MONITORING ACCESS TO PIPED WATER AND SANITATION INFRASTRUCTURE IN AFRICA AT DISAGGREGATED SCALES USING SATELLITE IMAGERY AND SELF-SUPERVISED LEARNING

Othmane Echchabi[*]
Mila – Quebec AI Institute
McGill University

Aya Lahlou[*]
Columbia University
Center for Learning the Earth with AI and Physics (LEAP)

Nizar Talty
Duke Kunshan University

Josh Malcolm Manto
Duke Kunshan University

Tongshu Zheng[†]
Duke Kunshan University

Ka Leung Lam[†]
Duke Kunshan University

ABSTRACT

Monitoring SDG 6 at policy-relevant spatial and temporal resolution remains difficult where household surveys are sparse or infrequent. We develop a satellite-based framework that links Afrobarometer rounds 7–9 to Sentinel-2 256×256 patches and self-supervised Vision Transformer embeddings for estimating household access to piped water and sanitation. We benchmark frozen-feature k-NN heads on DINOv3-sat, DINO, and DINOv2, selecting k on validation data and reporting held-out test accuracy, macro-recall, and macro-F1. DINOv2 performs best, reaching 84.13% accuracy for piped water and 87.17% for sanitation. For national reporting, we aggregate patch predictions with HRSL population weights and compare country-level estimates with WHO/UNICEF JMP statistics. Agreement is strong for piped water ($\mathrm{R}^2 = 0.92$) and moderate for sanitation ($\mathrm{R}^2 = 0.72$); in countries without Afrobarometer coverage, weighted MAE is 9.5 percentage points for piped water and 10.7 points for sanitation. These results indicate that satellite-based models can provide timely, screening-level evidence to complement—not replace—official SDG reporting, with greatest uncertainty where label definitions and survey coverage are weakest.

1 INTRODUCTION

Reliable SDG 6 monitoring requires geographically disaggregated estimates of drinking-water and sanitation access, yet many countries rely on expensive and infrequent survey or census programs (United Nations, 2015b;a). This creates temporal and subnational information gaps between reporting cycles and local planning needs.

Satellite imagery offers frequent, continent-scale observations, but piped networks and sanitation infrastructure are often only indirectly visible in optical data (Oshri et al., 2018; Persello et al., 2022; Yeh et al., 2020). Accordingly, we frame the problem as learning transferable proxies from built-environment patterns rather than direct infrastructure detection.

This workshop paper makes three contributions. First, we build a compact SDG 6 pipeline that links Afrobarometer labels, Sentinel-2 patches, and self-supervised Vision Transformer (ViT) representations. Second, we provide held-out model comparisons across DINOv3-sat, DINO, and DINOv2 using a common frozen-feature k-NN evaluation protocol (Caron et al., 2021; Oquab et al., 2023;

[*]These authors contributed equally to this work.

[†]Corresponding authors: tongshu.zheng@dukekunshan.edu.cn, kaleung.lam@dukekunshan.edu.cn.

