SocioSAT: Transfer Learning for High-Resolution Socio-Economic Data

4th JDC Research Conference on Forced Displacement, June 5, 2026 - Bangkok, Thailand

Presenter: Steven Ndung'u

Global Data Service (GDS), UNHCR

UNHCR The UN Refugee Agency

THE WORLD BANK

AI & GLOBAL DEVELOPMENT LAB

CHALMERS UNIVERSITY OF TECHNOLOGY

LINKÖPING UNIVERSITY

Context and Statement of the Problem

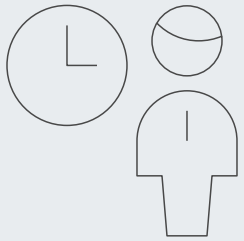
Data challenges on mapping for Refugees & Host communities

Core Problem

Lack of timely, granular socio-economic data on forcibly displaced and host community populations which constrains the design and implementation of effective support and development programs.

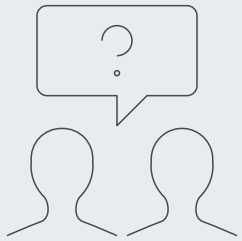
Limited Geographic Coverage

Data collection is limited in scale, hindering understanding of inaccessible areas.



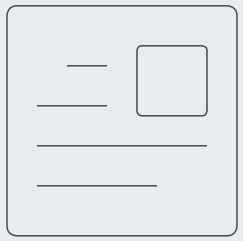
Temporal Data Gaps

Significant time lags between surveys, sometimes up to 10 years.



Insufficient Data

Lack of sufficient essential data on displaced and host communities hinders impact assessment and program design.



Limited Media Coverage

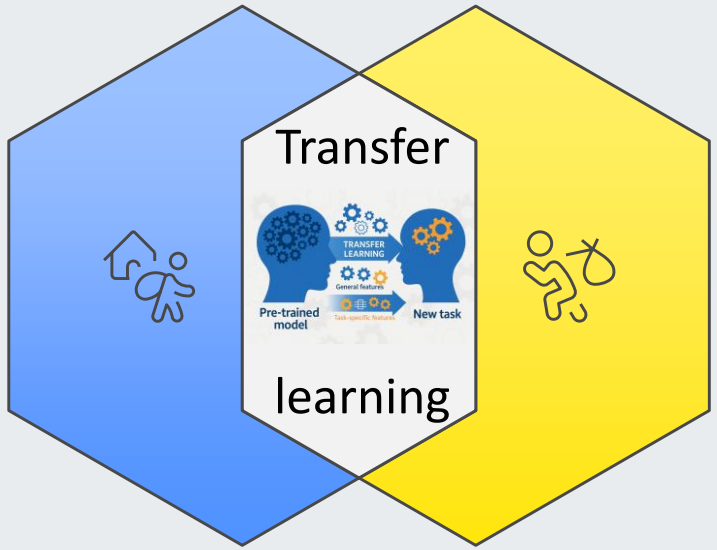
Cameroon, the world's most neglected displacement crisis in 2024 [was mentioned 15× less in the media (compared to Ukraine)].

Motivation of the Project

Key Question

Can socioeconomic-predictive features learned from national DHS samples transfer to refugee settings, given different and diverse settlement patterns and socio-economic distributions?

The Demographic and Health Surveys (DHS) Program → The Forced Displacement Survey (FDS) Program



National DHS Samples

Diverse household data

- **1.2M** households
- **140** DHS surveys
- **36** African countries
- **1990 – 2020** years

Large-sample household survey data *with broad spatial and temporal coverage*

Refugee/Host Data

Unique settlement patterns

- **10000** households
- **1** FDS survey round
- **3** African countries
- **2023 – 2025** years

Small-sample and scarce household survey data *with limited spatial and temporal coverage*

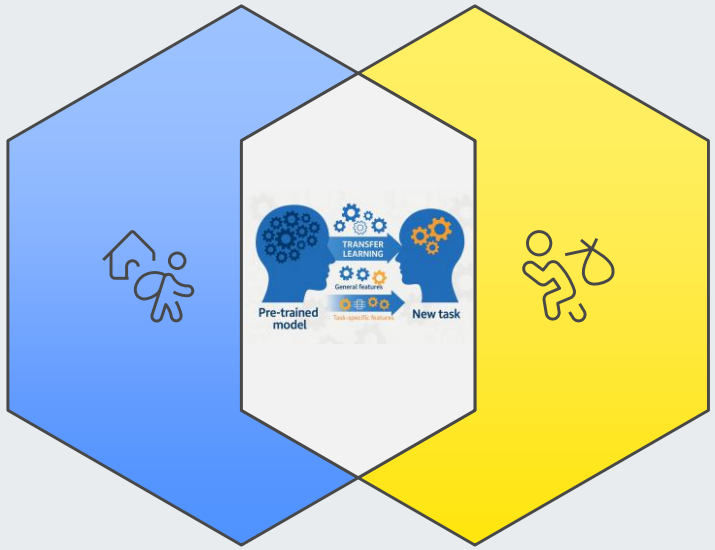
Motivation of the Project

Key Question

Can socioeconomic-predictive features learned from national DHS samples transfer to refugee settings, given different and diverse settlement patterns and socio-economic distributions?

National DHS Samples

Diverse household data



Refugee/Host Data

Unique settlement patterns

- The project aims to **adapt Earth Observation and Machine Learning (EO-ML**) models to estimate socio-economic conditions **between survey rounds and over time**.
- The overarching goal is to generate a household socioeconomic proxy **at the camp level (areas with substantial spatial variability in livelihoods)** from imagery to support prioritization and monitoring.

NB: Refugee camps are often remote and overcrowded, with mostly temporary shelters, limited access to jobs and education, heavy reliance on aid, and elevated health risks.

Why this Project Matters

1. **Prioritizing humanitarian and development interventions.** Quickly localize areas of deteriorating conditions to inform programming and protection response.

2. **Evidence-based survey investment.** Inform targeted survey investment by distinguishing where lightweight monitoring suffices from where full household data collection is warranted, maximizing return on diminishing collection budgets.

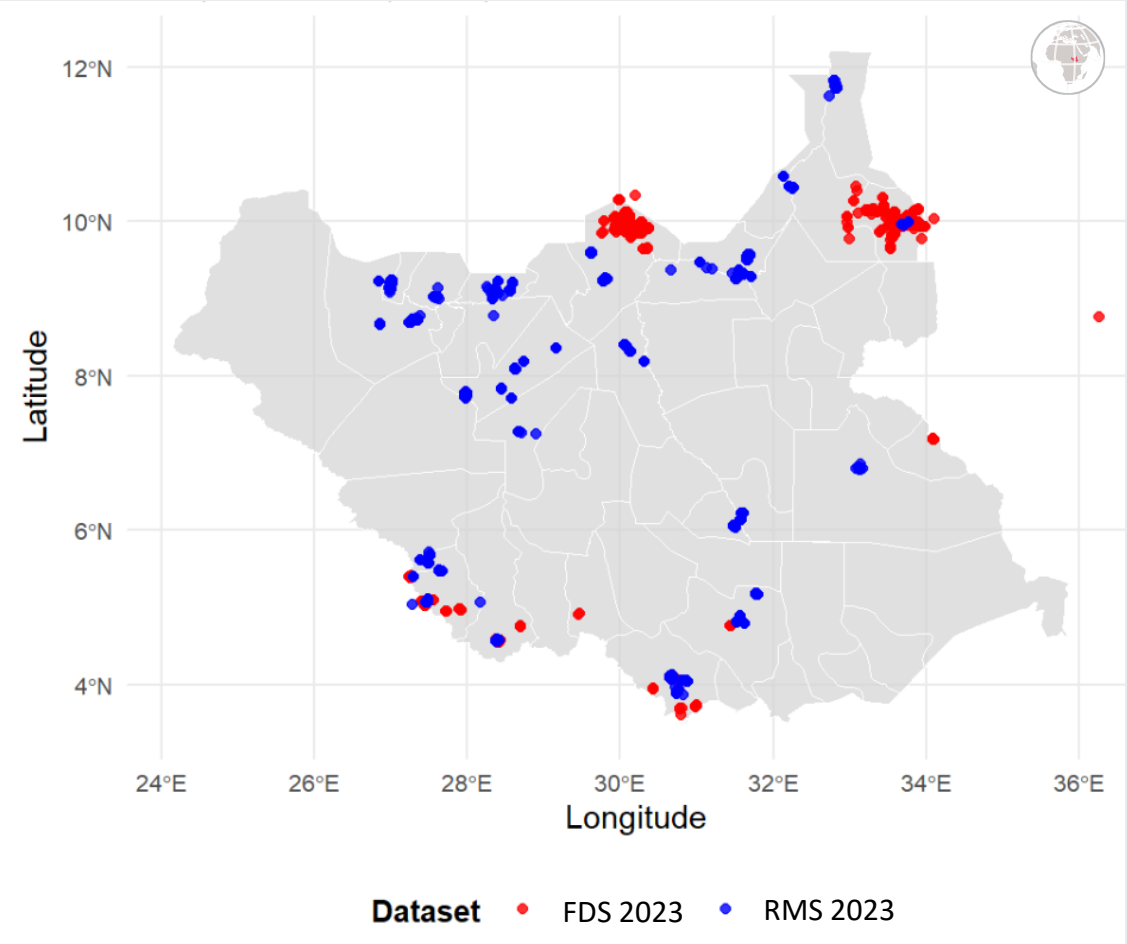
3. **Use and reuse of existing data and model assets.** Leverages open EO data, existing model weights, and prior surveys to produce large-area estimates from minimal new data collection.

