Small area estimation for SDGs

Haoyi Chen
Coordinator, Inter-Secretariat Working Group
on Household Surveys

unstats.un.org/iswghs

Outline

- ❑Inter-Secretariat Working Group on Household Surveys (ISWGHS) & IAEG-SDGs
- ❑Why Small area estimation (SAE) and what is it?
- ❑Key elements of SAE
- ❑ISWGHS and IAEG-SDGs collaboration: Toolkit on using SAE for SDG indicators
- ❑Capacity building activities on SAE

ISWGHS

The ISWGHS: a primer

- Established in 2015 under the aegis of the UNSC
- Objectives:
  - Improve coordination of household surveys
  - Advance cross-cutting survey methodology
  - Enhance communication and advocacy
- Governance
  - Membership: 11 international agencies + 8 (rotating) member states
  - Secretariat: UN Statistics Division
  - Current co-chairs: WB and UNW
- Work through time-bound Task Forces, led by and with contribution from members and non-member experts.

ISWGHS

Inter-agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs)

The 2030 Agenda for Sustainable Development

- A global blueprint for people, planet, prosperity , peace and partnerships, now and in the future
- 17 Goals, 169 targets and “Leaving no one behind” principle

The IAEG-SDGs :

- Composed of 28 Member States (and representatives of regional commissions, regional and international agencies and CSOs are observers)
- Developed the global indicator framework for SDGs (**231 indicators)**

IAEG-SDGs workstream on data disaggregation:

- Compilation of existing guidelines and methodologies on data disaggregation
- Preparation of Handbook on data disaggregation for SDGs
- Task Force on Small Area Estimation (joint with ISWGHS)

ISWGHS

Small area estimation: why and what is it?

❑SDGs: Leave no-one behind

- Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics.

❑Limitations of household surveys

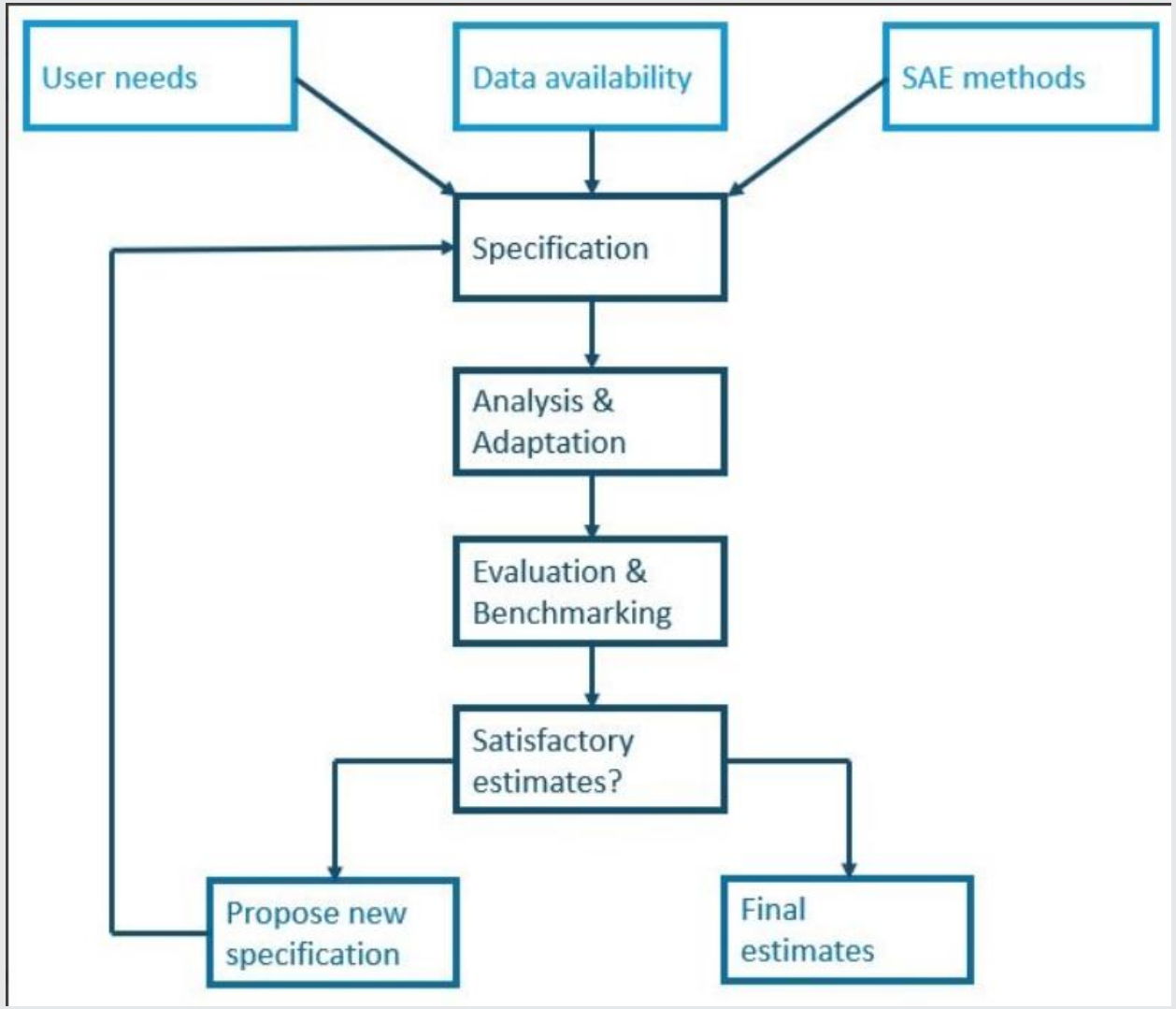
- Cross-classification (i.e., geographic, demographic) usually leads to small sample sizes even for very large surveys, hence acceptable accuracy cannot be obtained through direct estimates

