**Small Area Estimation of Poverty in Somalia: A Fay Herriot Model Approach**

Ifeanyi Edochie

September 2024

### Introduction

Household surveys are often representative at the national level or at the level of the first administrative division (region/state level). National Statistical Offices and government entities can benefit from poverty estimation at a higher level of resolution, such as the district or district level. This note describes the small area estimation (SAE) methodology implemented to estimate poverty rates in Somalia at the district level. SAE is a statistical method that can be used to improve the reliability of survey estimates by combining survey data with geographically comprehensive auxiliary data, such as census when available or geospatial, remotely sensed data. In Burundi, we show that SAE generates poverty estimates that are sufficiently precise to report at the district level instead of the regional level. This has the potential to improve the targeting and evaluation of interventions intended to achieve poverty reduction in the future. Ideally, SAE combines survey data with household-level data from a recent census. Countries often aim to collect census data every 10 years. However, many African countries take more years between consecutive censuses, and indeed Somalia’s last census was conducted in 1986. Therefore, this exercise relies on contemporaneous geospatial data derived from a variety of sources. (Battese, Harter, and Fuller 1988) were the first to use geospatial satellite data in the context of crop production. Other studies (Georganos et al. 2019), (Chi et al. 2022) have used satellite imagery to predict wealth indices, and (Van Der Weide et al. 2023) used satellite imagery to predict monetary poverty in Malawi. In this note, we present the approach that models poverty rates at the district level in Somalia using the model of (Fay III and Herriot 1979). The area level model approach allows us to relate district level direct survey poverty rates to auxiliary variables (geospatial indicators) to estimate poverty rates in all districts within Somalia. (Seitz 2019) provides district level poverty rates in the Central Asia region using Fay Herriot modelling approach using auxiliary geospatial data. The World Bank has employed the SAE methodology extensively to estimate poverty and other socioeconomic indicators of interest at more granular levels and continues to produce these estimates in combination with other non-monetary measures of poverty. At this point, SAE has been applied in a wide variety of contexts across many developing countries. This note is subdivided as follows. In section 2, we present survey data (specifically the household consumption data) and why SAE is necessary for district level poverty estimation in Somalia. We also present the Fay-Herriot model as described by (Fay III and Herriot 1979). Section 3 describes the geospatial databases sourced and indicators created as well as the model selection process employed. Sections 4 and 5 describes the FH model results and the poverty maps for the country.

### The Data

For Somalia, the 2022 Somalia Integrated Household Budget Survey (SIHBS) is representative at the regional level. The development of the SIHBS-22 sampling frame followed a stratified multi-stage probability cluster sample design. Urban and rural areas followed a three-stage stratified cluster sample design, while in nomadic areas the design was a two-stage stratified cluster sample design. The primary sampling units (PSUs) were selected with a probability proportional to the number of dwelling structures which constituted the sampling frame. The secondary sampling units (SSUs) for rural and urban areas were selected with a probability proportional to the number of listed households which constituted the frame. The ultimate sampling units (USUs) for rural, urban, and nomadic areas were randomly selected from listed households in the cluster.

District level poverty rate estimates computed from this survey will be insufficiently precise and unreliable for publication. Table 1 illustrates why it is necessary to use SAE to report poverty rates at more disaggregated geographic levels in Somalia. We use the mean coefficient of variation (CV) as a standardized measure of precision (i.e. the square root of the estimated mean square error divided by the poverty rate). Differing thresholds for mean or median CVs, often ranging from 0.1 to 0.3, have been applied by National Statistics Offices to determine if statistics are sufficiently reliable to report. The median and mean direct CVs in Somalia at the district level are approximately 0.18 and 0.23. While this is within the acceptable range of reliability for some countries, it is not considered reliable enough to publish by the Somalia National Statistics Office. At the regional (admin 1) level, the mean and median CVs are approximately 0.099 and 0.098, which is considered publishable.