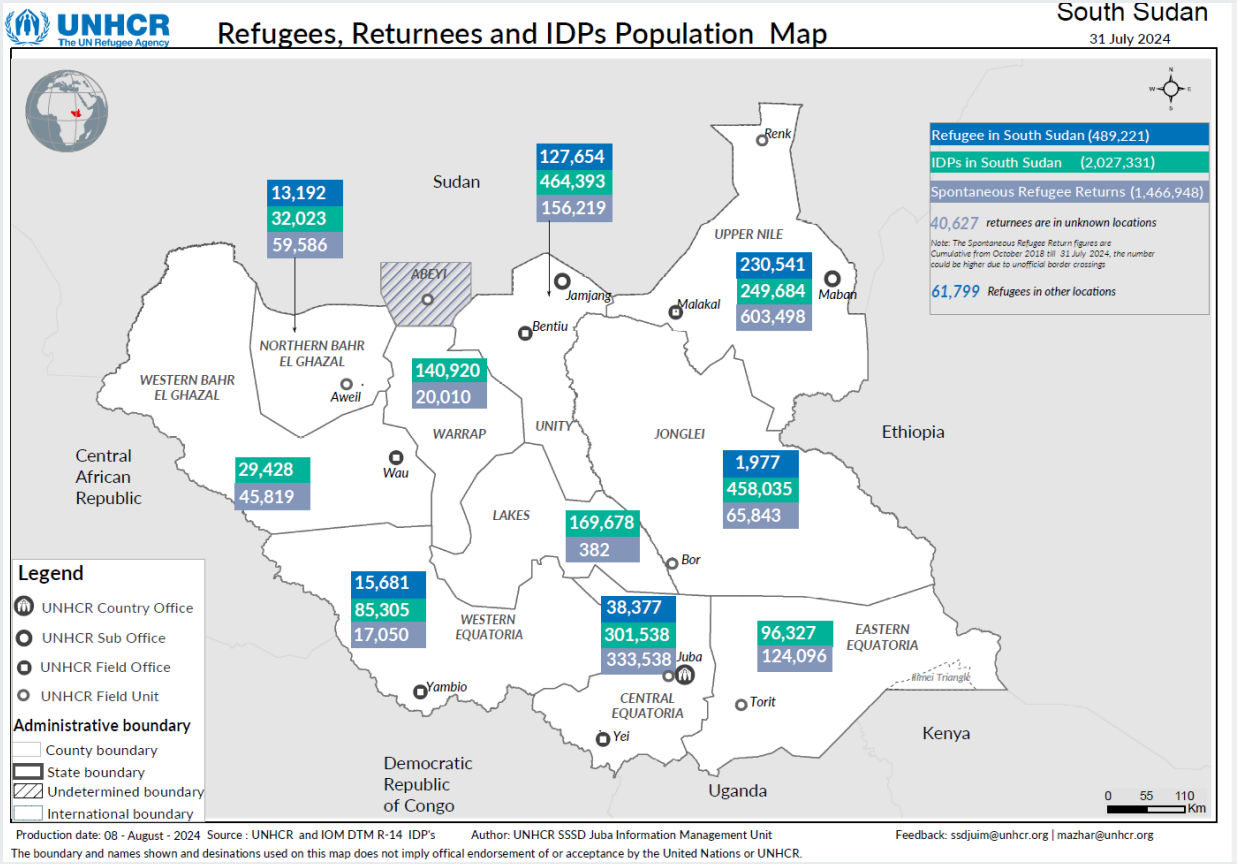
4. **Harmonization to bridge forced displacement data scarcity.** Unlock the combined use of multi-source, multi-temporal datasets across locations, as demonstrated by harmonizing RMS and FDS in South Sudan.

Data sets & socio-economic index

Data sets

Spatial coverage of validation data sets





RMS - Results Monitoring Survey | FDS - Forced Displacement Survey

Refugee settlement patterns differ markedly by region: **96%** of refugees in the North live in camps, compared with **59%** in the South. The remainder reside in settlements.

Data alignment

The FDS and RMS datasets share significant structural similarities, making them ideal candidates for domain adaptation.

FDS Data

Comprehensive, established survey data.

Shared Structure

Questions

Value Labels

RMS Data

Complimentary survey with similar structure.

A concrete example is the living conditions index variables , which are transferable indicators across both surveys.

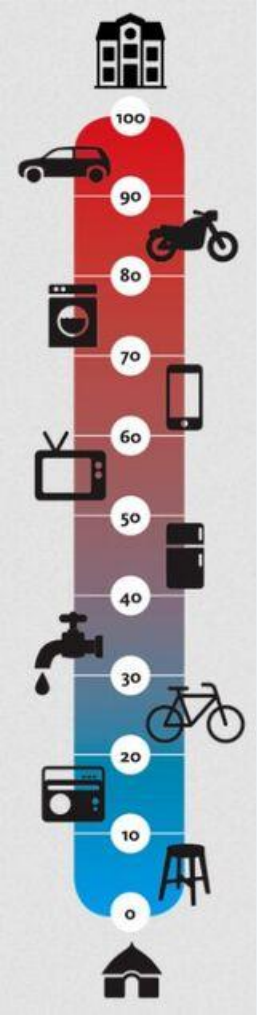
The RMS data set is closely related to the FDS data in terms of question structure (semantic) and value labels.

We utilize both **visible/observable** and **abstract features.** Abstract features connect raw, observable data to **bigger socio-economic outcomes** like health, income, and human development.

Socio-economic index

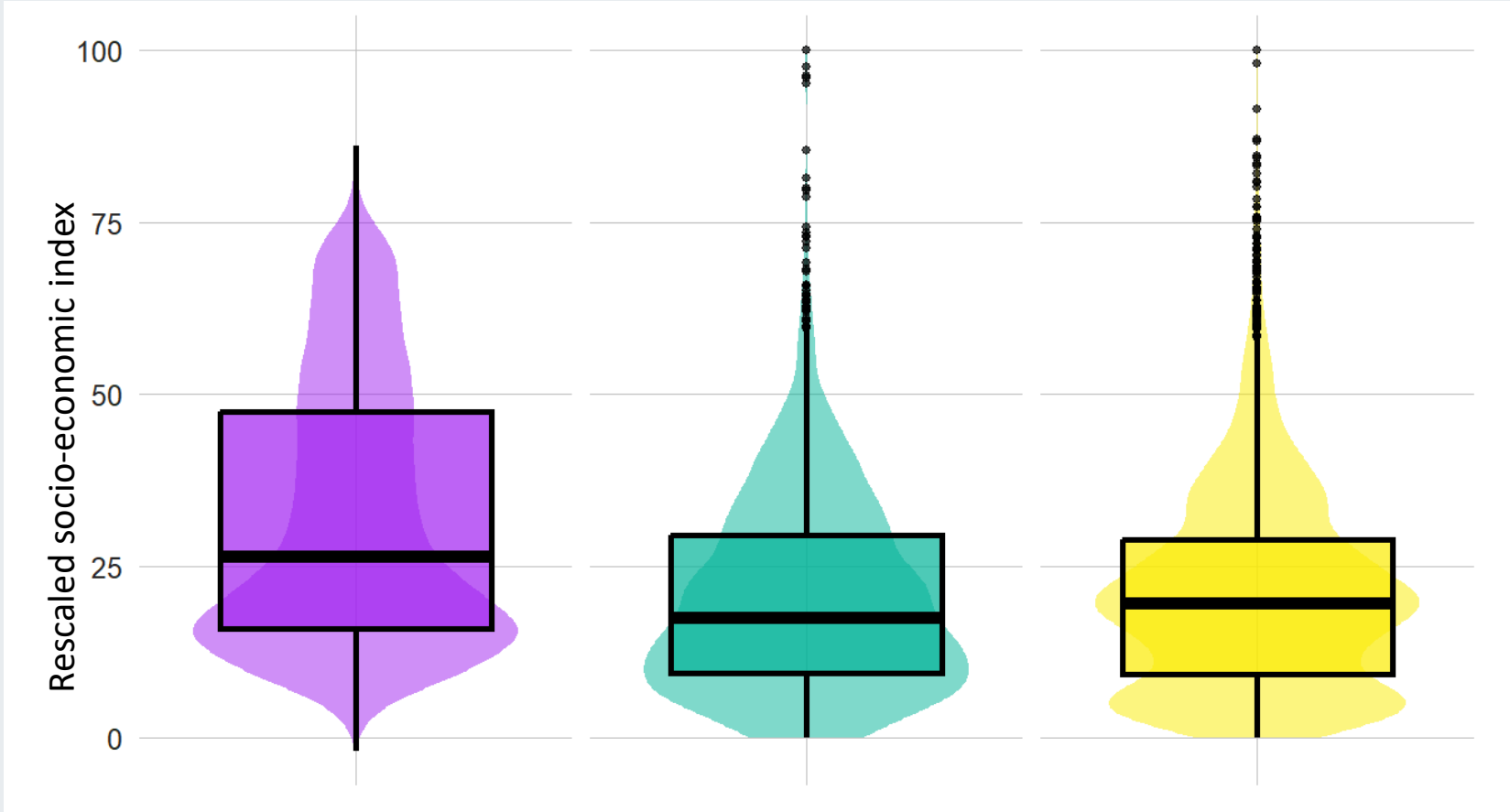
Constructing the socio-economic index

- Use household assets, housing quality, and access to services as input variables.
- Standardize all asset indicators (so they are on a comparable scale) - measured on different scales.
- Apply Principal Component Analysis (PCA)
- Take the **first principal component** as the underlying *'socio-economic dimension'* interpreted as a **latent socio-economic factor** – it captures the largest common variance across all these indicators.
- The raw scores are then linearly rescaled to a 0 –100 index



Ref: Leopold, S., Khan, A., and Brock, P. (2026). *Deriving an asset-based wealth index from refugee and host community survey data*. FDS Methodological Paper No. 2. Copenhagen, Denmark: United Nations High Commissioner for Refugees.

