Differential Privacy Guarantees in Small Area Estimation

Soumojit Das[1] Jörg Drechsler[2]

Privacy and Public Policy Conference 2026

[1]Washington State University

[2]IAB Germany

Introduction

Statistical Agencies Face a Dilemma

- Agencies release estimates for small geographic areas (counties, school districts)
- Small area estimation (SAE) uses hierarchical models that **pool information** across areas
- Current practice: ad-hoc disclosure review with **no formal privacy guarantees**

The Operational Challenge
Staff at research data centers struggle to assess disclosure risk for SAE outputs.

Example: SAIPE estimates are used in allocating millions of dollars in federal funding annually.

The Fay-Herriot Model

Sampling model:

$$y_i \mid \theta_i \sim N(\theta_i, \sigma^2_{y_i})$$

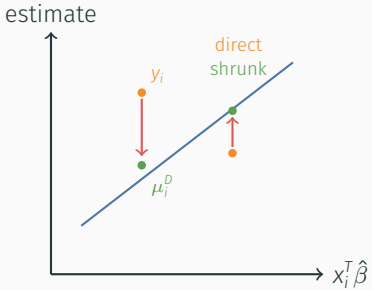
Linking model:

$$\theta_i \sim N(x_i^T \beta, \sigma^2_v)$$

Posterior mean (EBLUP):

$$\mu_i^D = (1 - B_i) y_i + B_i \cdot x_i^T \hat{\beta}$$

where $B_i = \sigma^2_{y_i} / (\sigma^2_{y_i} + \sigma^2_v)$



Key insight: Small areas (high $\sigma^2_{y_i}$) shrink *more* toward the model prediction
The model itself introduces randomness — can we quantify this as privacy?

Our Question & Roadmap

Does the Bayesian Fay-Herriot model provide inherent differential privacy guarantees — without adding any noise?

Today I'll show you:

1. **Background:** DP refresher — pure DP, Rényi DP, and zCDP
2. **Main Result:** The Bayesian Fay-Herriot mechanism satisfies ρ-zCDP with

$$\rho = S_y^2/(2\sigma_{\text{min}}^2)$$

3. **ACS PUMS Empirical Analysis:** Ground-truth validation on 2,462 PUMAs
4. **Implications:** What this means for practitioners

Background

Differential Privacy in 60 Seconds

Intuition:

- Changing one person's data shouldn't change the output "too much"
- Parameter ε measures "how much"
- Smaller ε = more private

(ε, δ)-DP:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

for all outputs S, all neighboring D, D'

Pure DP: when $\delta = 0 \implies$ pure DP — which is a stricter and often untenable notion for practitioners

Why This Fails for Us

Gaussian mechanisms have unbounded support $\Rightarrow$ no finite ε works for pure DP.

We need a relaxation that handles Gaussian noise naturally.

The Relaxations We Use: RDP and zCDP

Rényi Divergence of order $\alpha > 1$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right]$$

zero-Concentrated DP: $\mathcal{M}$ satisfies ρ-zCDP if $D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \rho \cdot \alpha$ for all $\alpha > 1$

Composition (Key Advantage)

$\rho_{\text{total}} = \sum_{i=1}^{m} \rho_i$

Converts to (ε, δ)-DP:

$$\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$$

$\Rightarrow \varepsilon \propto \sqrt{m}$, not m

Why zCDP?

- Tight composition for our case of application — multi-area release
- Clean conversion to (ε, δ)-DP

The Posterior Is Already Randomized

$$\tilde{\theta}_i \sim N(\mu_i^D, \sigma_i^2)$$

Key Insight
Drawing from the posterior adds noise centered at the data-dependent mean — **structurally identical to the Gaussian mechanism in DP**, but the noise variance is determined by the model, not a privacy budget.

