Project title

Strengthening the capacity of Jordan's Department of Statistics in terms of compilation, analysis and reporting of statistical data in line with International and European best practices

Twinning Reference Number: JO 21 ENI ST 01 22

Component 2. Small Area Statistics (SAS)

Michele D'Alò

Objectives component 2

Objective of the project:

The goal of this component is the development of SAE methods to be applied to some SDG indicators on poverty, and their possible extention to other statistical areas within Jordan DoS.

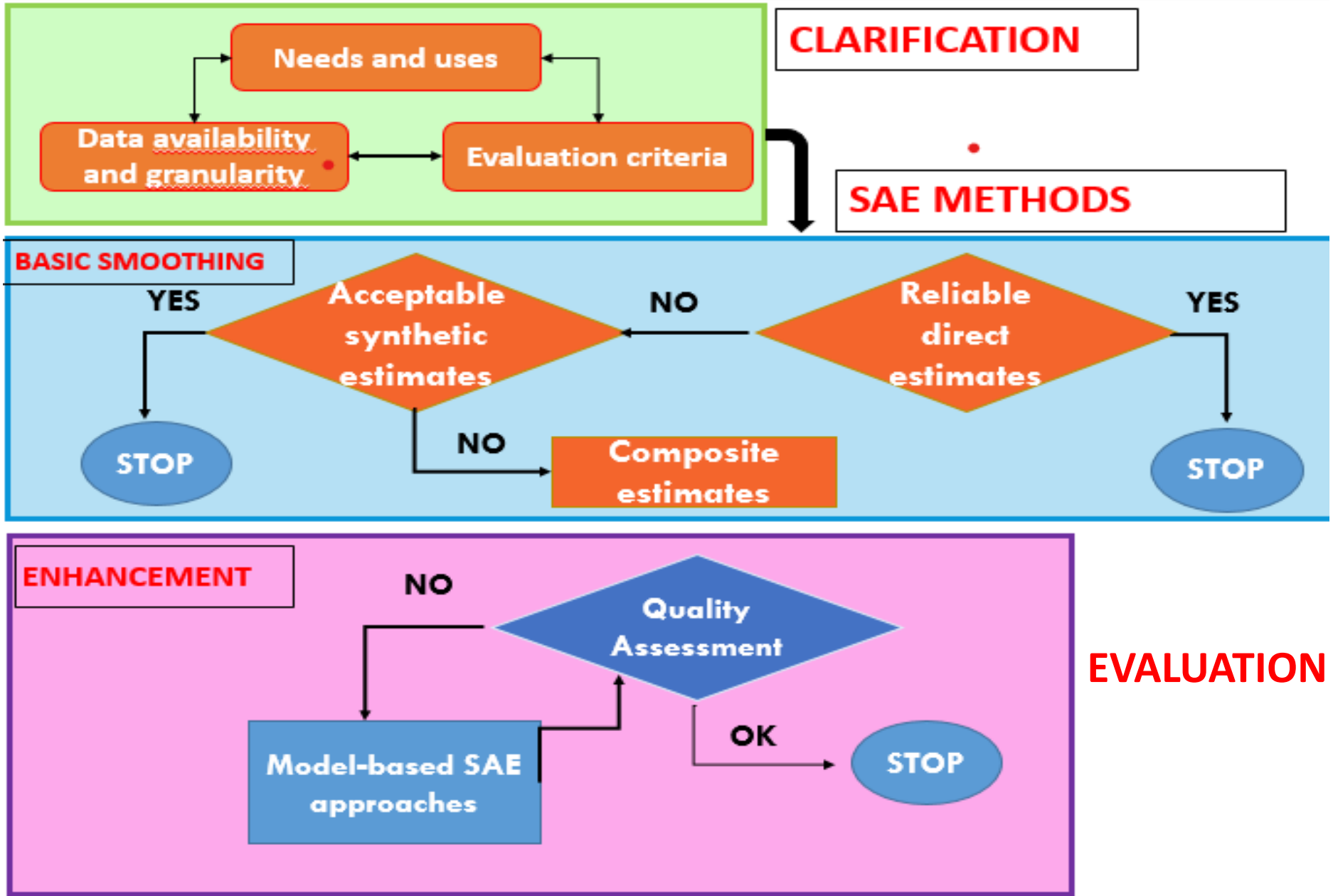
Objective of this mission:

Drafting together the first rolling work plan for activities on small area statistics for the coming two years.