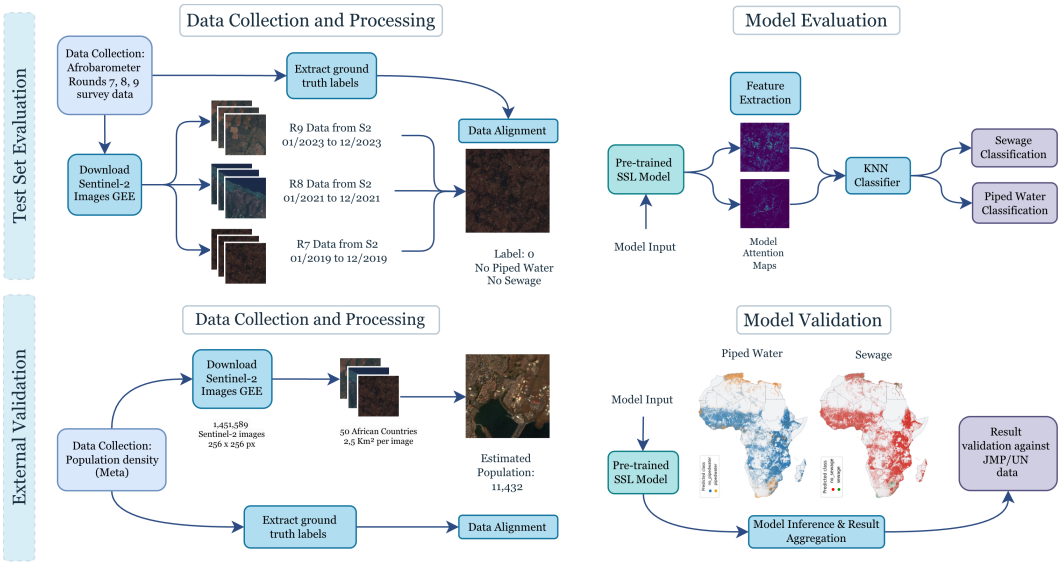


Figure 1: Workflow integrating survey labels, satellite imagery, self-supervised embeddings, and population-weighted aggregation for SDG 6 monitoring.

Siméoni et al., 2025). Third, we validate population-weighted national aggregates against JMP statistics and evaluate transfer to countries with no Afrobarometer coverage (Drusch et al., 2012; WHO/UNICEF Joint Monitoring Programme, 2024).

2 Data and Methods

Survey labels. We use georeferenced Afrobarometer rounds 7–9 (2019, 2021, 2023) and harmonize questionnaire responses into binary outcomes for household access to piped water and sewage/sanitation services (Afrobarometer, 2023).

Remote-sensing inputs. For each survey location, we retrieve cloud-filtered Sentinel-2 imagery and extract 256×256 patches using a Google Earth Engine processing workflow (Drusch et al., 2012; Gorelick et al., 2017).

Representation learning and classification. We evaluate DINOv3-sat (public satellite checkpoint) and DINO/DINOv2 models pretrained in our study, then train softmax-weighted k-NN heads on frozen embeddings. Model selection uses a validation sweep over $k \in \{5, 10, 20, 50, 100, 200\}$. Primary held-out metrics are accuracy, macro-recall, and macro-F1.

Population-weighted aggregation. To align patch predictions with SDG reporting, we compute country-level estimates using HRSL population weights (Facebook Connectivity Lab & Center for International Earth Science Information Network (CIESIN), Columbia University, 2016):

$$\hat{Y} = \frac{\sum_i p_i \hat{y}_i}{\sum_i p_i},$$

where $\hat{y}_i$ is predicted access for patch i and p_i is the corresponding population. External validation compares $\hat{Y}$ with JMP indicators 6.1.1 and 6.2.1 (WHO/UNICEF Joint Monitoring Programme, 2024). Additional implementation settings are summarized in Appendix Section B and Appendix Table 7.

Figure 1 summarizes the workflow.

3 RESULTS

3.1 HELD-OUT PERFORMANCE

Table 1 reports held-out test-set performance using the value of k selected on validation data for each model–task pair. DINOv2 performs best on both outcomes, reaching 84.13% accuracy for piped water and 87.17% for sanitation. Relative to DINOv3-sat, this corresponds to gains of 2.70 percentage points (piped water) and 2.52 points (sanitation). Sensitivity to k is reported in Appendix Tables 3 and 4, and a broader benchmark including Galileo and Prithvi-EO-2.0 is provided in Appendix Table 5.

Table 1: Held-out test-set performance (%) for each model and task using the value of k selected for each model–task pair. Recall and F1 are macro-averaged across classes. Best overall per task is shown in bold; second best is underlined.

Task	Model	Backbone	Pretrain data	Best k	Accuracy	Recall	F1
Piped Water	DINOv3	ViT-L/16	sat (orig.)	10	81.43	81.22	81.19
	DINOv2	ViT-L/8	ours	200	**84.13**	**83.76**	**83.85**
	DINO	ViT-B/8	ours	100	<u>83.81</u>	<u>83.42</u>	<u>83.52</u>
Sewage	DINOv3	ViT-L/16	sat (orig.)	10	84.65	81.06	81.92
	DINOv2	ViT-L/8	ours	200	**87.17**	**84.29**	**85.00**
	DINO	ViT-B/8	ours	100	<u>86.98</u>	<u>84.38</u>	<u>84.88</u>

3.2 NATIONAL-SCALE VALIDATION AGAINST JMP

We run continent-scale inference on 1,453,663 Sentinel-2 patches and aggregate predictions with HRSL population weights. Against JMP country statistics, piped-water estimates show strong correspondence ($R^2 = 0.92$), while sanitation agreement is weaker ($R^2 = 0.72$), likely reflecting indicator-definition mismatch between survey-derived labels and JMP sanitation categories. Coverage and distribution diagnostics across Afrobarometer rounds are provided in Appendix Figure 3.

