

IMPROVING ESTIMATES OF HUMAN CAPITAL OUTCOMES IN DEVELOPING COUNTRIES BY INTEGRATING SURVEY AND GEOSPATIAL DATA

Lanu Kim (KIST), Peter Lanjouw (VUA), Josh Merfeld (UQ),
David Newhouse (WB), Michael Weber (WB)



WORLD BANK GROUP

LSMS conference
December, 2025

Background

1. **New advances in accessibility of geospatial data offer new possibilities for data integration with surveys**
 - Burke et al, 2022, Newhouse et al, 2024, Watmough et al, 2025
 - Small area estimation (SAE) is a key application
2. **Evaluations generally find encouraging results for geospatial SAE applied to poverty and wealth**
 - Chi et al, 2022, Newhouse, 2024, Van der Weide et al, 2024, Zheng et al, 2026
3. **Less evidence demonstrating effectiveness of approaches to human capital**
 - Two existing papers (Daod et al, 2019, Head et al, 2017)
 - These papers apply Convolutional Neural Networks for prediction at highly granular levels
 - Find mixed results, depending on the indicator

What's new in this paper?

1. Different application

- Tests approach to meso-level (i.e. district level) and representative (i.e. state level)
- Can data integration techniques improve standard statistics in addition to enabling more granular statistics?

2. Simpler prediction methods

- Uses features instead of raw imagery
- Uses linear mixed model
- Allows for uncertainty estimation using well-established method

What's new in this paper?

3. Evaluation with design-based simulations

- Based on repeated samples from georeferenced census data
- Can compare with direct survey estimates to demonstrate improvement
- Allows for evaluation of Empirical Bayesian prediction methods

4. Propose a novel filtering criteria based on predictive power and normality to identify cases that are likely to perform poorly

- Criteria are that Lasso selects at least one variable for model and
- Marginal area $R^2 > 0.05$, Skewness < 5 , and Kurtosis < 50

Use georeferenced census data from four countries

	Madagascar	Malawi	Mozambique	Sri Lanka
<i>Sub-area</i>				
Name	Fokantany	Enumeration area	Enumeration area	GN Division
Sample	1,056	768	552	984
Population	14,412	18,700	67,239	13,989
<i>Target area</i>				
Name	Commune	Traditional Authority	Locality	DS Division
Sample	592	286	320	284
Population	1,524	420	1,282	331
<i>Representative area</i>				
Name	Region	District	Province	Province
Sample and Population	23	28	11	25

Test 23 indicators total across four countries

	Madagascar	Malawi	Mozambique	Sri Lanka
Health				
Fertility	Yes	Yes		
Mortality	Yes	Yes		
Handicapped		Yes		
Basic Sanitation	Yes	Yes		Yes
Basic Water	Yes	Yes		Yes
Education				
Primary graduate	Yes	Yes	Yes	Yes
Secondary graduate	Yes	Yes	Yes	Yes
Jobs				
LFP	Yes	Yes		
Employed	Yes	Yes		
Number of indicators	8	9	2	4

Geospatial predictor variables

Variable	Source
Pollution (CO2)	Copernicus
Land classification	Copernicus
Night time lights	VIIRS
NDVI	MODIS
Temperatur, precipitation, wind	TERRACLIMATE
Population density	Worldpop
Distance to major cities	Manual calculation
MOSAICS	Rolf et al (2021)

Over 4,000 candidate variables total

Methods

- Use design-based simulations based on census
 - Draw 100 two-stage samples designed to mimic household surveys
 - Treat sub-areas as clusters, assume eight households per cluster
 - Generate estimates, then compare with “truth” derived from full census
 - Examine Pearson correlation (but also Rank Correlation and Mean Absolute Error)

Sample size	Madagascar	Malawi	Mozambique	Sri Lanka
<i>Sub-areas</i>	1,056	768	552	984
<i>Households</i>	8,448	6,144	4,416	7,872

