Combining information from multiple data sources using statistical modeling and methods

Partha Lahiri
plahiri@umd.edu

Joint Program in Survey Methodology & Department of Mathematics
University of Maryland College Park

33rd Morris Hansen Lecture

March 30, 2026

Morris Hansen's Contributions

- Morris Hansen is widely regarded as the most influential survey statistician of the 20th century.
- Innovation at the U.S. Census Bureau
  1. Total survey design
  2. Importance of research in statistical agencies
- Dissemination of knowledge through his 1953 book (with Hurwitz and Madow) — Sample Survey Methods and Theory (Vol. 1 and Vol. 2) — and participation in numerous professional committees
- Leadership in the Census Bureau, Westat, and Professional Associations

Figure: Morris Hansen

Some Specific Examples of Hansen's work

- Played a key role in promoting sample surveys at the Census Bureau
  - 1937 Enumerative Check Census
  - 1940 U.S. Decennial Census of Population
  - Benifits of very large PSUs in a multistage sample when costs/administrative issues are taken into account.
- Development of sampling theory
- Development of Models and Theories for the Analysis of Nonsampling Errors (e.g., the Census Bureau model of survey error)
- His efforts led to purchase of the first computer for statistical purposes
- Development of optical scanning equipment, the introduction of self-enumeration and mail in both demographic and economic censuses.

International Association of Survey Statistician (IASS)

Inaugural President of IASS, 1973-1977: Morris Hansen

Some References

- An Interview by James O'Brien on June 22, 1983
- Some History and Reminiscences on Survey Sampling (1985), 53 minutes Amstat video
- Olkin, I. (1987) A Conversation with Morris Hansen, Statistical Science, 2, 162-179.
- Waksberg, J. & Goldfield, E.D. (1997) Morris Howard Hansen December 15, 1910-October 9, 1990 (1997) International Statistical Review, 87-96.

Nonresponse & Coverage Errors

An example: Hansen et al. (1953, pp 483-486)

To estimate the median number of radio stations heard during the day for over 500 counties of the USA (small areas).

Two different survey data used:

- Mail Survey
- Personal Interview Survey

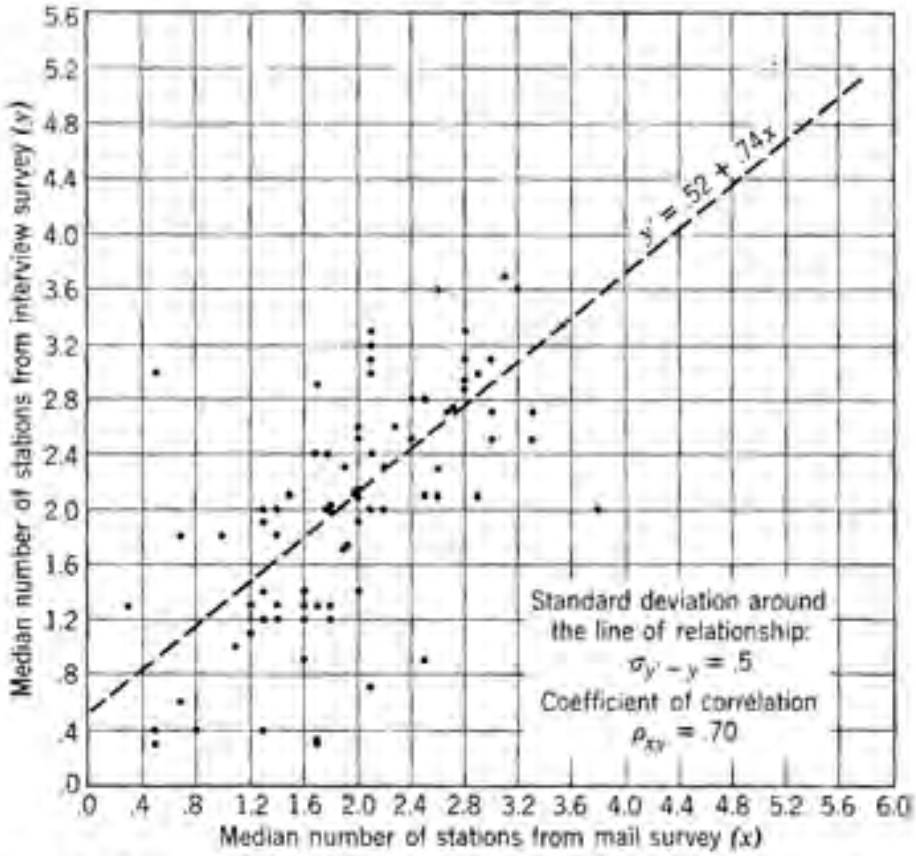


FIG. 2. Comparison of median numbers of stations heard during the day without trouble, as estimated from mail and interview surveys in selected primary sampling units.

Hansen et al. suggested the following:

- For counties with sample from the personal interview survey, use y_i.
- For counties with sample from only mail survey, use estimate $\hat{y}_i = .52 + .74x_i$

Linkage or Mismatch Errors

The Challenge: Why Probabilistic Linkage?

Reality: Observations can be recorded in multiple files, and the matching status between records can be unknown, due to the lack of unique and error-free identifier, such as Social Security Number (SSN).

- **Common Data Issues:** Typos (Smith vs. Smyth), nicknames (Robert vs. Bob), missing middle initials, outdated addresses, etc.

- **The Statistical Problem:** To estimate the probability that two records $\alpha \in A$ and $\beta \in B$ represent the same entity, given a comparison vector of agreement/disagreement across multiple matching fields.

The Fellegi-Sunter Framework (1969)

Two latent states for a pair (a, b):

- M: The pair is a true match.
- U: The pair is a non-match (unmatched).

The probabilities:

- **m-probability (m_i):** $P(\text{field } i \text{ agrees} \mid M)$.
- **u-probability (u_i):** $P(\text{field } i \text{ agrees} \mid U)$.

The agreement weight for field i:

$$w_i = \log_2 \left(\frac{m_i}{u_i} \right)$$

Scoring and Decision Rules

For every record pair, we sum the weights across all compared fields (Name, DOB, ZIP, etc.):

$$\text{Total Score} = \sum_{i=1}^{k} w_i$$

- **High Score:** Likely Match.
- **Low Score:** Likely Non-match.
- **The "Gray Area":** Scores in the middle represent "Potential Matches" that require manual clerical review.

Statistical Inference: The choice of thresholds involves a trade-off between False Positives and False Negatives.

Estimation with linked data

Most statistical models (like Linear Regression) assume that auxiliary variables X and outcome variable Y come from the same entity.

- **The Data Split:** Imagine X variables (e.g., Education) are in File A, and Y variables (e.g., Income) are in File B.
- **The Linkage Gap:** Because we don't have a perfect ID, we link them using Probabilistic Record Linkage (PRL).
- **The problem:** If we run a regression on these "best guess" links, the ordinary least squares estimators (OLS) of regression coefficients (β) will be **biased**.

How can we correct bias in OLS when we know some of our linked records are probably wrong?

Some references: Scheuren and Winkler (1993), Lahiri and Larsen (2005), Han and Lahiri (2019).

Integrating Linkage and Analysis

- Instead of picking the "top" match and throwing away the rest, use the probabilities of all potential matches.
- If Record a has a 90% chance of matching Record b and a 10% chance of matching Record c, the model accounts for both possibilities.
- Developed a way to "un-bias" the regression coefficients by factoring in the probability that a link is a mismatch.

Why It Matters: Accuracy Over Convenience

- Ignoring linkage error can lead to incorrect scientific conclusions.
- Turn record linkage into a transparent statistical procedure.
- Provide reasonable standard error estimates that account for the uncertainty of the match itself.

We shouldn't treat linked data as 'truth'; we should treat the link itself as a random variable.

Selected References

Topic 1: Books/Review

[1] Binette, O. & Steorts, R. C. (2022). *(Almost) All of Entity Resolution*. Science Advances.