❑Small area estimation

- Small area: domain where sample size is too small for reliable direct estimation
- Direct estimator: estimate based on sample data alone
- Small area estimation: Modeling, “borrow strength” from other data sources

ISWGHS

Framework for production of small area estimates



Source: Tzavidis et. al. (2018)

User needs

❑ Items to consider:

1. Data needed for key policies and funding decisions
   - *In 2009, the law of the Fondo Común Municipal (FCM) required the Ministry to provide poverty rate estimates every 2 years for all comunas in the country. Funding to all comunas will be allocated based on such data.* (Chile)
   - Improving America's Schools Act: "*the number of children aged 5 to 17, inclusive, from families below the poverty level on the basis of the most recent satisfactory data, ..., available from the Department of Commerce*" (US)

2. What types of indicators?

Type	Description	Example
Total	The sum of values.	3.3.5 Number of people requiring interventions against neglected tropical diseases.
Mean/Average	The sum of values divided by the number of values.	8.5.1 Average hourly earnings of employees.
Proportion	Fraction of population with specific characteristic.	1.1.1 Proportion of the population living below the international poverty line.
Rate	Ratio between two quantities in different units.	3.9.2 Mortality rate attributed to unsafe water.

3. Disaggregation dimensions
   - # categories

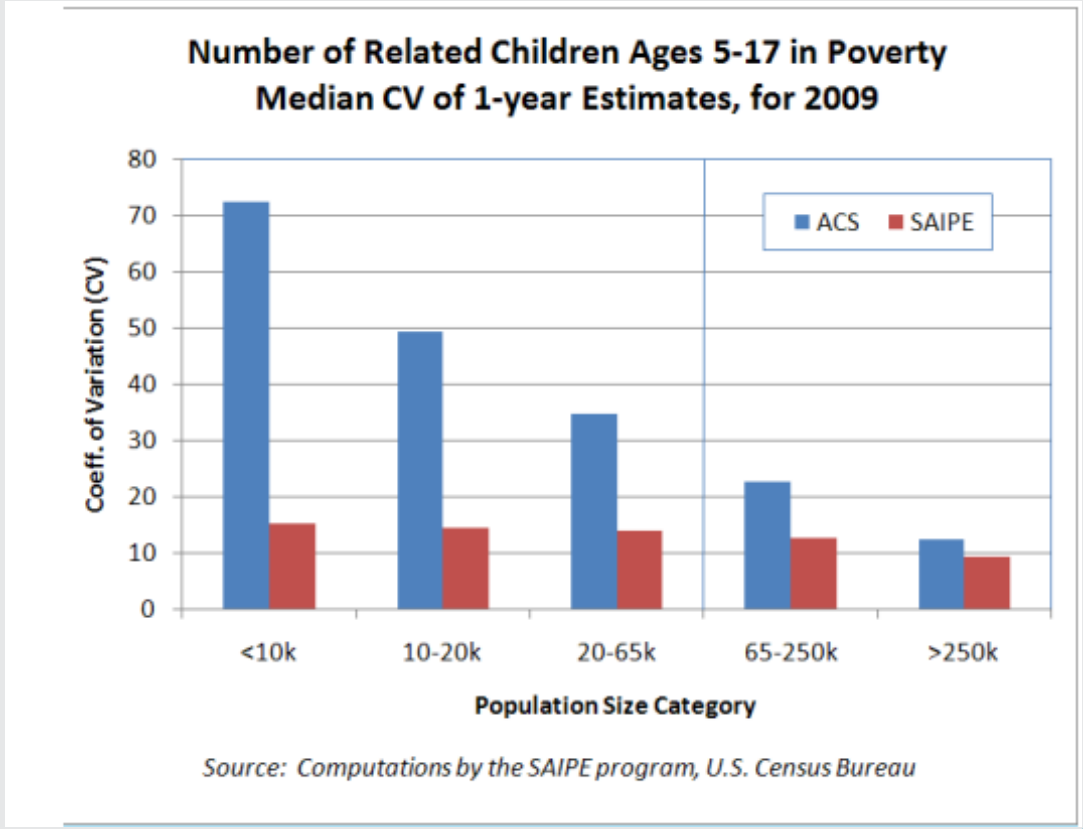
ISWGHS

SAIPE: poverty mapping in the US

Poverty statistics produced by SAIPE

- All people in poverty (state and county)
- Children under age 18 in poverty (state and county)