Estimation Method

- Linear Empirical Best Predictor model

Predict mean values at sub-area level, then aggregate to target area
Straightforward to implement with R Povmap package

$$G(\hat{Y}_{ras}) = x'_{ras}\beta + \delta_r + \eta_{ra} + \varepsilon_{ras}$$

r=representative area, a = target area, s = subarea

\hat{Y}_{ras} is average proportion at sub-area level (i.e. share of EA with basic sanitation)

$$G(\hat{Y}_{ras}) = \arcsin\left(\sqrt{\hat{Y}_{ras}}\right)$$

x'_{ras} are geospatial predictor variables

δ_r are representative area dummy variables

$\eta_{ra} \sim N(0, \sigma_\eta^2)$ is a random area effect conditioned on the survey data

$\varepsilon_{ras} \sim N(0, \sigma_\varepsilon^2)$ is an independent error term

Model selection

- Use LASSO for model selection

$$G(\hat{Y}_{ras}) = x'_{ras}\beta + \delta_r + \varepsilon_{ras}$$

Regional dummies δ_r are not penalized and “forced” to be selected

LASSO selects most predictive variables without overfitting to sample

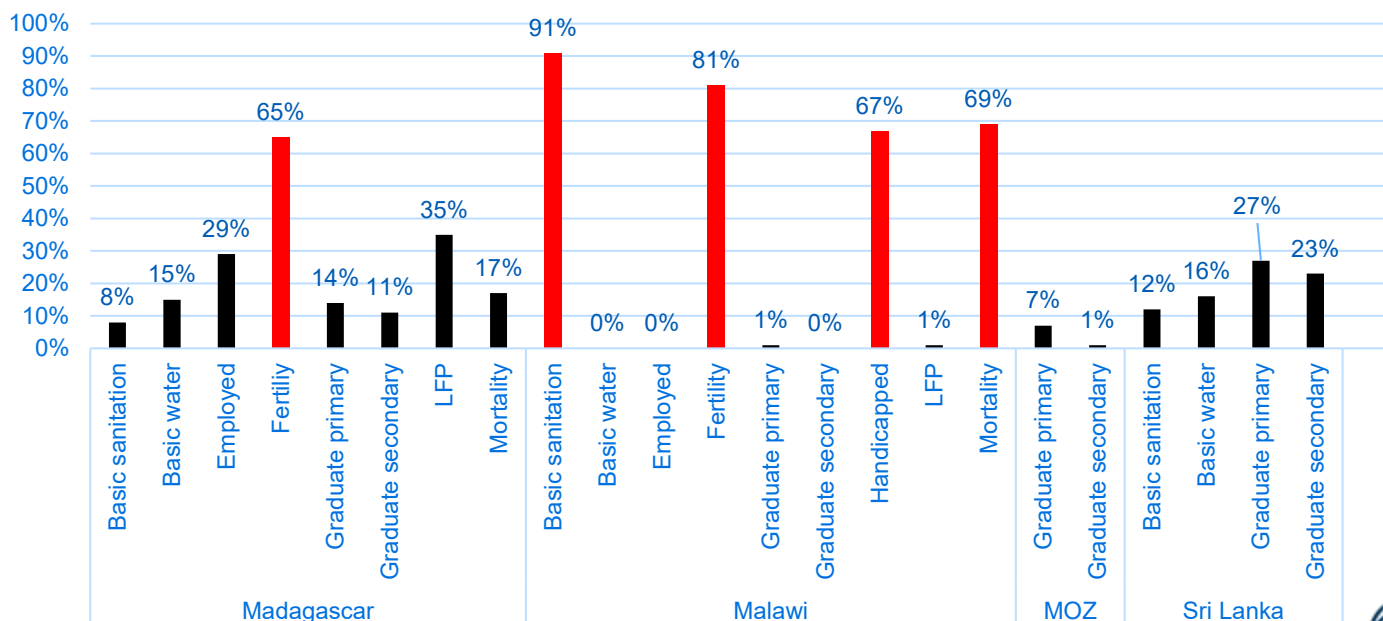
LASSO sometimes selects zero predictor variables

Common for basic sanitation, fertility, handicapped, mortality in Malawi and Fertility in Madagascar

In these cases, geospatial variables are not minimally predictive of outcomes

Census-based predictors also fail in these cases – sample not large enough to support SAE

Proportion of samples for which no geospatial predictors selected



Use “Marginal area R²” for filtering cases that are not predictive

- Definition of marginal area R²

$$MAR^2 = \text{cor}(\bar{y}_{ra}, \bar{x}'_{ra}\hat{\beta})^2$$

$$\bar{y}_{ra} = \frac{\sum_{s=1}^{S_a} w_{ras} y_{ras}}{\sum_{s=1}^{S_a} w_{ras}}$$