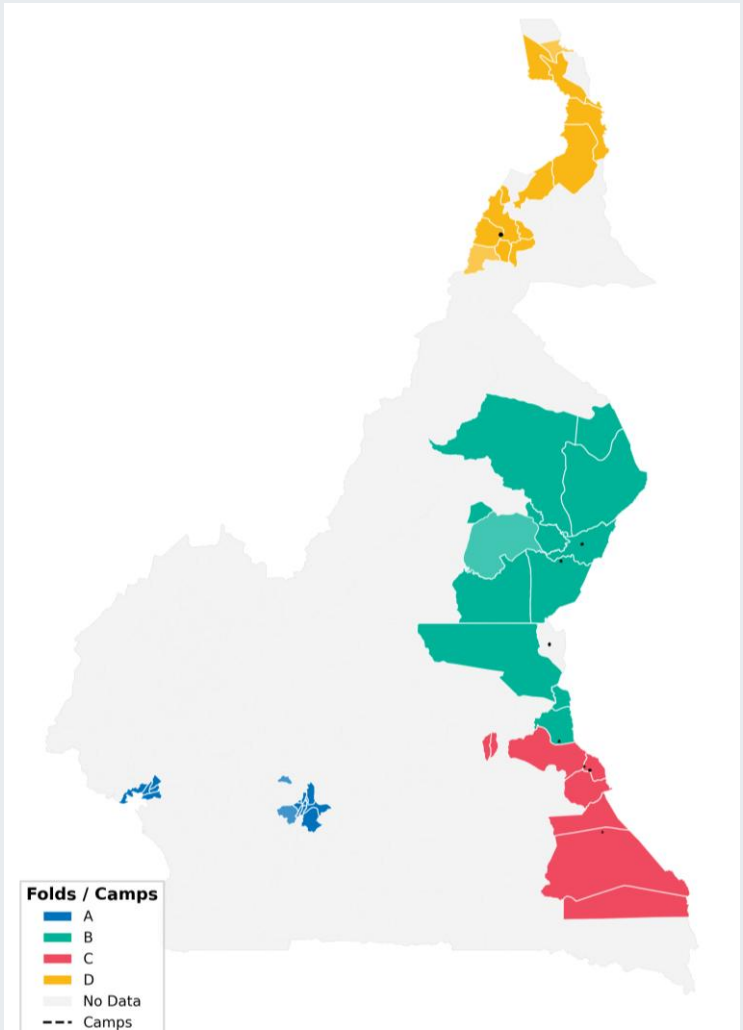
Socio-economic index

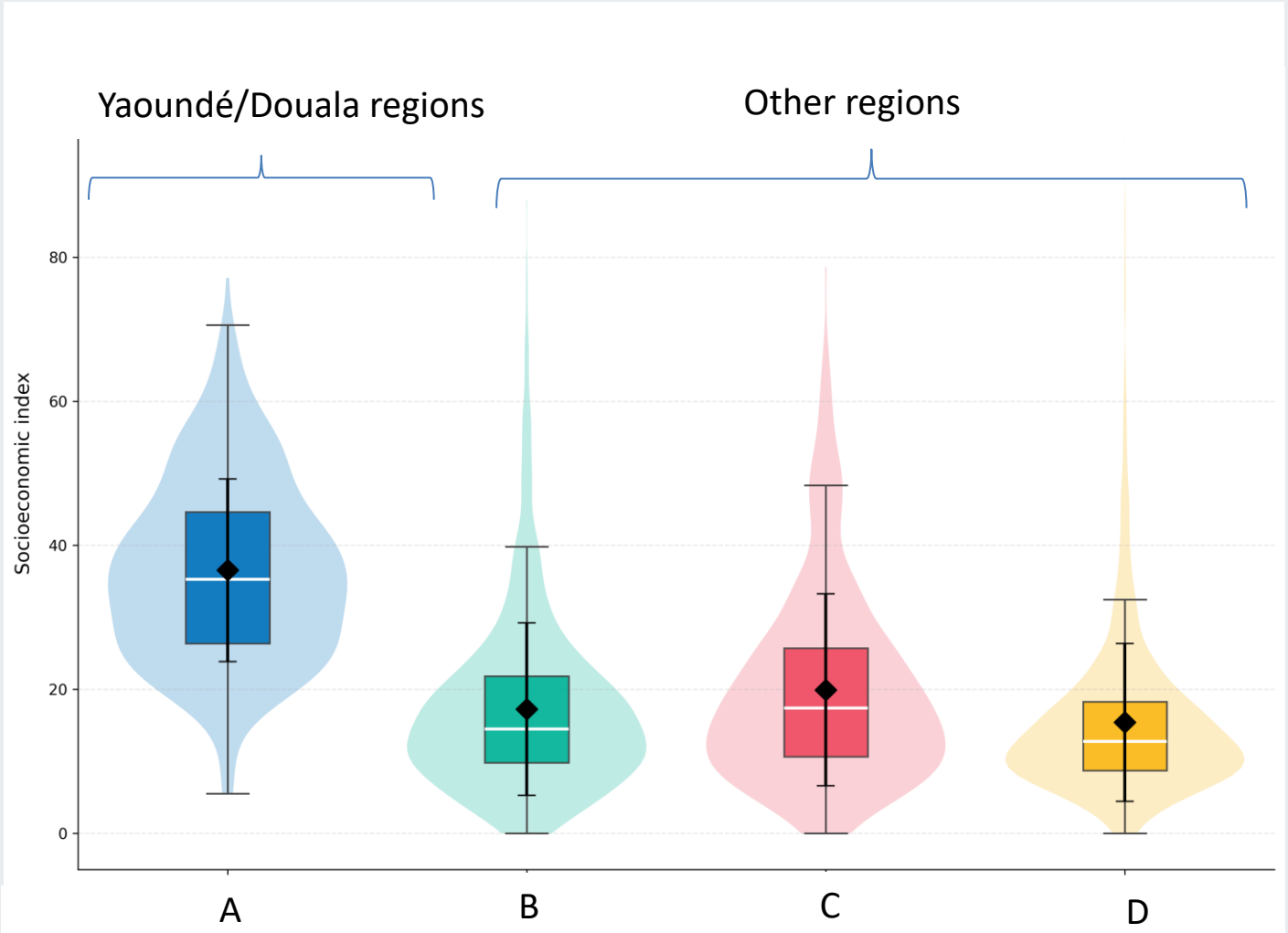


- There is **high congruence** between the index constructed using wealth variables and the index constructed without wealth variables.
- Around 80% of the refugees in our sample **fall below the DHS median wealth score**.

Socio-economic index

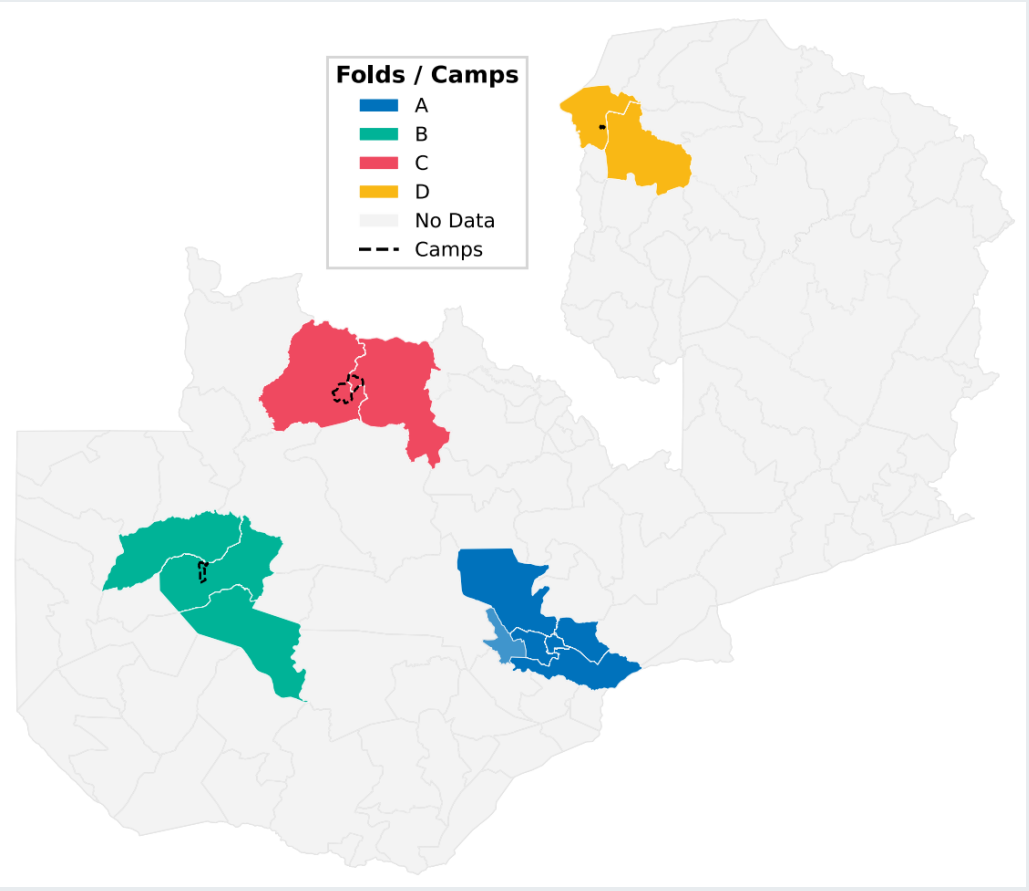
FDPs living in rural areas, in and around camps, have lower welfare than those in the Yaoundé/Douala regions

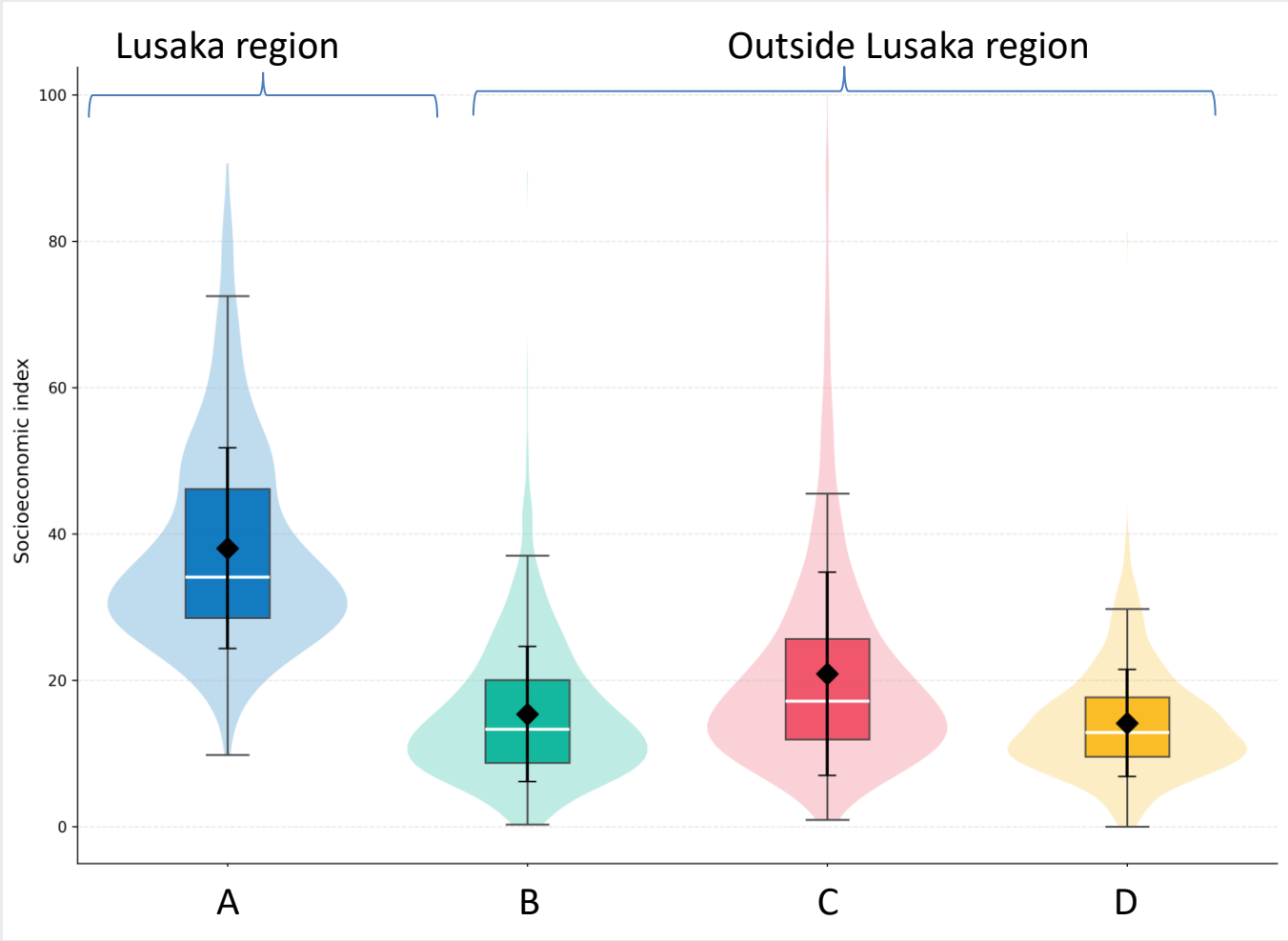




Socio-economic index

FDPs living in the capital, Lusaka, and its surrounding areas have better welfare than those living outside the capital.