We additionally evaluate transfer to countries without Afrobarometer coverage. For piped water (11 countries; 193.7M people), population-weighted MAE is 9.5 percentage points. For sanitation (6 countries; 165.7M people), MAE is 10.7 points. Table 2 summarizes aggregate population coverage by error threshold: piped-water estimates fall within 10 and 15 points of JMP for 115.7M (59.7%) and 121.4M (62.6%) people, while sanitation reaches 159.7M (96.4%) within 15 points. Country-level breakdown and error-distribution diagnostics are reported in Appendix Table 6 and Appendix Figure 5; Appendix Figure 4 provides the no-survey MAE scatter.

Table 2: Aggregate external-validation performance in countries without Afrobarometer survey data. Population-weighted MAE is computed against JMP national estimates. Population coverage indicates the number of people living in countries where the absolute error is within a given threshold.

Task	Countries	Population	Weighted MAE	Within 10 pp	Within 15 pp
Piped water	11	193.7M	9.5	115.7M (59.7%)	121.4M (62.6%)
Sanitation	6	165.7M	10.7	58.8M (35.5%)	159.7M (96.4%)

4 DISCUSSION AND CONCLUSION

Our results show that self-supervised satellite representations can recover meaningful SDG 6 access signals at continental scale, particularly for piped water. The framework is computationally lightweight and can be updated as new imagery becomes available, making it suitable for between-survey monitoring.

Construct validity and indicator mismatch. The sanitation task remains intrinsically harder because survey-derived binary labels and JMP service definitions are not perfectly aligned. As a result,

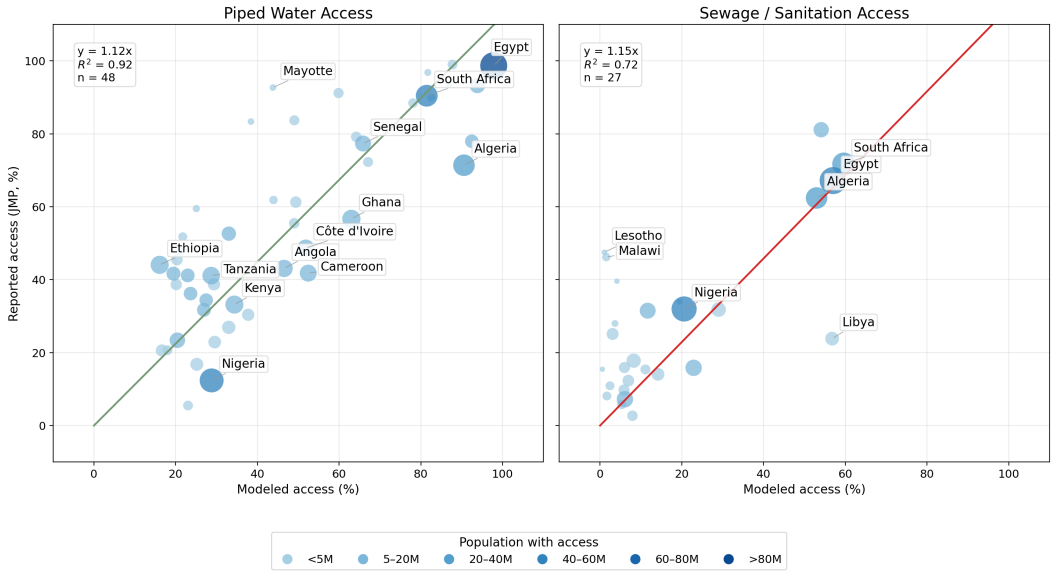


Figure 2: Comparison of model-derived national access estimates vs. JMP references for piped water and sanitation.

high classification performance does not automatically imply equally strong agreement with official sanitation indicators.