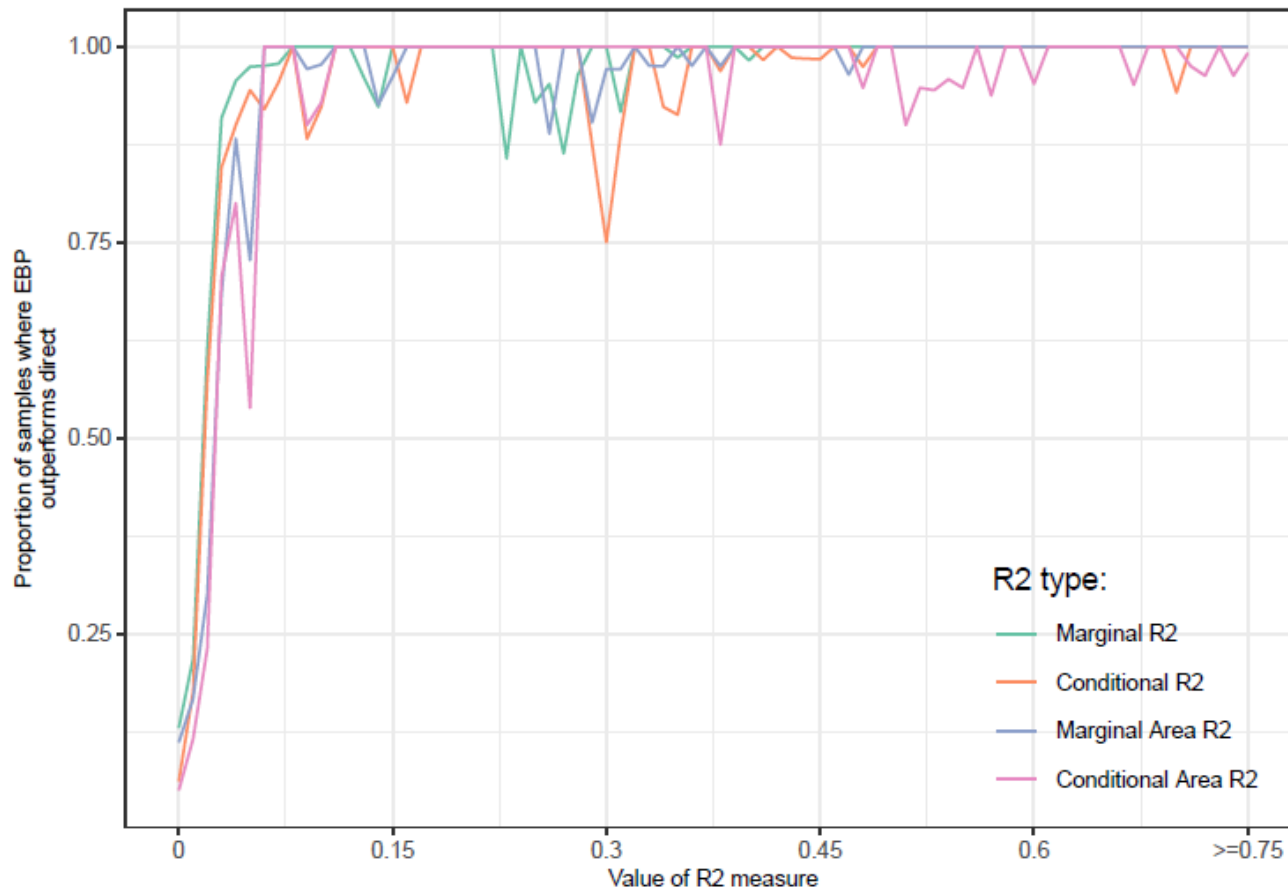
$$\bar{X}_{ra} = \frac{\sum_{s=1}^{S_a} w_{ras} x_{ras}}{\sum_{s=1}^{S_a} w_{ras}}$$

Where w_{ras} are sample weights

- A measure of model accuracy at predicting variation at the target area level
 - This is the variation that matters because we are aggregating predictions to that level
- Also impose skewness and kurtosis thresholds because estimation method assumes normal error terms
- Thresholds selected by empirical examination

