**Small areas estimation in conflict areas**

**Interest**

Bayesian statistical methods assume that additional information makes better estimations of the probabilities. Robust estimates at area level from surveys require big sample sizes. However, households’ surveys are often nationally representative or at least regionally representative. Most of samples at the commune level are small. Besides, depending on the country context, data are missing in some commune level. Based on the EHCVM WAEMU household standard surveys in Burkina, Mali and Niger, x%, y%, and z% of the communes are missing, respectively. This leads to a key issue of estimation in this area, added to the small sizes in some communes.

Since 1970, Fay-Herriot introduced the small are estimation approach which use additional information to improve estimates in areas with small or no sample sizes. Over the last years, census have been widely used to apply small area estimation methods to improve estimates for small areas (regions, communes…) where microdata are not available, inaccessible or only for a small size. Now, remote sensing offer alternatives to collect socioeconomic, geographic and conflicts data at a very fine level of disaggregation.

**The estimation issue**

Sample surveys are often used to produce estimation about socioeconomic characteristics of population of interest at the area representative level required. When direct estimates at a less disaggregated level (small areas) are needed, the quality of the estimates must be improved. To do so, the approach of smalls area estimation which aim to propose sufficient precise direct estimation in these areas. When some small areas are missing due to a selection problem (access to the area due to conflict or to flooding, excessive cost to access…), direct estimates at the aggregate level are biased due to selection bias. The level of bias depends on the extent of the selection. However, when the sample sizes are small, direct estimates are biased by the non-convergence of the estimates and higher variability. Besides, National Statistical Offices and other organizations require reasonable coefficient of variation to publish indicators. When the estimates at a disaggregated level don’t achieve the consistency and robustness required, improvement is suitable. It is worth noting that the term “small areas” is used to illustrate areas whose samples sizes are not large enough to lead to robust direct estimates.

**The census issue**

Census data access is often reserved for a third party. However, nowadays, data at area level are often available thanks to the open access of geospatial data, and socioeconomic data (infrastructures, health, conflict…). Out of the access issue, census data are not available each year. Most of the time, the sample year does not match with the census year. This leads to consider the off-census year for the needs of the Small Area Estimation (SAE). This leads to the use of outdated information.

**The census issue**

* **Census are expensive and not accessible,**
* **Census are not available each year, and administrative areas changed over the years**
* **beneficial to supplement the available data by using geospatial information about the small areas.**

**Literature review**

**Data description**

EHCVM household survey from 2020 to 2021 in Burkina faso

* Survey

The Harmonized Survey on Household Living Conditions (EHCVM) is a joint initiative of the UEMOA Commission and the World Bank. It is conducted as part of the Program for the Harmonization and Modernization of Household Living Conditions Surveys within the UEMOA member states (PHMECV). The Program's objective is to support the UEMOA Commission in strengthening the capacities of its member states to conduct household living conditions surveys that meet harmonized regional standards, and to make the collected microdata accessible to the public. The specific objectives include, among others, producing indicators on poverty and living conditions, and creating harmonized databases that allow for analyses to inform public policies. The beneficiaries are the eight UEMOA countries: Benin, Burkina Faso, Côte d’Ivoire, Guinea-Bissau, Mali, Niger, Senegal, and Togo. Apart from the PHMECV beneficiary countries, Cameroon, Chad, Guinea, and Congo have also conducted a similar survey using the same methodologies and have participated in various technical works. The first edition of the survey took place in 2018/2019, and the second, which is the subject of this document, in 2021/2022. The survey includes four components: a household component, a community component, a price component, and a component on non-standard units (NSU).

* ACLED data

|  |  |
| --- | --- |
| **Variable Name in the Database** | **Definition** |
| CDI | Conflict diffusion Indicator |
| CDI\_\* |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| CEI | Conflict Exposed Indicator (5 km) |
| CEI\_\* |  |
|  |  |
|  |  |

* Geospatial data

Landsat 9-C2-SR 2022 : 70,238 observations for settlements that cover all the 17 prefectures, 72 subprefectures and 171 communes.