Generalization and spatial bias. Model predictions may partially rely on settlement morphology and urban–rural cues rather than direct observation of infrastructure. Performance can therefore degrade under distribution shift (e.g., atypical settlement forms, sparse coverage, or different cloud regimes), especially in countries without survey supervision.

Implications for SDG workflows. These maps are most appropriate as prioritization and diagnostics tools: identifying potential hotspots, tracking broad trends, and targeting field verification. They should complement, not replace, official survey-based reporting and domain-specific validation. Additional spatial and qualitative diagnostics are provided in Appendix Figures 6, 7, 8, and 9.

Future work should prioritize uncertainty quantification, stronger domain adaptation for no-survey countries, and richer multimodal ground truth (e.g., utility network and service-quality data) to improve causal interpretability and policy readiness.

REFERENCES

Afrobarometer. Afrobarometer surveys: Rounds 7–9 (2019–2023), 2023. Accessed 2026-02-07.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2011.11.026. URL https://www.sciencedirect.com/science/article/pii/S0034425712000636. The Sentinel Missions - New Opportunities for Science.

Facebook Connectivity Lab and Center for International Earth Science Information Network (CIESIN), Columbia University. High resolution settlement layer (hrsl), 2016.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Véronique Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Dimitris Samaras, Gabriel Synnaeve, Hervé Jegou, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Barak Oshri, Annie N. Hu, Peter Adelson, Xiaodong Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 616–625, 2018.

Claudio Persello, Jan Dirk Wegner, Ronny Hänsch, Devis Tuia, Pedram Ghamisi, Mila Koeva, and Gustau Camps-Valls. Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):172–200, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint*, 2025. URL `https://arxiv.org/abs/2508.10104`.

United Nations. Data for development: A needs assessment for SDG monitoring and statistical capacity development, 2015a. Accessed 2026-02-07.

United Nations. Goal 6: Clean water and sanitation, 2015b. Accessed 2026-02-05.

WHO/UNICEF Joint Monitoring Programme. Jmp country estimates for wash (households), 2024. Accessed 2026-02-05.

Christopher Yeh, Antonio Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1):2583, 2020.

A Extended diagnostics and qualitative analysis

This appendix reports additional analyses that support the main text while keeping the same narrative style as the paper. Rather than listing figures alone, we explain what each diagnostic shows and how it relates to model behavior, transfer, and SDG 6 reporting utility.

A.1 Model-selection sensitivity across k

In the main text, we report the best-performing k-NN configuration for each backbone. To show robustness of that choice, Tables 3 and 4 summarize test performance across the full k sweep.

For piped water, DINOv2 improves steadily from $k = 5$ to $k = 200$ and reaches the best test accuracy (84.13%), while DINOv1 follows a similar but slightly weaker trend. DINOv3-sat is comparatively flatter and peaks at smaller k values. For sanitation, the same ordering holds: DINOv2 remains strongest and reaches 87.17% at $k = 200$, with DINOv1 close behind and DINOv3-sat lower across settings. Overall, the ranking reported in the main paper is stable to reasonable k choices.

Table 3: Test accuracy (%) for piped water (PW-s) across k in the k-NN head.

Model	$k = 5$	$k = 10$	$k = 20$	$k = 50$	$k = 100$	$k = 200$
DINOv3-sat	80.91	81.43	81.50	80.94	80.33	–
DINOv2	81.79	82.65	83.36	83.90	84.06	84.13
DINOv1	81.72	82.09	82.85	83.52	83.81	–

Table 4: Test accuracy (%) for sanitation (SW-s) across k in the k-NN head.

Model	$k = 5$	$k = 10$	$k = 20$	$k = 50$	$k = 100$	$k = 200$
DINOv3-sat	84.18	84.65	84.11	84.18	83.74	–
DINOv2	85.45	85.97	86.39	86.94	87.10	87.17
DINOv1	85.32	85.80	86.29	86.73	86.98	–

Table 5: Additional held-out test-set results for selected foundation models on piped water and sewage access. For each model–task pair, the value of k was selected based on validation performance, and the corresponding held-out test metrics are reported (in %).