“Marginal area R^2 ” more systematically related to improvement than other candidate R^2 measures

Figure 1: Proportion of samples in which EBP improves on direct estimates by different R^2 measures



Filtering process

	Number of samples	Number of indicators
Initial data	2300	23
LASSO selects at least one variable	1907	23
Marginal area $R^2 > 0.05$ and Skewness < 5 and Kurtosis < 50	1710	19

Data integration improves accuracy on average

Average correlation with truth (across all indicators)	Direct estimates	Model-based estimates
Small areas (i.e. districts)	0.70	0.82
Representative areas (i.e. states)	0.86	0.91

- For small area estimates, of 1710 samples that pass filter, 1706 (99.7%) improve on direct estimates
- For representative area estimates, 1314 samples out of 1710 (90%) improve on direct estimates

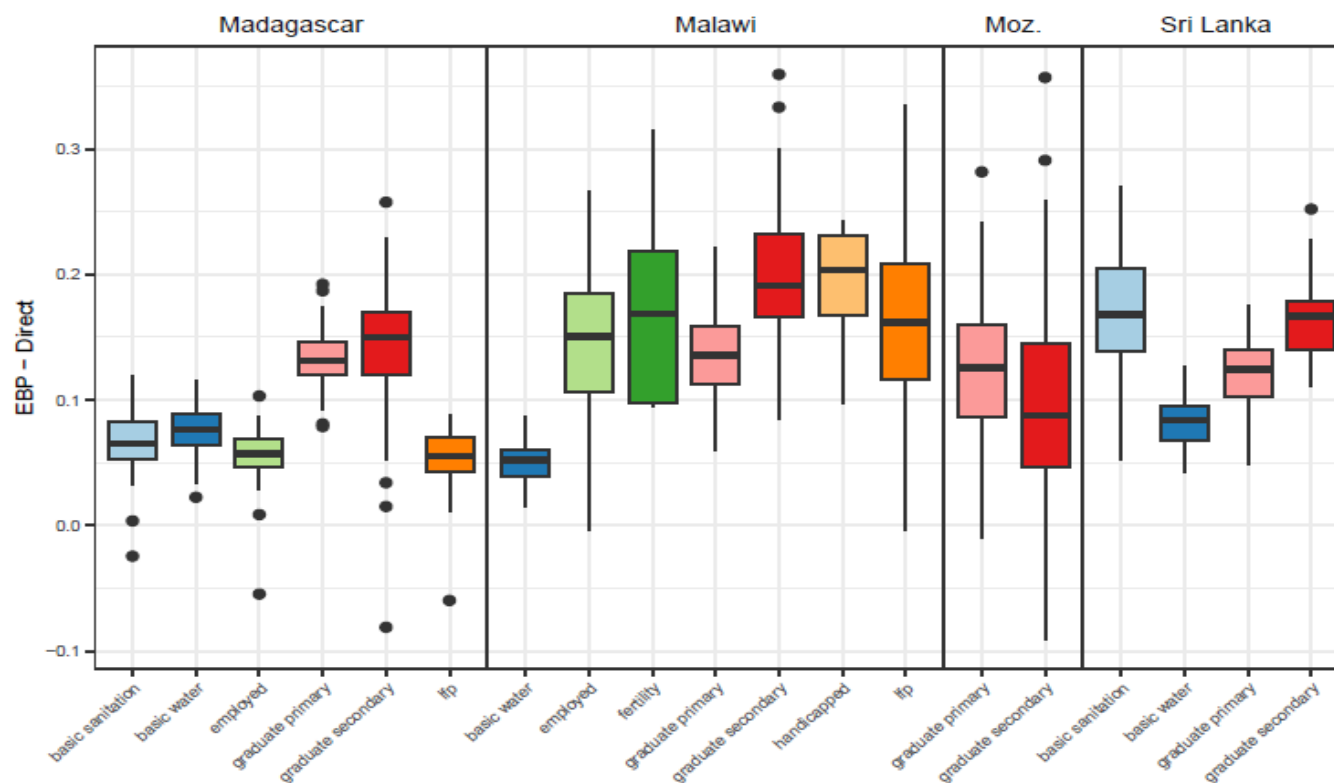
Coverage rates also improve

- Especially for small area estimates
 - Direct estimates underestimate uncertainty
 - Especially when using standard cluster-robust variance estimator