Table 1: Descriptive Statistics

| **Indicator** | **Estimate** |
| --- | --- |
| Population (in millions) | 16.9 |
| Population Number of HHs (in millions) | 13.6 |
| Sample # of HHs | 6,221 |
| Poverty Rate (IPL, $2.15/day) | 0.5135 |
| Latest Census Year | 1986 |
| Number of Regions | 17 |
| Region Median CV | 0.09800725 |
| Region Mean CV | 0.09745189 |
| Number of Targets (Population) | 74 |
| Number of Targets (Sample) | 48 |

We utilize freely available geospatial data for this small area estimation exercise since the last census carried out is outdated (from 1986). The goal of the SAE exercise is to estimate more reliable district level poverty rates in Somalia by using a Fay Herriot model based on relating the target area direct estimate poverty rates and district level geospatial indicators. Given that any recent developments in Somalia might not be captured by its 15-year-old census, it would be difficult to make a case for area poverty rates estimated using the 1986 census particularly to guide current policy interventions.

### The Methodology

(Corral et al. 2022) recommends implementing an area level Fay Herriot level model with geospatial indicators for poverty mapping. Imagine a finite population for Somalia, , that consists of households that are subdivided into districts with sizes . A random sample of households can be drawn from the district (i.e., . The Fay-Herriot (FH) model comprises of two levels. The first is a sample model which assumes a direct survey estimator:

is design unbiased for the small area parameter, the population indicator of interest, in this case, the poverty rate each district, . We assume a sample error with the usual independently and normally distribution properties.

In the second level, a linking model is assumed to relate to auxiliary variables via a linear regression. Both levels of the model together are presented as follows:

The empirical best linear unbiased estimators (EBLUP) are computed with by weighted least squares regression. The EBLUP of is obtained by substituting the variance parameter with an estimate. The resulting estimator can then be written as:

The EBLUP/FH estimator can be understood as a weighted average of the direct estimator and a regression synthetic part. The estimated shrinkage factor puts more weight on the direct estimator when the sampling variance is small and vice versa. Areas for which no direct estimation results are called out-of-sample domains. For those domains the prediction reduces the regression-synthetic component (Molina and Rao 2010). This model is then fit via a restricted maximum likelihood (REML) method.

This method is widely used by the Small Area Income and Poverty Estimates (SAIPE) program of the US census bureau and has been thoroughly validated in (Corral Rodas, Henderson, and Segovia 2023). This approach improves the error efficiency rates over the direct estimates at the target area level. Inter-area unexplained heterogeneous area effects are accounted for within the model. Section 3.3 in (Corral et al. 2022) provides a full list of pros and cons of the Fay-Herriot modelling approach. We use the R package “povmap” to estimate this model in R.

### The Geospatial Data and Model Selection Process

The process leading up to model selection involves sourcing freely available geospatial indicators that might be correlated with household welfare and poverty. The geospatial features were sourced at native resolution and then zonal statistics were computed at the target area level (districts). Table 2 shows all the geospatial features and the data sources employed.

Table 2: EBP Model (Regression Results)

| **Feature** | **Source** | **Original Data Resolution** | **Year** |
| --- | --- | --- | --- |
| Built-settlement extent area | WorldPop Building Footprints | 1km | 2001 - 2020 |
| Gridded Population & Density | WorldPop Gridded Population Counts & Density | 90m | 2020 |
| Share of area planted by crop for banana, beans, cassava, maize, sesame seed, sorghum, sugar cane, temperature fruit,  tropical fruit, vegetables | IFPRI Spatial Production Allocation Model (SPAM) | 10km | 2009, 2017, 2020 |
| Production quantity for each crop for banana, beans, cassava, maize, sesame seed, sorghum, sugar cane, temperature  fruit, tropical fruit, vegetables | IFPRI Spatial Production Allocation Model (SPAM) | 10km | 2009, 2017, 2020 |
| % production as a total crop production for banana, beans, cassava, maize, sesame seed, sorghum, sugar cane,  temperature fruit, tropical fruit, vegetables | IFPRI Spatial Production Allocation Model (SPAM) | 10km | 2009, 2017, 2020 |
| Standardized precipitation evaporation index, 12 month | Global SPEI database, version 4.03 | 0.5 degrees | 2020 |
| Drought exposure, Drought hazard, Drought risk index, Drought vulnerability | (Carrao et al. 2018) | 0.5 degrees | 2000-2014 |
| Drought hazard, risk for irrigated agricultural systems | Drought risk for rainfed, irrigated agric. systems aggregated as an average per polygon based on the data from (Meza et al. 2020) | 1 km | 2020 |
| Percent of area (with Vegetation Index below 40) for the Gu season (April - June) | STAR - Global Vegetation Health Products | 1 km | 2017-2022 |
| Average travel time in nearest urban areas with 5000km, 20000km and 50000km | Computed based on population data from WorldPop and accessibility data from (Nelson et al. 2019) | 1 km | 2019 |

