Ph.D. in "Life Course Research" – Socio-Demographic curriculum

Academic Year 2026-2027

Analysis of survey data by Small Area Estimation methods

Prof. Monica Pratesi
University of Pisa
monica.pratesi@unipi.it

Prof. Stefano Marchetti
University of Pisa
stefano.marchetti@unipi.it

Dr. Francesco Schirripa Spagnolo
University of Pisa
francesco.schirripa@unipi.it

Three mornings and one afternoon laboratory | Teaching language: English | Software: R

Objectives

The course develops a coherent approach to producing estimates for population domains from survey data. It moves from direct estimation using survey weights to small area estimation for unplanned domains, and then to the integration of probability surveys with non-probability, citizen and administrative data. Particular attention is given to the assumptions, uncertainty and quality of domain estimates.

Examples draw on the Istat Multipurpose survey system (Indagini Multiscopo), especially Aspects of Daily Life, and on large non-probability datasets. A common application follows an indicator of digital inclusion, such as Internet use, across demographic and territorial domains. The exact indicator, age range and domain definitions will follow the documentation of the selected survey edition.

By the end of the course, participants will be able to:

– Define a target population, domain and estimand, distinguishing planned domains from domains not explicitly supported by the sample design.
– Compute weighted domain totals, means and proportions, with design-based uncertainty measures.
– Identify domains for which direct estimates are unreliable and select an appropriate area-level or unit-level SAE approach.
– Assess how auxiliary information, model assumptions and diagnostics affect the quality of small area estimates.
– Explain and apply introductory methods for integrating probability and non-probability data, and assess the contribution of citizen and administrative sources.
– Produce a reproducible comparison of domain estimates and communicate their limitations.

Program

Proposed timetable: mornings 09:00–12:30, with a 15-minute break; laboratory on Day 3, 14:00–17:30, with a 15-minute break. Total contact time is 13 hours excluding breaks. Teaching combines lectures, worked examples, short discussions and a guided laboratory.

Day 1 morning

Direct domain estimates using survey weights

09:00–10:30 – From the sample design to the domain estimand

– Target populations and finite-population totals, means and proportions; domain membership and denominators. Planned and unplanned domains; geographical and demographic cross-classifications.
– Inclusion probabilities and design weights; nonresponse adjustment, post-stratification and calibration. Horvitz–Thompson totals and Hájek means and proportions.
– Design-based domain estimation: use domain indicators while retaining information about the full sample design. Distinguish survey weights from ordinary regression weights.

10:45–12:30 – Precision and practical interpretation

– Variance estimation by linearization or replication; confidence intervals, standard errors, coefficients of variation and effective sample sizes. Interpret CVs cautiously for near-zero estimates.
– Small domain samples, unstable weights and empty domains. Quality flags based on sample size and uncertainty, with thresholds justified by the intended use.
– Worked Multiscopo example: compare unweighted and weighted estimates for demographic groups and examine what happens as domains become more detailed.

Session outcome: a domain definition and a direct-estimate quality table that motivate the need for SAE.

Day 2 morning

Small area estimation for unplanned domains of study

09:00–10:30 – Borrowing information across domains

– Direct, synthetic and composite estimators; model-assisted versus model-based inference. A small area is defined by insufficient information for reliable direct estimation, not simply by geographical size.
– Area-level Fay–Herriot model: direct estimates, sampling variances and area covariates; empirical best linear unbiased prediction and shrinkage. Auxiliary data must cover the target domains.
– Treatment of unstable sampling variances and proportions near boundaries; introductory discussion of transformations and appropriate generalized models.

10:45–12:30 – Prediction and quality assessment

– Unit-level mixed models and empirical prediction; requirements for population covariates or suitable population summaries. Brief introduction to models for binary outcomes.
– Sampled versus unsampled domains: predictions for empty domains depend more strongly on model assumptions. Distinguish conditional model precision from design-based precision.
– Mean squared error estimation, residual diagnostics, validation and benchmarking to reliable totals. A worked example compares direct and SAE estimates and their uncertainty.

Session outcome: a justified SAE specification and a diagnostic plan. M-quantile methods and machine learning are introduced as extensions rather than core laboratory requirements.

Day 3 morning

Integrating probability and non-probability data for domain inference

09:00–10:30 – Sources and inferential assumptions

– Roles of a probability reference survey, a large non-probability sample and administrative information. Separate auxiliary-variable enrichment from estimation when the outcome is observed only in a non-probability source.
– Citizen data: participation, relevance, coverage and measurement quality. Citizen involvement does not itself imply a non-probability design; evaluate how each dataset was generated.
– Administrative data: target-population alignment, coverage, definitions, reference periods, linkage and duplicates. Administrative counts are valid calibration controls only when their population and concepts match the estimand.
– Selection bias and lack of overlap within domains. Explain conditional exchangeability, positivity and measurement comparability; a large number of records does not eliminate selection bias.