Coverage rate	Direct estimates	Model-based estimates
Small areas	63.9%	89.1%
Representative areas	67.4%	71.8%

Large heterogeneity in improvement (in Pearson correlation) over direct estimates

- Even within outcomes: Accuracy of LFP improves more in Malawi than in Madagascar
- Malawi and Sri Lanka appear generally conducive to geospatial SAE
- Madagascar least so, though estimates still improve on average



Results suggest optimism regarding geospatial SAE for human capital indicators

- Tests show significant improvement in accuracy on average
 - Even small amounts of improvement matter a lot
- Effectiveness is heterogeneous across indicators and contexts
- Differences likely due to differences in strength of relationship between outcomes and population density
 - Satellite data capture population density extremely well
 - Stronger relationship with educational attainment and LFP than mortality, basic water, employment
- Filtering criteria is promising for weeding out cases that don't work
- Applying these techniques is essentially costless
 - GeoLink and Povmap R packages make it simpler
- We should be doing this routinely

Relatively small increases in correlation can coincide with large increases in precision

Poverty case studies	Increase in correlation	Relative increase in effective sample size
Mexico	0.8 to 0.86	4.4
Sri Lanka	0.73 to 0.88	3.2
Tanzania	0.77 to 0.88	6.7

Caveats and future extensions

- Could incorporate more predictors such as buildings and machine learning embeddings to improve predictive performance
- Can test machine learning models (i.e. XGboost)
 - Does not condition on survey data like EBP
 - Can use all candidate variables
 - More flexible functional form
 - Avoids normality assumptions
 - Outperforms EBP in many settings for geospatial SAE and uncertainty can be accurately estimated (Merfeld et al, 2025)

Caveats and future extensions

- **More sophisticated foundational models may improve accuracy**
 - Learn from massive amounts of labeled images
 - Better for more granular predictions
 - Foundational models and fine-tuning procedures are improving rapidly but interpretability remains a key challenge
 - Evidence is mixed on whether vision transformer models outperform simpler machine learning models
 - Depends on training data and features (Zheng et al, 2026)

Thank you!

Additional outcomes indicators: Rank correlation

Table A1: Rank correlation

Country	Variable	Small area (in sample only)			Representative area		
		Direct	EBP	EBP-Direct	Direct	EBP	EBP-Direct
Madagascar	basic sanitation	0.718	0.788	0.070	0.860	0.911	0.051
Madagascar	basic water	0.698	0.763	0.065	0.760	0.892	0.131
Madagascar	employed	0.643	0.708	0.065	0.774	0.844	0.070
Madagascar	graduate primary	0.648	0.773	0.125	0.770	0.841	0.071
Madagascar	graduate secondary	0.444	0.669	0.225	0.502	0.743	0.241
Madagascar	lfp	0.630	0.698	0.068	0.778	0.855	0.077
Malawi	basic water	0.787	0.819	0.033	0.803	0.824	0.021
Malawi	employed	0.492	0.664	0.172	0.746	0.834	0.088
Malawi	fertility	0.345	0.512	0.167	0.627	0.670	0.042
Malawi	graduate primary	0.662	0.819	0.157	0.799	0.857	0.058
Malawi	graduate secondary	0.564	0.760	0.195	0.632	0.804	0.172
Malawi	handicapped	0.311	0.456	0.146	0.490	0.407	-0.083
Malawi	lfp	0.460	0.642	0.182	0.740	0.836	0.096
Mozambique	graduate primary	0.749	0.892	0.144	0.828	0.854	0.026
Mozambique	graduate secondary	0.698	0.862	0.164	0.671	0.639	-0.032
Sri Lanka	basic sanitation	0.524	0.747	0.223	0.787	0.887	0.100
Sri Lanka	basic water	0.753	0.843	0.089	0.863	0.899	0.036
Sri Lanka	graduate primary	0.707	0.864	0.157	0.800	0.901	0.101
Sri Lanka	graduate secondary	0.681	0.849	0.168	0.797	0.880	0.083
Average	all	0.632	0.768	0.136	0.759	0.847	0.088

Note: Measures of accuracy are averages across all 100 independent samples drawn from the respective country census. These are simple, unweighted averages across areas. Direct refers to direct survey estimates and ebp refers to SAE. Averages are weighted based on the number of non-zero simulations.