Variable Index for the Landsat 9-C2-SR 2022 georeferenced

|  |  |
| --- | --- |
| **Variable Name in the Database** | **Definition** |
| MNDWI | Modified Normalized Difference Water Index |
| BRBA | Band Ratio for Built-up Area |
| NBAI | Normalized Built-up Area Index |
| NDSI | Normalized Difference Snow Index |
| VARI | Visible Atmospherically Resistant Index |
| SAVI | Soil Adjusted Vegetation Index |
| OSAVI | Optimized Soil Adjusted Vegetation Index |
| NDMI | Normalized Difference Moisture Index |
| EVI | Enhanced Vegetation Index |
| NDVI | Normalized Difference Vegetation Index |
| NDBI | Normalized Difference Built-up Index |
| SR | Simple Ratio |
| ARVI | Atmospherically Resistant Vegetation Index |
| UI | Urban Index |

Reference: Landsat 9 C2 SR - 2022 (median composite)

|  |  |  |
| --- | --- | --- |
| **Description** | **Spatial Resolution** | **Temporal Resolution** |
| Travel Time to Nearest Healthcare Facility with Access to Motorized Transport | 1km | 2019 |
| Walking-only Travel Time to Nearest Healthcare Facility without Access to Motorized Transport | 1km | 2019 |
| Travel Time to Cities | 1km | 2015 |

|  |  |
| --- | --- |
| **Description** | **Spatial Resolution** |
| Proportion of Children 2 to 10 years of age showing, on a given year, detectable Plasmodium falciparum parasite | 5km |
| Number of deaths from Plasmodium falciparum per 100,000 population during a defined year 2000-2022 | 5km |

* Nighttime data

Monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB).

Compute a linear fit over the series of values at each pixel, keeping the y-intercept (offset) and slopes(scale).

|  |  |
| --- | --- |
| **Indicator** | **Description** |
| offset | the y-intercept (offset) of monthly average radiance composite images (2015 - 2023) |
| scale | the slopes (scale) of monthly average radiance composite images (2015 - 2023) |

**Methodology**

* **Traditional approaches**

**FAY-HERRIOT MODEL**

**Where**

**Linking model**

**Sampling model**

**Best Linear Unbiased Predictor**

Weighted combination of direct estimator d and regression synthetic estimator .

It gives more weight to when sampling variance small ( reliable).

It gives more weight to the synthetic estimator when large large ( unreliable) or small ( reliable).

The BLUP is a theorical estimator which depends on  **, and .**

Approaches such as the maximum likelihood, the restricted maximum likelihood (REML), and the Prasad-Rao (1990) method of moments have been used to estimate the and allow to estimate the empirical best linear unbiased predictor (EBLUP).

When the target indicator is not necessarily linear, the best predictor

* Machine learning approach

**RESULTS AND DISCUSSION**

* **RESULTS**
* Selection bias test

A screenshot of a math program

Description automatically generated

|  |  |
| --- | --- |
|  | (1) |
| Table x: xxxxxxxxxx | dummy |
| dummy |  |
| acled\_ei\_erv | 4.738\*\*\* |
|  | (4.29) |
|  |  |
| acled\_ei\_protests | 2.116\* |
|  | (1.98) |
|  |  |
| acled\_ei\_battles | -1.325\*\* |
|  | (-2.69) |
|  |  |
| acled\_ei\_vac | -29997.8\*\*\* |
|  | (-3.44) |
|  |  |
| \_cons | 0.372\*\*\* |
|  | (4.35) |
| *N* | 351 |

*t* statistics in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

* Fay – Herriot model results
* Machine learning model results
* **DISCUSSION**

In many practical situations, however, the aim is to estimate parameters for domains that contain only a small number of sample observations.

The term “small areas” is used to describe domains whose sample sizes are not large enough to allow sufficiently precise direct estimation. When direct estimation is not possible, one has to rely on alternative, model-based methods for producing small area estimates. Such methods depend on model specification as well as on the availability of population level auxiliary information related to the variable of interest, and are commonly referred to as indirect methods (Rao, 2003).

The underlying theory is referred to as the small area estimation (SAE), and SAE techniques aim at producing reliable estimates based on such small sample sizes by using the model “linking” the small areas to “borrow strength” from the sample data from other small areas, see for example, Pfeffermann (2002) and Rao and Molina ( 2015).

The term “small areas” is used to describe domains whose sample sizes are not large enough to allow sufficiently precise direct estimation.

The Fay-Herriot model (Fay and Herriot, 1979), is a widely used area level model that assumes area-specific survey estimates are available, and that these follow an area level linear mixed model with independent area random effects.

Unlike the Fay-Herriot model, this approach implicitly assumes simple random sampling with replacement within each area and ignores the survey weights.

One approach to incorporating such spatial information in SAE modelling is to extend the random effects model to allow for spatially correlated area effects using, for example, a Simultaneous Autoregressive (SAR) model.

Poverty rate is a key indicator of economic development and human welfare.