We begin by transforming all indicators as necessary to minimize the risk of divergence in model parameter estimation. For indicators with values greater than 100, we take the natural logarithm. We have avoided feature scaling to avoid excessive distortion or loss of information for the scaled variables.

The geospatial data listed under the previous header was used to construct candidate features at the grid and target area level. In addition, we include regional dummy variables. In all, we created ~157 potential geospatial candidate indicators. Using all these features in the linear mixed model risks potentially leads to overfitting the survey sample and generates poor out-of-sample estimations.

Next, we employ a Least Absolute Shrinkage and Selection Operator (LASSO) model selection approach, which selects a predictive set of variables while avoiding model overfitting. In particular, we use the rlasso function with the hdm R package (Bach, Chernozhukov, and Spindler 2018) which allows for selecting the optimal lambda parameter by minimizing the Bayesian Information Criterion (BIC) based on coefficients that are not shrunk towards zero. We employ this approach because we are using the LASSO to select variables for use as predictors in the Fay Herriot Model. The dependent variable under the model selection is the district (target area) level direct poverty rate.

### Fay Herriot Model Estimation Results

The FH model is implemented without any transformation to the district poverty rates. The final selected model includes covariates share of the population within a 2km grid as well as the share of total production that is tropical fruit production. The regression results are as follows:

Table 3: FH Model (Regression Results)

| **Variables** | **Coefficients** | **Standard Error** |
| --- | --- | --- |
| Intercept | 0.499\*\*\* | 0.032 |
| Production share in Tropical Fruit Production | 0.038\*\* | 0.013 |
| Average share of population within 2km grids | -0.0039\*\* | 0.002 |

The regression coefficients have the expected sign. The share of the population within the 2km grid is decreasing in poverty i.e. districts with more densely populated grids appear to be more well off (i.e. less poor). This may indicate that higher density areas attract individuals with more income or earning potential which is in line with economic literature.

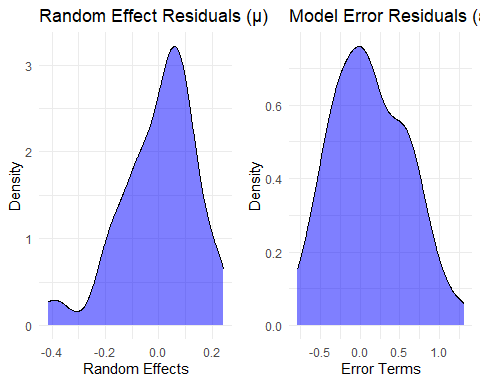
Several assumptions are made in this model which needed to be verified. The Fay Herriot model equals 28.5% with an adjusted of 20.4%which is typical for the FH model particularly with only 48 in-sample districts used in the regression of only geospatial features. We assume independent normal distributions for the area effects as well as error terms. The table shows the skewness and kurtosis which should be approximately 0 and 3 for normally distributed random variables.

