**Small Area Estimation of Poverty in Namibia: The Census EBP Approach**

Ifeanyi Edochie

2024-06-20

## **0.1 Introduction**

The Republic of Namibia, a rapidly urbanizing country in southwestern Africa with a small population of 3 million people (2023), faces poverty and inequality levels that are high for its income level. The country borders Angola and Zambia (north), Botswana (east), South Africa (south), and the Atlantic Ocean (west), lying within the Namib and Kalahari (semi-arid) deserts, and is one of the driest countries of sub-Saharan Africa. Its economy depends on agriculture, tourism, and mining (particularly diamonds and metals). Strong economic growth, improvements in labor incomes and educational attainment, and the expansion of social protection programs, helped significantly reduce poverty from 37.5 in 2003/04 to 17.4 percent in 2015/16 (the last available poverty data; using the official poverty line). Despite these improvements, Namibia continues to face one of the highest levels of income inequality in the world, with a per adult equivalent consumption Gini coefficient of 57.6 percent in 2015.

Namibia made considerable progress since independence, with steady increases in access to education, water, sanitation, and electricity. However, access remains much weaker in certain parts of the country and especially in historically disadvantaged communities. In addition, the country’s relatively high poverty rate (for its income level) remains strongly correlated with lagging human capital and poor access to basic services. Deep inequalities remain in the distribution of resources, opportunities, and income. The country experiences prolonged droughts and growing water insecurities, with urbanization and economic growth increasing the demand for water. The poorest population, particularly in rural areas, are particularly vulnerable to weather-related shocks.

Poverty maps are useful to help visualize and understand the distribution of poverty at a much more granular level, which allows for better targeted policymaking. The national poverty rate, or even poverty rates by region, can hide large intraregional differences in poverty. By showing the incidence of poverty at very disaggregated levels, in the case of Namibia at the 107 constituency levels, poverty maps can help improve targeting to reduce poverty and inequality, improve public service delivery, and strengthen crisis response. This report presents the constituency-level poverty map of Namibia using the small area estimation (SAE) technique based on the (latest available) 2015/16 Namibia Household Income and Expenditure Survey (NHIES) and the Namibia 2011 Population and Housing Census. The combination of the NHIES data (which captures a sample of household data for measuring poverty at more aggregated levels) and the Census data (which collects information from every household but does not include expenditure data for poverty measurement) allows for the estimation of monetary poverty indicators at the constituency level in Namibia.

The Unit Level Small Area Estimation methodology

We estimate constituency level headcount poverty rates using a one-fold nested error empirical best predictor (EBP) household (unit) level model Tzavidis et al. (2018). We predict household welfare, i.e. log per capita consumption, as a linear function of the survey variables/features at the household level. The assumption is that per capita household welfare is a function of survey indicators, a random area (constituency) specific effect, and a household level idiosyncratic error term.

More formally, assume that is the design matrix containing p auxiliary features. The model is defined by:

where represents the log household per capita welfare, is a transposed vector of household level features and is a vector of the regression coefficients (both and have a length of , including the intercept). and refer to the random commune specific effects and unit level error terms, respectively, which are assumed to be , independently identically distributed random variables. The constituency effect is conditioned on the sample data .

The empirical EBP estimation process works as follows:

1. Use the household survey to fit the above-described model to estimate , , parameters. Restricted Maximum Likelihood (REML) is used to compute , using the nlme (Pinheiro et al. 2018) R package. The parameters are used to calculate the shrinkage factor

* where
* is the sample weight for household and is therefore the ratio fo the squared sum of weights to the square of the sum of weights for each constituency, c.

2) For in-sample target areas, is estimated L = 100 times as follows where and , the expected value of the area effect given the sample data. For the out-of-sample communes, there is no sample data to condition on, so and . Therefore, for these communes, and the estimates are purely synthetic predictions.

1. We take the exponential of

* to transform all monte carlo simulations back to the original scale. We apply the poverty line (6249 Namibian dollars, which is the upper bound poverty line) to for each of the simulations to estimate poverty rates for each simulation, and then average the estimated poverty rates over the simulations.

Mean square error, a measure of uncertainty, is estimated through the parametric bootstrap procedure implemented in the povmap package (González-Manteiga et al. 2007). See Kreutzmann et al. (2019) for more details.