Additional outcomes indicators: Mean Absolute Error

Table A1: Rank correlation

Country	Variable	Small area (in sample only)			Representative area		
		Direct	EBP	EBP-Direct	Direct	EBP	EBP-Direct
Madagascar	basic sanitation	0.718	0.788	0.070	0.860	0.911	0.051
Madagascar	basic water	0.698	0.763	0.065	0.760	0.892	0.131
Madagascar	employed	0.643	0.708	0.065	0.774	0.844	0.070
Madagascar	graduate primary	0.648	0.773	0.125	0.770	0.841	0.071
Madagascar	graduate secondary	0.444	0.669	0.225	0.502	0.743	0.241
Madagascar	lfp	0.630	0.698	0.068	0.778	0.855	0.077
Malawi	basic water	0.787	0.819	0.033	0.803	0.824	0.021
Malawi	employed	0.492	0.664	0.172	0.746	0.834	0.088
Malawi	fertility	0.345	0.512	0.167	0.627	0.670	0.042
Malawi	graduate primary	0.662	0.819	0.157	0.799	0.857	0.058
Malawi	graduate secondary	0.564	0.760	0.195	0.632	0.804	0.172
Malawi	handicapped	0.311	0.456	0.146	0.490	0.407	-0.083
Malawi	lfp	0.460	0.642	0.182	0.740	0.836	0.096
Mozambique	graduate primary	0.749	0.892	0.144	0.828	0.854	0.026
Mozambique	graduate secondary	0.698	0.862	0.164	0.671	0.639	-0.032
Sri Lanka	basic sanitation	0.524	0.747	0.223	0.787	0.887	0.100
Sri Lanka	basic water	0.753	0.843	0.089	0.863	0.899	0.036
Sri Lanka	graduate primary	0.707	0.864	0.157	0.800	0.901	0.101
Sri Lanka	graduate secondary	0.681	0.849	0.168	0.797	0.880	0.083
Average	all	0.632	0.768	0.136	0.759	0.847	0.088

Note: Measures of accuracy are averages across all 100 independent samples drawn from the respective country census. These are simple, unweighted averages across areas. Direct refers to direct survey estimates and ebp refers to SAE. Averages are weighted based on the number of non-zero simulations.

References

- Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), eabe8628.
- Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3), e2113658119.
- Daoud, A., Jordán, F., Sharma, M., Johansson, F., Dubhashi, D., Paul, S., & Banerjee, S. (2023). Using satellite images and deep learning to measure health and living standards in India. *Social Indicators Research*, 167(1), 475-505.
- Head, A., Manguin, M., Tran, N., & Blumenstock, J. E. (2017, November). Can human development be measured with satellite imagery?. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development* (pp. 1-11).
- Merfeld, J. D., Dang, H.A., and Newhouse, D. (2025). Improving estimates of mean welfare and uncertainty in developing countries. *World Bank*.

References

- Newhouse, D. (2024). Small Area Estimation of Poverty and Wealth Using Geospatial Data: What Have We Learned So Far?. *Calcutta Statistical Association Bulletin*, 76(1), 7-32.
- Newhouse, D., Ramakrishnan, A., Swartz, T., Merfeld, J., & Lahiri, P. (2025). Small area estimation of monetary poverty in Mexico using satellite imagery and machine learning. *Oxford Bulletin of Economics and Statistics*, 87(6), 1158-1172.
- Van Der Weide, R., Blankespoor, B., Elbers, C., & Lanjouw, P. (2024). How accurate is a poverty map based on remote sensing data? An application to Malawi. *Journal of Development Economics*, 171, 103352.
- Watmough, G. R., Brockington, D., Marcinko, C. L., Hall, O., Pritchard, R., Berchoux, T., ... & Seth, S. (2025). A perspective on the interpretability of poverty maps derived from Earth Observation. *Science of Remote Sensing*, 100298.
- Zheng, Z., T. Wu, R. Lee, D. Newhouse, T. Kilic, M. Burke, S. Ermon, D. Lobell (2026). Dynamic, High-Resolution Wealth Measurement in Data-Scarce Environments, *Journal of Development Economics*, 179 (doi.org/10.1016/j.jdeveco.2025.103691)

References