Practical note:

- Standard practice releases point estimates (posterior mean) $\Rightarrow$ deterministic, no privacy
- Our proposal: release a *single posterior draw* instead
- Utility cost is negligible — just reflects honest uncertainty

Assumptions for Formal Guarantees

Our analysis requires:

1. **Variance components known or fixed**
   - σ_y^2 (sampling variance) and σ_v^2 (model variance) treated as fixed
   - In practice: use plug-in estimates — bounds remain valid (Proposition 3)

2. **Auxiliary information is public**
   - Covariates x_i from Census tables, administrative records
   - Not derived from the confidential survey data

3. **Release posterior draw, not posterior mean**
   - Point estimates are deterministic $\Rightarrow$ infinite ε
   - Single draw provides randomization for DP

These are standard SAE assumptions; we make them explicit for DP analysis.

Main Results

Theorem 1: Expected Privacy Loss

Theorem 1 (Expected Privacy Loss)
For the Bayesian FH mechanism releasing $\tilde{\theta}_i \sim N(\mu_i^D, \sigma_i^2)$, the expected privacy loss equals the KL divergence:

$$\mathbb{E}[\mathcal{L}(\tilde{\theta}_i)] = D_{KL}(P_D \| P_{D'}) = \frac{(\mu_i^D - \mu_i^{D'})^2}{2\sigma_i^2}$$

Bounding the KL divergence:

- The posterior mean shift $|\mu_i^D - \mu_i^{D'}| \leq S_y$ (sensitivity)
- Using minimum posterior variance $\sigma^2_{\min} = \min_i \sigma_i^2$:

$$\boxed{D_{KL}(P_D \| P_{D'}) \leq \frac{S_y^2}{2\sigma^2_{\min}}}$$

Key insight: For Gaussians with equal variance, $D_\alpha = \alpha \cdot D_{KL} \Rightarrow$ this bound extends to all Rényi orders

Theorem 2: zCDP Guarantee for Fay-Herriot

Theorem 2 (Rényi DP Guarantee)
The Bayesian Fay-Herriot mechanism satisfies $(\alpha, \varepsilon_\alpha)$*-RDP for any* $\alpha > 1$*:*

$$\varepsilon_\alpha \leq \frac{\alpha S_y^2}{2\sigma^2_{\min}}$$

Since the bound is linear in α*, the mechanism satisfies* ρ*-zCDP with* $\rho = S_y^2/(2\sigma^2_{\min})$.

Sensitivity under complex surveys (Hájek estimator with weights w_{ir}):

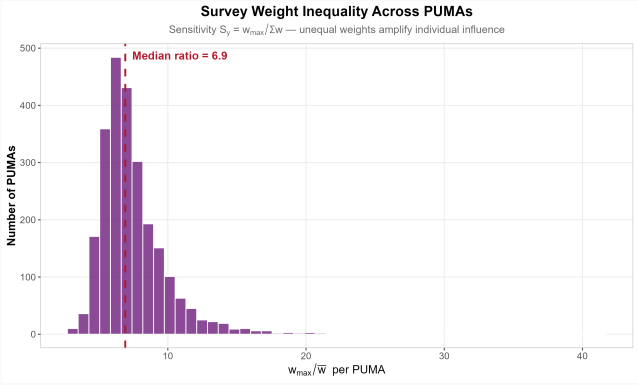
$$S_{y_i} = \frac{w_{\max,i}}{\sum_r w_{ir}}$$

Special case (equal weights / SRS): $S_y = 1/n$

Key insight: Privacy improves with sample size ($\varepsilon \propto 1/n$) and with more shrinkage (larger $B_i \Rightarrow$ smaller ρ_i)

ACS PUMS Empirical Analysis

ACS PUMS: Data & Sensitivity Under Complex Surveys



Data: 2022 ACS 1-year PUMS

3.19M records, 2,462 PUMAs

Binary outcome: poverty indicator

Fay-Herriot model:

$\hat{\sigma}_v^2$ (REML)	2.73×10^{-3}
Median shrinkage B	0.101
Sample sizes	368–3,911