Task	**Model**	**Backbone**	**Pretraining data**	**Params (M)**	k	**Accuracy**	**Recall**	**F1**
Piped water	DINOv3	ViT-B	LVD-1689M	86	10	80.65	80.69	80.49
	DINOv3	ViT-L	SAT-493M	300	10	**81.43**	**81.22**	**81.19**
	Galileo	ViT-B	Galileo-pretrain	86	5	69.27	69.01	68.94
	Prithvi-EO-2.0	ViT-L	HLS time series	300	10	77.91	77.59	77.71
	Prithvi-EO-2.0	ViT-H	HLS time series	600	10	77.69	77.30	77.34
Sewage	DINOv3	ViT-B	LVD-1689M	86	20	**84.85**	**81.32**	**82.16**
	DINOv3	ViT-L	SAT-493M	300	10	84.65	81.06	81.92
	Galileo	ViT-B	Galileo-pretrain	86	5	76.81	71.10	72.04
	Prithvi-EO-2.0	ViT-L	HLS time series	300	10	82.88	80.91	79.05
	Prithvi-EO-2.0	ViT-H	HLS time series	600	10	82.59	80.33	79.22

A.2 Country-level agreement and data coverage diagnostics

Country-level agreement remains consistent with the main-text validation results: piped-water estimates track JMP more closely than sanitation, while sanitation displays broader dispersion attributable to indicator-definition mismatch.

Figure 3 provides an explicit coverage diagnostic across Afrobarometer rounds. The figure shows that both sample density and outcome range vary substantially by country and round, which helps

explain why model uncertainty is heterogeneous rather than uniform. Countries represented across multiple rounds contribute more stable supervision, while sparsely observed countries are more sensitive to temporal and distributional shift.

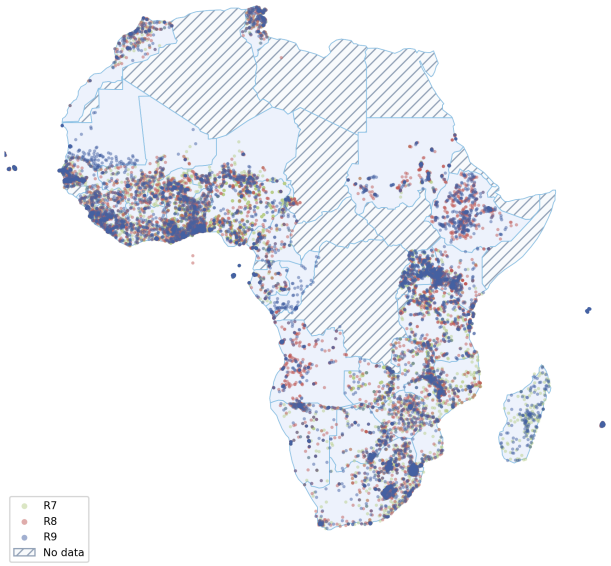


Figure 3: Country-level coverage and access scatter across Afrobarometer rounds (R7–R9), used to contextualize data availability and distribution shift.

A.3 Generalization to countries without survey labels

Figure 4 reports weighted MAE against JMP for countries excluded from survey training labels. This plot separates countries where transfer is already close to JMP from countries that would benefit from local calibration. Most errors concentrate in a small set of countries, indicating that performance gaps are driven by specific outliers rather than broad model collapse.

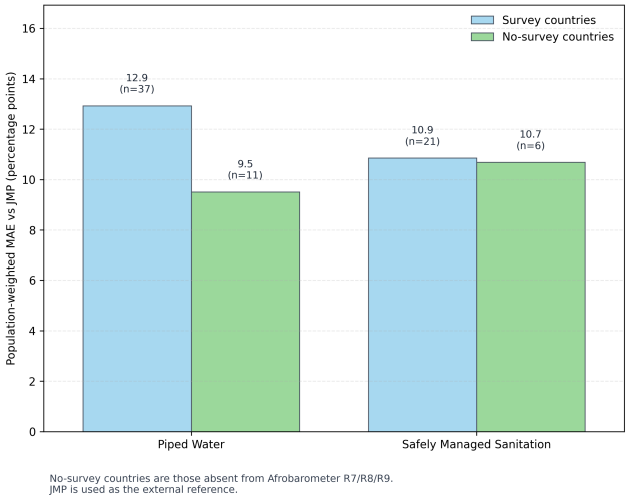


Figure 4: Weighted MAE versus JMP for countries not seen in survey-based training.