Objective of this presentation

Try to highlight important points to consider for the development of small area statistics plan with reference to relevant EU and international best practices.

Steps to compute SAS



Target parameters

When developing an SAE implementation plan it is important to carefully examine the characteristics of the variables of interest for which disaggregated estimates need to be produced

Indicators can have different functional forms that needs to be considered when choosing the appropriate SAE approach. **Examples:**

Type	Description	Example
Total	The sum of values	Number of people with ownership or secure rights over agricultural land
Mean/Average	The sum of values divided by the number of values	Average agriculture income, average labour productivity, average crop yield
Proportion	Fraction of the population with specific characteristics	Proportion of individuals experiencing moderate or severe food insecurity
Rate	Ratio between two quantities in different units	Mortality rate attributed to unsafe water

"The definition of the target parameter to be estimated is of vital importance. The target parameter needs to be well-supported by the available data. An increasing complexity of the indicators of interest simultaneously increases the granularity of the data that are needed for the estimation "

Data disaggregation and the SDGs

The key dimensions for disaggregation include:

- Characteristics of the individual or household such as sex, ages, income, disability, religion, ethnicity, Economic activity and so on
- Spatial dimensions such as metropolitan areas, urban and rural, or districts.

Many of the SDG indicators, like labor, health, poverty and other dimensions of wellbeing, can all be derived from household surveys.

➢**Sample surveys:** designed to provide broad-area estimates with high level of accuracy and precision. E.g. survey sample sizes are typically calculated such that estimates of the parameter of interest are reliable enough at national or regional level, or at certain highly aggregated level.

➢In **planned domains**, the estimation of indicators is based on **direct estimators**, i.e. estimators solely based on survey data from the considered domain.

- The direct estimator cannot be obtained for some areas due to missing survey information on non-sampled areas.

Assessment of data availability

- Surveys
- Censuses

TRADITIONAL

- Administrative records
- Big data

"INNOVATIVE"

- Surveys are generally carried out to estimate various indicators, but the sample size allow to estimates parameters only at an high level of aggregation;
- Census do not have any sample size problem, but the information are collected less frequently and on more limited number of item respect to surveys;
- Administrative records (registers) are innovative sources already in use to produce official statistical information;
- Big data are still experimental , and still do not in use to produce official statistics.

Step 2. Specification of SAE methods

1: Basic smoothing (ESSnetSAE (2012) proposes to start with a triplet of estimates:

- Production of direct estimates and their variances
- Production of other indirect estimators (synthetic and composite) from a design-based perspective
- Assessment: CV is below the required threshold. For instance Statistics Canada uses three categories of reliability for the Labour Force Survey: no release restriction for a CV ≤ 16.5% , added warning when 16.5% < CV≤ 33.3% and otherwise, the data is not recommended for release