## **0.2 Data Sources**

### 0.2.1 Namibia Household Income and Expenditure Survey (2015/2016)

The Namibia Household Income and Expenditure Survey (NHIES) is a household-based survey that collects data to measure the living conditions of the Namibian population, including poverty indicators. Information is collected on the actual patterns of consumption and income, and a range of other socio-economic indicators including demographic factors, housing and utilities, access to services, health, education, main source of income, household indebtedness and ownership and access to assets. The NHIES 2015/16 is a representative sample of 10,368 households from 864 primary sampling units (PSUs). The primary sampling frame used for this survey is a list of PSUs based on the 2011 Population and Housing Census Enumeration Areas (EAs). A PSU can be one EA, part of an EA, or more than one EA. A secondary sampling frame for each of the selected PSUs was created for the purpose of selecting the sample households through a listing procedure. The data was collected over a 12-month period consisting of 24 survey rounds, between March 27, 2015, and March 21, 2016. The survey was divided into four quarters and each quarter was further divided into survey rounds. A sample of 12 households were selected within each selected PSU. The survey is representative at the national level, by urban-rural, and by 14 regions (!Karas, Erongo, Hardap, Kavango East, Kavango West, Khomas, Kunene, Ohangwena, Omaheke, Omusati, Oshana, Oshikoto, Otjozondjupa, and Zambezi).

### 0.2.2 Namibia 2011 Population and Housing Census

The 2011 Population and Housing Census was undertaken in August 2011 (from August 28 to September 15, with August 28 marked as the reference period or “Census Day”) and is the main source for demographic and socio-economic statistics in the country. It provides information on the size, distribution, composition and other social and economic characteristics of the population as well as on household and housing amenities. The population characteristics include the population’s spatial distribution, age and sex composition, marital status, education, literacy, economic activity, orphanhood, and disability. The household and housing conditions include household size, housing amenities, household ownership, and the quality of housing. The 2011 Census reported a total population of 2,113,077 Namibians in 465,400 households. The population was split 43 percent urban (903,434 people) and 57 percent rural (1,209,643). The largest urban center was the capital of Windhoek with a population of 325,858 people. In 2011, Namibia was divided into 13 regions: Caprivi (called Zambezi in the 2015/16 NHIES), Karas (changed to !Karas in the NHIES), Erongo, Hardap, Kavango (split into East and West in the NHIES), Khomas, Kunene, Ohangwena, Omaheke, Omusati, Oshana, Oshikoto, and Otjozondjupa. These regions were further subdivided into 107 constituencies in 2011 (which became 121 in 2013 based on the fourth Boundaries Delimitation and Demarcation Commission). There were 53 proclaimed towns and villages in Namibia in 2011.

Table 1: Comparison of Coefficient of Variation (CV) between EBP and Direct Estimates

| **Area** | **EBP CV** | **Direct CV** |
| --- | --- | --- |
| CV for Area: Omusati | 0.0019 | 0.0621 |
| CV for Area: Zambezi | 0.0012 | 0.3646 |
| CV for Area: Oshana | 0.0021 | 0.3606 |
| CV for Area: Oshikoto | 0.0007 | 0.2515 |
| CV for Area: Otjozondjupa | 0.0075 | 0.2877 |
| CV for Area: Erongo | 0.0016 | 0.7145 |
| CV for Area: Hardap | 0.0011 | 0.0904 |
| CV for Area: Karaso | 0.0043 | 0.3450 |
| CV for Area: Kavango | 0.0015 | 0.1249 |
| CV for Area: Khomas | 0.0097 | 0.6373 |
| CV for Area: Kunene | 0.0058 | 1.0423 |
| CV for Area: Ohangwena | 0.0052 | 0.4227 |
| CV for Area: Omaheke | 0.0010 | 0.2976 |

Table 2: Basic Information on Units and Regions

| **Indicator** | **Census** | **Survey** |
| --- | --- | --- |
| Number of Households | 465,447 | 9,920 |
| Number of Regions | 13 | 13 |
| Number of Target Areas | 104 | 103 |

## 

## **0.3 The Variable Selection Algorithm**

As presented in the previous sectionschapters, the aim is to estimate a model of per capita household consumption for the survey and then useing the model to predict the distribution of per capita welfare within the census. Consequently, selecting the optimal set of indicators from the household survey is essential to the poverty mapping exercise. We create a set of harmonized variables between the NHIES 2015/16 survey and the 2011 census, surveys i.e. 290 variables with the same definitions between both census and survey data. These include indicators for household asset ownership (dummy variables for household ownership of cars, internet, fridges, number of rooms, computer, phone, cellphone, land phone, sewing machine, washing machine, stove etc), education (household literacy rate, highest education attainment), disability (sight, hearing, walking and selfcare disability dummies), among others.