- **Related children aged 5-17 in poverty** (state, **county***, and school district)
- Children under age 5 in poverty (states only)
- Supports the Elementary and Secondary Education act of 1965 (reauthorized by the Every Student Succeeds Act of 2015)

United States Census Bureau



Source: Carolina Franco, 2021. ECLAC-UNSD Webinar on Poverty Mapping Using Small Area Estimation Techniques

ISWGHS

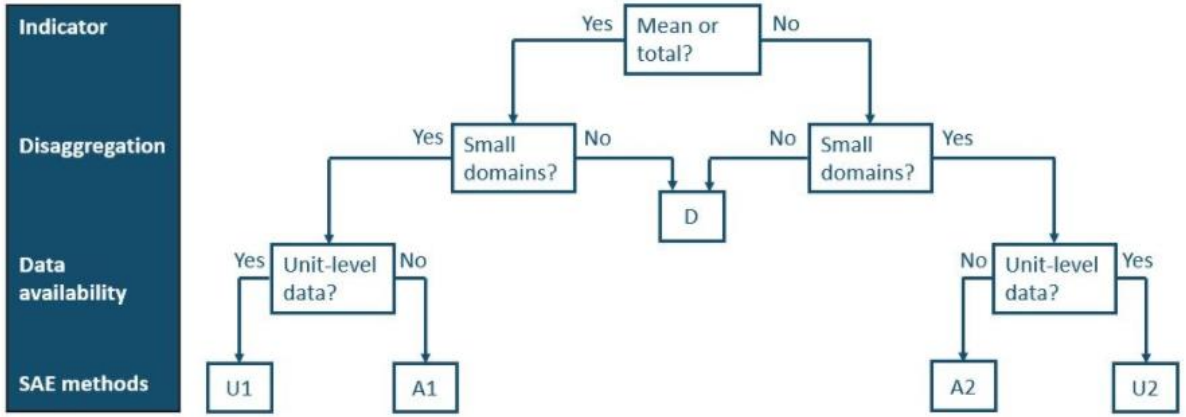
Data availability

- Possible sources:
  - Population censuses
  - Administrative sources
  - Surveys (same or different)
  - Other data: commercial data, satellite data, cellphone data etc…
- Requirement:
  - Data availability and accessibility
  - Quality: coverage, accuracy, timeliness, how often are data updated, availability of auxiliary variables; predicting power
  - Time frame

Name of the auxiliary variables	Institution responsible for data collection	Frequency of publication of the data

SWGHS

SAE specification



- Area level model (A1, A1)
- Unit level model (U1, U2)

8.5.2 Unemployment rate

R Code

User needs

Goal: Employment is often a key against hunger and extreme poverty. Thus, the identification of groups without employment could be of interest in order to counteract their unemployment with specialized programs.

Indicator of interest: The unemployment rate defined as the number of unemployed persons divided by the total number of persons in the working age population. The unemployment rate is a proportion describing the fraction of the labor force with the characteristic to be unemployed and has a value between 0 and 1. In the example, the working age is defined between 15 and 74.