Table 4: Assessing Normality Assumptions

| **Model R²** | | | | **(Error Term) ε** | | **(Random Effect) μ** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Marginal** | **Conditional** | **AIC** | **BIC** | **Skewness** | **Kurtosis** | **Skewness** | **Kurtosis** |
| 0.204 | 0.286 | -14.43 | -6.94 | 0.287 | 2.507 | -0.807 | 3.786 |

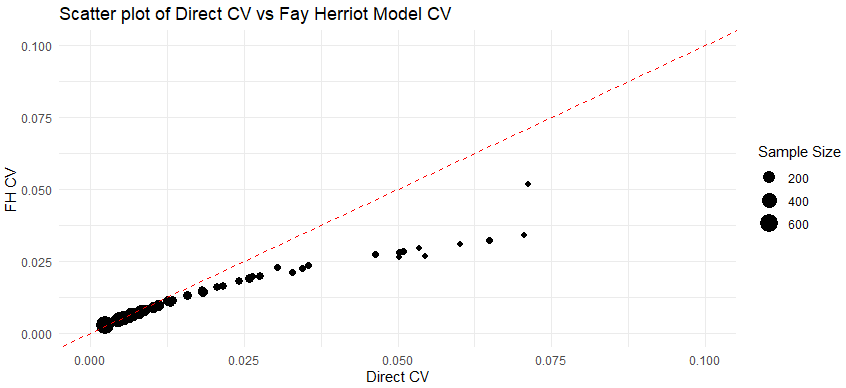
The normality assumptions proposed in the method section matter for the noise estimates but the empirical bayes methodology ensures that the poverty estimates are unbiased. The residual analysis suggests that the skewness and kurtosis of the idiosyncratic and district level area effects match the normality assumptions. However, there appears to be few outliers within the error term normal density plot. The residual plots for both the random error and idiosyncratic errors can be found below:

Figure 1: Fay-Herriot Residual Plots



We also attempt to compare the efficiency gains of the Fay Herriot modelling exercise by comparing the coefficients of variation (mean square errors divided by the target area poverty rates). On average, the ratio of FH model estimate CVs to direct CVs is 0.76. The plot below shows the relationship between both CV types.

Figure 2: Precision Gains from Fay Herriot Modelling

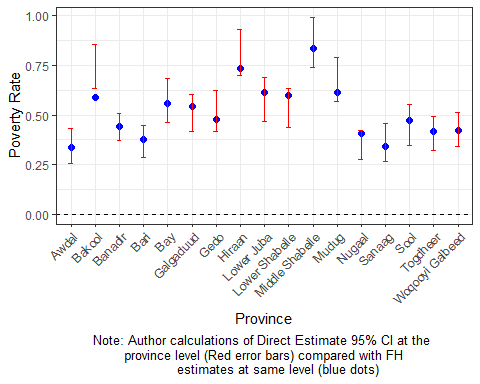


The results, as expected, show that the real value of the FH model appears to be in the precision gains made in target areas with a small sample size.

### Poverty Maps

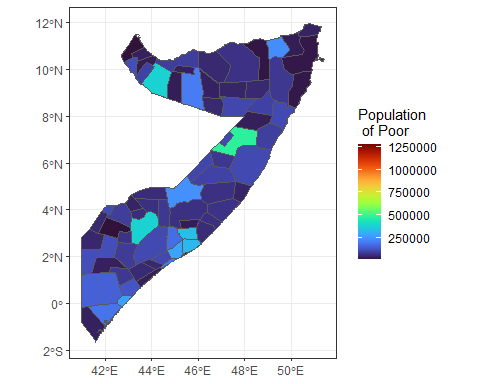
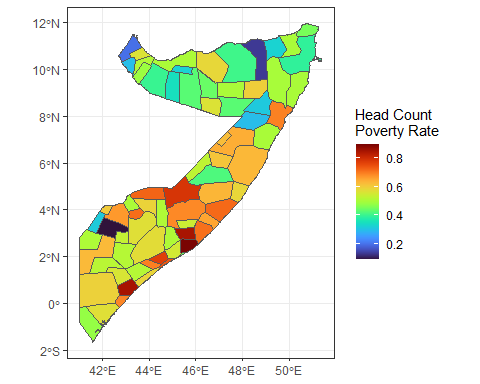
As a final check, the FH poverty rates at the district level are aggregated to the regional level to compare against the direct estimates. The regional level is the highest level of resolution at which survey design specifies national representativeness. The direct estimates in Figure 3 are shown as 95% confidence intervals (in red) which are plotted in comparison with Fay Herriot poverty estimates.