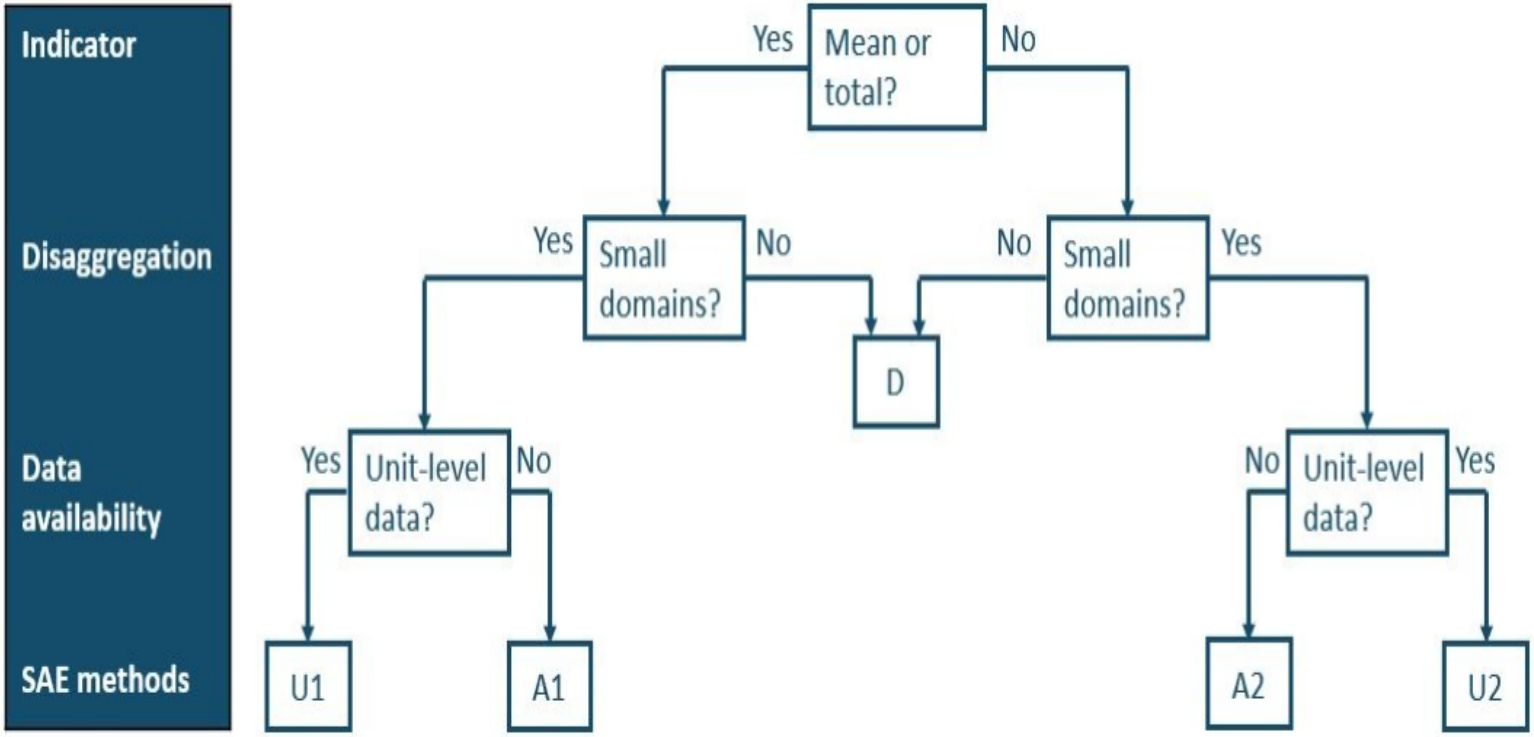
2: Enhancement through model-based (SAE) approaches The model specification will strong depends on the

- The characteristics of the indicator;
- The disaggregation of the small domains of interest and thus the corresponding sample sizes
- The data availability determines which SAE methods could actually be applied.

Importantly, some SAE methods are applicable to one form of variable, but not to others. For other type of indicators different SAE method might be necessary.

2. SAE methods

Initial specification depending on input factors



D: Direct estimation approaches.

A1: The basic area-level model and its extensions for means and totals including, e.g., the spatial-correlation and robust models.

A2: Area-level models for other indicators such as ratios. These can either use transformations or a non-linear model specification.

U1: The basic unit-level model and its extensions for means and totals, including robust models.

U2: Extended unit-level models such as the ELL and the EBP approaches.

3. Quality assessment of model based estimates

- ✓ Model building, model diagnostics, sensitivity analysis and validation play a central role in model-based SAE.

- ✓ **This phase includes the proper model specification; the checking of the model assumptions and how to deal with departures from these assumptions**

- ✓ **After producing small area estimates using model based methods, their properties must be evaluated.**

Reliability of small area estimates (Mean Square Error)

- **Precision** measures the closeness of the estimates to each other (variance or standard error)
- **Accuracy** measures the closeness of the estimates to the true value of the parameter being estimated (bias).

Evaluation of set small area estimates

- Simulation studies can be conducted to evaluate the methods used to produce small area estimates;
- Comparison with direct estimates and benchmarking properties can be evaluated;
- Thematic analysis by users/experts are also important.