[2] Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.

[3] Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data Quality and Record Linkage Techniques. Springer.

Topic 2: Early papers

[4] Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association (JASA).

[5] Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84, 414-420.

[6] Newcombe, H. B., et al. (1959). Automatic Linkage of Vital Records. Science.

Topic 3: Statistical Analysis with linked data

[7] Han, Y. and Lahiri, P. (2019), Statistical Analysis with Linked Data , International Statistical Review, 87, S139-S157.

Selected References

[8] Lahiri, P. and Larsen, M. (2005), Regression analysis with linked data , Journal of the American Statistical Association, Vol 100, 222-230.

[9] Scheuren, F., and Winkler, W. E. (1993), Regression Analysis of Data Files That Are Computer Matched, Survey Methodology, 19, 39–58.

Topic 4: Bayesian Entity Resolution/Scalable Probabilistic Record Linkage.

[10] Marchant, N. G., Kaplan, A., Elazar, D., Rubinstein, B. I., & Steorts, R. C. (2021) d-blink: Distributed End-to-End Bayesian Entity Resolution, Journal of Computational and Graphical Statistics, 30, 406-421.

Topic 5: Privacy-Preserving Record Linkage (PPRL)

[11] Mirel, L. B., Resnick, D.M., Aram, J., and Cox, C.S.. (2022) A Methodological Assessment of Privacy Preserving Record Linkage Using Survey and Administrative Data. Statistical Journal of the IAOS 38, 413–21. `https://doi.org/10.3233/SJI-210891`.

Selection and Measurement Error Biases

Nonprobability surveys: Recent works

- Special issues of Survey Methodology and Calcutta Statistical Association Bulletin on statistical research in nonprobability sampling.
- Review papers - Rao (2021); Wu (2022); Beaumont (2020); Kalton (2023).
- Recent literature focus on integrating nonprobability with probability (reference) surveys (usually not having response):
  - **Matching and mass imputation** - Rivers (2007); Kim, Park, Chen, and Wu (2021).
  - **Inverse propensity score weighting (IPW)** - Lee and Valliant (2009); Valliant and Dever (2011); Wang, Valliant, and Li (2021); Beaumont, Bosa, Brennan, Charlebois, and Chu (2024).
  - **Doubly robust methods** - Chen, Li, and Wu (2020).
- Pfeffermann (2015) and Kim and Morikawa (2023) consider only nonprobability surveys for inference.
- Huge body of work on improving representativeness, **but not much focus on measurement error in nonprobability surveys**.

Selected References

Topic 1: Review Papers

[1] Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? Survey Methodology, 46(1), 1–29.

[2] Kalton G. (2023). Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day, Comments and Discussions : Pfeffermann D., Lehtonen R., Gershunskaya J., Lahiri P., Münnich R., C. Statistics and Transition New Series, 24, 1–45.

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

[4] Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, Statistics Canada, 48(2), 5237–5250.

Topic 2: Reducing selection bias considering only nonprobability survey

[5] Kim, J. K. and Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples . Calcutta Statistical Association Bulletin, 75(1), 8–27.

[6] Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. Journal of Survey Statistics and Methodology, 3(4), 425–483.

Selected References

Topic 3: Reducing selection bias by combining probability and nonprobability surveys

[7] Beaumont, J. K., Bosa, K., Brennan, A., Charlebois, J., & Chu, K. (2024). Handling nonprobability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. Survey Methodology, Statistique Canada, 50(1), 77–106.

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

[9] Elliott, M., & Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. Survey Methodology, 33(2), 211–215.

[10] Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society: Series A (Statistics in Society), 184(3), 941–963.

[11] Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods & Research, 37(3), 319–343.

[12] Rivers, D. (2007). Sampling for web surveys. In Joint Statistical Meetings, Volume 4. American Statistical Association, Alexandria, VA.

Selected References

[13] Savitsky, T., Williams, M., Gershunskaya, J., & Beresovsky, V. (2023). Methods for combining probability and nonprobability samples under unknown overlaps. Statistics in Transition new series, 24, 1–34. https://doi.org/10.59170/stattrans-2023-061

[14] Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. Sociological Methods & Research, 40(1), 105–137.

[15] Wang, L., Valliant, R., & Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. Statistics in Medicine, 40(4), 5237–5250.

Topic 4: Reducing selection bias and measurement error by combining probability and nonprobability surveys

[16] Kennedy, C., Mercer, A., & Lau, A. (2024). Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith. Survey Methodology, Statistique Canada, 50(1), 3–21.

[17] Sen, A., and Lahiri, P. (2026). Improving measurement error and representativeness in nonprobability surveys. To appear in Survey Methodology; arxiv paper

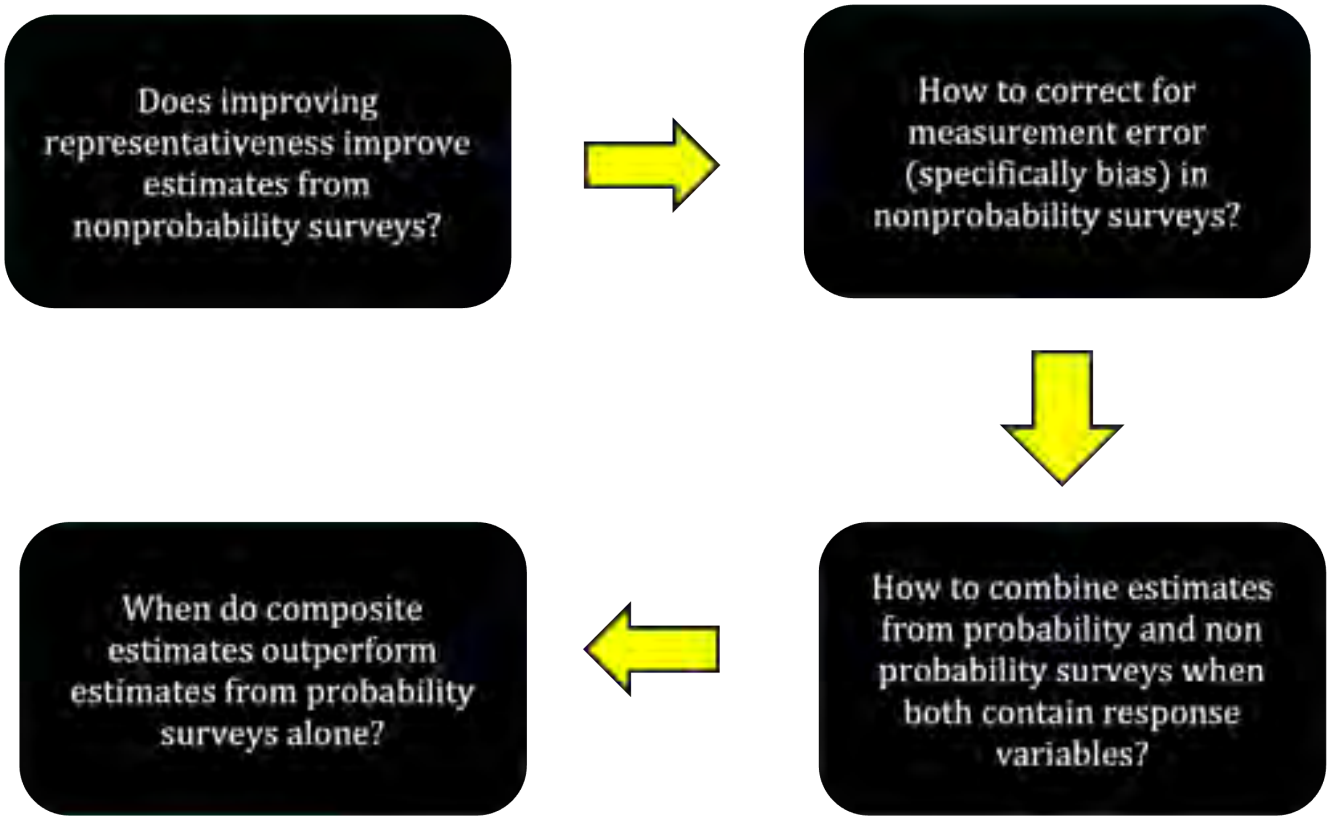
Topic 5: Combining probability and nonprobability surveys for Small Area Estimation

[18] Nandram, B., & Rao, J. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. Statistics and Applications, 22, 343–374.