Figure 3: Region Fay-Herriot Estimates



It should be noted that all the model based regional estimates appear to lie within the direct estimates confidence intervals except in the Bakool region.

The poverty map corresponding to the 2022 for Somalia at the district level in the subsequent figures:



Upon review of the initial estimates in collaboration with SNBS, 3 districts were identified as having unrealistically low estimates of poverty as the surrounding districts within each region had significantly higher poverty rates. These districts were Garbahaarey, Lassqoray and Zeylac in Gedo, Sanaag and Awdal regions respectively. Three benchmarking approaches were implemented in attempts to solve the problem:

i) First, the raking benchmarking method iteratively adjusts district estimates until convergence is reached with the regional poverty rate. However, the FH model regional poverty rates are all within [X-X] percent of the direct estimates, as a result this had little effect in changing the district poverty rates.

ii) Next, the ratio method adjusts the district estimates using a constant factor.

iii) Finally, a method which incorporates the MSE estimates was also applied.

All three methods had minimal effect on the district poverty rates as they depend on the precision of sampling in the specific districts. The decision was made to treat all three districts as out of sample, which resulted in poverty rates more aligned with the neighboring districts. The results of this exercise can be found in the supplementary section.

### Appendix

Figure 5: Correlation between FH Model Estimates and Direct Estimates at District Level

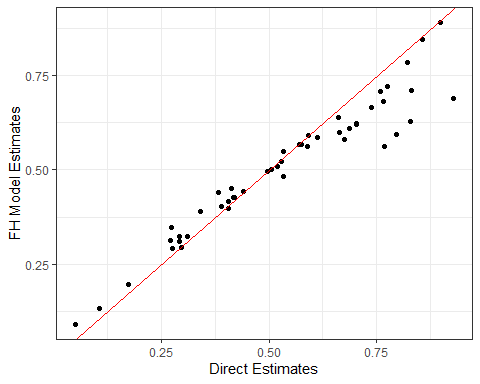


Table 5: Regional FH model estimates with confidence bounds

|  | **Survey** | **FH Estimate 95% Confidence Intervals** | |
| --- | --- | --- | --- |
| **Region** | **Direct Estimate** | **Lower Bound** | **Upper Bound** |
| Awdal | 0.343 | 0.254 | 0.432 |
| Woqooyi Galbeed | 0.427 | 0.340 | 0.515 |
| Togdheer | 0.406 | 0.322 | 0.490 |
| Sool | 0.450 | 0.348 | 0.553 |
| Sanaag | 0.363 | 0.266 | 0.460 |
| Bari | 0.366 | 0.285 | 0.447 |
| Nugaal | 0.349 | 0.276 | 0.422 |
| Mudug | 0.679 | 0.568 | 0.789 |
| Galgaduud | 0.509 | 0.416 | 0.602 |
| Hiraan | 0.815 | 0.700 | 0.931 |
| Middle Shabelle | 0.866 | 0.739 | 0.992 |
| Banadir | 0.440 | 0.374 | 0.505 |
| Lower Shabelle | 0.534 | 0.436 | 0.631 |
| Bay | 0.572 | 0.462 | 0.681 |
| Bakool | 0.743 | 0.632 | 0.855 |
| Gedo | 0.521 | 0.419 | 0.624 |
| Lower Juba | 0.578 | 0.468 | 0.689 |

Supplementary Section: Re-estimating FH Model without Garbahaarey, Lassqoray and Zeylac in SIHBS sample