Disaggregation dimension: The required disaggregation dimensions are sex, age, geographic location (urban/rural), and disability status. The example will consider the dimensions and show some limitations and challenges.

Data availability

Information about the employment status is available in the survey of the working population (syntheticSurvey2). The data further contains variables for the different disaggregation dimensions, The aggregated data sets contain variables that could help to explain the unemployment rate at different domain levels.

Load data sets › Expand source

First lines of data sets › Expand source

Specification

The indicator of interest is a proportion (**Mean or total? → No**). The goal is to try different disaggregation dimensions. For the disaggregation dimensions sex, age, geographic location (urban/rural), and disability status direct estimates are obtained assuming that the sample sizes are large enough (**Small domains? → No**). For a combination of the dimensions age group and geographic location, as well as sex and administrative region, area-level models are applied, assuming that the precision could be improved compared to the direct estimates (**Small domains? → Yes**) and considering that only aggregated auxiliary data is available (**Unit-level data? → No**).

The sample sizes are relatively large for all single disaggregation dimensions (except for unknown disability status). The combinations of disaggregation dimensions partly lead to smaller sample sizes.

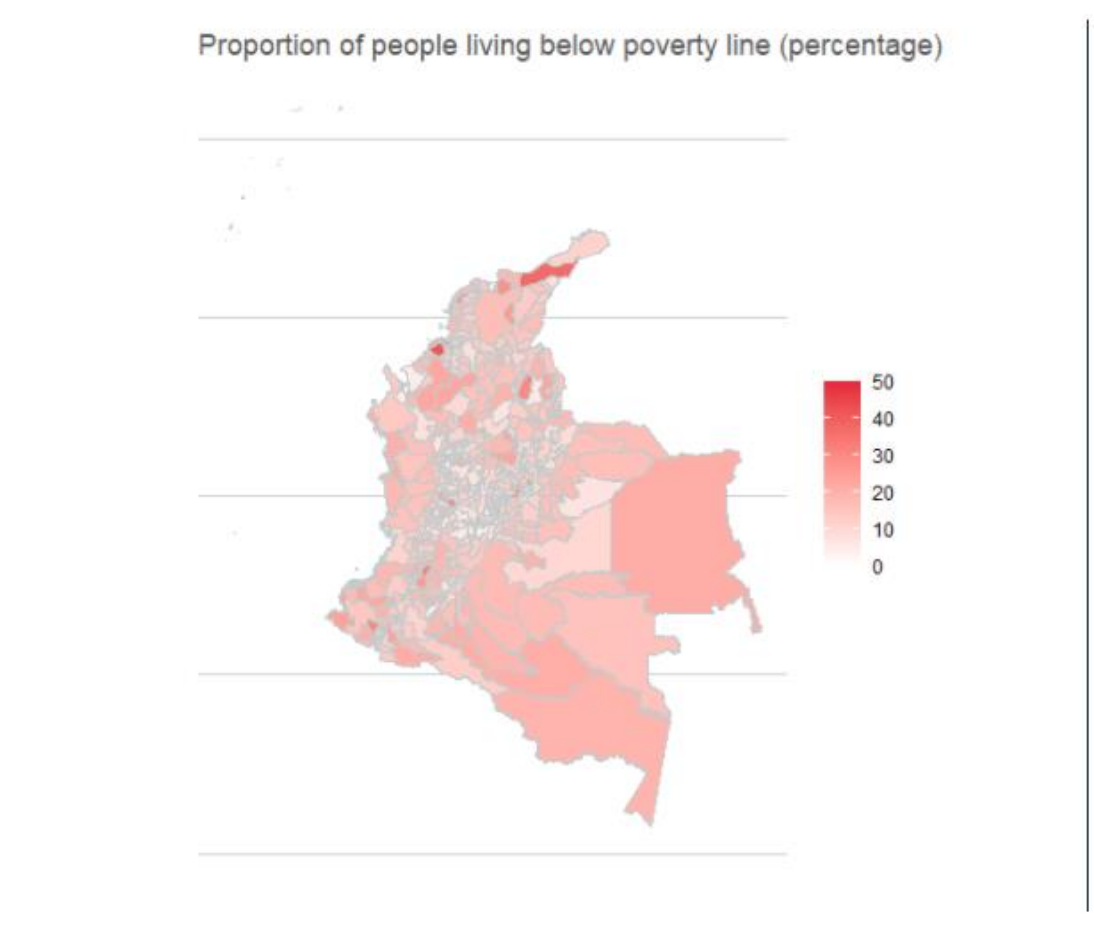
Source: Toolkit for using Small Area Estimation for the SDGs: wiki platform

