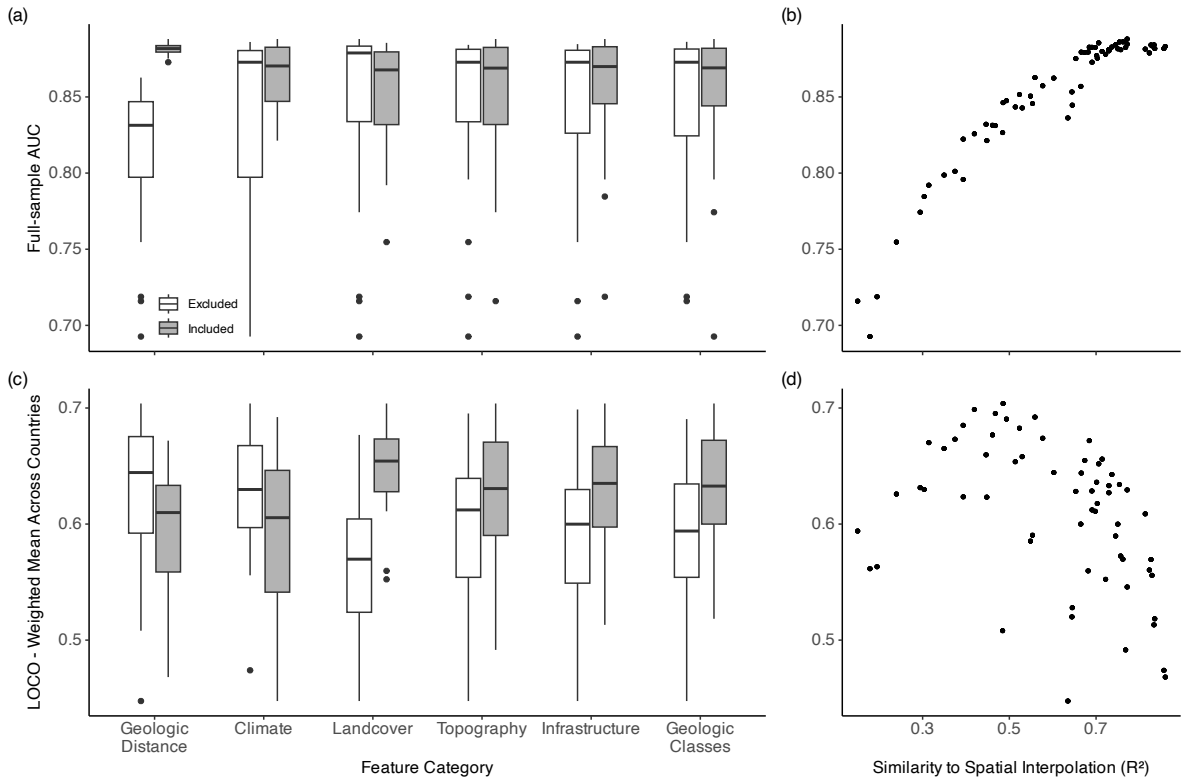

Figure D.1: Distribution of AUC values when feature categories are excluded from (white) or included in (grey) the feature set for the model trained on geographic information only. Fig. D.1(a) plots performance for models trained and tested on the full sample. Fig. D.1(c) plots performance for the out-of-domain experiment, in which target countries for testing are held out of the training set. Figures D.1(b) and D.1(d) plot full sample and out-of-domain performance, respectively (y-axis), against the correlation between the model's predictions and those from a purely spatial interpolated model (x-axis) (Methods Section 7.4.2). In these scatter plots, each point represents one of 64 models trained on geographic information in which zero or more categories of features are omitted. All models trained with coordinates are excluded, as they were intended for diagnostic purposes only.