Figure 5 complements MAE by showing population coverage as a function of tolerated absolute error. The sanitation curve rises steeply and saturates earlier, meaning a larger fraction of the popu-

lation lies within relatively small error thresholds. The piped-water curve improves more gradually, indicating heavier error tails in selected countries. Read together with Figure 4, this confirms that transfer quality is usable at screening scale but uneven across national contexts.

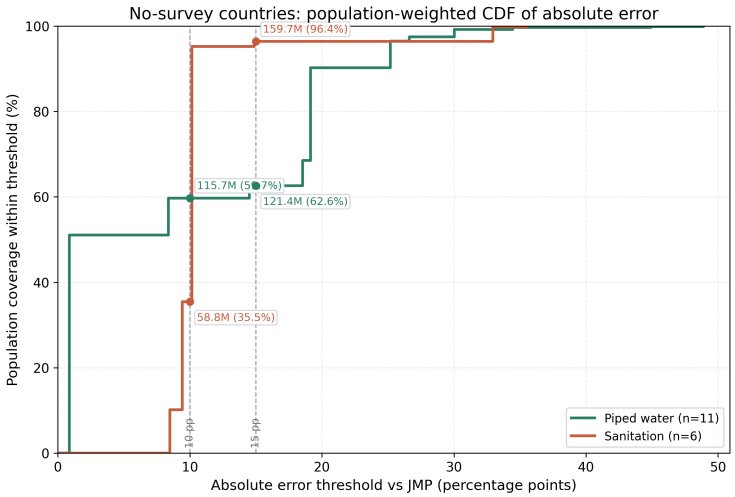


Figure 5: Population-weighted cumulative distribution function (CDF) of absolute error between model-derived estimates and JMP statistics for countries without survey data. The x-axis shows the absolute error threshold (percentage points), and the y-axis shows the share of population covered within that threshold. Results are shown for piped water and sanitation.

Table 6: Country-level external-validation results for countries without Afrobarometer survey coverage. Absolute error is computed as the absolute difference between the model-derived estimate and the corresponding JMP national statistic. Here, pp denotes percentage points.

Task	Country	Population (M)	Prediction (%)	JMP (%)	Abs. error (pp)
Piped water	Egypt	98.9	97.82	98.69	0.87
Piped water	Chad	16.8	25.17	16.79	8.38
Piped water	Libya	5.7	92.46	77.96	14.50
Piped water	Burundi	11.4	20.13	38.67	18.54
Piped water	Algeria	42.0	90.54	71.39	19.15
Piped water	Rwanda	12.1	20.31	45.49	25.18
Piped water	Guinea-Bissau	2.0	21.18	47.81	26.63
Piped water	Eritrea	3.4	21.73	51.76	30.03
Piped water	Comoros	0.9	25.06	59.50	34.44
Piped water	Djibouti	0.3	38.42	83.36	44.94
Piped water	Mayotte	0.3	43.78	92.67	48.89
Sanitation	Chad	16.8	2.43	10.91	8.48
Sanitation	Algeria	42.0	52.98	62.41	9.43
Sanitation	Egypt	98.9	57.02	67.17	10.15
Sanitation	Guinea-Bissau	2.0	0.56	15.44	14.88
Sanitation	Libya	5.7	56.78	23.83	32.95
Sanitation	Djibouti	0.3	4.11	39.60	35.49

A.4 Spatial predictions and map interpretation

Figures 6 and 7 connect observed survey geography to model-derived continental predictions. The key question is whether broad spatial structure in survey outcomes is preserved after extrapolation to unsampled locations.

Figure 6 summarizes survey-derived access patterns. It highlights strong within-country variation and regional contrasts, which motivates a disaggregated mapping approach rather than country-level averages alone.