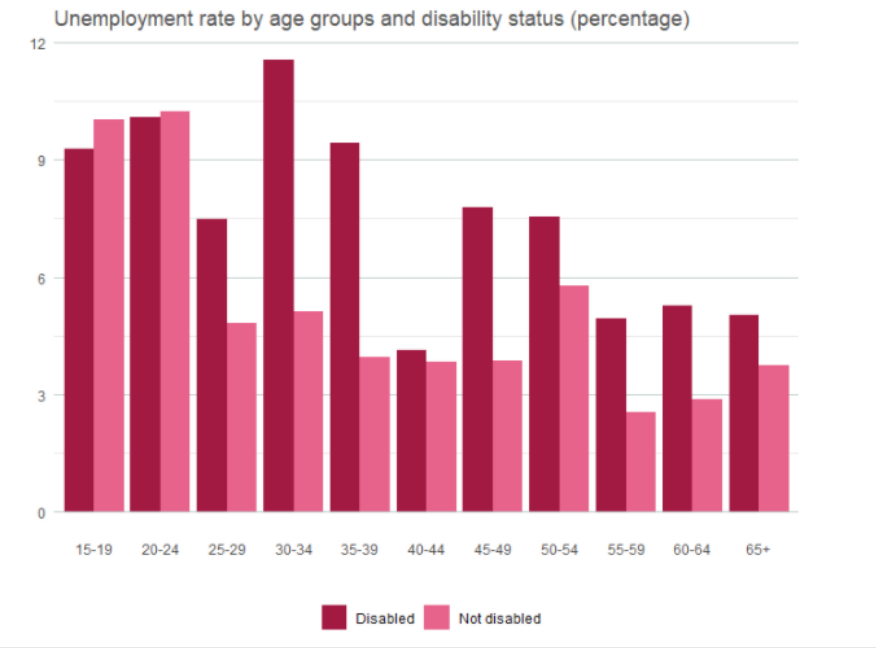
Key methodological aspects

- ❑ Two main models:
  - Area level model
  - Unit level model
- ❑ Looking for the right covariates:
  - Theoretical basis: how correlated are with the outcome indicator?
  - Experts' opinion: subject-matter experts; data producers; modeling specialists; GIS experts
- ❑ Model validation:
  - Assumptions? Or using transformation?
  - Quality of the estimates: CV comparison?
    - ISTAT: CV≤15% for domains; CV ≤18% for small domains
    - Statistics Canada: CV ≤16.5% no release restriction; 16.5%<CV ≤33.3% add warnings; >33.3% not recommended for release
- ❑ Evaluation:
  - Benchmarking: comparing with direct estimates
  - External and independent review: with experts + local government consultation (do the estimates make sense?)

ISWGHS

Communicating SAE - results





ISWGHS

Communicating SAE - methods

- Audience
  - Users who would like to have some general information about the methods and how data can be used
  - Interested in specific models, assumptions and validation
- What to communicate?
  - Why using Small area estimation?
  - What is the method? How is it different from the direct estimates from surveys
  - Technical reports for more sophisticated users

Expand | Example: Explain why small area estimation is used

Synthetic estimation produces estimates for domains where survey data are insufficient, by borrowing strength from other data sources. The other data sources (known as auxiliary data or covariates) are available on an area basis and for all areas in the target population. At the level of these small areas, sample survey sources are not generally available, so the covariate data are usually from some administrative system or a previous census.

Source: Income estimates for small areas in England and Wales, technical report: financial year ending 2018, UK Office for National Statistics, 2020.

ISWGHS

Improve national capacity on using SAE for regular production

SAE Toolkit for SDGs, UNSD working with experts on SAE, under the guidance of the IAEG-SDGs and ISWGHS