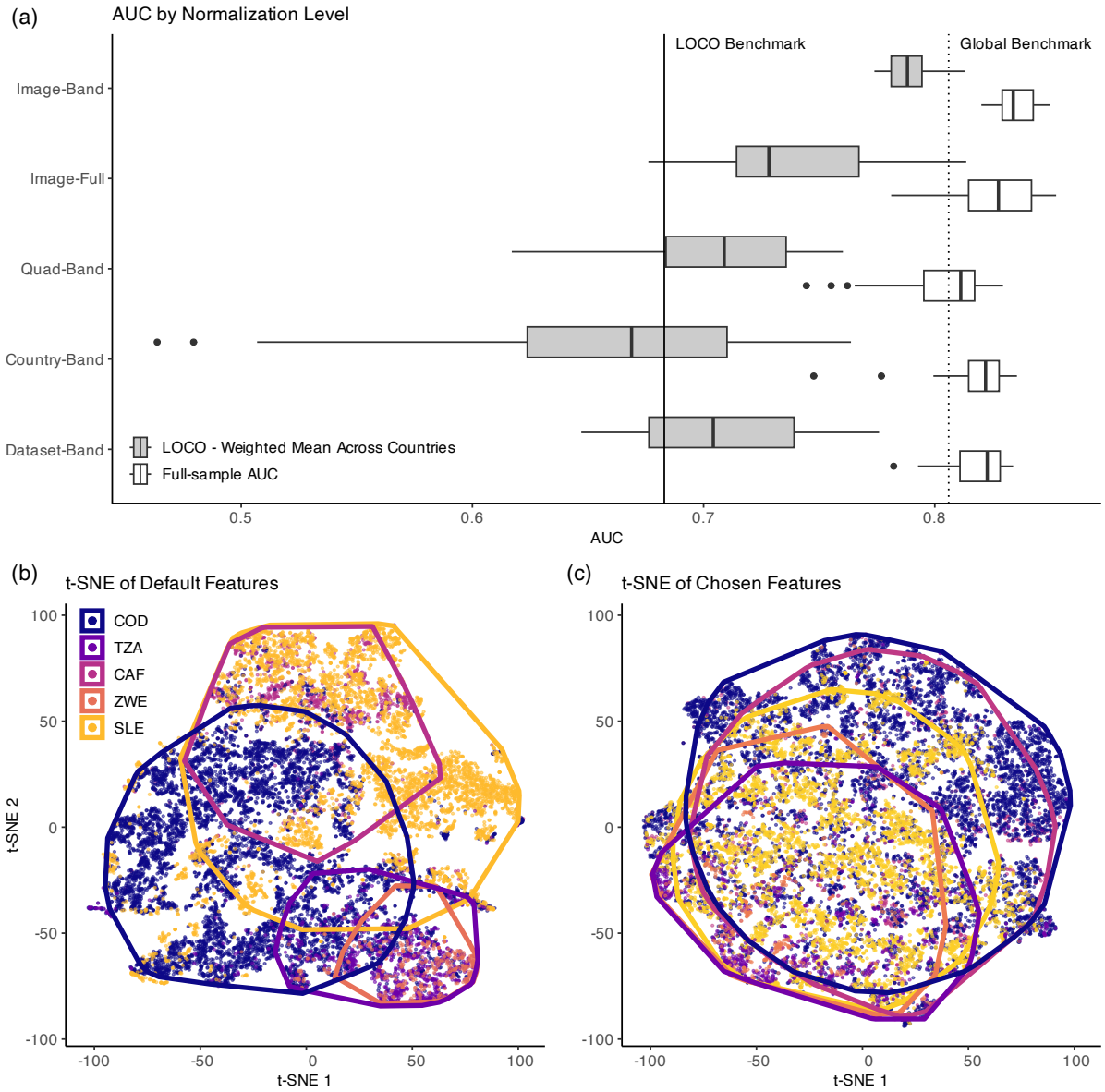

Figure D.2: Model performance for different image pre-processing choices. Fig. D.2(a) plots the distribution of model performance for the full-sample and out-of-domain experiment using different reference groups to scale the pixel values. The dashed and solid vertical lines correspond to performance for the MOSAIKS defaults. Figures D.2(b) and D.2(c) use t-SNE to visualize the 4,000 RCF generated by the MOSAIKS defaults versus our preferred pre-processing choices, respectively. We color observations and include minimum convex polygons (i.e., polygons with interior angles less than 180 degrees that cover 90% of points) by country.

