Integrating Latent Class Models with Multiple Systems Estimation for Population Counts

Dan Weinberg

Center for Statistical Research & Methodology
U.S. Census Bureau
Joint Work with Tom Mule

August 4, 2026

Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the author and not those of the U.S. Census Bureau.

Motivation

- Individuals may be found in multiple administrative lists.
- Coverage may differ by database and demographics.
- Some individuals are not found on any list.
- Characteristics (e.g., state) may differ by database.

Research Question

Can we estimate population counts by state and demographic group?

Fictional Data

Three person-level databases: {D1, D2, D3}

ID	Age	D1 Found	D2 Found	D3 Found	D1 State	D2 State	D3 State
P001	Adult	Yes	Yes	Yes	State1	State1	State1
P002	Child	Yes	Yes	No	State1	State1	–
P003	Adult	Yes	Yes	No	State1	State2	–

Red indicates cross-database disagreement for the same person.

General Modeling Strategy

- Treat *true state* as a latent class.
- Obtain state proportions from latent class modeling.
- Fit triple-system capture-recapture model to estimate national population totals.
- Combine to obtain state counts.

Modeling Details

Procedure (per state j)

1. **Relabeling into in-/out-of-state.** Recast observed counts into two categories:

   $$\text{In-state} \quad \text{vs.} \quad \text{Out-of-state.}$$

2. **Latent class fit.** Fit a latent class model on the data; obtain posterior membership probabilities for each individual:

   $$\pi_m^{(j)} = P(\text{State} = j \mid \text{observed state labels in record } m).$$

3. **Age-specific aggregation.** For age group $a \in \{\text{Child}, \text{Adult}\}$, aggregate and normalize by the number of records in that age group:

   $$P(\text{State} = j \mid a) \; = \; \frac{1}{|S_a|} \sum_{m \in \mathcal{S}_a} \pi_m^{(j)}.$$

   Here, S_a is the set of people in age group a and $|S_a|$ is its count.

1. **Capture–recapture sizing (national).** Using the posterior-based proportions within age group a, fit a capture–recapture model to estimate the *national* size attributable to state j:

$$n_j(a) \quad \text{for } a \in \{\text{Child}, \text{Adult}\}.$$

Average across T states to obtain national totals:

$$n(\text{Child}) = \frac{1}{T}\sum_{i=1}^{T} n_i(\text{Child}), \qquad n(\text{Adult}) = \frac{1}{T}\sum_{i=1}^{T} n_i(\text{Adult}).$$

2. **Coherence normalization across states.** Posterior-derived probabilities may not sum to 1 within each age group. Apply normalization:

$$P'(\text{state} = j \mid a) \; = \; \frac{P(\text{state} = j \mid a)}{\sum_{i=1}^{T} P(\text{state} = i \mid a)}.$$

3. **Final state counts.** Multiply adjusted probabilities by national totals:

$$n(j, a) \; = \; P'(\text{state} = j \mid a) \; n(a).$$

United States® Census Bureau

Simulation Setup

- Simulate N people in T states with known proportions.
- For each state, divide into children/adults with known proportions.
- People may or may not be captured in each of D databases. Capture rates depend on database and age group.
- If captured, a person's state may have measurement error.
- Measurement error only depends on database.
- Incorrect classifications are placed in other states proportional to state population.

Simulation Setup

- Total population: $N =$1,000,000.
- Number of states: $T = 16$.
- State proportions: Vary from 0.2% to 12%.
- Adult proportions: Vary from 72% to 82%.
- Number of databases: $D = 3$.
- Measurement error by database: $1\%, 2\%, 3\%$.

United States® Census Bureau

Simulation Setup

- Capture probabilities:

Database	Children	Adults
A	0.9	0.9
B	0.1	0.9
C	0.5	0.5

- Database A: High coverage for both groups.
- Database B: Low coverage for children, high coverage for adults.
- Database C: Moderate coverage for both groups.

Latent Class Model: 3 Binary Items, 2 Classes

Model

- Individuals respond to 3 binary manifest items:

$$Y_1,\ Y_2,\ Y_3 \in \{1, 2\}.$$

- Two latent classes:

$$C \in \{1, 2\}, \qquad P(C = c) = \pi_c, \ \ \pi_1 + \pi_2 = 1.$$

- Class–conditional response probabilities:

$$\theta_{jck} = P(Y_j = k \mid C = c), \quad \theta_{jc1} + \theta_{jc2} = 1.$$

Model Fitting

- We use the R package poLCA to fit the model (Linzer and Lewis, 2026).
- We preserve NA values.
- Large state estimates were used as starting points for small states.

Triple–System Estimation (TSE) for Population Size

Model: We fit a main effects *log-linear model* for the 7 observed capture-pattern cell counts:

$$\log(\mu_{abc}) = \lambda + \lambda_A a + \lambda_B b + \lambda_C c,$$

where $a, b, c \in \{0, 1\}$ are indicators of presence in their respective databases.
The fitted count for the unobserved $(0, 0, 0)$ cell is given by

$$\hat{\mu}_{000} = \exp(\lambda)$$

and the estimation for total population is

$$\hat{N} = \sum_{\text{observed cells}} n_{abc} \; + \; \hat{\mu}_{000}.$$

where the n_{abc} are the observed counts.

Model Fitting: We use the R package Rcapture (Rivest and Baillargeon, 2022) to fit the model on the capture pattern counts.

United States® Census Bureau

Results

Table: National Results

Metric	Adults	Children	Total
True Value	762,242	237,758	1,000,000
Estimate	762,193	237,739	999,932
Percent Error (%)	-0.006	-0.008	-0.007

Summary: National totals by age group have little error.

Results

Table: States 1-14 Counts

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
% Error (Adults)	-0.1	-0.1	0.1	0.1	0.0	0.0	0.0	-0.1	0.0	0.1	0.0	-0.1	0.0	0.1
% Error (Children)	0.1	0.0	0.4	-0.2	0.1	-0.1	0.5	-0.4	-0.3	-0.4	0.1	0.5	-0.3	0.0
% Error (Total)	-0.1	-0.1	0.1	0.0	0.0	0.0	0.1	-0.1	0.0	0.0	0.0	0.0	-0.1	0.1
State Share (%)	12	10	10	10	10	9	8	7	6	5	4	3	2.8	2

Summary:

- Adult counts are within 0.1%.
- Child counts are within 0.5%.
- State counts are within 0.1%.

Results

Table: States 15 and 16 Counts

	S15	S16
% Error (Adults)	-0.1	0.6
% Error (Children)	-0.5	0.8
% Error (Total)	-0.2	0.6
State Share (%)	1	0.2

- Counts are still within 1%.
- In our original method, these were close to 5%.
- On investigation, some class-conditional probabilities were estimated close to zero.
- We then used large state estimates as starting points in the EM algorithm for small states, improving the estimates greatly.

Conclusions

- Our method produces accurate estimations at all levels.
- Convergence issues arise for small states, but can be mitigated through specifying starting point of EM algorithm.
- Future work:
  - Improve method for small states.
  - Explore sensitivity of method to modeling assumptions.
  - Develop uncertainty measures.

References

Drew Linzer and Jeffrey Lewis. *poLCA: Polytomous Variable Latent Class Analysis*, 2026. URL https://CRAN.R-project.org/package=poLCA. R package version 1.6.0.2.

Louis-Paul Rivest and Sophie Baillargeon. *Rcapture: Loglinear Models for Capture–Recapture Experiments*, 2022. URL https://CRAN.R-project.org/package=Rcapture. R package version 1.4-4.

Thank you!

Contact: daniel.weinberg@census.gov