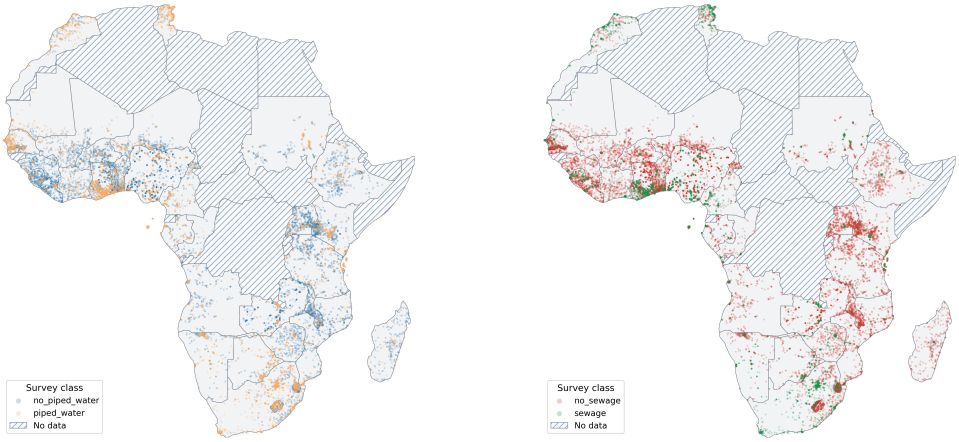


Figure 6: Survey-derived access maps across Africa for piped water (left) and sanitation (right).

Figure 7 shows continent-scale predictions from the trained model. The inferred surfaces reproduce many large-scale gradients visible in survey data while filling gaps in unobserved areas. At the same time, local discontinuities and transition zones indicate where prediction confidence should be treated cautiously and prioritized for follow-up validation.

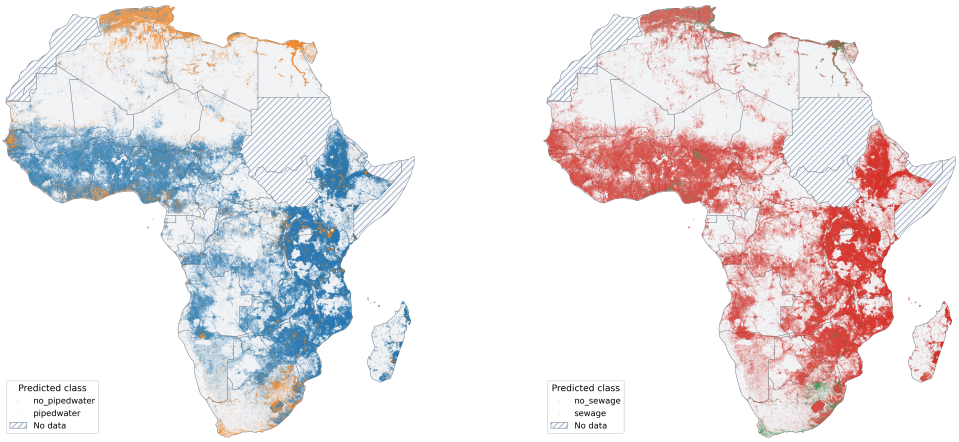


Figure 7: Model-predicted continental access maps for piped water (left) and sanitation (right).

Figure 8 provides service-specific diagnostics that make cross-task differences easier to inspect. Piped-water maps appear more spatially coherent and stable, whereas sanitation maps show greater local variability, consistent with weaker country-level JMP agreement. This visual contrast mirrors the quantitative gap reported in the main text.