10:45–12:30 – Estimation and critical evaluation

– Calibration and propensity-based adjustment using a probability reference sample; incorporate the reference survey weights and inspect balance and extreme weights.
– Outcome modelling and mass imputation onto the probability sample; conceptual introduction to doubly robust estimation and the assumptions it still requires.
– Use administrative and citizen-derived area covariates within SAE; assess the error in constructed auxiliary variables. Compare this use with directly pooling records from different sources.
– Uncertainty from sampling, estimated weights, model fitting and linkage. Sensitivity to unobserved selection, domain-specific coverage gaps and incompatible measurements. Responsible interpretation and disclosure protection.

Session outcome: an integration strategy specifying the contribution of each source, identifying assumptions, diagnostics and remaining threats to validity.

Day 3 afternoon laboratory

From weighted survey estimates to integrated domain estimates

14:00–14:25 – **Define and prepare.** Inspect the data dictionary, outcome eligibility, domains and auxiliary information. Specify the survey design and document the origin of every input.

14:25–15:10 – **Produce direct estimates.** Compute weighted domain proportions, standard errors and confidence intervals. Report sample sizes, assess weight variability and identify weakly supported domains.

15:10–15:45 – **Fit an SAE model.** Fit a supplied area-level workflow to the direct estimates and sampling variances. Inspect shrinkage, uncertainty and diagnostics; compare results with direct estimates.

15:45–16:00 – Break.

16:00–16:45 – **Integrate data.** Use a guided calibration or propensity-adjustment workflow for an opt-in sample. Compare naive and adjusted domain estimates, check overlap and weight balance, and explore a scenario with unobserved selection. An outcome-model alternative is supplied as an extension.

16:45–17:30 – **Evaluate and communicate.** Assemble a comparison table and plot, discuss why methods differ, and write a short quality statement. In simulated data, use known population values to examine bias; in real data, do not treat a noisy direct estimate as ground truth.

Case study design

The common teaching case estimates an Internet-use proportion for age-by-sex-by-region domains, refining domain definitions only where the accessible data permit. The probability-survey component uses Multiscopo data or a documented teaching extract. An opt-in citizen survey or a large online dataset illustrates selective participation; administrative population counts provide potential calibration totals and SAE predictors.

The instructor prepares aligned variables and reference periods before the course. Real non-probability data may be used when access, provenance and comparability are established. Otherwise, clearly labelled simulated opt-in records illustrate selection driven by observed and unobserved characteristics. An optional example using public citizen-generated records can demonstrate auxiliary information without treating record counts as population prevalence.

Detailed geographical identifiers and sample-design variables depend on the microdata access route. If they are unavailable, a synthetic complex-survey teaching dataset will support valid design-based demonstrations; the course will not present an unjustified simple-random-sampling variance as the actual survey variance.

Laboratory materials and assessment

Instructor materials: a documented survey extract or synthetic equivalent, aligned domain covariates and population controls, a large opt-in teaching dataset, and annotated R scripts. Suggested tools include survey for design-based estimation and sae or emdi for selected SAE workflows. Installation instructions and tested scripts are circulated before the course.

Participant output: a reproducible R script or notebook, a table comparing direct, SAE and adjusted non-probability estimates with appropriate uncertainty measures, and a brief explanation of the method preferred for each domain. Assessment is formative and considers correct estimand definition, weighting, diagnostics and interpretation. The lab uses prepared code so that the emphasis remains on inference rather than programming from scratch.

Suggested literature

Rao, J. N. K. and Molina, I. (2015). Small Area Estimation, 2nd edition. Wiley. Selected introductory and area-level modelling chapters.

Pratesi, M. (ed.) (2016). Analysis of Poverty Data by Small Area Estimation. Wiley. Selected introductory and applied chapters.

Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, 48(2). https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X202200200002

Istat. Aspects of daily life: public use micro.stat files. Read the selected edition’s questionnaire, weights and access documentation. https://www.istat.it/en/microdata/aspects-of-daily-life/

United Nations Statistics Division. Copenhagen Framework on Citizen Data. Background for discussing participation and data quality. https://unstats.un.org/UNSDWebsite/citizen-data/story-details/Copenhagen_Framework_Nov2025

Requirements

Designed for doctoral students, researchers and analysts working with social and official statistics. Participants should know basic sampling concepts, statistical inference and multiple regression, and be able to read and modify simple R scripts. Familiarity with mixed models is helpful but not required.