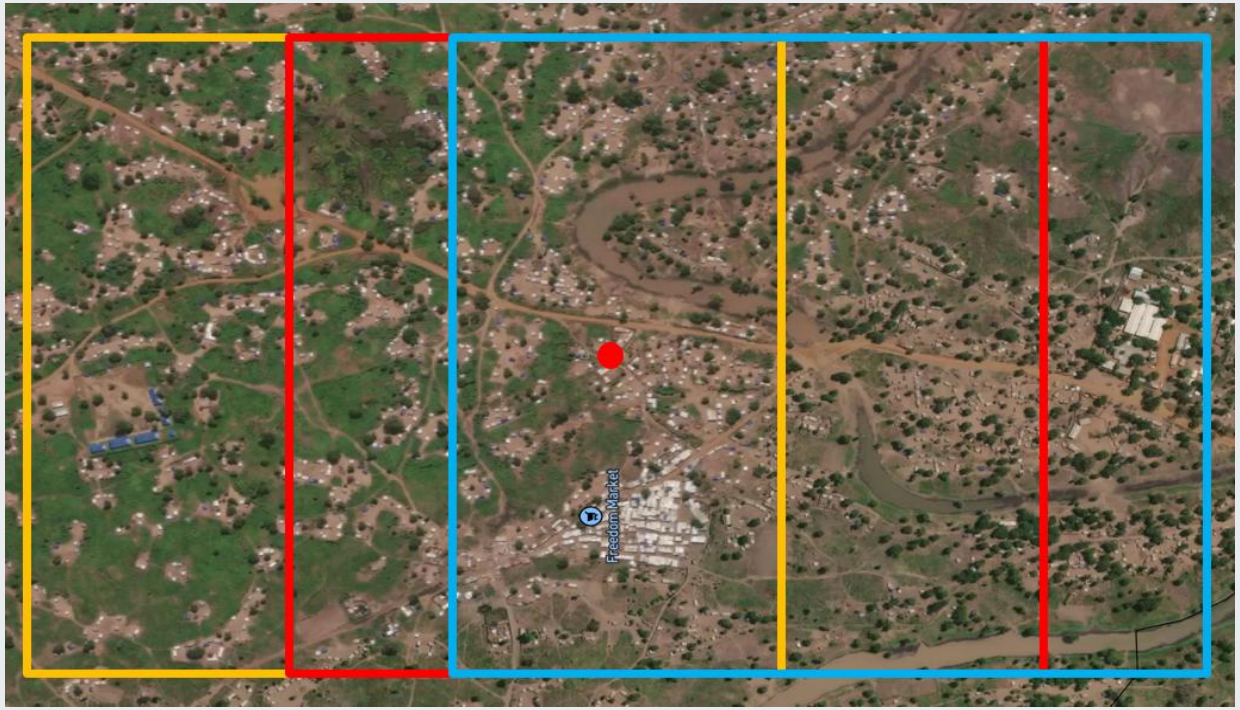




Integrating EO with survey data

Integrating EO with survey data

- The model requires spatial resolution of spatial size of 6.72 km X 6.72 km for a single cluster of forced displaced people.
- One grid (cluster) translates to a *single observation.*
- Household survey data: South Sudan (FDS ~3000 ; RMS ~5500), Cameroon (FDS ~4000) and Zambia (FDS ~4000)

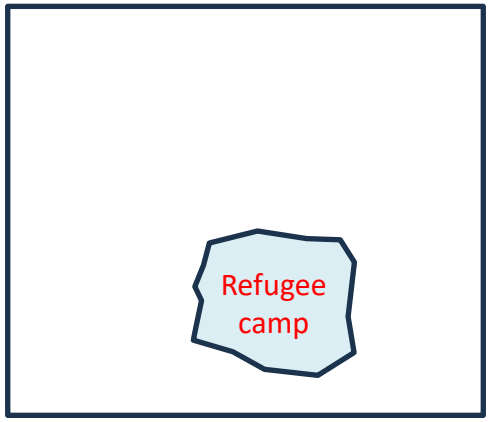


Integrating EO with survey data

Spatial sliding-window data augmentation strategy

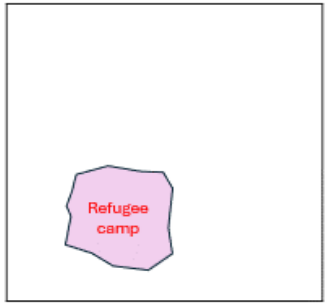
Data augmentation creates training samples that preserve geophysical meaning while building invariance to clouds, illumination, and seasonal shifts through **crops**, **rotations**, **flips**, **mask-aware erasing**, **brightness adjustments**, and **noise**.

- In this project: for dense areas like refugee camps, we're developing grid sampling strategy that balances quantity with spatial variation.



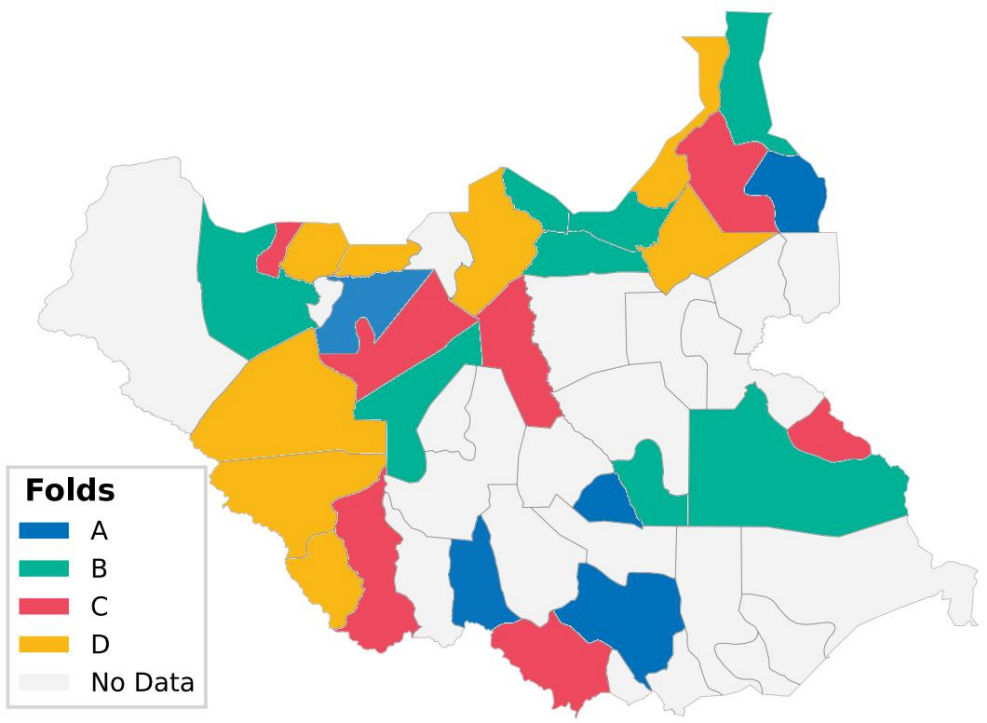
6.72 km

6.72 km



EO - Data preparation

‘Stratified’ Cross-Validation design



Experiment	Test Set	Validation Set	Training Set
Fold 1	Group A	Group D	Groups B, C
Fold 2	Group B	Group A	Groups C, D
Fold 3	Group C	Group B	Groups A, D
Fold 4	Group D	Group C	Groups A, B

Note: The allocation ensures that for any given fold, the Test, Validation, and Training sets are spatially disjoint.

- Create 4-fold CV with data under 2 folds, for instance, A and B for **training**, C for **validation**, and D set apart for **testing**.
- The spatial coverage of A, B, C and D are **non-overlapping** to avoid *data leakage*; split across region. Key aspects to consider;
  - ✓ **Level of development representation:** Urban, peri-urban, and rural areas diversity.
  - ✓ **Population groups representation:** Refugee and host populations are stratified across four folds, so each validation set reflects demographic diversity while preserving spatial independence.
  - ✓ **Ecological/climate zones representation:** The geographic areas distribute across different ecological and climate zones.