Sen, A. and Lahiri. P. Improving measurement error and representativeness in nonprobability surveys. Accepted in *Survey Methodology*. arXiv preprint

Combining probability and nonprobability surveys

- *Goal*: Reduce measurement and sampling biases in nonprobability surveys (*nps*) through data integration with probability surveys (*ps*).
- *Research Questions* :



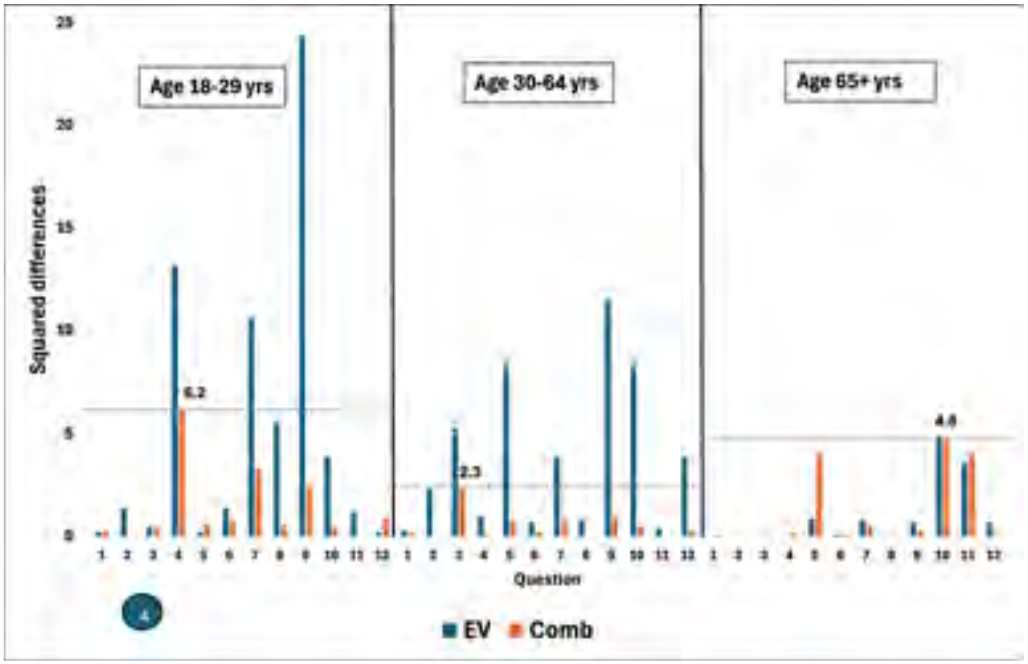
Nonprobability surveys : Introduction

- *Motivating example*: *Pew Research Center* 2021 benchmarking study to determine accuracy of online surveys on general population estimates for all U.S. adults as well as key demographic subgroups.
- Study compared 3 ps (ABS samples P_1, P_2, P_3) and 3 *nps* (opt-in samples O_1, O_2, O_3) with large govt. datasets (NHANES, NHIS, ACS, CPS, etc.) using benchmarking.
- Subgroups $18-29$ *year-old* and *Hispanic adults* in *nps* were identified as *bogus respondents* - replying 'Yes' regardless of the question.
- These subgroups in *nps* have high Mean Absolute Error (MAE) values than the same groups in ps - indicating measurement error in nps.
- To improve estimates from *nps* in terms of both *representativeness* and *measurement error*, we propose composite estimators from bias-corrected *nps* and *ps* estimates, considering two cases:

Case 1: When bias is computed from alternative sources.
Case 2: When bias-corrected estimates are predicted using modeling.

Improving nonprobability surveys: Method 1

- In Case 1, when more than 2 surveys (1 ps and 1 nps) are available, we estimate bias (ϵ) at question $\times$ age group level using O_1, O_2 and P_1, P_2.
- The proposed composite estimator (Comb) is a weighted combination of bias-corrected IPW from O_3 with survey-weighted estimate from P_3.
- When compared with existing composite estimator of Elliot and Haviland (2007), proposed estimator shows improved performance in terms of squared differences with benchmarks.



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

Improving nonprobability surveys: Method 2

- In Case 2, when other sources of information about bias are not available, we fit a machine learning (ML) model on P_3.
- Using fitted parameter estimates from ML model, we predict the unit level responses in O_3 (or equivalently, the probabilities of responding 'Yes').
- We then use a similar IPW method on the predicted probabilities to finally produce a composite estimator of *nps* and *ps*.
- We fit Random Forest (RF) and Gradient Boosting Model (GBM), choose GBM as final model based on AUC.

Table: Details of the predictive modeling on *ps* P_3 of 4912 respondents, with the aim of obtaining bias-corrected estimates from *nps* O_3.

Data structure	Question × respondent level data
Size (Train + test)	59 thousand observations split into train (80%) and test (20%)
Response variable type	Binary (Refusals are not considered)
Auxiliary variables (no. of levels)	Question (12), age (3), race-ethnicity (5), education (3), gender (3), region (4)
Predictive models (AUC, Accuracy)	Random Forest (0.88, 0.85), Gradient Boosting (0.92, 0.85)

Nonprobability surveys: SAE connections

- We see that the proposed composite estimator is not uniformly better than the estimator from ps alone, as ps has an adequate sample size (close to 5k).
- To investigate if the performance of composite estimator can be further improved, we generate ps of smaller sample sizes from P_3 using a stratified sampling procedure with proportional allocation.
- We calculate MSD values for the smaller samples from P_3, keeping the original nps O_3 intact.
- We see that as sample size of ps decreases the variance of ps estimator increases, as a result the composite estimator outperforms ps estimator.
- But upon reducing sample size of P_3, we encounter SAE problems - some subgroups do not have any observations.
- In such cases, small area modeling techniques are required (currently under investigation).

Large Sampling Variance/No Sample: SAE

Selected References

Topic 1: Review Papers

[1] Erciulescu, A. L. Franco, C. and Lahiri, P. (2021), Use of Administrative Records in Small Area Estimation, Administrative Records for Survey Methodology, Chapter 10, 231-267 Wiley Series in Survey Methodology, eds. Chun, A.Y., Larsen, M.D., Durrant, G., Jerome P. Reiter.

[2] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

[3] Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation*, 2nd Edition. Wiley.

Topic 2: Survey to survey imputation

[4] Das, S., Basu, A., Lahiri, P. & Sengupta, S. (2022), Bayesian synthetic prediction of state level poverty using Indian Household Consumer Expenditure Survey Data, Statistical Journal of the International Association for Official Statistics (IAOS), 38, 1325-1335.

[5] Himanshu, Lanjouw, P., & Schirmer, P. (2025). Poverty Decline in India after 2011-12: Bigger Picture Evidence. Economic and Political Weekly, 60(15), 36-46.

[6] Newhouse, D., & P. Vyas (2022) Estimating Poverty in India without Expenditure Data: A survey to survey imputation approach, Economic and Political Weekly, v. 57 no. 48.

