Small Area Estimation Beyond Traditional Data: Opportunities, Challenges, and Emerging Applications

Monica Pratesi

Department of Economics and Management, University of Pisa
Centro ASESD Camilo Dagum
monica.pratesi@unipi.it

Framework

Context

The data environment is changing:

- surveys, census and administrative registers
- satellite imagery, mobile phone data, social media, web and platform data
- ML- and AI-generated features

Growing demand for reliable statistics at fine geographical and population-subgroup levels

- Direct survey estimators are often unstable or unavailable in small domains
- SAE addresses this problem by borrowing strength from auxiliary information

→ Focus: **how alternative data and ML extend, but also challenge, SAE**

The Small Area Estimation Problem

- Finite population partitioned into D small areas
- Target: area-specific parameter, e.g. mean, total, rate

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j \in U_d} Y_{dj}$$

- Direct estimator uses only the area-specific sample
- Problem:
  - small n_d
  - large sampling variance
  - no observations in some areas
- SAE predicts the unobserved part of the population using models and auxiliary information

What is a SAE Model Today?

- Classical distinction:
  - **Area-level models**: direct estimates + area-level covariates
  - **Unit-level models**: unit-level survey data + population covariates
- Contemporary SAE is broader:
  - spatial and spatio-temporal models
  - robust and semiparametric methods
  - multivariate and functional SAE
  - data-integration and ML-assisted frameworks
- SAE is evolving from model specification to a structured and comprehensive inferential framework.

Outline

- Alternative data sources as covariates in SAE models.
- Selection bias of alternative data sources.
- SAE and Machine Learning.
- Where do we go from here?

Alternative data sources as covariates in SAE models

Alternative Data as Covariates

- Common setting:
  - Y **from probability surveys**
  - X from administrative, Big Data or alternative sources
- Examples:
  - GPS mobility data for poverty estimation in Tuscany (Marchetti et al., 2015)
  - Google Trends as functional covariates in spatial Fay–Herriot models (Porter et al., 2014)
  - Mobile phone data for estimating literacy rates in Senegal (Schmid et al., 2017)
  - Satellite and geospatial predictors for poverty, health and environmental indicators
- Main issue: alternative covariates are often **proxies**, not error-free variables

When Auxiliary Variables Are Measured with Error

- Traditional SAE often assumes: X_d fixed and known, while with alternative data, we may only observe:

$$W_d = X_d + \eta_d$$

- Measurement error may arise from:
    - image classification
    - text processing
    - GPS tracking
    - linkage procedures
    - ML-generated features
- Ignoring this error may lead to biased coefficients and underestimated uncertainty

Reference

Ybarra and Lohr-type models are directly relevant when auxiliary information is measured with error.

Selection bias of alternative data sources

When the Target Variable Comes from Alternative Data

- More challenging setting:
  - Y **from non-probability or alternative sources**
  - X from surveys, administrative data and/or Big Data
- Example:
  - web-scraped or platform-based information used to define the outcome
- Main problem:
  - the data-generating mechanism is unknown
  - inclusion may depend on the outcome or related characteristics
  - selection bias becomes central

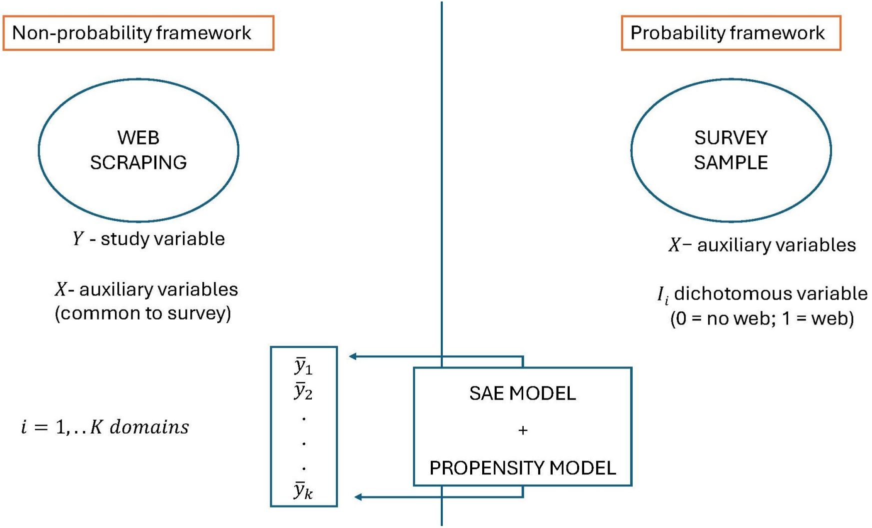
Overcoming Selection Bias

- Alternative data sources are usually non-probability sources
- Correction requires a link to credible population information
- Main strategies:
  - weighting and calibration (Baker et al., 2013; Marella, 2023)
  - propensity-score adjustment
  - data integration with probability samples
  - doubly robust estimation
  - multilevel regression and poststratification (Si, 2025)