Model Architecture

Vision Transformer Architecture

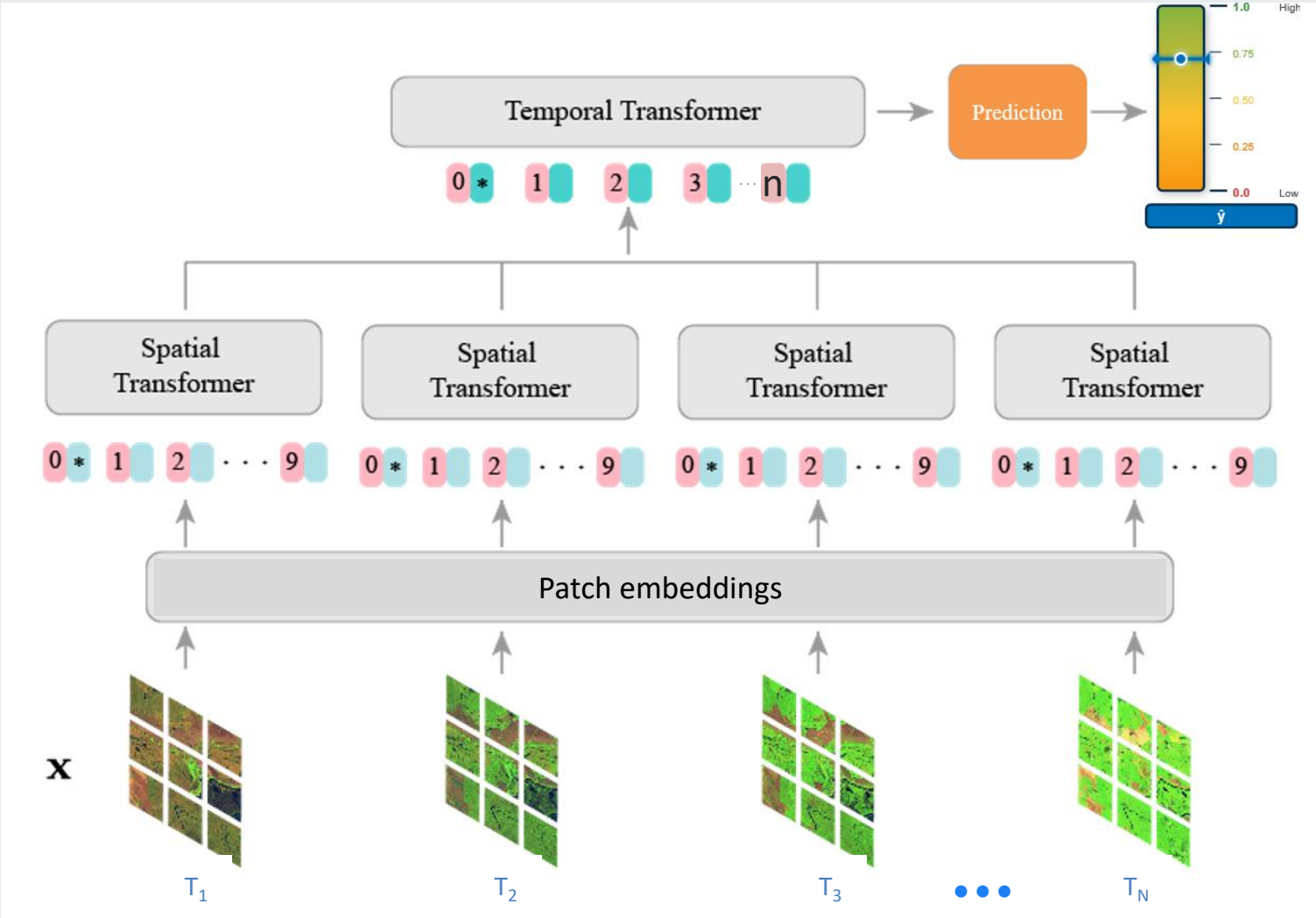
Tempero-Spatial Vision Transformers (TSVIT)

Multi-modal data

Nightlights radiance &

Landsat multispectral bands

Building footprints



Model Performance

Model Results

South Sudan Results - Country specific performance

Spatial CV fine-tuning improves substantially over zero-shot transfer

Model	MAE	MSE	RMSE	R^2	Corr.
A	6.70	78.12	8.84	0.33	0.59
B	**5.37**	**45.00**	**6.71**	**0.49**	**0.70**
C	8.10	107.92	10.39	0.33	0.58
D	10.47	182.88	13.52	0.02	0.17

Best fold: B (lowest MAE/RMSE; highest R^2 & correlation)

- Direct **zero-shot** transfer of the DHS-trained model to FDS fails the refugee data, underscoring the need for transfer learning.

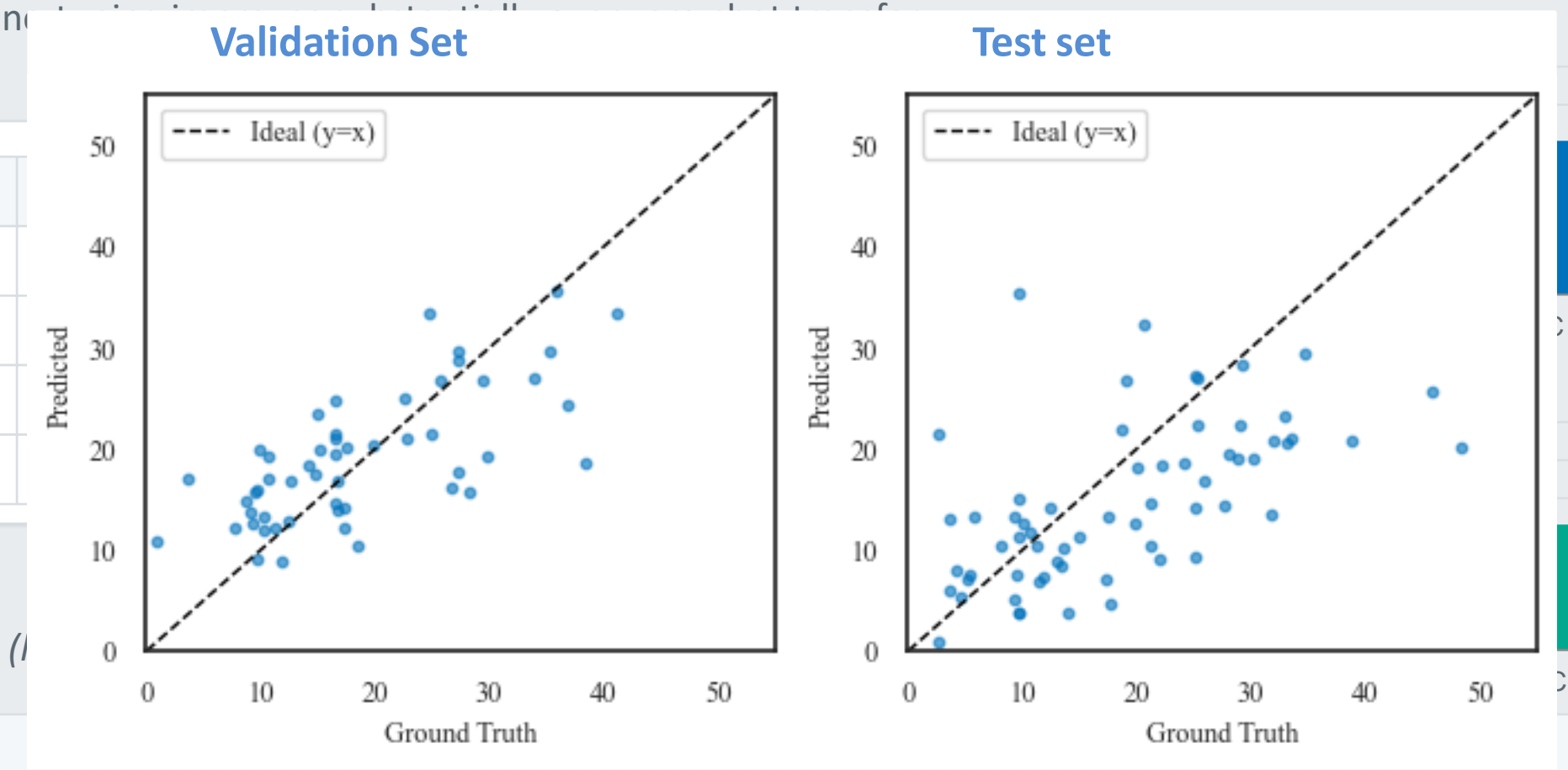
Model Results

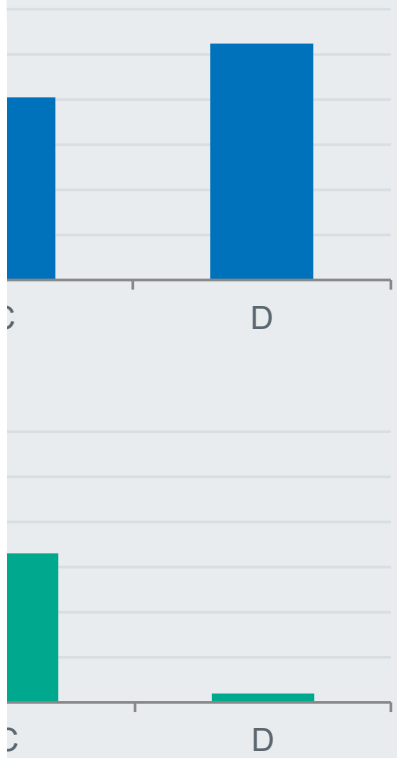
South Sudan Results - Country specific performance

Spatial CV fin

Model
A
B
C
D

Best fold: B (



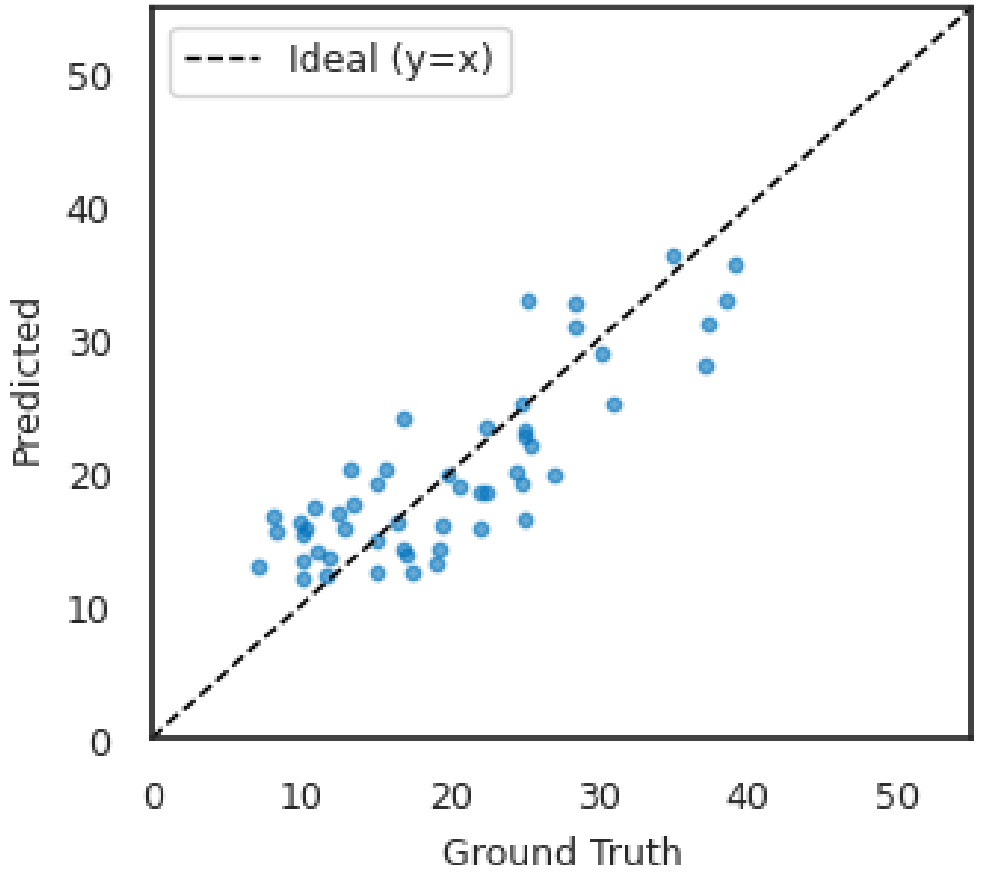


- Direct **zero-shot** transfer of the DHS-trained model to FDS fails the refugee data, underscoring the need for transfer learning.