Country	Population	Urban Center		Total
		Inside	Outside	
CAF	Population Count	1,384,737	3,937,572	5,322,309
	Population in Cells Labeled Mining			
	Administrative Labels	0%	1%	1%
	Ensemble Predictions	95%	61%	70%
		[73–98%]	[35–75%]	[45–81%]
COD	Population Count	38,390,940	54,466,547	92,857,487
	Population in Cells Labeled Mining			
	Administrative Labels	1%	1%	1%
	Ensemble Predictions	63%	66%	65%
		[57–68%]	[61–69%]	[60–69%]
SLE	Population Count	2,685,238	5,414,381	8,099,620
	Population in Cells Labeled Mining			
	Administrative Labels	5%	7%	6%
	Ensemble Predictions	25%	35%	32%
		[24–29%]	[32–39%]	[29–35%]
TZA	Population Count	12,449,391	49,039,975	61,489,366
	Population in Cells Labeled Mining			
	Administrative Labels	< 1%	< 1%	< 1%
	Ensemble Predictions	8%	17%	15%
		[4–13%]	[9–24%]	[8–22%]
ZWE	Population Count	2,937,921	12,725,235	15,663,155
	Population in Cells Labeled Mining			
	Administrative Labels	0%	< 1%	< 1%
	Ensemble Predictions	58%	22%	29%
		[37–71%]	[11–31%]	[16–38%]
Other	Population Count	197,845,740	471,054,821	668,900,561
	Population in Cells Labeled Mining			
	Ridge Predictions	17%	7%	10%
		[6–41%]	[2–18%]	[3–25%]

Table D.1: Estimated population exposed to ASM activity Table reports the total population located within 0.01° grid cells estimated to contain ASM activity, within and outside urban centers. Cells containing ASM activity are computed based on administrative labels (second row) and our machine learning predictions (third row). Predictions and clipping thresholds are constructed identically to those mapped in Figure 4. Population estimates are from the Global Human Settlements Layer from the 2020 epoch at 3 arc second resolution, with urban delineations computed from the (GHSL UCDB layer) classification. For the first five in-sample countries listed, we construct bounds (in square brackets) by setting the clipping thresholds to $\widehat{q} \pm 1.96 \times \sqrt{\widehat{q}(1-\widehat{q})/N}$, where $\widehat{q}$ is the proportion of cells in each country that contain ASM, estimated using a uniform-at-random sample of labeled training data. For the remaining countries, clipping threshold bounds are a 95% prediction interval around $\widehat{q}$, which we estimate using a linear model that relates the share of cells with ASM to per capita employment in ASM (see Methods 7.5 for details).

References

1. Maus, V. *et al.* A global-scale data set of mining areas. *Scientific Data* **7,** 179 (2020).
2. Couttenier, M., Di Rollo, S., Inguere, L., Mohand, M. & Schmidt, L. Mapping artisanal and small-scale mines at large scale from space with deep learning. *PLOS ONE* **17,** 1–13. https://doi.org/10.1371/journal.pone.0267963 (Sept. 2022).
3. Nava, L., Cuevas, M., Meena, S. R., Catani, F. & Monserrat, O. Artisanal and Small-Scale Mine Detection in Semi-Desertic Areas by Improved U-Net. *IEEE Geoscience and Remote Sensing Letters* **19,** 1–5 (2022).
4. Gallwey, J., Robiati, C., Coggan, J., Vogt, D. & Eyre, M. A Sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in Ghana: Applying deep learning to shallow mining. *Remote Sensing of Environment* **248,** 111970. ISSN: 0034-4257. https://www.sciencedirect.com/science/article/pii/S0034425720303400 (2020).
5. Malik, K., Robertson, C., Braun, D. & Greig, C. U-Net convolutional neural network models for detecting and quantifying placer mining disturbances at watershed scales. *International Journal of Applied Earth Observation and Geoinformation* **104,** 102510. ISSN: 1569-8432. https://www.sciencedirect.com/science/article/pii/S0303243421002178 (2021).
6. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597 [cs.CV].
7. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal Loss for Dense Object Detection* 2018. arXiv: 1708.02002 [cs.CV].