SAE software

Overview of R packages for small area estimation

Package	Model type		Vignettes/Paper
	Unit	Area	
sae	×	×	• Basic direct and indirect estimators in sae • R package sae: Methodology • sae: An R Package for Small Area Estimation
emdi	×	×	• The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators • A Framework for Producing Small Area Estimates Based on Area-Level Models in R
rsae	×	×	• Robust Small Area Estimation: a Vignette
hbsae	×	×	
JoSAE	×	×	• Vignette of the JoSAE package
BayesSAE		×	
saery		×	
mme		×	• mme: a package for small area estimation with multinomial mixed models
smallarea		×	• An overview of Fay Herriot model with our package smallarea
saeRobust		×	• Notes on the fixed point framework
msae		×	
MIND	×		

Overview of Stata packages for small area estimation

Package	Model type		Vignettes/Paper
	Unit	Area	
sae	X		• sae: A Stata Package for Unit Level Small Area Estimation
FHSAE		X	
fayherriot		X	• The fayherriot command for estimating small-area estimators • A practical guide for the computation of domain-level estimates with the Socio-Economic Panel (and other household surveys)

Overview of further packages for small area estimation

Package	Model type		Vignettes/Paper
	Unit	Area	
PovMap	X		• Using PovMap2: A user's guide

Overview of SAS packages for small area estimation

Package	Model type		Vignettes/Paper
	Unit	Area	
StatCan prototype	X	X	• Development of a small area estimation system at Statistics Canada
Mukhopadhyay and McDowell	X	X	• Small Area Estimation for Survey Data Analysis Using SAS® Software

Overview of Python packages for small area estimation

Package	Model type		Vignettes/Paper
	Unit	Area	
samplics	X	X	• Area level model: empirical best linear unbiased predictor (EBLUP) • Unit level model: empirical best linear unbiased Prediction (EBLUP)

4. Communicating SAE methods and results

Fruitful communication and dissemination of small area estimation results requires: knowledge about the target audience (e.g. policy makers, researchers, users in statistical offices);

The important aspects that should be communicated with users includes:

- Methods applied;
- Input data and models used;
- Assumptions and validation of followed process;
- Quality assessment of the estimates, including a possible comparison of the results over the time;
- Indication on how to interpret the results and ways to (or not to) use the results.

Some International SAE application on SDG's (GOAL 1)

Goal 1. End poverty in all its forms everywhere

Case studies

Poverty mapping is one of most common applications in small area estimation. Many examples are available for the proportion of population living below the international or national poverty line (indicators 1.1.1 and 1.2.1).

World Bank applications

The World Bank proposed a poverty mapping process that was conducted in several countries. Based on surveys and additional data sources, various poverty and inequality estimates such as the Foster-Greer-Thorbecke poverty estimates and the Gini coefficient were derived.

The report More than a pretty picture - Using poverty maps to design better policies and interventions published in 2007 shows case studies for the countries **Albania, Bolivia, Bulgaria, Cambodia, Yunnan Province (China), Ecuador, Indonesia, Mexico, Morocco, Sri Lanka, Thailand** and **Vietnam** that describe all poverty mapping steps and also lessons learned. Hence this can be a good starting point for a new poverty mapping study.

In 2005, the World Bank provided technical assistance to the **Philippine** national statistical system to leverage on small area estimation techniques to produce municipality- and city-level poverty statistics. The Philippine Statistics Authority conducts the Family Income and Expenditure Survey (FIES), which is the main source of official poverty statistics in the country, every three years. The small area estimation technique used in the Philippines is based on the ELL method. It entails regressing (log) per capita household income from the FIES with auxiliary information from the FIES, the Labor Force Survey, and the Census of Population and Housing. The model regressors include survey-obtainable variables such as educational attainment of the household head and other household characteristics, and census-derivable information like average family size in a village, and other village-level information. Since small area poverty statistics became available in 2005, numerous government agencies have used these data as inputs for formulating and implementing poverty reduction programs. For example, the Philippine Department of Social Welfare and Development (DSWD) used the estimates to identify poor municipalities for its National Household Targeting System for Poverty Reduction (NHTS-PR) data collection.

Indicators	Disaggregation dimension	Data availability	Estimation approach	Model
1.1.1/1.2.1	Spatial	Unit-level survey and auxiliary data	Model-based estimation	ELL

Poverty estimation in Chilean comunas

To improve fund allocations among comunas, the Chilean Ministerio de Desarrollo Social (in the following the ministry) is required to provide poverty estimates for all 345 comunas in Chile which is the smallest territorial entity. After the evaluation of various options, the ministry decided to combine the National Socioeconomic Characterization Survey (CASEN), which is Chile's official data source for poverty statistics, with relevant administrative records. Since 2011, model-based poverty statistics are obtained for Chilean comunas.