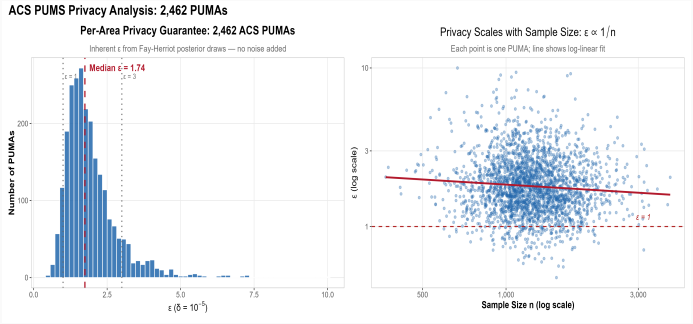
Sensitivity Matters

$S_{y_i} = w_{\max,i} / \sum w_{ir}$

Median $w_{\max}/\bar{w} = 6.9$

$\Rightarrow$ Most influential person in a typical PUMA has $\sim 7\times$ the average survey weight

ACS PUMS: Privacy Results



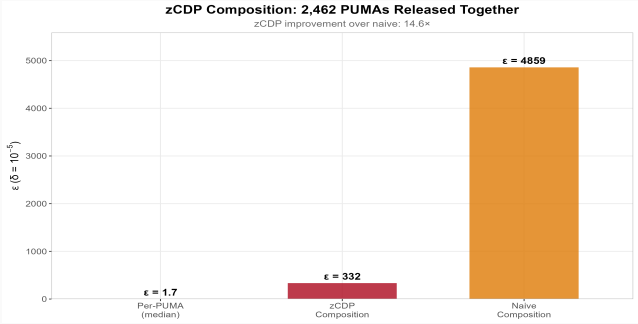
Per-PUMA privacy ($\delta = 10^{-5}$):

Statistic	ε
Minimum	0.48
10th percentile	1.10
Median	**1.74**
90th percentile	3.08
Maximum	10.0

- **63.6%** of PUMAs: $\varepsilon < 2$
- **89.0%** of PUMAs: $\varepsilon < 3$

Confirms: $\varepsilon \propto 1/n$ (right panel)

zCDP Composition: Releasing All 2,462 PUMAs



Composition under zCDP:

$$\rho_{\text{total}} = \sum_{i=1}^{m} \rho_i = 230$$

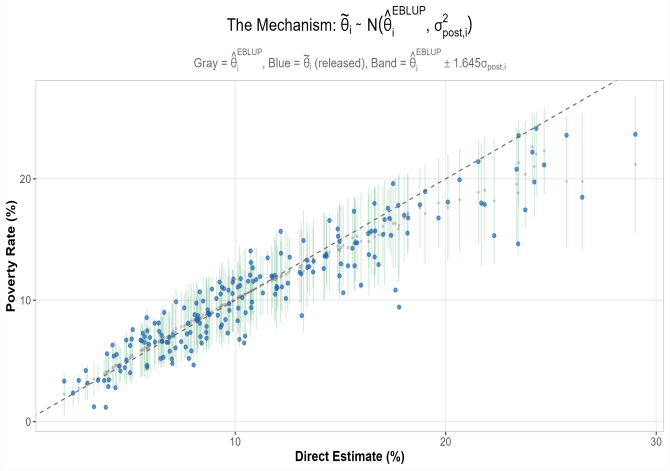
$$\varepsilon_{\text{total}} = \rho + 2\sqrt{\rho \ln(1/\delta)}$$

Method	ε
Per-PUMA (median)	1.7
zCDP composed	**332**
Naive $\sum \varepsilon_i$	4,859

$\Rightarrow$ **14.6× improvement** over naive

This is the $\sqrt{m}$ scaling: $\sqrt{2462} \approx 50$, so composed ε grows much slower

The Mechanism in Action



The mechanism:

$$\tilde{\theta}_i \sim N\left(\hat{\theta}_i^{\text{EBLUP}}, \sigma^2_{\text{post},i}\right)$$