### The National Statistical Office in Somalia flagged the aforementioned areas as possessing abnormally low poverty rates. Upon further inspection, it appears these districts make up 3 of the largest 4 areas in terms of coefficient of variation (CV). See table below:

Table S1: Sample District CV for the 10 top districts within SIHBS

|  |  |
| --- | --- |
| **District** | **Direct Sample CV** |
| Garbahaarey | 0.9997 |
| Zeylac | 0.6424 |
| Owdweyne | 0.4287 |
| Laasqoray | 0.4189 |
| Caynabo | 0.4119 |
| Lughaye | 0.3842 |
| Caluula | 0.3515 |
| Baki | 0.3213 |
| Iskushuban | 0.3179 |
| Sheikh | 0.3011 |

Finally, we remove the 3 districts flagged by the SNBS and re-iterate the entire modelling exercise previously described including both the model/variable selection process and the FH model estimation. The results are as follows:

Table S2: FH Model (Regression Results), without SNBS concern districts

|  |  |  |
| --- | --- | --- |
| **Variables** | **Coefficients** | **Standard Error** |
| Intercept | 0.5468\*\*\* | 0.0000 |
| Share of the harvest in all crops (cassava) | -0.0474\*\* | 0.0084 |
| Hiraan (Regional Dummy) | 0.2839\*\* | 0.0086 |
| Middle Shabelle (Regional Dummy) | 0.3370\*\*\* | 0.0004 |

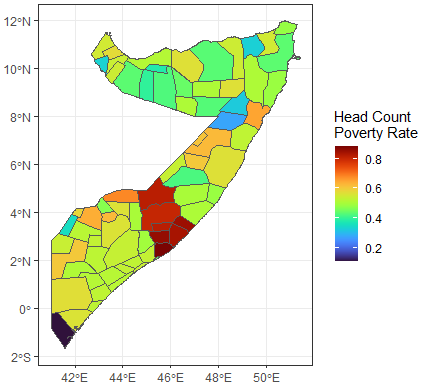
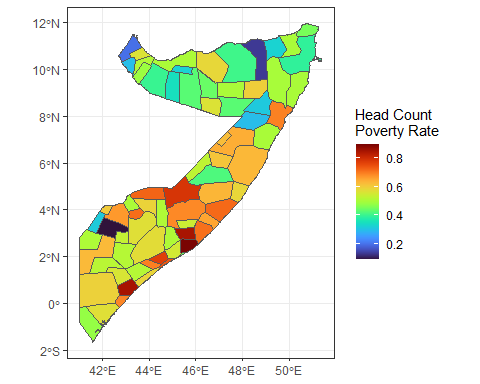
The table above shows 3 variables selected. Living in Hiraan and Middle Shabelle appear to increase the incidence in poverty. This makes sense considering that the areas are the poorest in the country. Cassava is a main cash crop in Somalia. A larger share of harvest area to cassava appears to reduce poverty rates would be consistent with economic intuition.

Table S3: Assessing Normality Assumptions, FH model without SNBS concerning districts

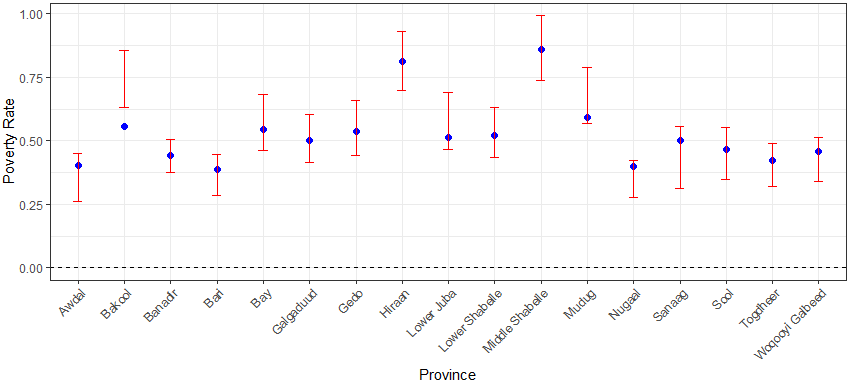
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model R²** | | | | **(Error Term) ε** | | **(Random Effect) μ** | |
| **Marginal** | **Conditional** | **AIC** | **BIC** | **Skewness** | **Kurtosis** | **Skewness** | **Kurtosis** |
| 0.266 | 0.404 | -30.93 | -21.90 | 0.327 | 2.608 | -0.312 | 3.040 |