[7] Sen A. and Lahiri P. (2025). Estimation of finite population proportions for small areas – a statistical data integration approach. Journal of Survey Statistics and Methodology. `https://doi.org/10.1093/jssam/smae049`

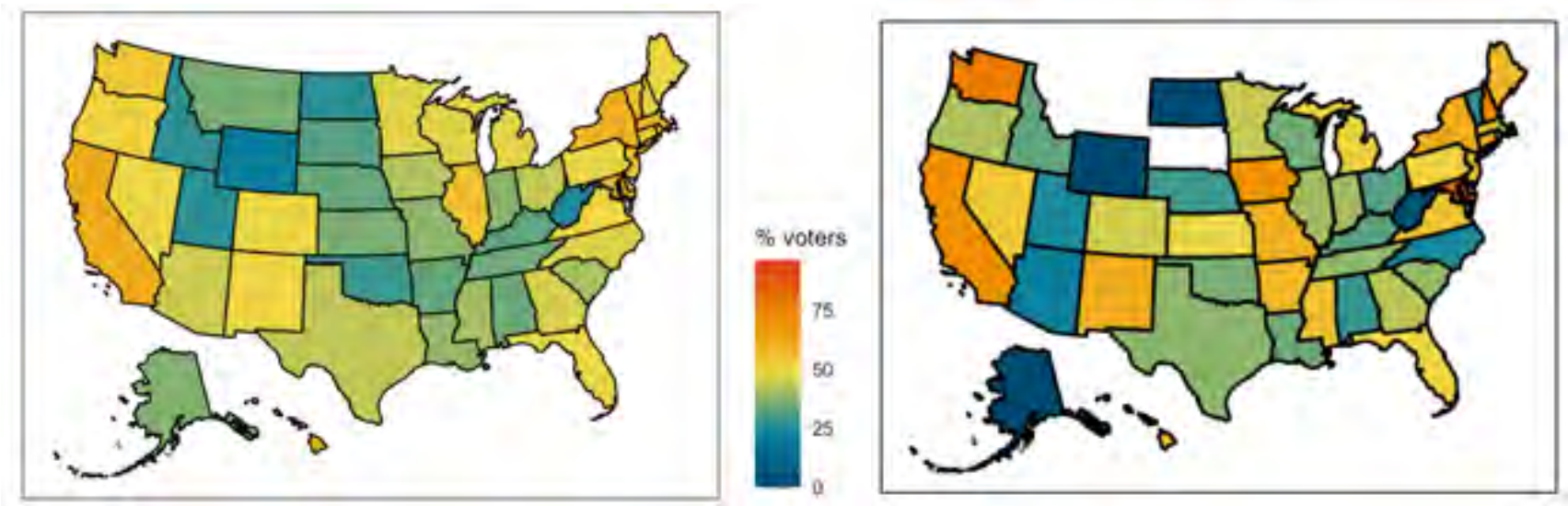
Sen, A. and Lahiri. P. "Estimation of finite population proportions for small areas – a statistical data integration approach." Journal of Survey Statistics and Methodology, Vol.13, No.3, 2025, pages: 309–332. arxiv preprint

Combining two probability surveys

- *Goal*: Produce reliable estimates with uncertainty quantification of finite population proportions for small areas using probability survey.
- *Issue*: Area sample sizes are small and even zero for some areas, leading to inaccurate estimates and uncertainty measures.
- *Motivating example*: Election prediction using Pew Research Center's October 2016 Political Survey – predict % of voters for Clinton for 50 states of the Unites States (U.S.) and District of Columbia (DC).

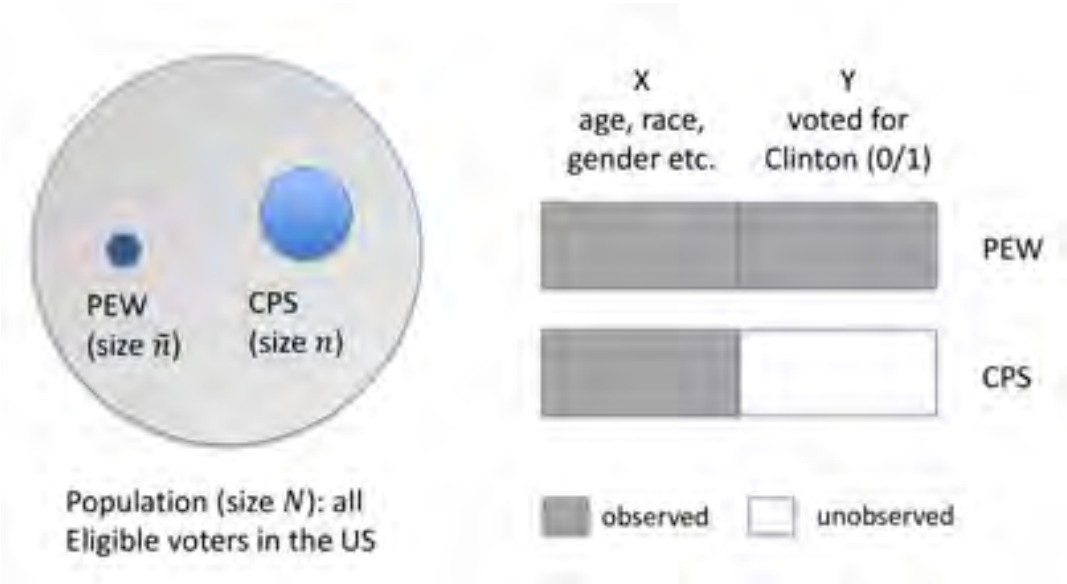
(a) Actual % of voters for Clinton

(b) Direct Estimate



Combining two probability surveys

- We borrow strength from alternate data sources - larger in size with enough samples for areas.
- We replace the finite population frame by the big sample (s), avoiding linking of the sample to the finite population frame.
- Combine multiple sources having common auxiliary variables - CPS 2016 Voting and Registration Supplement with common demographics (age, race, etc.).



Hierarchical Model

- We approximate area-level population means ($\bar{Y}_i$) by the survey-weighted proportions ($\bar{Y}_{iw},\ i = 1, \cdots, m$) from s, appealing to the law of large numbers as area sample sizes n_i are large.

- Since, $\bar{Y}_{iw}$ is unknown, we estimate it by the following multi-level **working model**, of which mixed logistic is a special case.

- For $i = 1, \cdots, m;\ j = 1, \cdots, N_i$,

$$\begin{aligned} \text{Level 1:} \quad & Y_{ij}|\theta_{ij} \overset{ind}{\sim} \text{Bernoulli}(\theta_{ij}) \\ \text{Level 2:} \quad & \theta_{ij} = H(x_{ij}, \beta, v_i) \\ \text{Level 3:} \quad & v_i \overset{iid}{\sim} N(0, \sigma^2), \end{aligned}$$

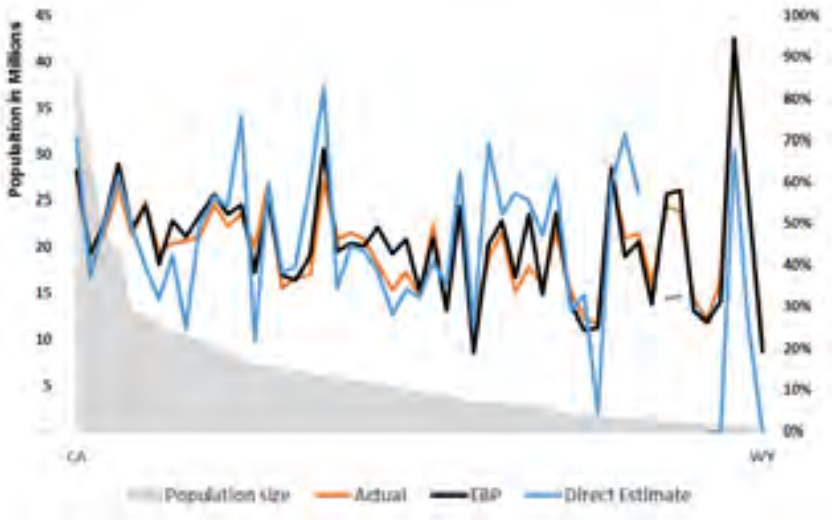
where
$H(\cdot)$ = known cumulative distribution function (cdf),
β = vector of unknown fixed effects,
v_i = random effect specific to the i^{th} area with unknown variance component σ^2.