Table 3: EBP Model (Regression Results)

| **Indicator.Class** | **Variables** |
| --- | --- |
| Household asset ownership status | Dummies for household ownership of cars, internet, fridges, number of rooms, computer phone, cellphone, landphone, sewing machine, washing machine, stove, bicycle, motorcycle, radio |
| Education | Household literacy rates, highest education attainment in household, whether ever attended school |
| Household-dwelling characteristics | Household size, floor type, wall type, roof type, proportion of household members in specific age category |
| Sanitation | Source of water, type of water pipe, garbage disposal type, toilet type |
| Energy | Type of fuel for cooking and lighting, heating source dummies |
| Location controls | Provincial dummies, sector (rural, urban) |

In predicting the welfare distribution of the census, a variable selection process is important to minimize both the error rate of the model and the risk of overfitting the model from including too many variables. As a result, we apply the Least Absolute Shrinkage Selector Operator (LASSO) algorithm method (Tibshirani 1996) as follows:

Consequently, high values of will increase the rate at which elements of the vector approach 0 and are thus eliminated from the selection model. We use the AIC selection criterion (Akaike 1974) to chooe the optimal yielding the selected variables highlighted in the subsequent model results section.

The step-by-step process of variable selection is explained in Table 4.

Table 4: Step-by-Step summary of variable selection process

| **Step** | **Process Summary** |
| --- | --- |
| Survey & Census GMD Variable Harmonization | The same set of variables (deemed potential predictors of per capita welfare) are created in the survey and census. We use the standard Global Monitoring Database 2.0 Harmonization dictionary for the feature creation process |
| Equivalent feature distribution check | We ensure that census and survey have similar distributions. In ideal circumstances, the survey is drawn from the census i.e. a census and survey in the same year. However, with the NHIES survey being from 2015 and the census in 2011, we perform a standard z-test to test that for each feature the population mean (census 2011) is not statistically different (p-value > 0.05) from the survey mean (NHIES 2015/16). We drop all variables for which we have to reject the (equal means) null hypothesis. |
| Model selection | Apply the LASSO algorithm with the ordinary least squares (OLS) estimation. The use of the OLS model in this case may be less than optimal being that the EB empirical model rather applies a linear mixed effects approach. |

## 

## **0.4 EBP Unit Model Estimation Results**

One advantage of using the linear mixed effect modelling approach over recent developments in machine learning in the ease of interpretation of the coefficient estimates. The estimation results of the model can be found in table 5, for the most part, appear consistent with expectations in terms of the direction of the signs of the estimates with a few exceptions.

Average welfare (measured as per capita household consumption) appears to be higher in Karas, Erongo, Hardap, Kavango, Khomas, Ohangwena, Omusati, Oshana, Oshikoto and Otjozondjupa provinces than in the Zambezi province (table 5. In terms of household asset ownerships, households with no cellphone are more likely to have lower consumption levels (thus are more likely to be poor). In a similar vein, households with internet access and a landphone consume more than their counterparts without any digital access. Ironically, households with no computer are also better off. This could be a result of inaccurate data records. There is also a positive correlation between the number of rooms in the household dwelling and the per capita consumption.

In terms of energy use, those households that power (lighting and heating) their homes with electricity and use kerosene for lighting also consume more. Unsurprisingly, households with no energy use are more likely to have lower consumption. With regards to dwelling characteristics of the household, households with natural/sand flooring have consumption.