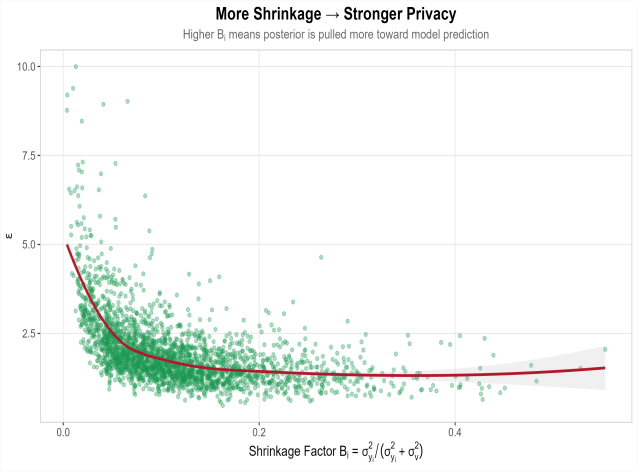
What we release:
One draw (blue) per PUMA

What we don't release:
The EBLUP (gray) — deterministic, no privacy

Band: $\hat{\theta}_i \pm 1.645\sigma_{\text{post},i}$

The noise *is* the privacy guarantee — and it reflects posterior uncertainty

More Shrinkage ⇒ Stronger Privacy



Why shrinkage helps:

$$\rho_i = \frac{S_y^2}{2\sigma_v^2 \cdot B_i}$$

Larger B_i (more shrinkage) ⇒ smaller ρ_i

Intuition:

More shrinkage → posterior pulled further from the raw data → harder to detect any ONE person's contribution

Conclusion

What We Showed

Theoretical result:

- The Bayesian Fay-Herriot mechanism satisfies ρ-zCDP with $\rho = S_y^2/(2\sigma_{\min}^2)$
- First formal DP analysis of Bayesian small area estimation

Empirical validation (ACS PUMS, 2,462 PUMAs):

- Median per-area $\varepsilon = 1.74$ ($\delta = 10^{-5}$); 89% of PUMAs with $\varepsilon < 3$
- Composed $\varepsilon = 332$ for all PUMAs
- Survey weight inequality matters: $S_y = w_{\max}/\sum w$, not $1/n$

We are NOT claiming:

- "Privacy for free" for all SAE — continuous/unbounded outcomes have weak guarantees
- Unit-level models are covered (area-level Fay-Herriot only)
- The composed ε is "small" — but it's quantifiable and didn't require adding noise

Agencies can quantify the inherent privacy of existing SAE releases

Thank You

Questions?

Soumojit Das
soumojit.das@wsu.edu

Backup Slides for Q&A

Notation Reference

Data & Model

D, D'	Neighboring databases
y_i	Direct estimate (area i)
x_i	Public covariates
θ_i	True area parameter
σ_y^2	Sampling variance
σ_v^2	Model variance
B	Shrinkage: $\frac{\sigma_y^2}{\sigma_y^2+\sigma_v^2}$

Posterior & Privacy

μ_i^D	Posterior mean under D
σ_i^2	Posterior variance (area i)
$\tilde{\theta}_i$	Posterior draw (output)
S_y	Sensitivity of y_i
$\sigma^2_{\min}$	$\min_i \sigma_i^2$
R_y	Range of outcome

Key Relationships

Mechanism	General sensitivity	SRS case
$\tilde{\theta}_i \sim N(\mu_i^D, \sigma_i^2)$	$S_y = \frac{w_{\max}}{\sum w}$	$S_y = \frac{1}{n}$

Why Pure ε-DP Cannot Be Achieved

The problem: Gaussian distributions have unbounded support

Likelihood ratio for Gaussians:

$$\frac{f_D(\tilde{\theta})}{f_{D'}(\tilde{\theta})} = \exp\left\{\frac{\tilde{\theta}(\mu^D - \mu^{D'})}{\sigma^2} - \text{const}\right\}$$

As $\tilde{\theta} \to \pm\infty$, the ratio is **unbounded** $\Rightarrow$ no finite ε works

Intuition:
With tiny probability, the posterior can output *any* value. For extreme outputs, likelihood ratios explode.