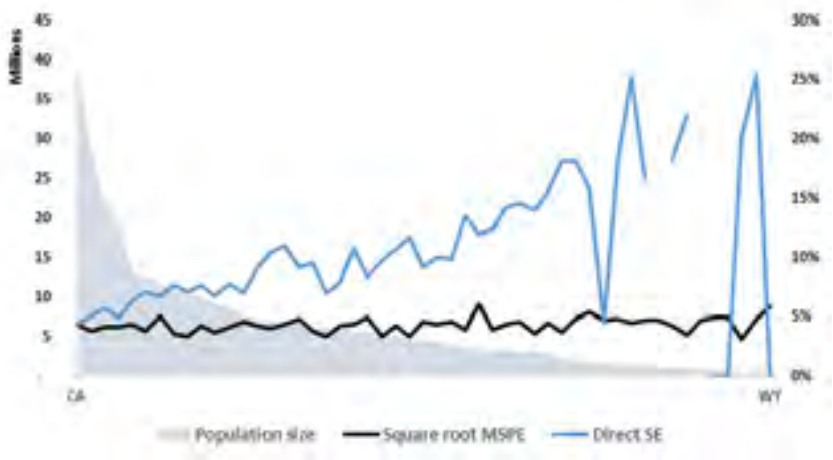
Point estimation and uncertainty

- We fit working model on the data from small sample ($\tilde{s}$) and estimate parameters $(\hat{\beta}, \hat{\sigma}^2)$.
- We propose an *Expectation Maximization* (EM) algorithm based *adjusted maximum likelihood* (AML) approach for model parameter estimation.
- AML avoids boundary and convergence issues faced with MLEs : in 22% of the bootstrap samples, we see $\hat{\sigma}^2 = 0$ or algorithm does not converge using R function `glmer`.
- Due to unknown random effects, EM approach provides $(\hat{\beta}, \hat{\sigma}^2)$, which are then used to predict unit-level responses in s.
  $\hat{Y}_{ij}^{EBP},\ j = 1, \cdots, n_i,\ i = 1, \cdots, m.$
- Using this *Empirical Bayes Prediction* (EBP) approach, we get weighted proportions $\hat{\bar{Y}}_i^{EBP}$, which are area-level point estimates.
- For uncertainty quantification we calculate *Mean Square Prediction Error* (MSPE) based on a parametric bootstrap method, proved to be consistent.
- Square root of $\widehat{\text{MSPE}}(\hat{\bar{Y}}_i^{EBP})$ can be treated as the standard error (SE) of EBP of $\bar{Y}_i$.

Results

Comparison of Direct and EBP estimates and their SE*, with actual values (in %) and populations sizes at state level : all 50 states and DC are shown in image and values of a few important states are in noted in table.





State	Population size	Actual	Direct	EBP	SE (Direct)	SE (EBP)
CA	39 M	61.5	70.5	61.1	4.3	5.2
FL	21 M	47.4	49.2	50.2	5.7	4.7
MD	6 M	60.3	83.2	67.9	7.0	4.5
MT	110 K	35.4		30.5		4.8
SD	895 K	31.7		28.9		5.1
AK	732 K	36.6	0	31.8	0	4.7
DC	670 K	90.9	68.2	95.4	20.2	2.4
WY	578 K	21.9	0	19.1	0	4.5

*Estimated SE of direct estimates (R package `survey`) and square root MSPE of EBPs (parametric bootstrap method using 500 replications) for the 50 U.S. states and DC.

Poverty Mapping: SAE

Selected References

Topic 1: Poverty measurement

[1] J. Foster, J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica, 761–766.

Topic 2: Review Papers

[2] Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation*, 2nd Edition. Wiley.

Topic 3: Microsimulation

[3] Das, S. & Chambers, R. (2024). Small Area Poverty Estimation under Heteroskedasticity, Journal of Survey Statistics and Methodology, 12, 369-403.

[4] C. Elbers, J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality, Econometrica, 71(1):7, 355–364.

[5] W. Gonz´alez-Manteiga, M. J. Lombard´ıa, I. Molina, D. Morales, and L. Santamar´ıa. Bootstrap mean squared error of a small-area eblup, Journal of Statistical Computation and Simulation, 78(5):443–462.

[6] I. Molina and J. N. Rao (2010). Small area estimation of poverty indicators. Canadian Journal of statistics, 38(3), 369–385.

[7] I. Molina, B. Nandram, and J. N. K. Rao (2014). Small area estimation of general parameters with application to29 poverty indicators: A hierarchical Bayes approach. The Annals of Applied Statistics, 8(2):852 – 885.

Selected References

Topic 4: Heteroscedastic mixed models

[8] Jiang and T. Nguyen. Small area estimation via heteroscedastic nested-error regression (2012). Canadian Journal of Statistics, 40:588–603.

[9] Kubokawa, S. Sugasawa, M. Ghosh, and S. Chaudhuri (2016). Prediction in heteroscedastic nested error regression models with random dispersions, Statistica Sinica, 26:465–492.

[10] P. Lahiri and N. Salvati (2023). A nested error regression model with high-dimensional parameter for small area estimation, Journal of the Royal Statistical Society Series B: Statistical Methodology, 85(2):212–239. arxiv preprint

[11] S. Sugasawa and T. Kubokawa (2017). Heteroscedastic nested error regression models with variance functions. Statistica Sinica, 27(3):1101–1123.

Topic 5: Evaluation

[12] Dang, H.A., M. Do, P. Lahiri, M. Gualavisi, D. Newhouse, T. Kilic, P. Lanjouw, and R. Van der Weide, forthcoming, Evaluating alternative approaches to small area poverty estimation with survey and census data, unpublished report.

FGT poverty measures; Foster et al. (1984)

- FGT poverty measures: $F_{\alpha i} = \frac{1}{N_i} \sum_{j=1}^{N_i} F_{\alpha ij}$,
- $F_{\alpha ij} = \left(\frac{z - E_{ij}}{z} \right)^{\alpha} I(E_{ij} < z),\ j = 1, \cdots, N_i;\ \alpha = 0, 1, 2,$
  with $I(E_{ij} < z) = 1$ if $E_{ij} < z$, otherwise, it equals to 0;
- E_{ij}: a suitable quantitative measure of welfare for individual j in small area i,
- z: a fixed poverty line (threshold).

Chen, Y., Lahiri, P. and Salvati. P. (2025) Empirical Best Prediction of Poverty Indicators via Nested Error Regression with High Dimensional Parameters. arxiv preprint .

Application: poverty mapping for Albania

- 2002 Living Standards Measurement Survey (LSMS) data: $3,591$ households
- 374 municipalities: 213 sampled municipalities and 161 out-of-sample municipalities.
- Auxiliary data: 2001 Census, which covers 726,895 households across Albania.
- E_{ij}: household monthly consumption expenditure. Y_{ij}: shifted logarithm transformation of E_{ij}.
- The poverty line: 4,891 Leks per month.