Table 5: Mixed effects regression model

| **Variable** | **Coefficients** | **Standard Error** |
| --- | --- | --- |
| Intercept | 10.033\*\*\* | 0.078 |
| Erongo Dummy (=1 if HH in Erongo Region) | 0.263\*\* | 0.093 |
| Hardap Dummy (=1 if HH in Hardap Region) | 0.382\*\*\* | 0.096 |
| Karas Dummy (=1 if HH in Karas Region) | 0.225\*\* | 0.095 |
| Kavango Dummy (=1 if HH in Kavango Region) | 0.343\*\*\* | 0.085 |
| Khomas Dummy (=1 if HH in Khomas Region) | 0.643\*\*\* | 0.088 |
| Kunene Dummy (=1 if HH in Kunene Region) | 0.109 | 0.094 |
| Ohangwena Dummy (=1 if HH in Ohangwena Region) | 0.587\*\*\* | 0.085 |
| Omaheke Dummy (=1 if HH in Omaheke Region) | 0.138 | 0.094 |
| Omuasti Dummy (=1 if HH in Omusati Region) | 0.350\*\*\* | 0.085 |
| Oshana Dummy (=1 if HH in Oshana Region) | 0.490\*\*\* | 0.089 |
| Oshikoto Dummy (=1 if HH in Oshikoto Region) | 0.475\*\*\* | 0.092 |
| Otjozondjupa Dummy (=1 if HH in Otjozondjupa Region) | 0.270\*\* | 0.091 |
| Household with no cellphone | -0.19\*\*\* | 0.027 |
| Household with no computer | 0.417\*\*\* | 0.025 |
| Energy Source is electricity | 0.075\*\* | 0.028 |
| Main floor material: Natural Earth/Sand | -0.17\*\*\* | 0.022 |
| Main floor material: Natural-Other | -0.13\*\*\* | 0.034 |
| Main floor material: rudimentary wood planks | 0.182 | 0.107 |
| Main cooking fuel is Firewood | -0.21\*\*\* | 0.027 |
| Main lighting fuel: Candles | -0.037 | 0.027 |
| Main lighting fuel: Kerosene | 0.136\*\* | 0.051 |
| No main lighting fuel | -0.17\*\* | 0.063 |
| Main heating source: Electricity | 0.040 | 0.026 |
| Main heating source: Kerosene | -0.15 | 0.103 |
| Household size | -0.18\*\*\* | 0.003 |
| Household has access to internet | 0.183\*\*\* | 0.020 |
| Household owns landphone | 0.331\*\*\* | 0.038 |
| Household marital status: Living together | -0.18\*\*\* | 0.022 |
| Roofing material: finished concrete | 0.072\*\* | 0.036 |
| No of rooms | 0.083\*\*\* | 0.005 |
| Household doesnt own a television | -0.19\*\*\* | 0.023 |
| Main toilet facility: Flush | 0.119\*\*\* | 0.028 |
| Main material for external walls: cement blocks | 0.076\*\* | 0.027 |
| Main material for external walls: other | -0.028 | 0.024 |
| Main source of drinking water: bottled water | 0.469\*\*\* | 0.103 |
| Main source of drinking water: Tanker truck | 0.121 | 0.096 |
| Household has piped water in premise | 0.150\*\*\* | 0.021 |

Note: All variables are 0-1 dummies with the exception of “household size” and “number of rooms”

Larger households are negatively correlated with per capita household consumption. Households with cohabitating relationship appear to be poorer than their married counterparts. Household welfare is also an increasing function of the quality of the roofing, flooring, wall material and sanitation facilities as well. In addition, households that can afford bottled water tend to have on average higher per capita consumption than those that do not.

We compare the distribution of household welfare between predicted and actual household consumption. See results in Figure 1.

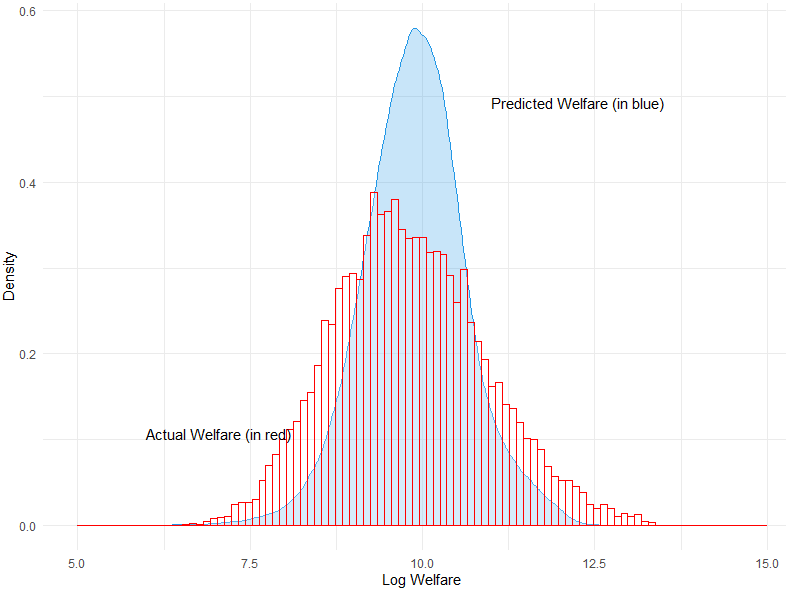


Figure 1: Comparing Welfare Density Distributions in Census and Survey

**0.5 Analysis of Variance**

Assessing the goodness of fit of the model estimation is important for understanding how well the selected set of variables explain the distribution of household consumption in the survey. In addition, the EB theoretical model assumes normality of the area random effect residuals and the idiosyncratic error terms.