The normality assumptions appear to have been met as both the skewness and kurtosis appear to centered around 0 and 3 respectively as economic theory postulates.

Figure S1: FH model poverty rates without SNBS areas (to the right) and with the entire SIHBS sample (to the left)



Finally, we present the results of regional estimates comparing FH model estimates to direct estimates. The result are as follows:

Figure S2: Regional FH model estimates, model without SNBS concerning districts 

### References

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2018. “Valid Simultaneous Inference in High-Dimensional Settings (with the Hdm Package for r).” *arXiv Preprint arXiv:1809.04951*.

Battese, George E, Rachel M Harter, and Wayne A Fuller. 1988. “An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data.” *Journal of the American Statistical Association* 83 (401): 28–36.

Carrão, Hugo, Gustavo Naumann, and Paulo Barbosa. 2016. “Mapping Global Patterns of Drought Risk: An Empirical Framework Based on Sub-National Estimates of Hazard, Exposure and Vulnerability.” *Global Environmental Change* 39: 108–24. https://doi.org/<https://doi.org/10.1016/j.gloenvcha.2016.04.012>.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua E Blumenstock. 2022. “Microestimates of Wealth for All Low-and Middle-Income Countries.” *Proceedings of the National Academy of Sciences* 119 (3): e2113658119.

Corral, Paul, Isabel Molina, Alexandru Cojocaru, and Sandra Segovia. 2022. *Guidelines to Small Area Estimation for Poverty Mapping*. World Bank Washington.

Corral Rodas, Paul, Heath Henderson, and Sandra Segovia. 2023. “Poverty Mapping in the Age of Machine Learning.” *Available at SSRN 4587156*.

Fay III, Robert E, and Roger A Herriot. 1979. “Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data.” *Journal of the American Statistical Association* 74 (366a): 269–77.

Georganos, Stefanos, Assane Niang Gadiaga, Catherine Linard, Tais Grippa, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, Sébastien Dujardin, and Moritz Lennert. 2019. “Modelling the Wealth Index of Demographic and Health Surveys Within Cities Using Very High-Resolution Remotely Sensed Information.” *Remote Sensing* 11 (21): 2543.

Meza, I., S. Siebert, P. Döll, J. Kusche, C. Herbert, E. Eyshi Rezaei, H. Nouri, et al. 2020. “Global-Scale Drought Risk Assessment for Agricultural Systems.” *Natural Hazards and Earth System Sciences* 20 (2): 695–712. <https://doi.org/10.5194/nhess-20-695-2020>.

Molina, Isabel, and J. N. K. Rao. 2010. “Small Area Estimation of Poverty Indicators.” *Canadian Journal of Statistics* 38 (3): 369–85. https://doi.org/<https://doi.org/10.1002/cjs.10051>.

Nelson, Andy, Daniel J Weiss, Jacob van Etten, Andrea Cattaneo, Theresa S McMenomy, and Jawoo Koo. 2019. “A Suite of Global Accessibility Indicators.” *Scientific Data* 6 (1): 266.

Seitz, William Hutchins. 2019. “Where They Live: District-Level Measures of Poverty, Average Consumption, and the Middle Class in Central Asia.” *World Bank Policy Research Working Paper*, no. 8940.

Van Der Weide, Roy, Brian Blankespoor, Chris Elbers, and Peter Lanjouw. 2023. “How Accurate Is a Poverty Map Based on Remote Sensing Data? An Application to Malawi.” *An Application to Malawi*.