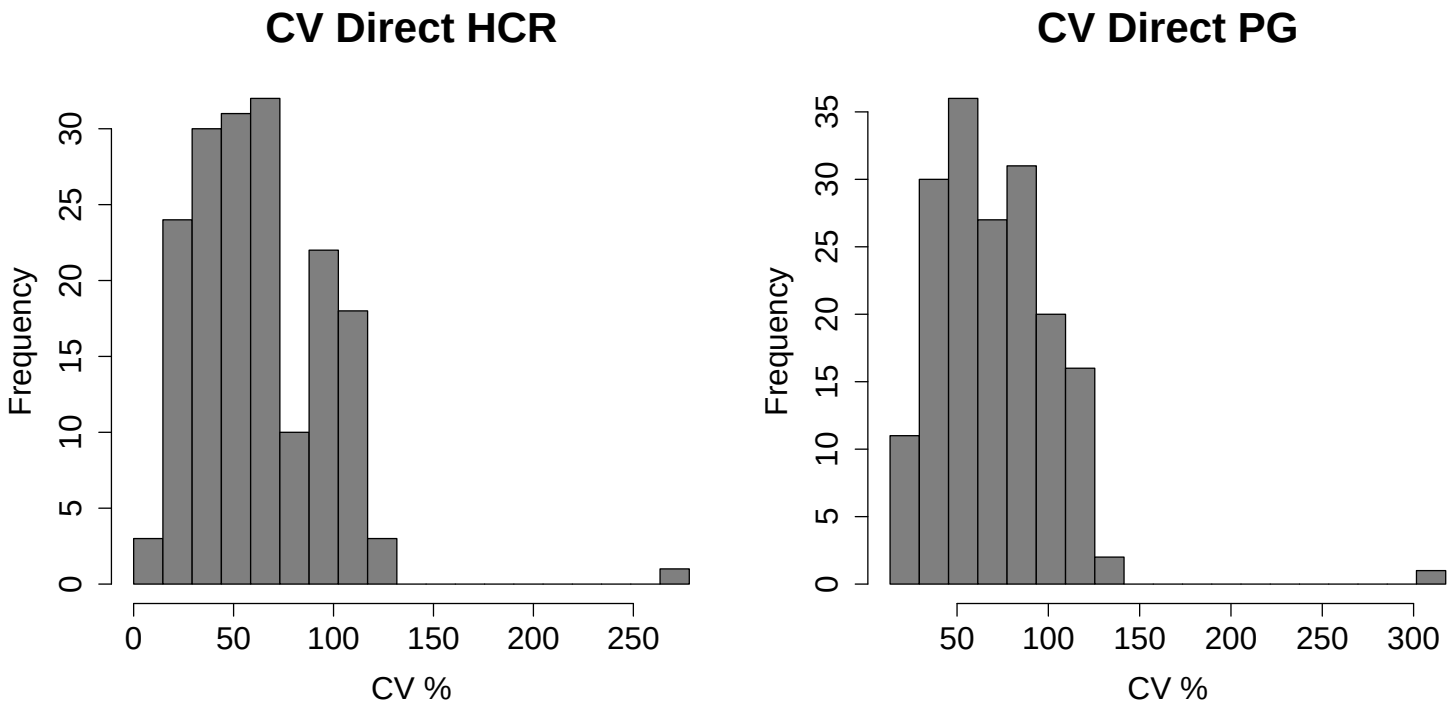


Figure: Histograms of percent estimated coefficient of variation of direct estimates for HCR and PG for all sampled municipalities using 2002 LSMS data.

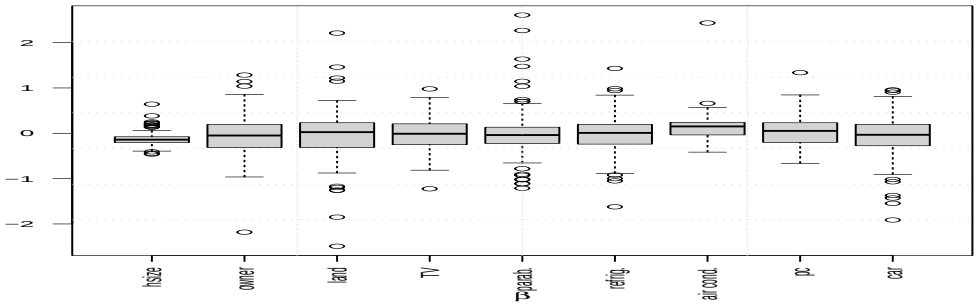


Figure: Distributions of estimated regression coefficients fitted for each of the 213 municipalities of Albania.

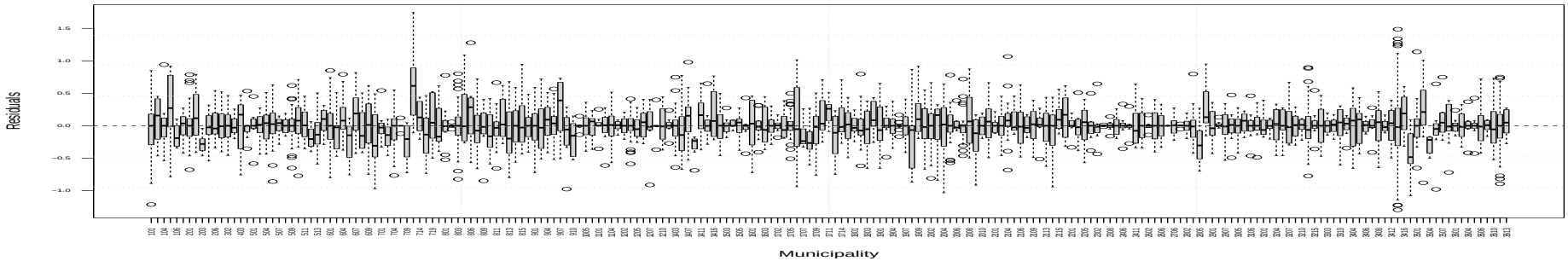


Figure: Distributions of the standardized residuals by municipality.

NER model with high-dimensional parameters

- An extension of the NER model:

$$Y_{ij} = \beta_0 + \mathbf{x}'_{ij}\boldsymbol{\beta_i} + v_i + e_{ij}, i = 1, \cdots, m; j = 1, \cdots, N_i.$$

- β_0 is a common intercept;

- $\boldsymbol{\beta}_i$ is a $p \times 1$ vector of fixed regression coefficients for i^{th} area;

- v_i and e_{ij} are all independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_i^2)$.

- Nested error regression with high dimensional parameter (NERHDP).

Model parameter estimation

- Generalized estimating equations (GEE) with area-specific tuning parameters
- Allows to borrow strength across areas
- When the tuning parameters are known, the model parameters can be consistently estimated

Especially, for out-of-sample areas

- Less synthetic small area estimates compared to existing methods are produced for out-of-sample areas

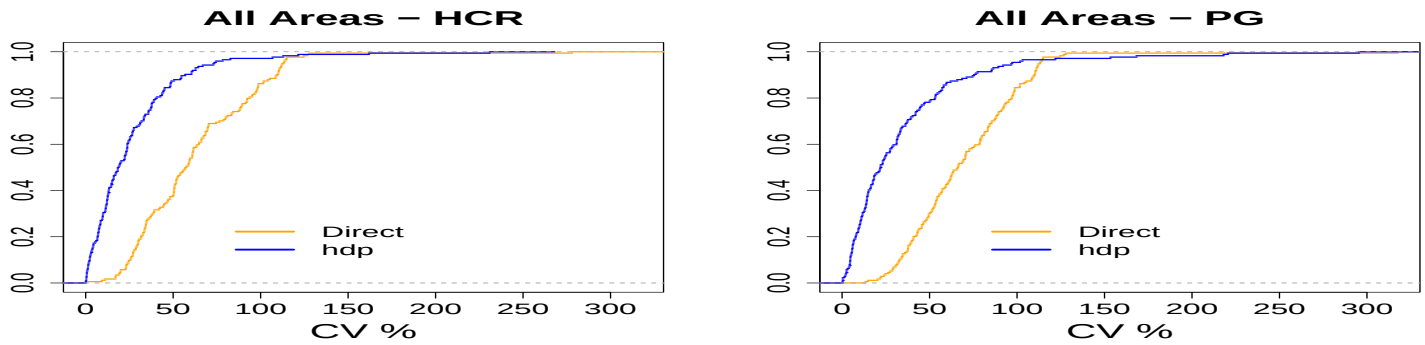


Figure: CVs empirical cumulative density functions for the CLS and direct estimator.

