EUPM Practitioners Guide

2025-02-12

Table of contents

Introduction

TODO

1 Small Area Estimations for the poverty mapping: an overview

As discussed with Nobuo, Danielle and Eduard, place for an practitioners overview chapter.

Some overview ideas:

- Make a decision tree of methodology given data availability.
- Make a table to take a stock of methods and Corresponding R funciton/packages that implement it. For example, see **?@tbl-methods**

Methods	`sae`	`emdi`	`SUMMER`
Spatial Fay-Herriot FH Miltivariate Autocorrelation	`sae::mseFH()`		`SUMMER::smoothArea()`

Place figures under `images` and use them in the text as follows making sure to refer the sources correctly. For example, Figure 1.1 is adapted from (Corral et al. 2022, 5).

One can also embed mathematical formulas following the latex syntax: $y = a + b \log x$. For more information, see quarto help on authoring.

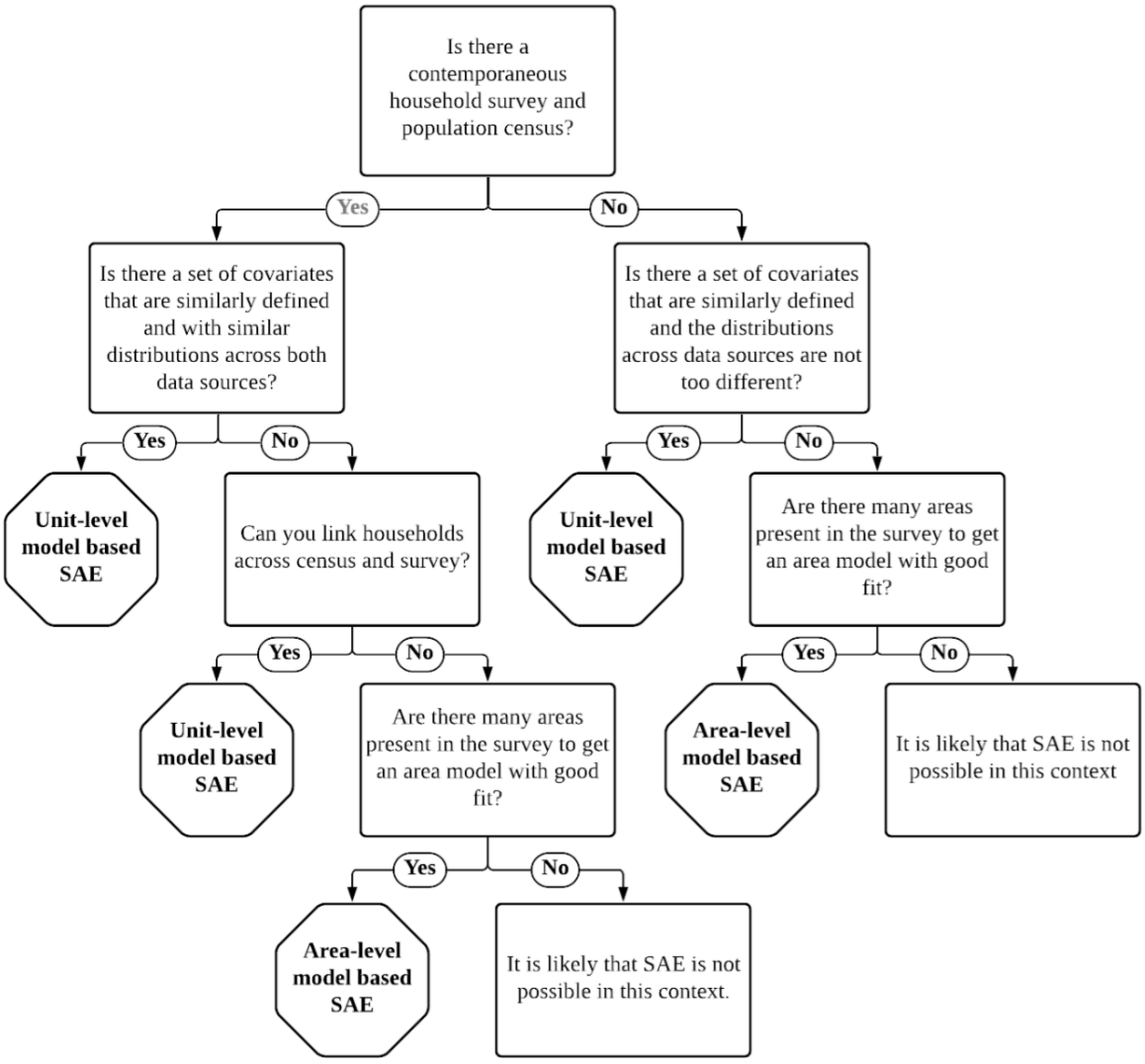

Figure 1.1: SAE decision tree

2 Data preparation

Introduction: this chapter provides a template for organizing the data flow in the EUPM analysis.