Model Results

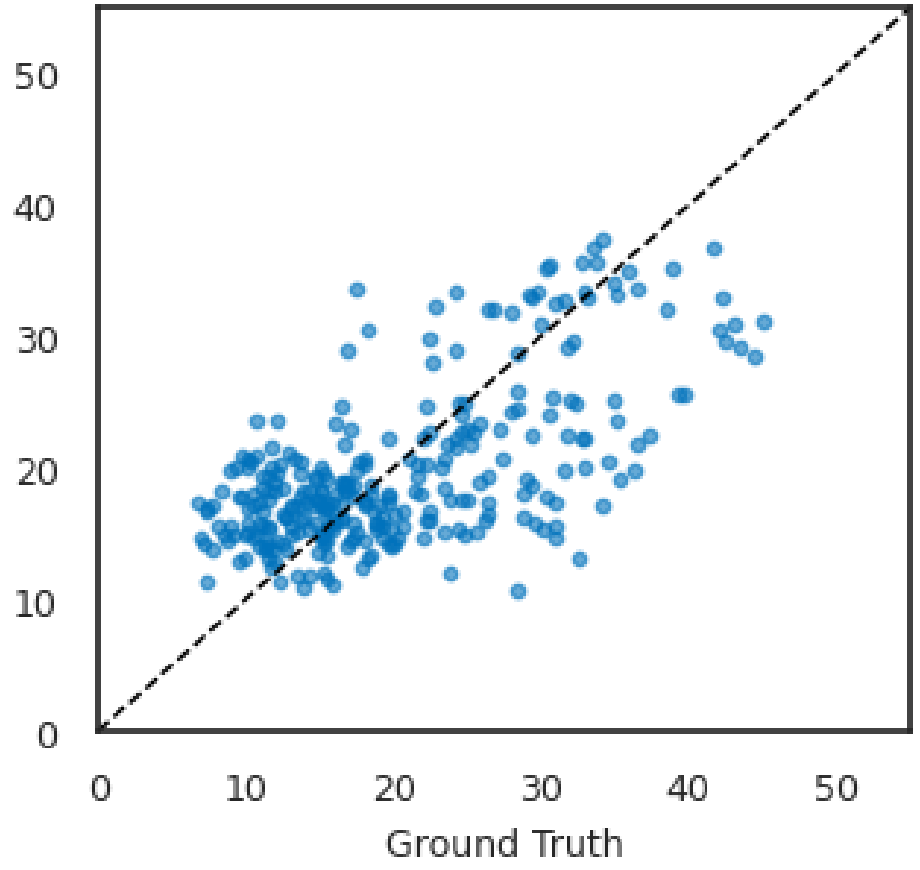
South Sudan / Cameroon / Zambia Results – Combined cross-country performance

Refugees Camp Areas



MAE	R^2	Corr.
4.13	0.68	0.83

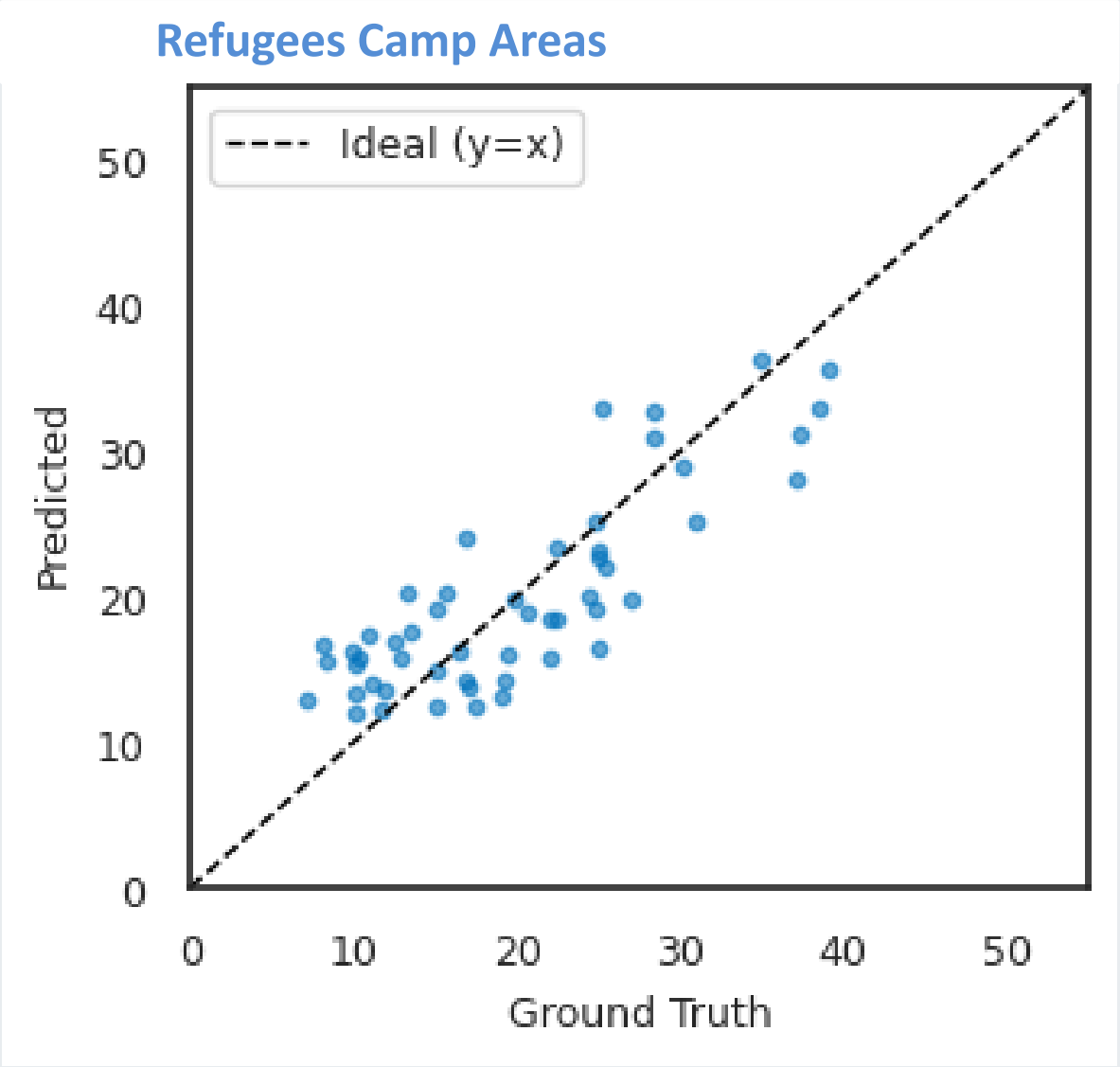
Host Comm. & Refugees outside the Camp



MAE	R^2	Corr.
5.19	0.45	0.67

Model Results

Performance in refugee settings

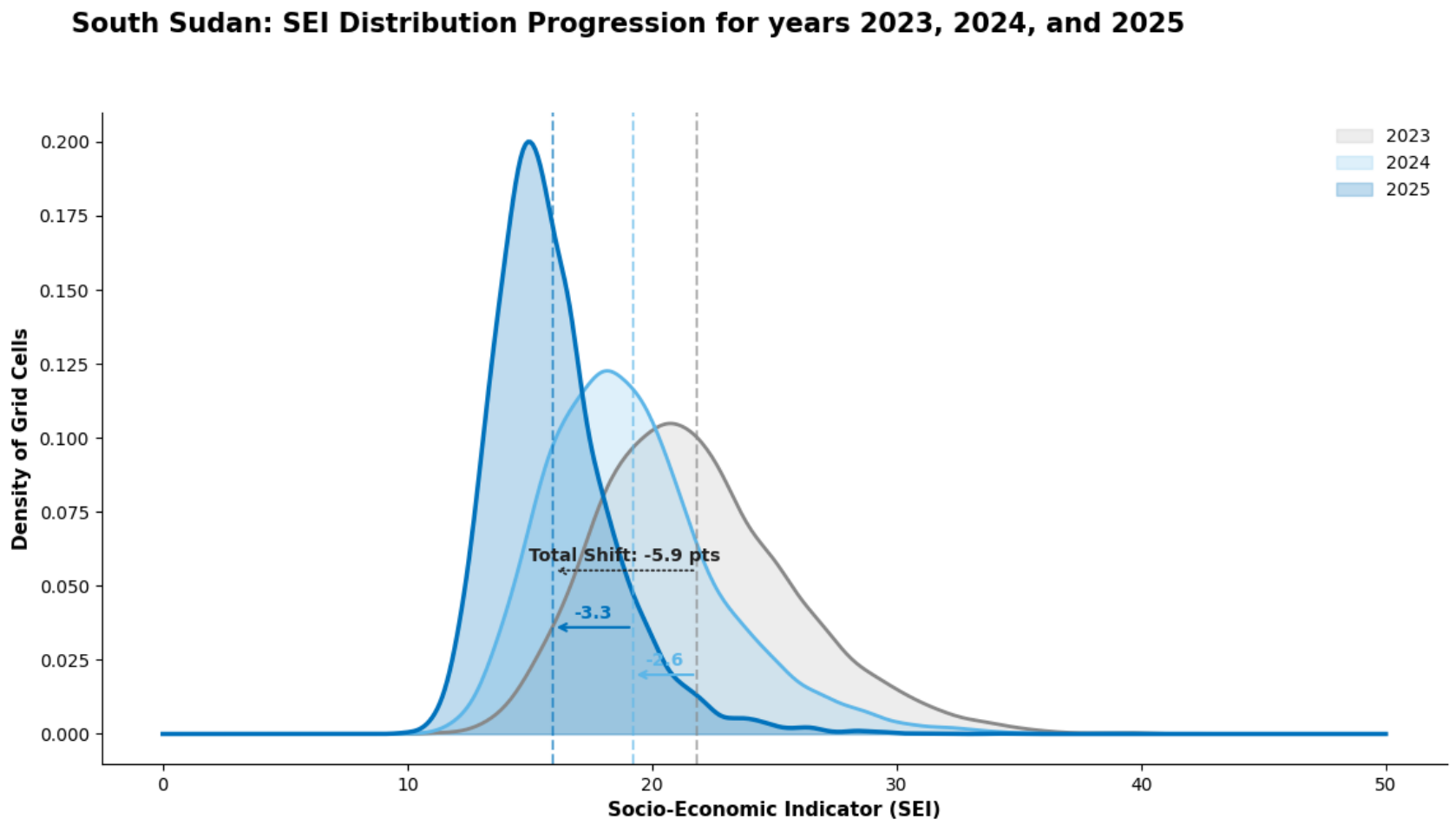


MAE	R^2	Corr.
4	0.7	0.8

Using the model e.g., Updating Socio-economic Insights 1-2 years after the survey

Socio-economic decline, South Sudan

Predictions from a fine-tuned earth-observation model, comparing 2023 baseline imagery with 2025 current imagery across South Sudan's 28 regions (payams) .