Albanian poverty maps

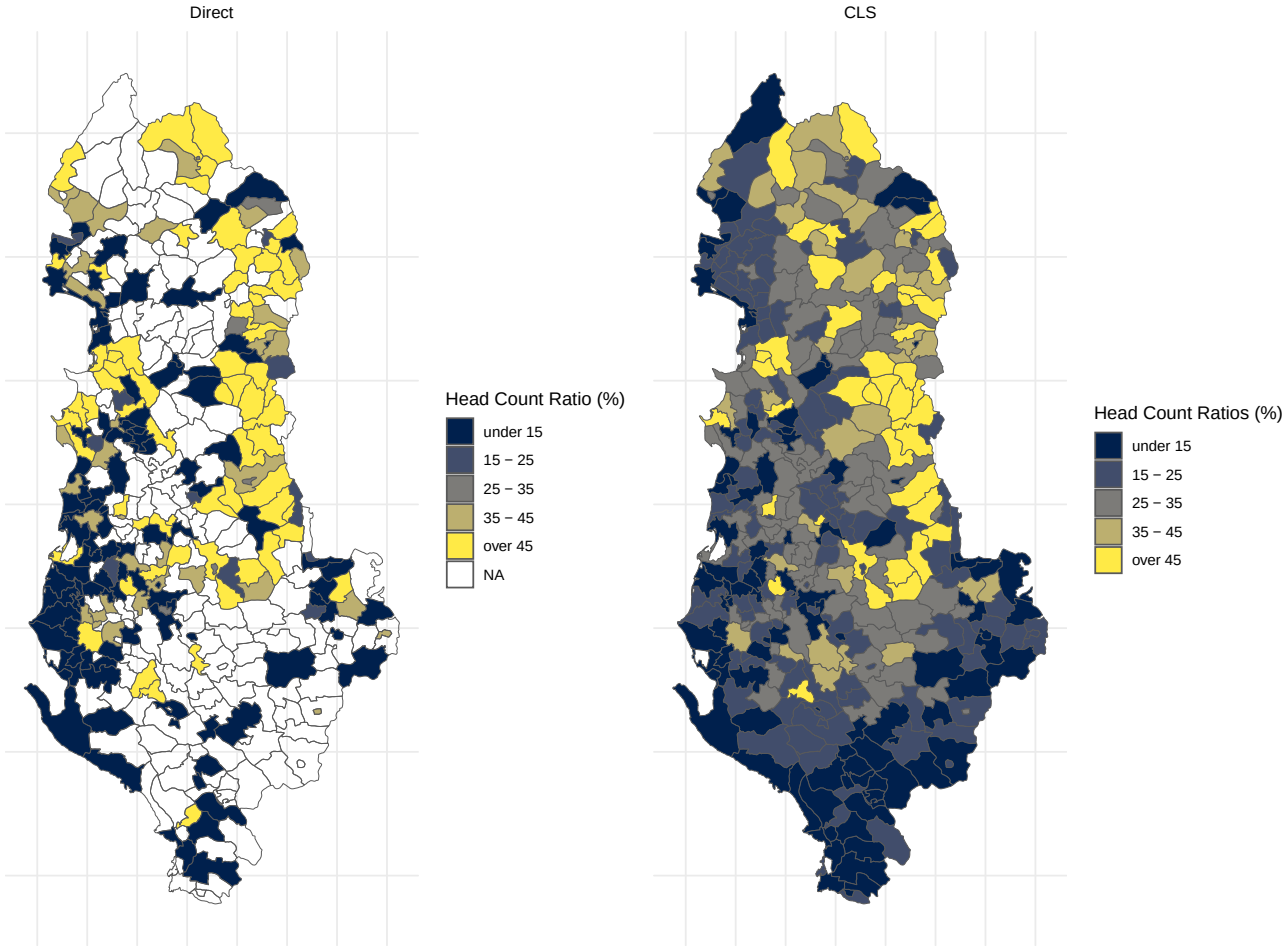


Figure: Municipality-level direct and CLS estimates of headcount ratios in Albania.

Albanian poverty maps

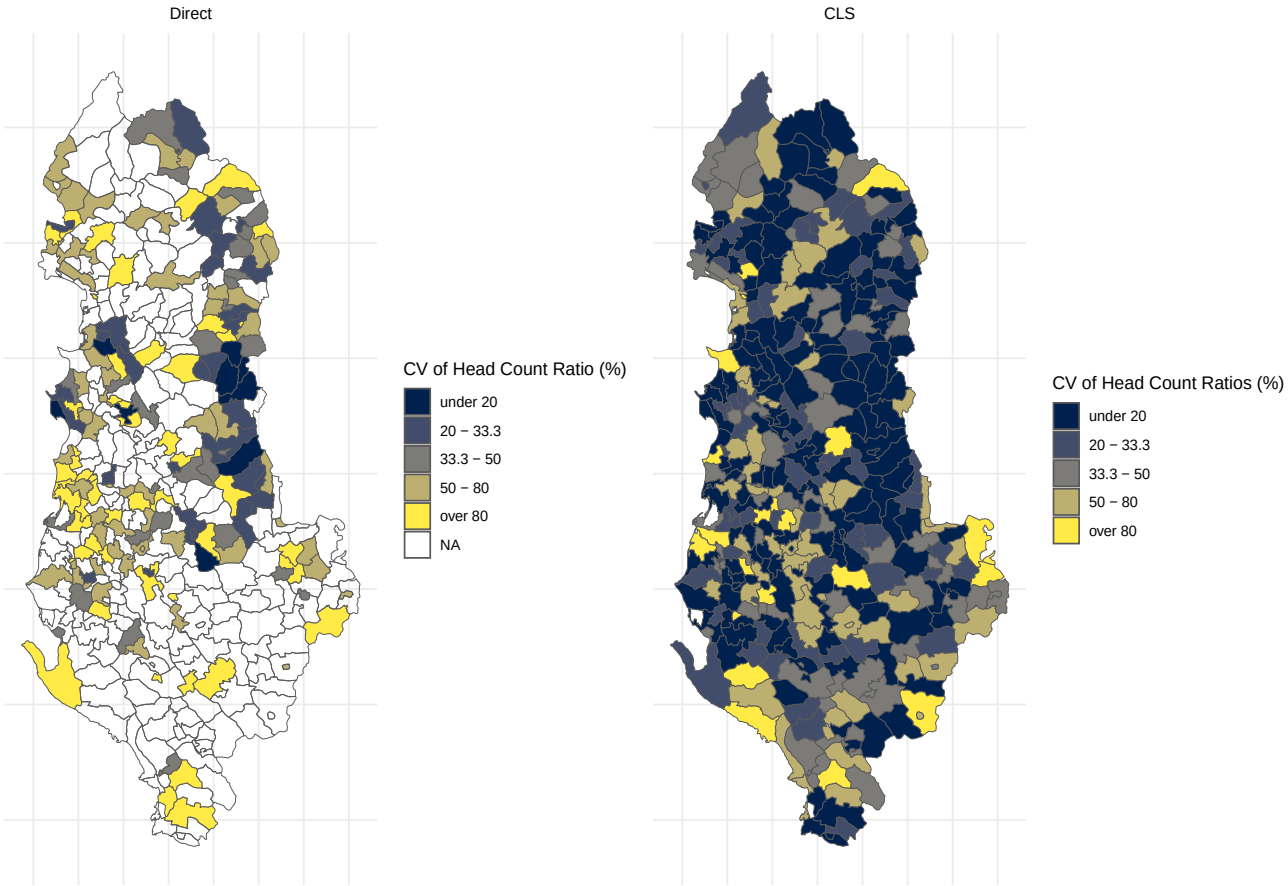


Figure: Municipality-level CV of headcount ratios for direct and CLS estimates in Albania.