2.1 Workflow and reproducibility principals

- Brief overview of the data workflow between raw, auxiliary and clean data types.
- Basic principles of reproducibility.

2.2 Geospatial data

1. Description of the basics of the GIS data preparation for countries at different admin levels. Key problems addressed.
2. Spatial validity of polygons.
3. Nested geospatial structure and non-intersecting boundaries.
4. Polygon -unique identifiers.
5. GIS boundaries harmonization over time.
    - Synthetic regions aggregation constant in time
6. Quality assurance of the administrative boundaries.

2.3 Raw data search, collection, and documentation

Description of the process of data search, collection and documentation that yields with a systematized, but unstructured data library

1. Variables search, and priority indicators.
2. Key challenges and considerations for data inclusion:

- Thematic relevance
- Time range available
- Territorial unit available
- Coverage and completeness

3. Details on specific data sources:
    - API-based data
    - Manually downloaded spreadsheets
    - Bulk downloads
    - Remote sensing and GIS-based data, zonal statistcs, etc.
4. Principals of data storage and systematization
5. Documenting collected data with metadata and it notes on search
    - Key validation requirements – source type, survey type

Country-based examples: use one country as an example.

2.4 Preparing auxiliary data

Adding structure to the raw data by transforming it into a normalized data set with columns: id, year, variable, value.

- creating and storing the auxiliary data
- data reproduction and version-control
- principals of the data quality assurance and quality control

Country-based examples: use one country as an example.

2.5 Clean and analysis-ready data

Getting meaningful and relevant indicators out of the data.

- Reshaping auxiliary data into the analysis-ready dataset.
- Computing relevant indicators: means, ratios, fractions, etc.
    - Group-wise operations by year, and across regions.
- Regression-data quality assurance: spatial and temporal completeness
- Adding data important data from elsewhere: SILK poverty estimates

2.6 Descriptive analysis

Key principals and examples of descriptive statistics.

- Examples of existing R functional for this.

3 Fay–Herriot (TODO)

Present examples of the FH model chronologically with additional explanations based on the Pauls' materials for the summer school.

4 Unit-Level models (TODO)

Reproducible guide to the practitioners on implementing the unit-level SAE models.

- Place all data necessary under `data` (we will consider later if it worth moving data and potential functions into a separate R Package.

4.1 Outline for the Unit Level EBP estimation in R

4.1.1 Introduction

unit level EBP estimation the whole game in a snapshot, we will also make it clear that this document will not provide a methodological/statistical primer in unit level poverty mapping but be focused on showing how to do this in R. For statistical details, please see Corral et al., 2022)

4.1.2 The Data

- here we introduce the data that will be used for the process (fake survey and census data from Estonia). End by mentioning that the full data creation process for the fake dataset can be found in the data-raw folder

4.1.3 Data Preparation for unit level model

- creating the variables that will correlate will household level welfare

4.1.4 Variable Selection

- Checking that each variable has similar distribution between the survey and census and dropping variables that do not meet (a function has been written to do this test better than the ebp_test_means() function in povmap)
- Dropping multicollinear variables (using the VIF method and complementing with correlation threshold method)

- Implementing variable selection under different welfare transformations (use wrapper functions that I have written for the variable selection using glmmLasso and GLMNET R packages)
- Cross-Fold Validating the variable selection process i.e. a plot to show how MSE for each lambda of glmnet is performed. (May also show how to do this with glmmLasso)

4.1.5 EBP Unit Level Model Estimation

- Start with a few notes on the pre-reqs needed to use the ebp() function in EMDI/povmap R packages i.e.
  - all target areas (domain argument) in the survey must be in the census
  - domain argument must be integer class
  - remove all missing observations in survey and census
- Implementation of the ebp() function call
- Detailed description of the ebp class object which is returned

4.1.6 Post Estimation Diagnostics

- Presenting the regression table estimates (use povmap::ebp_reportcoef_table() and then translate into a flextable which can be rendered in Word, PDF or HTML)
- Checking that all model assumptions hold (normality assumptions for the miu and epsilon terms), using povmap::ebp_normalityfit() to present skewness and kurtosis for both errors. Then show a function that uses ebp object to plot the distribution of the errors and compute the kolmogrov-smirnov test statistics. We can also include the shapiro-wilks which will break down for larger sample sizes but is equally well known. Another function that produces q-q plots for miu and epsilon terms from ebp object.

- Check on model validity: Create a plot to show how poverty rates vary at each ventile i.e. at 5% poverty line increments. This is to show how to check the quality of model prediction without the added bias of out of sample prediction
- Computing MSE and CVs and computing the statistical gains made from performing small area estimation i.e. in practical terms, how much bigger would the survey have to be the get the same level of precision that SAE now gives us with the census data.
- Final validation: EBP estimates vs Direct Estimates (supposedly truth) at the highest resolution level that is considered nationally representative, this is usually the regional level in Africa.
- Plotting the poverty map using the ebp object and the shapefile

References

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