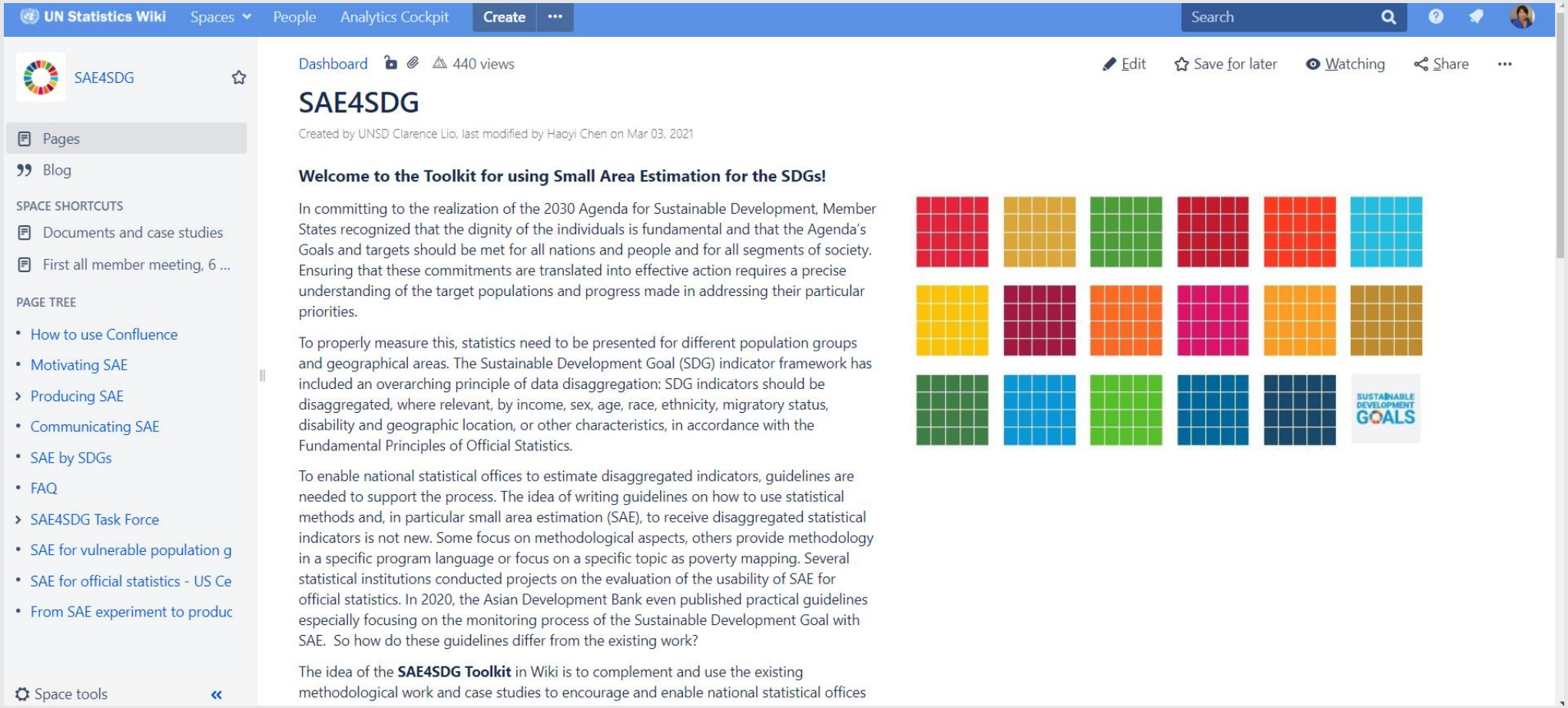
- ❑Using SAE methods to improve SDG data availability for vulnerable population groups – requested by IAEG-SDGs
- ❑Offering practical guidance and country case studies
- ❑Guiding on the enabling environment for using SAE for official data production
- ❑Providing a space for partners to document and disseminate their SAE methodologies: transparency

eLearning on SAE, UNSD in collaboration with UN-ECLAC & UNFPA

- ❑a set of eLearning courses on small area estimation for SDGs data disaggregation
- ❑Organized in 10 modules