Albanian poverty maps

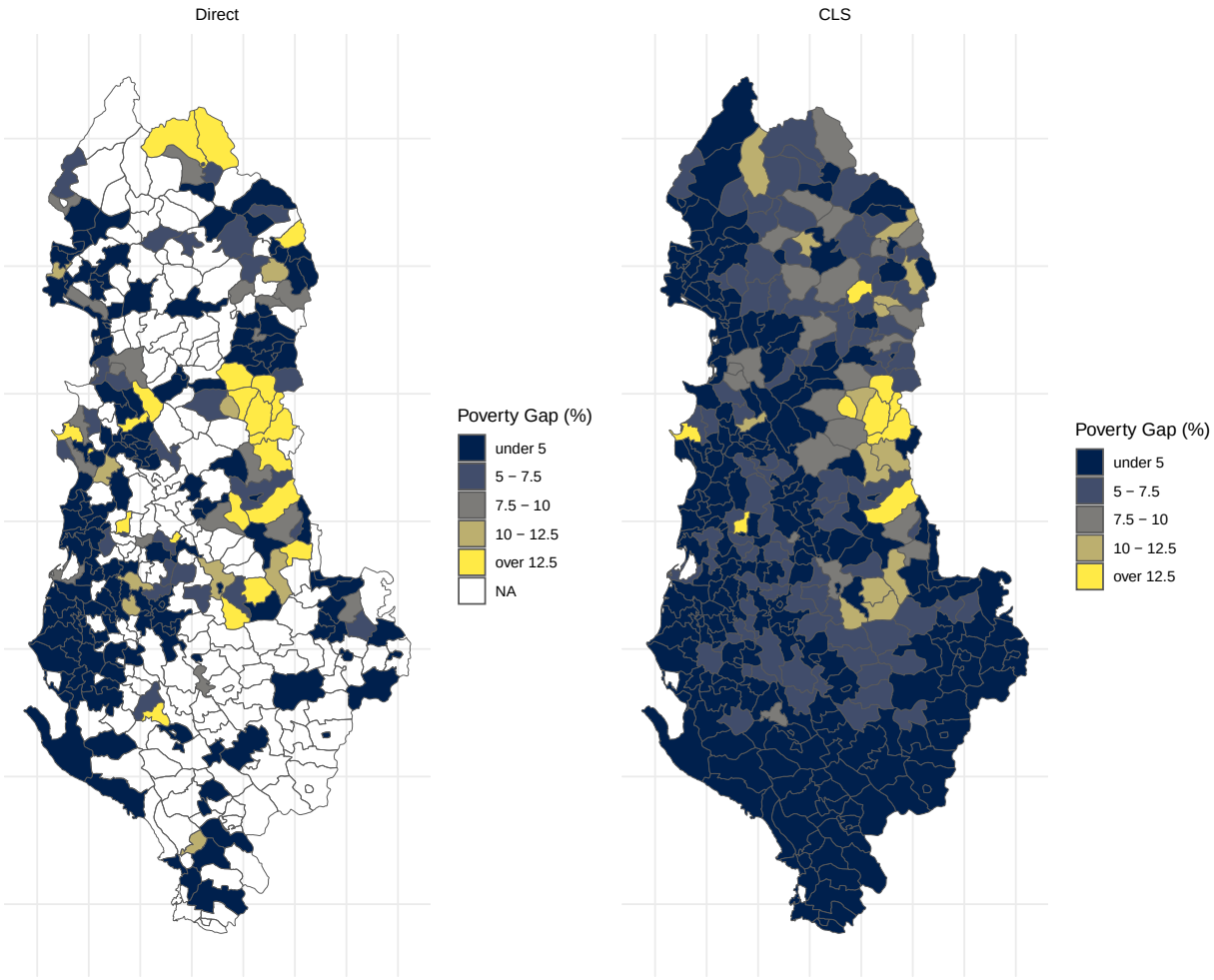


Figure: Municipality-level direct and CLS estimates of poverty gaps in Albania.

Albanian poverty maps

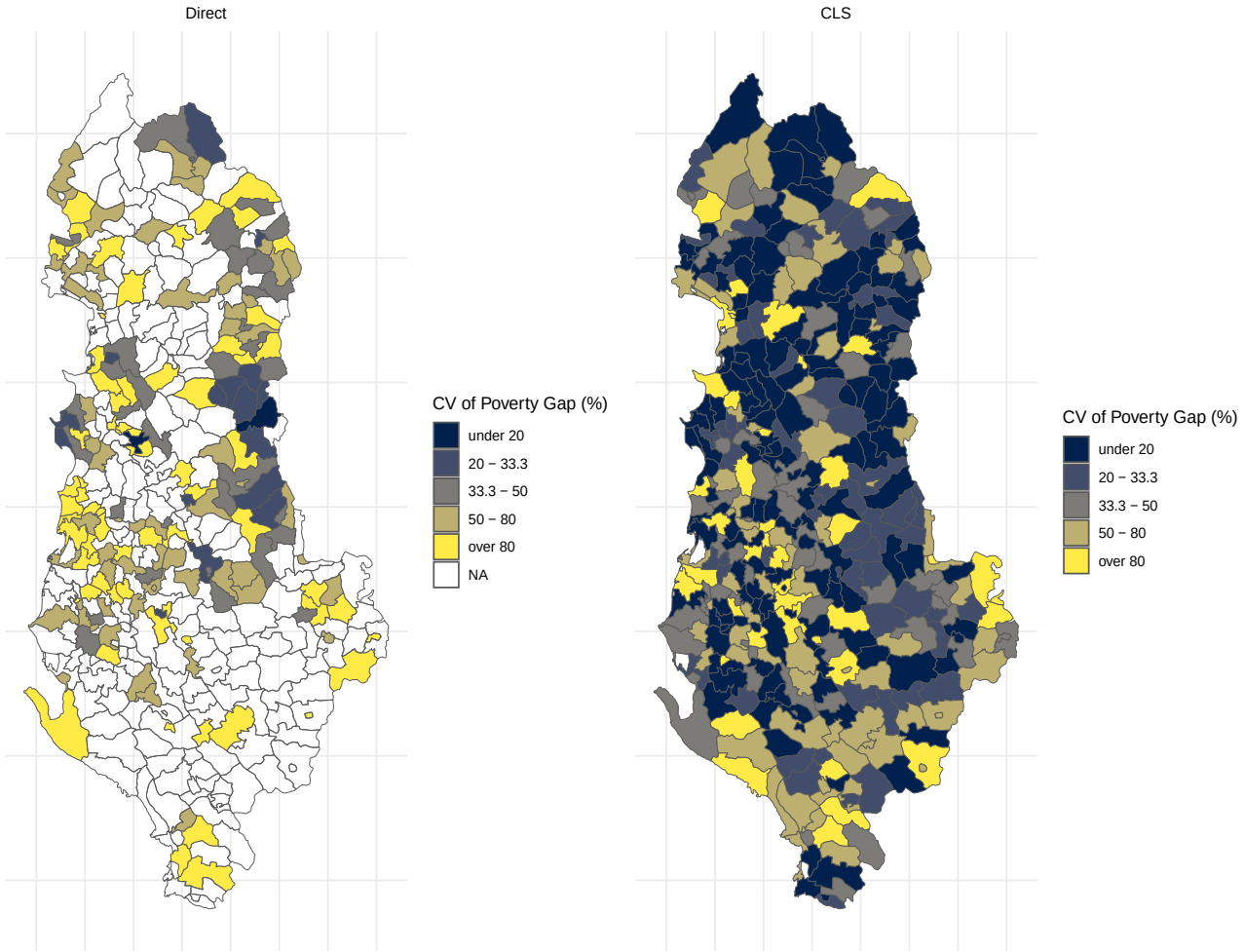


Figure: Municipality-level CV of poverty gaps for direct and CLS estimates in Albania.

Concluding Remarks

- Morris Hansen suggested combining information from different data sources as early as in late 1930's.

- We discussed a few situations where combining information from multiple data sources is needed. For more examples, see Statistics in Transition New Series, Special Issue, Aug 2020 .

- By combining information from multiple data sources using statistical models, we can potentially cut costs and improve efficiency.

- Can AI help?

- Research collaboration between survey organizations and academia.

A Quote from George Box

Statisticians, like artists, have the bad habit of falling in love with their models.

But as George Box said

Essentially, all models are wrong, but some are useful.

Thank You!