88 %

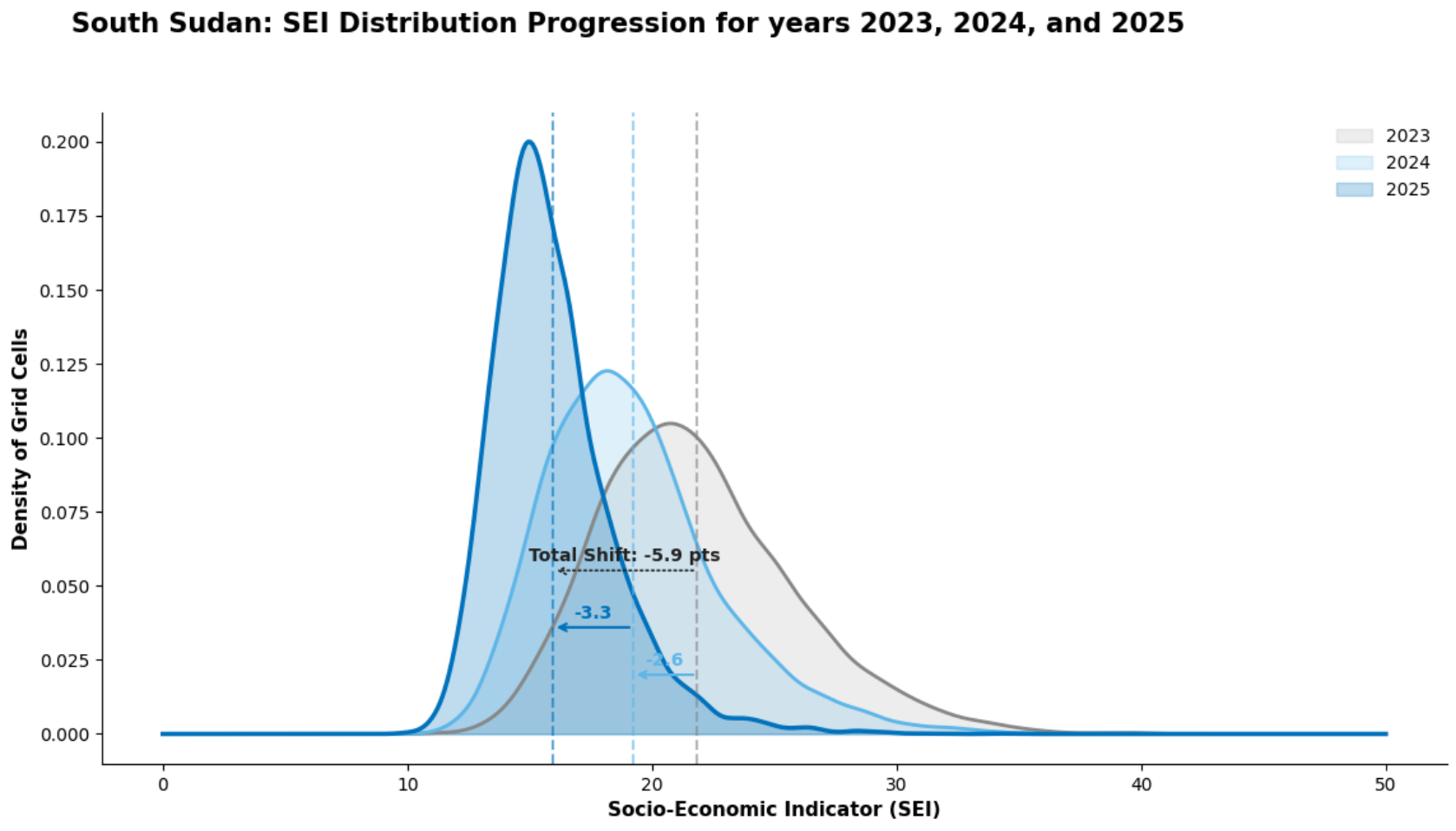
of grid cells declined

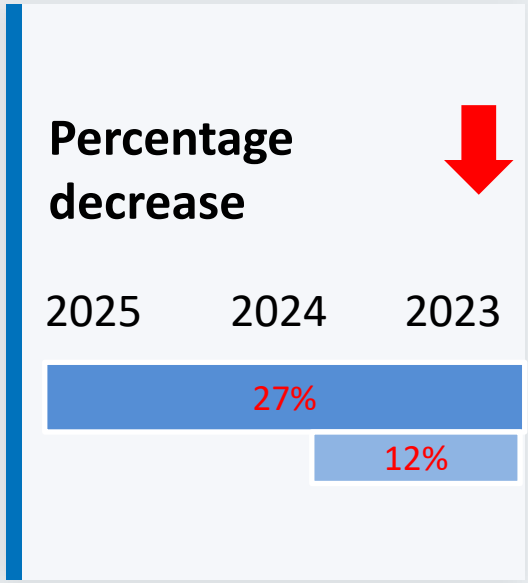
Grids cells of areas covering (Payams)

- refugees living inside camps,
- refugees living outside camps, and
- host-communities

Socio-economic decline, South Sudan

Predictions from a fine-tuned earth-observation model for 2023, 2024, and 2025.





Ground validation, South Sudan

Documented ground realities

Four converging shocks during the prediction window help explain the widespread, regionally differentiated decline the model detects.

Sudan war spillover

Over **1.2 M** arrivals from Sudan since April 2023 - straining services in already-poor areas.

Intercommunal violence

UN MISS documented a **43%** increase in violence incidents in Q2 2024 vs 2023, concentrated in Jungoli, Warap, Unity and Western Equatoria - regions our model flags as most affected.