Application

Schirripa Spagnolo et al. (2025): non-probability data assisted by SAE methods for SDG indicators in Italy.

Cross the borders by SAE models!



Doubly Robust Estimation

The paper of Schirripa Spagnolo et al. (2025) builds on Kim and Wang (2019), which proposed a doubly robust (DR) estimator

- It combines:
  - Propensity score model (to reweight non-probability sample)
  - Outcome model (predict target variable)
- Uncertainty via bootstrap
- $\mathbf{Z}$ covariates used to estimate propensity: P(unit in the sample be in the webscraping$|\mathbf{Z}$)

Application

Giusti et al. (2025): Digital poverty of Italian teachers: an application at the small area level in Italy by using a doubly robust estimator

Doubly Robust Estimation: advances

- Strong assumption in Schirripa Spagnolo et al. (2025): we can observe the big data sample inclusion indicator, from the probablistic sample
- On going work: Estimate propensity score by employing a pseudo-log-likelihood with area specific random-effects following Chen et al. (2020)

SAE and Machine Learning

Transition to Machine Learning

- Alternative data often require transformation before entering SAE models
- Machine Learning can help by:
  - extracting features from images, text and digital traces
  - modelling nonlinear relationships
  - handling high-dimensional auxiliary information
- But the inferential question remains:
  - Are estimates reliable for small domains?
  - Is uncertainty properly quantified?
  - Are bias and data quality issues explicitly addressed?

Next Step

SAE and Machine Learning: opportunities, limits and uncertainty.

Three Ways to Combine SAE and ML

1. **ML upstream**: feature extraction before SAE
    - satellite images, text, web data, digital traces
    - extracted features used as auxiliary variables
2. **ML in the prediction stage**
    - Random Forests, Gradient Boosting, XGBoost, BART
    - flexible modelling of $f(X)$
3. **Hybrid ML–SAE frameworks**
    - combine algorithmic prediction and statistical borrowing of strength
    - e.g. Mixed Effects Random Forests (Krennmair and Schmid, 2022)

From Parametric SAE to Flexible SAE

Traditional unit-level SAE:

$$Y_{ij} = X_{ij}'\beta + u_i + e_{ij}$$

ML-assisted SAE:

$$Y_{ij} = f(X_{ij}) + u_i + e_{ij}$$

- $f(X)$ is learned from the data using flexible algorithms
- Area effects u_i remain important when residual between-area heterogeneity persists
- If area effects are omitted, small area estimates may be biased

Interpretation

ML can replace the linear regression component, but not the need to model small-area heterogeneity.

Main Risks of ML in SAE

- Good predictive accuracy does not guarantee good small-area inference
- Main risks:
  - overfitting, especially with small samples
  - weak performance in some domains despite good global performance
  - instability across areas or time
  - poor interpretability
  - non-representative training data
  - underestimated uncertainty

Warning

In SAE, the relevant question is not only: “Does the model predict well?” It is: “Does it produce reliable domain estimates?”

Uncertainty Estimation Remains Central

- In SAE, uncertainty is part of the estimate
- ML-assisted SAE introduces additional sources of uncertainty:
  - sampling variability
  - model uncertainty
  - tuning uncertainty
  - feature-construction uncertainty
  - selection bias in training data
  - linkage or integration error
- Standard ML metrics do not capture all these components
- Bootstrap methods are useful, but only if the full pipeline is resampled

Take-home Messages: SAE and ML

- ML is useful for flexible modelling and feature extraction
- SAE remains essential for:
  - borrowing strength across domains
  - modelling area heterogeneity
  - correcting bias
  - quantifying uncertainty
- Predictive accuracy alone is not enough
- Future SAE should evaluate ML using:
  - domain-level bias
  - stability across areas
  - transportability
  - quality of MSE or uncertainty estimates

Where do we go from here?

Where do we go from here?