Table 6: Assessing Normality Assumptions

| **Model R²** | | **(Error Term) ε** | | **(Random Effect) μ** | |
| --- | --- | --- | --- | --- | --- |
| **marginal** | **conditional** | **skewness** | **kurtosis** | **skewness** | **kurtosis** |
| 0.563 | 0.579 | 0.512 | 4.321 | 0.028 | 3.001 |

In table 6, our results show reasonably high marginal and conditional R-square values i.e. around 56-58% of the variation in the per capita consumption can be explained by the EB model. A quick glance at the random effect normality estimates (skewness = 0.028, kurtosis = 3.001) shows kurtosis and skewness appear to be approximately 3 and 0 respectively. However, the idiosyncratic error term may tend towards a leptokurtic distribution due to the heavier tails and the sharper peak as seen with the kurtosis values around 4. We formally test the normality of both the idiosyncratic and random effect errors using the Kolmogrov-Smirnov test. Testing the normality of residuals is crucial to validate the assumptions underlying the EBP model and ensure the reliability of statistical inferences drawn from our analysis.

The Kolmogorov-Smirnov (KS) test is a non-parametric statistical test used to measure the closeness of a distribution to any distribution of choice, in our case a normal distribution. It quantifies the maximum distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. The KS test is particularly useful because it makes no assumptions about the distribution of the data, providing a robust method to assess the goodness-of-fit for continuous variables. In this study, we employ the KS test to evaluate the normality of both residuals from our unit-level Empirical Best Prediction (EBP) model. The Kolmogorov-Smirnov (KS) test statistic is given by:

where: - is the Kolmogorov-Smirnov test statistic. - denotes the supremum (the maximum value). - is the empirical distribution function (EDF) of the sample. - is the cumulative distribution function (CDF) of the reference distribution.

The test statistic measures the maximum absolute difference between the empirical distribution function of the observed data and the cumulative distribution function of the theoretical distribution. A large value of indicates a significant difference between the observed and expected distributions, leading to the rejection of the null hypothesis that the sample comes from the specified distribution.

Figure 2: Kolmogrov-Smirnov Normality Test (Random Effect and Standard Error residuals)

A comparison of normalized results

Description automatically generated

The MSE is an absolute measure that combines both the variance and the square bias of the estimator. It is not easily comparable across constituencies with different poverty rates especially when the rates vary significantly. Therefore, we present the coefficients of variation for each constituency. This is useful for a couple of reasons:

* 1. A measure of spread presented as the ratio of variance (i.e. CV) to the poverty rate is dimensionless. This allows for easy within and across group comparisons regardless of the poverty rate.
  2. This provides policy makers and non-technical stakeholders are more intuitive measure of dispersion with which they may set standards. A higher CV always indicates less precision and a national statistical office, as a rule of thumb, may decide to merge areas for which CVs are greater than 0.3, for instance.
  3. For the above reasons, the CV has been widely adopted in the context of SAE to assess the reliability of estimates as it helps in identifying areas were estimates are less trustworthy.

We compute CVs for both the direct head count rates in the survey and for the poverty rates estimated via EBP in the census. The direct poverty rates from the household survey at the target area level are expected to be highly imprecise due to low sample sizes. There are 3 direct head count rate CVs presented using 3 methodologies. Method 1 uses a calibrated (or naive) bootstrapping to estimate the mean squared error (MSE). The bootstrapping method calibrates each resampled dataset on auxiliary information using the direct function from the sae package. Calibrated bootstrapping has been shown to improve bias correction compared to naive bootstrapping, especially in the context of complex survey designs. For more details, refer to (Rao and Wu 1988). The second type employs the Horvitz-Thompson variance estimation technique to calculate MSE. Here, each household is assigned a probability of selection based on the sampling scheme. The sae::direct function is used for this computation. This method accounts for the sampling design and ensures accurate estimation of variance in survey contexts. Method 3 adjusts the naive calibrated MSE using the design effect. The design effect is calculated with the svydesign function from the survey package, which helps to account for the increased variability introduced by the complex survey design. This adjustment provides a more realistic measure of variability for the MSE. For the EBP head count CV, we simply divide the mean square estimate with the poverty rate.