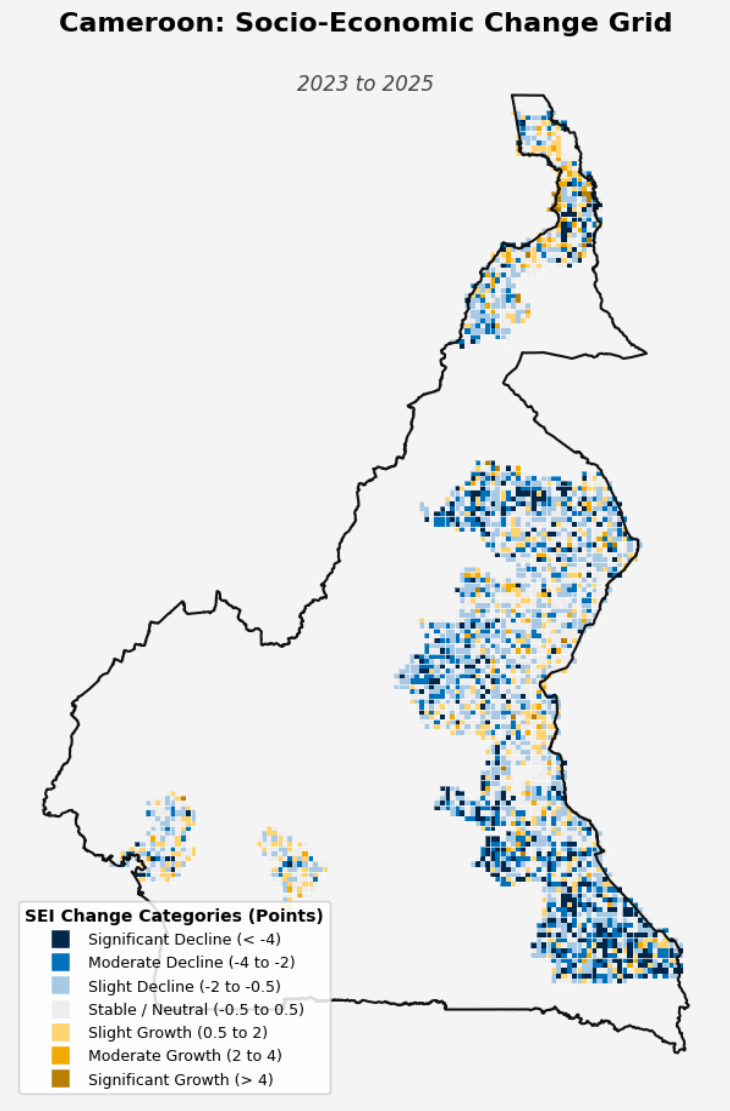
Economic disruption

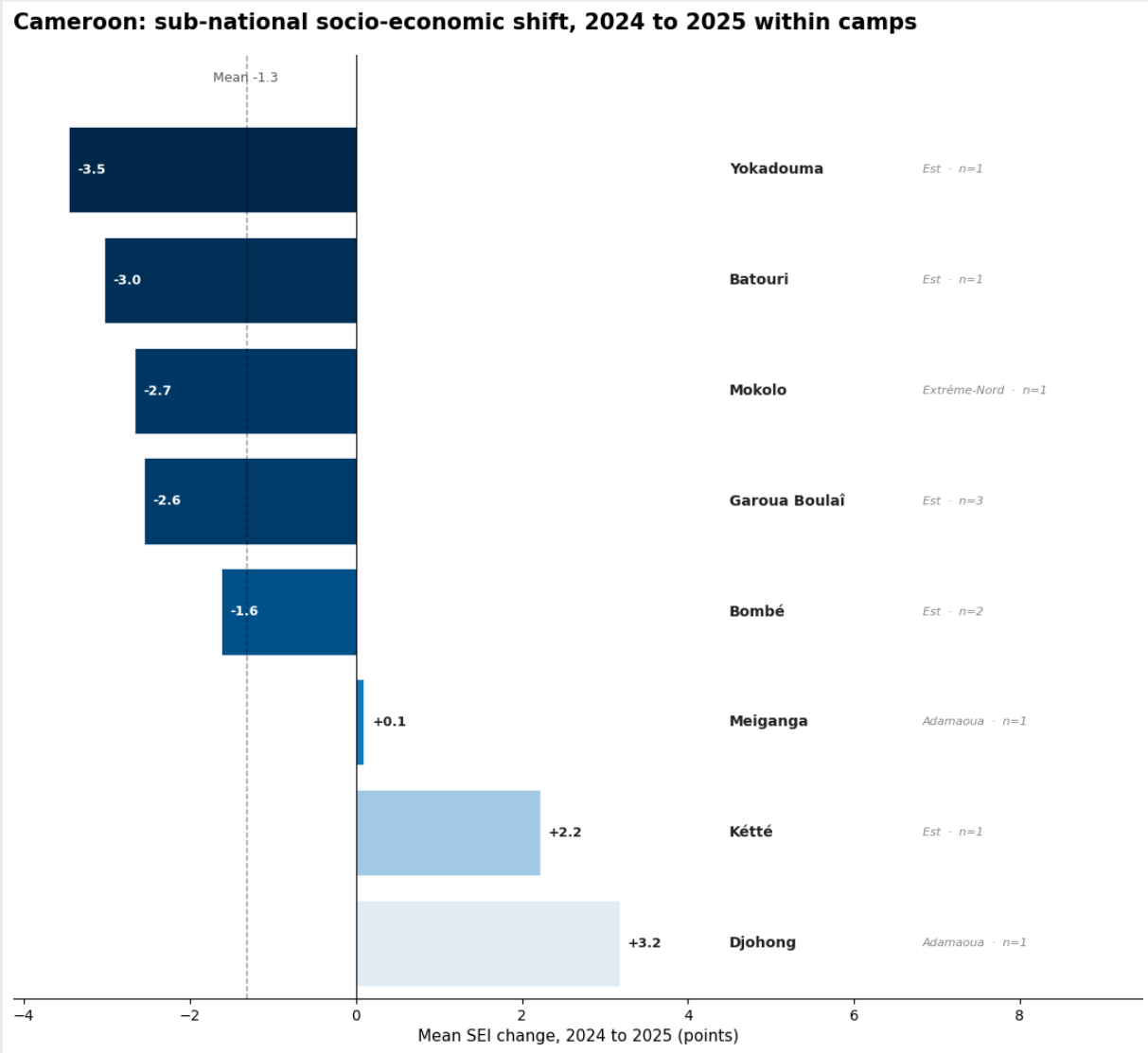
Pipeline carrying 70% of oil exports inoperable since Feb 2024; inflation climbed past **107** % by mid-year, eroding household purchasing power nationwide (IMF, Oct 2024).

Compounding climate shocks

Four consecutive years of flooding; drought-like conditions affecting ~36 % of the population; 2nd most natural-hazard-vulnerable country (INFORM 2024).

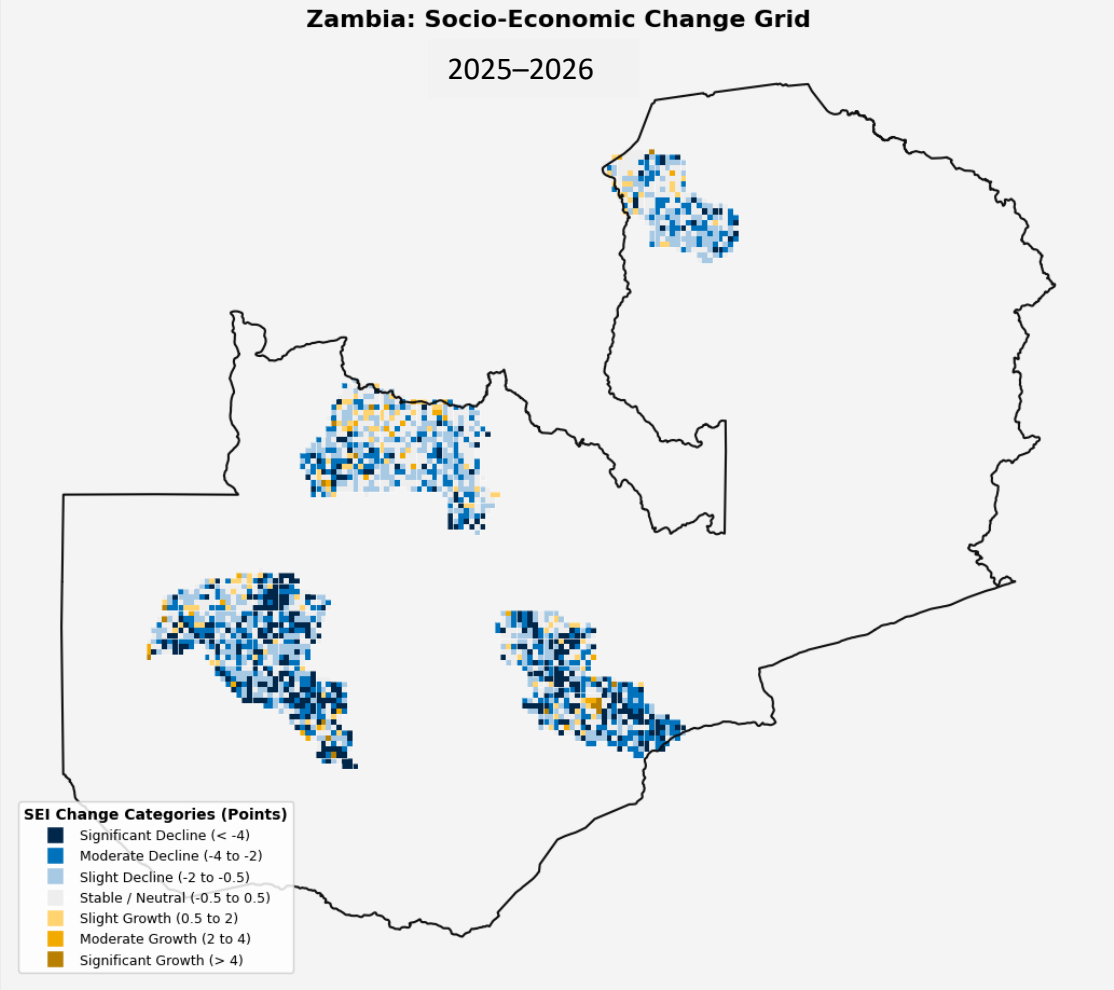
Socio-economic decline, Cameroon

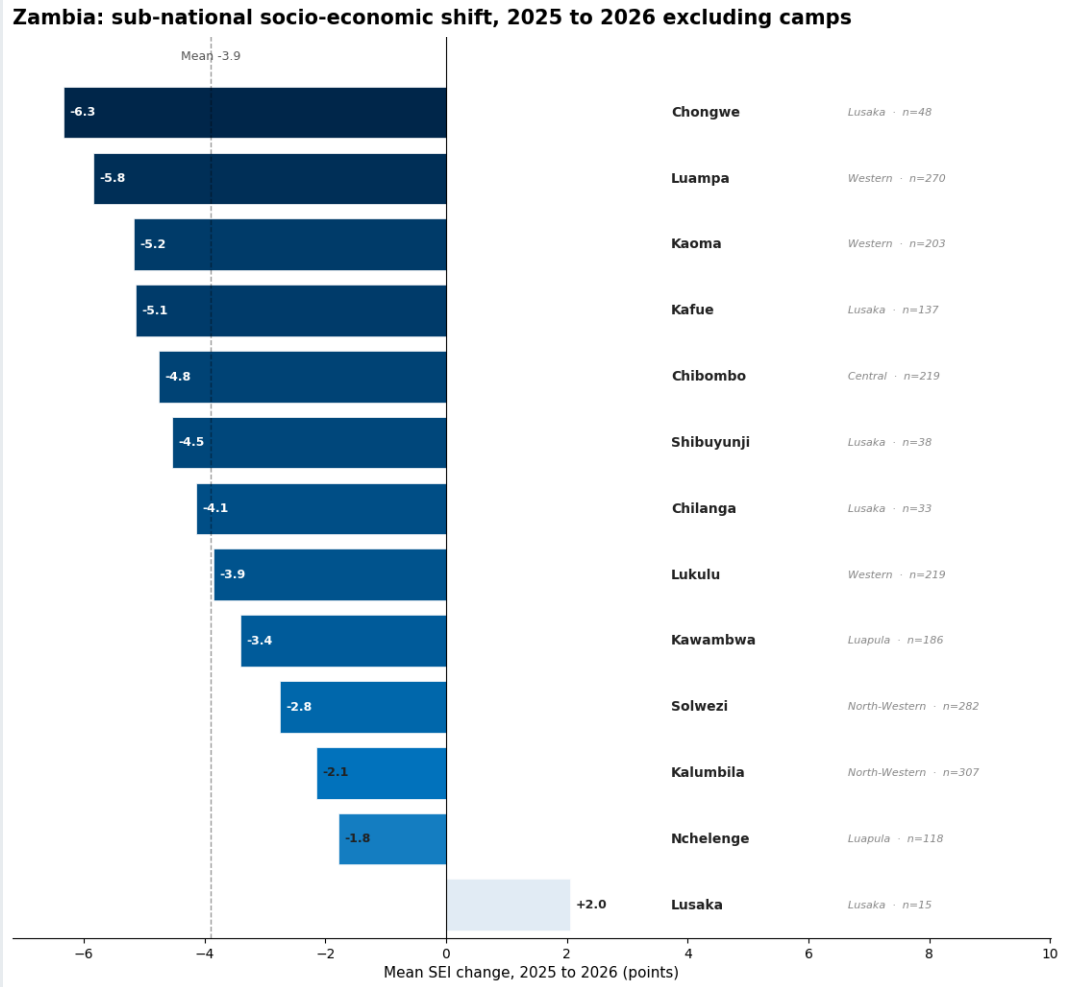




Better-connected, urbanized sites gained while remote Eastern camps declined.

Socio-economic decline, Zambia





Lusaka rises while its peri-urban ring (Chongwe, Kafue) declines

Contributions

Advancing socioeconomic measurement for forcibly displaced populations

- **First systematic transfer-learning evaluation from DHS to forced-displacement settings.** Demonstrates that features learned from large-scale national surveys can be adapted to refugee and host populations, recovering reliable camp-level predictions where direct transfer fails.
- **Use and reuse of existing data assets.** Leverages value from open Earth observation data, existing model weights, and prior surveys (DHS, RMS, FDS) to produce large-area estimates from minimal new data collection.
- **Continuous monitoring between survey rounds.** Generates timely estimates across the multi-year intervals between costly household surveys, so socioeconomic change can be tracked continuously rather than observed only at survey points.
- **Evidence-based survey prioritization.** Directs scarce survey budgets to where representative data collection yields the highest return, reserving lightweight monitoring elsewhere.

Looking ahead - translating outputs into outcomes. Operational validation, integrated uncertainty in outputs delivered to operations, and using the model to prioritize the next round of household surveys.

Follow project updates:

Appendix

Socio-economic decline, South Sudan

Predictions from a fine-tuned earth-observation model for 2023, 2024, and 2025 – for the camp areas.