- The future of SAE is not a choice between:
  - traditional statistics
  - machine learning
  - alternative data
- The key challenge is to integrate them within a coherent inferential framework
- New data and ML are useful only if SAE preserves:
  - bias control
  - model validity
  - uncertainty quantification
  - interpretability for policy use

Assumptions of SAE challenged by today data environment

- **Representativeness**: alternative data sources are generated through self-selection, platform participation, device ownership or access to digital infrastructure $\Rightarrow$ Selection bias.
- **Measurement quality**: new variables may not be simply observed; they may be constructed, often through several pre-processing and modelling stages.
- **Model specification**: alternative data may produce highly nonlinear, context-dependent and temporally unstable relationships
- **Independence**: Alternative data often carry complex dependence structures (spatial autocorrelation, temporal dependence. . .)
- **Separability of sampling error and model error**: When the target or covariates come from alternative data sources, the distinction between sampling error, processing error, measurement error and linkage error becomes blurred

Quality, Ethics and Transparency

- **Quality**
  - alternative data may suffer from coverage, measurement and processing error
  - ML-generated variables may add algorithmic uncertainty
- **Ethics**
  - privacy, GDPR and data access constraints
  - sensitive information may emerge through linkage
- **Transparency**
  - data production processes may be proprietary or opaque
  - FAIR principles are not always satisfied

A Modern SAE System

- A contemporary SAE framework may include several modelling layers:
  - **Measurement layer**: direct estimator, target proxy or non-survey outcome
  - **Covariate layer**: census, administrative, alternative or ML-generated variables
  - **Latent heterogeneity layer**: area effects, spatial or temporal dependence
  - **Uncertainty layer**: sampling error plus data-source and model uncertainty

Key Point

The defining feature of SAE is not a specific estimator, but the organized treatment of information borrowed across areas, sources and modelling layers.

Future Direction 1: Treat Auxiliary Data as Stochastic

- Traditional SAE often treats auxiliary variables as fixed and error-free
- This assumption is weaker when X comes from:
  - satellite imagery
  - mobile phone data
  - social media
  - web scraping
  - AI-generated indicators
- Future SAE should explicitly model:
  - measurement error
  - processing error
  - temporal instability
  - algorithmic uncertainty

Future Direction 2: Selection Bias as a First-Order Problem

- Alternative data sources are rarely probability samples
- Inclusion may depend on:
  - platform participation
  - digital access
  - device ownership
  - user behaviour
  - socio-economic characteristics
- SAE with alternative data requires anchoring to credible population information
- Possible strategies:
  - calibration
  - propensity-score adjustment
  - doubly robust estimators
  - integration with probability samples

Future Direction 3: Data Integration as Part of the Model

- Modern SAE often combines surveys, administrative data and alternative sources
- Data integration introduces additional uncertainty:
    - linkage error
    - statistical matching assumptions
    - spatial mismatch
    - temporal misalignment
    - inconsistent definitions
- These errors should not be treated only as preprocessing issues
- They should enter the uncertainty discussion of SAE

Future Direction 4: How Should ML Be Evaluated in SAE?

- Predictive accuracy alone is not enough
- ML-assisted SAE should be evaluated using SAE-relevant criteria:
  - domain-level bias
  - stability across areas
  - robustness to small samples
  - transportability across contexts
  - interpretability
  - quality of uncertainty estimates

Key Point

A model that predicts well globally may still fail for small domains.

Future Direction 5: Uncertainty Remains the Core

- SAE estimates are incomplete without measures of uncertainty
- Alternative data and ML can produce sharper-looking maps, but sharper maps are not necessarily more reliable
- Future methods must account for:
  - sampling error
  - model error
  - measurement error
  - selection bias
  - linkage error
  - ML feature-construction uncertainty

Warning

Without credible uncertainty estimation, SAE risks exchanging statistical reliability for algorithmic opacity.

AI and the Next Phase of SAE

- AI can contribute to SAE by transforming unstructured data into structured information
- Examples:
  - text classification
  - image feature extraction
  - synthetic data generation
  - indicators derived from foundation models
- However, AI-derived variables depend on:
  - training data
  - model architecture
  - algorithmic choices
  - black-box processing pipelines
- Treating AI-generated variables as fixed inputs may underestimate uncertainty

A Multidisciplinary Agenda

The future of SAE requires collaboration among:

- statisticians and survey methodologists: inference, sampling, uncertainty
- ML researchers and computer scientists: algorithms, feature extraction, data pipelines
- domain experts: interpretation and policy relevance
- legal and governance experts: privacy, access and ethical use

Implication

Modern SAE is no longer only a modelling problem; it is also a data-quality, governance and infrastructure problem.