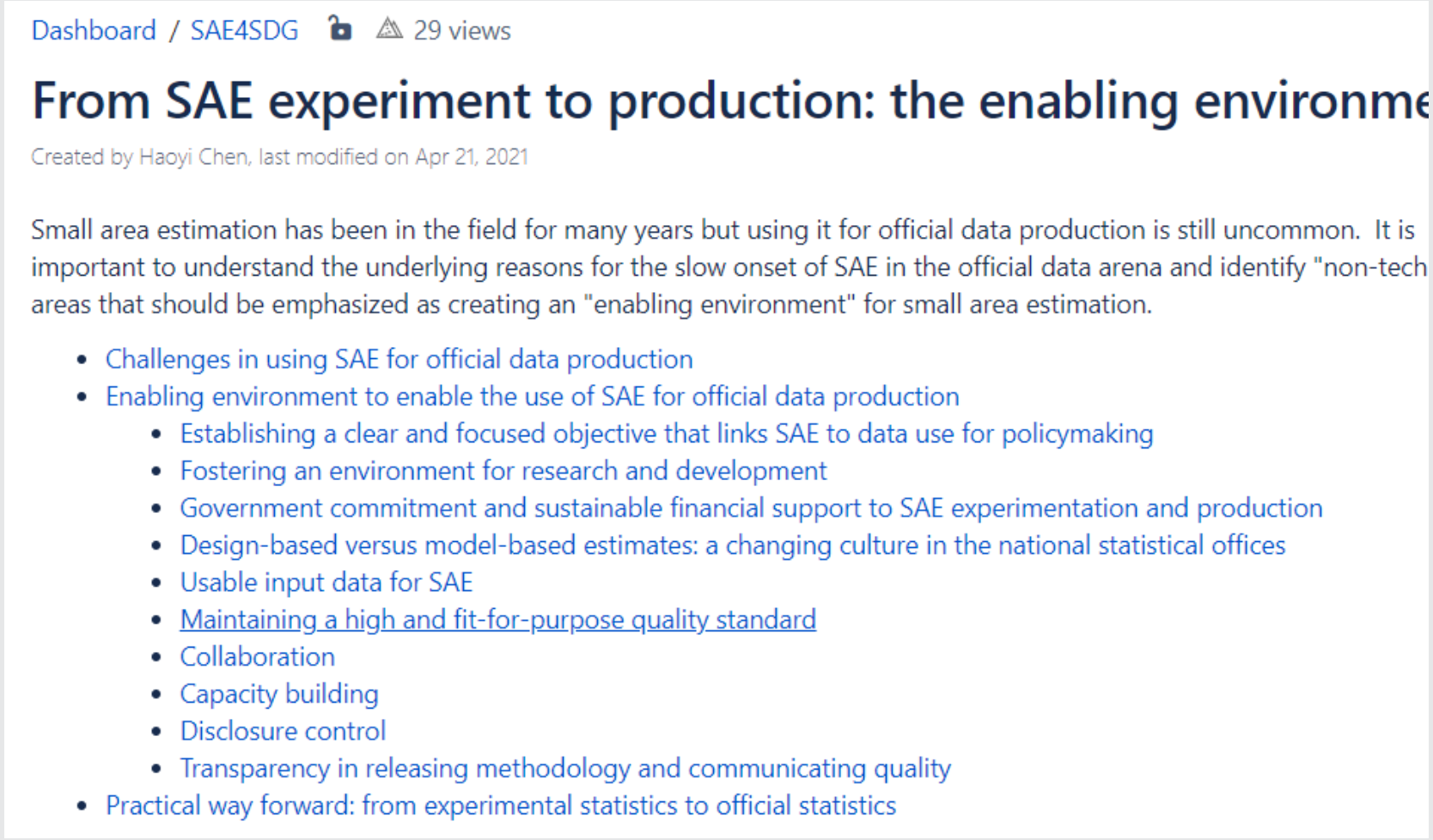
ISWGHS

ISWGHS-IAEG-SDGs collaboration: work modality



ISWGHS

SAE for official statistics



Dashboard / SAE4SDG 29 views

From SAE experiment to production: the enabling environme

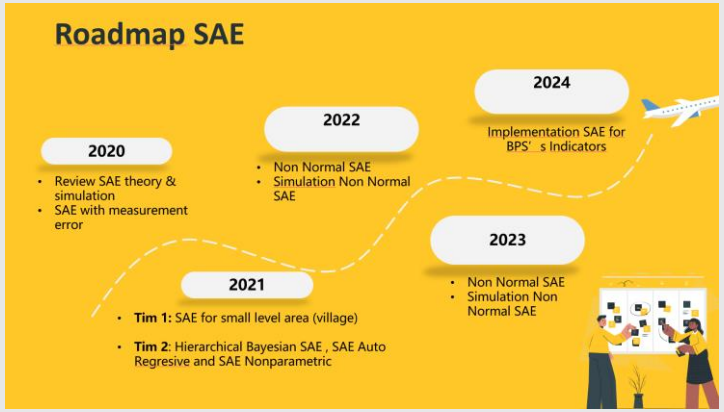
Created by Haoyi Chen, last modified on Apr 21, 2021

Small area estimation has been in the field for many years but using it for official data production is still uncommon. It is important to understand the underlying reasons for the slow onset of SAE in the official data arena and identify "non-tech areas that should be emphasized as creating an "enabling environment" for small area estimation.