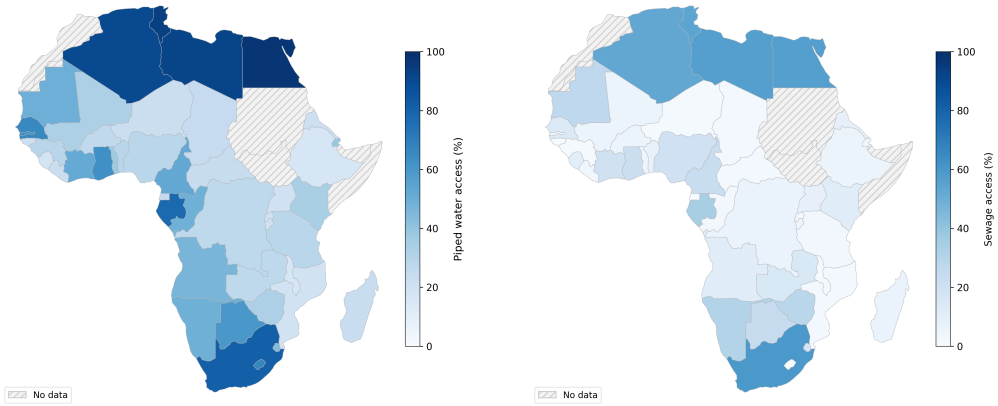


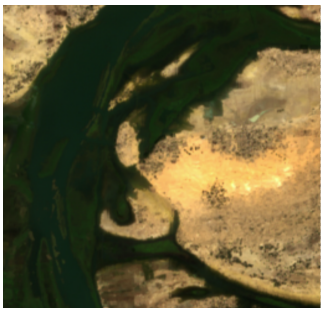
Figure 8: Service-specific spatial diagnostics for piped water (left) and sanitation (right).

A.5 Qualitative patch-level examples

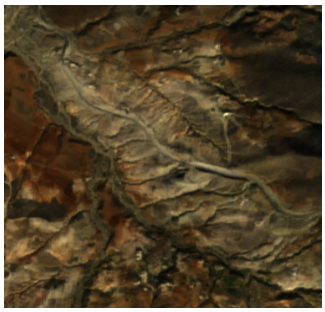
Figure 9 presents representative Sentinel-2 patches grouped by regions without and with piped-water access. The objective is not to claim direct visual detection of pipes, but to illustrate recurring contextual cues (settlement compactness, road geometry, and built-up texture) that likely mediate the learned signal. The contrast between rows also underscores a central caveat: proxy features can improve prediction while still encoding socioeconomic and urban–rural correlates, so qualitative evidence should be interpreted alongside external validation metrics.

Regions without piped water access

15°01’03.4”N 23°36’58.1”W
Cabo Verde



16°13’33.6”N 3°13’28.2”W
Mali

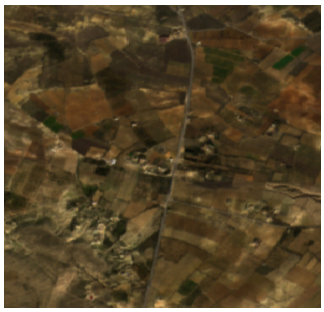


35°07’52.7”N 4°30’55.5”W
Morocco

Regions with piped water access

13°28’41.9”N 16°31’37.3”W
The Gambia

14°55’03.8”N 23°30’06.3”W
Cabo Verde



35°24’59.6”N 9°07’27.9”E
Tunisia

Figure 9: Examples of Sentinel-2 patches from regions without and with piped water access.

B Implementation Details

This appendix provides additional implementation details and supplementary results referenced in the main text.

Table 7: Implementation details for pretraining and held-out evaluation.

Stage	Model	Key settings
Pretraining	DINO	8 NVIDIA A40 GPUs; distributed training; ViT-Base; patch size 8; 300 epochs; batch size 16 per GPU; learning rate 3×10^{-4}; mixed precision (FP16).
Pretraining	DINOv2	8 NVIDIA A40 GPUs; distributed training; ViT-Large; patch size 8; 200 epochs; 20 warmup epochs; batch size 16 per GPU; centering set to Sinkhorn-Knopp; training configuration defined by project overrides together with official DINOv2 default settings.
Held-out evaluation	DINO	Teacher checkpoint used for embedding extraction; $k \in \{5, 10, 20, 50, 100\}$; cosine similarity; softmax-weighted k-NN; temperature 0.07; batch size 64; 4 workers.
Held-out evaluation	DINOv2	Teacher checkpoint used for embedding extraction; $k \in \{5, 10, 20, 50, 100, 200\}$ during model selection; cosine similarity; softmax-weighted k-NN; temperature 0.07; batch size 128; 4 workers; resize 256 and crop 224.
Held-out evaluation	DINOv3	Public pretrained checkpoint; $k \in \{5, 10, 20, 50, 100\}$; cosine similarity; softmax-weighted k-NN; temperature 0.07; batch size 64; 4 workers.
Held-out evaluation	Galileo	Public pretrained checkpoint; $k \in \{5, 10, 20, 50, 100\}$; cosine similarity; softmax-weighted k-NN; temperature 0.07; batch size 64; 4 workers; bf16 inference; input resolution 10 m; bands B2/B3/B4.