Final Consideratons

- Alternative data, ML and AI do not replace SAE
- They make the inferential role of SAE even more central
- The key challenge is not only to predict better, but to produce reliable small-domain estimates
- Future SAE should integrate:
  - richer data sources
  - flexible algorithms
  - explicit bias correction
  - credible uncertainty quantification

Main Message

More data and more flexible algorithms are useful only if embedded in a framework for valid inference.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013) Summary report of the AAPOR task force on non-probability sampling *Journal of Survey Statistics and Methodology, 1*(2), 90–143.

Chen, Y., Li, P., & Wu, C. (2020) Doubly robust inference with nonprobability survey samples *Journal of the American Statistical Association, 115*(532), 2011–2021

Chi, G., Fang, H., & Zhao, Z. (2022) Geospatial machine learning for poverty prediction and mapping: A systematic review *Computers, Environment and Urban Systems, 94*

Giusti, C., Pratesi, M., Schirripa Spagnolo , F., Salvati, N., D'Agostino, A., & Calzone, S. (2025) Digital poverty of Italian teachers: an application at the small area level in Italy by using a doubly robust estimator *SAE Conference 2025*

Kim, J. K., & Wang, Z. (2019) Sampling techniques for big data analysis *International Statistical Review, 87*(2), 177–191

Krennmair, P., & Schmid, T. (2022) Flexible domain prediction using mixed effects random forests *Journal of the Royal Statistical Society Series C: Applied Statistics, 71*(5), 1865–1894

Krennmair, P., Würz, N., & Schmid, T. (2025) Analysing opportunity cost of care work using mixed effects random forests under aggregated auxiliary data *Journal of the Royal Statistical Society Series C: Applied Statistics*

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. & Gabrielli, L. (2015) Small area model-based estimators using big data sources *Journal of Official Statistics, 31*(2), 263–281

Marella, D. (2023) Adjusting for selection bias in nonprobability samples by empirical likelihood approach *Journal of Official Statistics, 39*(2), 151–172.

Porter, A. T., Holan, S. H., Wikle, C. K., & Cressie, N. (2015) Spatial Fay–Herriot models for small area estimation with functional covariates *Spatial Statistics, 10*(2), 27–42

Pratesi M. (2016) Analysis of Poverty Data by Small Area Estimation. *John Wiley & Sons, Ltd.*

Pratesi, M. , Moretti, A. (2026) Small Area Estimation in the Era of Alternative Data, Machine Learning and AI: a Conceptual Synthesis *Annual Review of Statistics and Its Application*

Reluga, K., Kong, D., Ranjbar, S., Salvati, N., van der Laan, M. (2025) The impact of job stability on monetary poverty in Italy: causal small area estimation *Journal of the American Statistical Association*

Rao, J. (2021) On making valid inferences by integrating data from surveys and other sources *Sankhya B 83*(1), pp. 242–272.

Schirripa Spagnolo, F., Bertarelli, G., Summa, D., Scannapieco, M., Pratesi, M., Marchetti, S., & Salvati, N. (2025) Inference for big data assisted by small area methods: an application on sustainable development goals sensitivity of enterprises in Italy *Journal of the Royal Statistical Society Series A: Statistics in Society, 188*(1), 27–45

Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T. (2017) onstructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal *Journal of the Royal Statistical Society Series A: Statistics in Society, 180*(4), 1163–1190

Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T. (2017) Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal *Journal of the Royal Statistical Society Series A: Statistics in Society, 180*(4), 1163–1190

Si, Y. (2025) On the use of auxiliary variables in multilevel regression and poststratification *Statistical Science, 40*(2), 272.

Tzavidis, N. (2025) Small area estimation in the era of machine learning and alternative data sources: opportunities, challenges and outlook *Journal of Official Statistics*

Ybarra, L. M., & Lohr, S. L. (2008) Small area estimation when auxiliary information is measured with error *Biometrika, 95*(4), 919–931.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M.(2020) Using publicly available satellite imagery and deep learning to understand economic well-being in Africa *Nature Communications, 11*(1), 2583

THANK YOU!

Monica Pratesi

Dipartimento Economia e Management
Università di Pisa

monica.pratesi@unipi.it