- Challenges in using SAE for official data production
- Enabling environment to enable the use of SAE for official data production
  - Establishing a clear and focused objective that links SAE to data use for policymaking
  - Fostering an environment for research and development
  - Government commitment and sustainable financial support to SAE experimentation and production
  - Design-based versus model-based estimates: a changing culture in the national statistical offices
  - Usable input data for SAE
  - Maintaining a high and fit-for-purpose quality standard
  - Collaboration
  - Capacity building
  - Disclosure control
  - Transparency in releasing methodology and communicating quality
- Practical way forward: from experimental statistics to official statistics

ISWGHS

SAE for official statistics – national examples



Example: United States SAIPE Program

In September 1994, the Congress passed the Improving America's Schools Act and signed it into law (PL 103-382). Title I of the law specifies the distribution of Federal funds to school districts based largely on "the number of children aged 5 to 17, inclusive, from families below the poverty level on the basis of the most recent satisfactory data, ..., available from the Department of Commerce."

This law further requires that in Fiscal Year 1997, the Secretary of Education use updated data on poor children for counties and, beginning in Fiscal Year 1999, updated data for school districts, published by the Department of Commerce, unless the Secretaries of Education and Commerce determine that the use of updated population data would be "inappropriate or unreliable."

It also directs the Secretary of Education to fund a National Academy of Sciences panel to provide advice on the suitability of the Census Bureau estimates for use in allocating funds.

Source: Small Area Income and Poverty Estimates (SAIPE) Program, Origins of the Project

Challenges in using SAE for official data production
From National Statistical Offices

- *"We did an experiment using small area estimation method for poverty but the results were not consistent with our own estimates so we did not pursue it again."*
- *"We do not have good input data source for SAE - census data are outdated and administrative data sources do not have good coverage and are lack of proper auxiliary variables."*
- *"SAE method is complicated and we are not comfortable with independently developing the method"*
- *"It is very difficult to convince the managers to use model-based estimates."*

Model-based estimates at Statistics Netherlands

In a more recent paper from Statistics Netherlands (Buelens, Wolf and Zeelenberg, 2016), a set of guidelines were provided that can be used to eval
interested in more details should refer to the original paper.

1. General principle. The general principle when using model based estimation in official statistics, is the principle that official statistics give a de
   a. Objectivity: data used to estimate the model should be related to the subject of the statistic of interest. The model should only be usec
      model, but estimation should not exceed the present.
   b. Reliability: failure of the model should not lead to changes in the (conclusions based on the) estimate of the statistical phenomenon. Th
2. The use of models.
   a. Goal. The goal of using model based estimation should be to estimate data that is not available, and as such to improve the overall est
   b. Data. Models are used to estimate missing data. Both for fitting the model as well as for the final estimation procedure, only data that
   c. Standard. Model based methods that are used at Statistics Netherlands should follow any general consensus in the literature on similar
   d. Model selection. Alternative models should be considered, in order to find the most appropriate model. With model selection, the aim

ISWGHS

Consultations

- ❑ Key SAE experts: consultation meeting organized by JPSM Technical Group on SAE, May 2021
- ❑ Emails and focus-group discussions
  - Australia, Canada, Chile, Colombia, Indonesia, Jamaica, Moldova, Philippines, South Africa, US, Viet Nam
- ❑ Next steps:
  - Approaching more countries and document the challenges/lessons learned
  - Finalise the first stage of the Toolkit; advocating the usefulness of SAE while underlining the important aspects to be considered
  - Organise small technical group discussion (countries + academic)

ISWGHS

eLearning course

- Course format
  - Reading materials
  - Recorded videos (50 videos with about 10-15 minutes for each video), organized in 10 modules
  - Evaluation materials including weekly computer-graded assessments, two mid-term projects, and a final project
  - R program language code that can be used for SAE modelling

- Self-paced learners:
  - Learn at their own pace
  - Access all the above learning materials
  - Machine graded weekly assessments
  - Access to projects, R script and data – not graded
- Guided learners
  - Guided learning and need to follow a fixed schedule that entails about 1-hour of work per day for ten weeks, reading assigned materials, watching course videos, and completing the assigned projects
  - 2-hour interactive workshop per week for ten weeks that will cover a summary of the weekly learning materials and instructions of R code that can be used for SAE modeling
  - Feedback and grading for all three projects

ISWGHS

Thank you

Haoyi Chen

chen9@un.org

ISWGHS