Figure 3: Comparison of EBP Head Count CV and Survey Design Effect CV

A graph with blue dots and lines

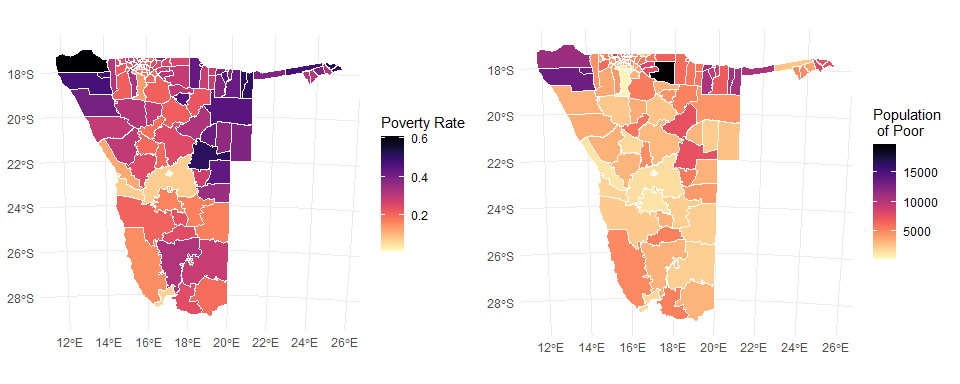
Description automatically generated with medium confidence

Figure 3 shows the precision gains achieved of EBP estimation over reporting direct poverty rates. On average, EBP headcount poverty rate CVs are 3.83 times less than the design effect CVs.

## **0.6 The Poverty Map**

The figure 4 shows the estimated head count poverty rates and the population of poor. Poverty appears to be concentrated towards to the north-east and north west corners of the country. On the the other hand, the southern most parts of the country appear to be least poor both in terms of headcount rates and population of poor.

Figure 4: Namibia Poverty Map: Headcount Rate & Population of Poor



We validate the predicted poverty rates by aggregating the poverty rates to the regional level and comparing the results to the direct estimates at the same level. See results in table 7.

Table 7: Regional Poverty Comparisons: Direct Survey vs Census EB estimates

| **Region** | **Direct Survey Estimates** | **EBP Model Estimates** |
| --- | --- | --- |
| Zambezi | 0.432 | 0.375 |
| Erongo | 0.047 | 0.105 |
| Hardap | 0.143 | 0.186 |
| Karaso | 0.112 | 0.190 |
| Kavango | 0.451 | 0.408 |
| Khomas | 0.084 | 0.080 |
| Kunene | 0.473 | 0.428 |
| Ohangwena | 0.205 | 0.240 |
| Omaheke | 0.398 | 0.346 |
| Omusati | 0.299 | 0.264 |
| Oshana | 0.159 | 0.175 |
| Oshikoto | 0.212 | 0.237 |
| Otjozondjupa | 0.247 | 0.244 |

References

Akaike, Hirotugu. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control* 19 (6): 716–23.

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.

Edochie, Ifeanyi, David Newhouse, Nikos Tzavidis, Timo Schmid, Elizabeth Foster, Angela Luna Hernandez, Aissatou Ouedraogo, Aly Sanoh, and Aboudrahyme Savadogo. 2024. “Small Area Estimation of Poverty in Four West African Countries by Integrating Survey and Geospatial Data.” *Journal of Official Statistics*.

González-Manteiga, Wenceslao, Marı́a José Lombardı́a, Isabel Molina, Domingo Morales, and Laureano Santamarı́a. 2007. “Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters Under a Logistic Mixed Model.” *Computational Statistics & Data Analysis* 51 (5): 2720–33.

Jiang, Jiming, and Partha Lahiri. 2006. “Mixed Model Prediction and Small Area Estimation.” *Test* 15: 1–96.

Kreutzmann, Ann-Kristin, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, and Nikos Tzavidis. 2019. “The r Package Emdi for Estimating and Mapping Regionally Disaggregated Indicators.” *Journal of Statistical Software* 91.

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>.

Pinheiro, J, D Bates, S DebRoy, D Sarkar, R Core Team, et al. 2018. “Nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-131.1.” *Computer Software). Retrieved from Https://CRAN. R-Proje Ct. Org/Packa Ge= Nlme*.

Rao, J. N. K., and C. F. J. Wu. 1988. “Bootstrap Inference for Sample Surveys.” *Statistical Science* 3 (3): 280–92.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88.

Tzavidis, Nikos, Li-Chun Zhang, Angela Luna, Timo Schmid, and Natalia Rojas-Perilla. 2018. “From Start to Finish: A Framework for the Production of Small Area Official Statistics.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 181 (4): 927–79.