Indicators	Disaggregation dimension	Data availability	Estimation approach	Model
1.2.1	Spatial	Unit-level survey and area-level auxiliary data	Model-based estimation	Arcsin-transformed area-level model

Small Area Income and Poverty Estimates (SAIPE) program by the U.S. Census Bureau

The SAIPE program produces small area estimates of income and poverty statistics for all school districts, counties, and states. The estimates are based on several data sources such as the American Community Survey and Federal Income Tax Returns. The produced indicators do not exactly follow the definition of the SDGs but the example is added since the SAIPE program is continuously improving their approach and the disaggregation is not only spatial but also by age groups.

Indicators	Disaggregation dimension	Data availability	Estimation approach	Model
	Spatial and age	Unit-level survey and area-level auxiliary data	Model-based estimation	Log-transformed area-level model

Mean income in Middle-layer Super Output Areas in England and Wales

Super Output Areas (SOAs) are a geographic hierarchy introduced for the reporting of small area estimates. The mean population of Middle-layer SOAs (MSOAs) ranges from a minimum of 5,000 to 7,200. To obtain different income estimates (e.g., equivalised and unequivalised) for the MSOAs, data from the Family Resources Data is combined with additional data including Census information, energy consumption and house price statistics using a linear mixed model.

The case study is explained on the homepage with more methodological details in the technical report.

Indicators	Disaggregation dimension	Data availability	Estimation approach	Model
	Spatial	Unit-level survey and area-level auxiliary data	Model-based estimation	Log-transformed unit-level model

ISTAT applications

- ✓ SAE to produce estimates of employment and unemployment rates for local labour market areas (LMA)

- ✓ Experimental SAE of a selection of labour market variables for cities and functional urban areas (FUA) based on a unit-level multivariate model. These are published as Italian sub-national statistics within the scope of an agreement between ISTAT and Eurostat.

The main indicators produced are:
Economically Active Population, total, and by sex
Economically Active Population aged 20-64, total, and by sex
Persons Unemployed, total, and by sex
Persons Employed aged 20-64, total, and by sex

Input data
The Italian Labour Force Survey (LFS) referring to year 2018 is the source of direct estimates whereas the Integrated System of Registers, in particular the Base Register of Individuals (RBI) and the Thematic Labour Registry (RTL), the Italian Ministry of Finance and the Revenue Agency provided auxiliary data (demographic, employment, social security and income information).

ISTAT case studies

Experimental studies to develop and propose SAE for SDG indicators

•Goal 1 - End poverty in all its forms everywhere

- Relative and Absolute Poverty indicators from EU-SILC and the Household Budget Survey (Laken indicators and others) for metropolitan cities and provinces.
- The challenge is how to choose the best method and the benchmarking approach.

•Goals 9 and 17 - Build resilient infrastructure, promote sustainable industrialization and foster innovation; Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development

- ICT indicators based on Multipurpose survey on everyday life aspect for metropolitan cities and provinces.
- Specific small area estimators can be applied like the design-based projection estimators, besides the classic SAE

- Goal 3 Ensure healthy lives and promote well-being for all at all ages
  - Health indicators based on European Health Interview Survey for Italian health districts/regions.
- Goal 5. Achieve gender equality and empower all women and girls
  - Direct and model-based estimates for indicators related to violence against women.

References

- Guidelines on small area estimation for city statistics and other functional geographies estimation.https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-19-011

- Guidelines on SAE ESSnet on SAE (2012) https://ec.europa.eu/eurostat/cros/system/files/WP6-Report.pdf

- Tzavidis et al. (2018), From start to finish: a framework for the production of small area official statistics J. R. Statist. Soc.A (2018) 181,Part4,pp.927–979

- UNSTATS wiki page/ The SAE Toolkit) https://urlsand.esvalabs.com/?u=https%3A%2F%2Funstats.un.org%2Fwiki%2Fdisplay%2FSAE4SDG%2FSAE4SDG&e=17c5563b&h=502428d8&f=n&p=y

THANK YOU FOR THE ATTENTION