Solution
Use **zCDP/RDP**:
Bounds *expected* privacy loss (KL divergence), not worst-case

ACS PUMS Analysis: Methodology

Data Source:

- 2022 ACS 1-year PUMS from IPUMS USA
- 3.19 million person records across 2,462 PUMAs
- PUMAs: $\geq$100,000 population, clean geographic alignment

For each PUMA, we compute:

1. **Direct estimate:** Weighted poverty rate using person weights (PWGTP)
2. **Sampling variance:** Successive difference replication with 80 replicate weights
3. **Weight statistics:** $w_{\max}$, $\sum w$ for exact sensitivity

Model: Intercept-only Fay-Herriot via REML (`sae::eblupFH`)

$$y_i = \mu + v_i + e_i, \quad v_i \sim N(0, \hat{\sigma}_v^2), \quad e_i \sim N(0, \sigma_{y_i}^2)$$

Key advantage: Exact microdata $\Rightarrow$ ground-truth sensitivity $S_{y_i} = w_{\max,i} / \sum w_{ir}$

Why Not Release Posterior Mean?

The issue:

- Posterior mean is deterministic $\Rightarrow$ technically infinite ε
- No privacy guarantee without randomization

Our proposal: Release a single posterior draw instead

Utility cost:

- Added variance = posterior variance σ_i^2
- This reflects *honest uncertainty* that point estimates suppress
- Negligible impact on downstream inference

Alternative: Add Gaussian noise calibrated to σ^2_{min} (equivalent)

Proposition 3: Robustness to Variance Estimation

Proposition (Robustness)
Let $\hat{\sigma}_v^2 = \sigma_v^2(1+\eta)$ with relative error η. Then:

$$\left| \frac{\hat{\rho} - \rho^*}{\rho^*} \right| \leq \frac{B \cdot |\eta|}{1 - |\eta|}$$

where $B = \sigma_y^2/(\sigma_y^2 + \sigma_v^2)$ is the shrinkage factor.

Corollary: When model variance dominates ($\sigma_v^2 \gg \sigma_y^2$), shrinkage $B \to 0$, and ρ becomes insensitive to variance estimation errors.

Practical example:

- Typical SAE: $B \approx 0.1$
- 20% variance error ($|\eta| = 0.2$)
- $\Rightarrow$ only $0.1 \times 0.2/0.8 = 2.5\%$ error in ρ

Comparison to Census 2020 TopDown

	Census TopDown	Our Approach
Total ε	≈ 17	inherent
Mechanism	Intentional noise	Model uncertainty
Scope	All geographic levels	SAE outputs only
Modification	Required	Minimal (draws vs. means)

Key differences:

- We analyze *existing* methods; Census *modified* outputs
- Not directly comparable budgets (different queries, different data)
- Complementary perspectives on statistical disclosure control

Our contribution: Formal guarantees for SAE without changing practice

Unit-Level Models

This paper: Area-level (Fay-Herriot) models only

Unit-level models:

- Sensitivity analysis is more complex
- Individual covariates create additional privacy leakage channels
- Posterior depends on individual-level data in more intricate ways

Future work:

- Extend framework to nested error regression models
- Analyze Battese-Harter-Fuller model
- Same principles should apply (posterior sampling provides randomization)

Conjecture: Unit-level models may offer *stronger* inherent privacy due to additional model layers

Simulation Study Highlights

Bound validity (Goal 1):

- KL bounds valid across all 80 configurations
- Tight when σ_v^2/σ_y^2 is large
- Conservative when shrinkage is high

Outcome range (Goal 3):

- Binary: $\varepsilon \approx 0.07$ ($n = 100$)
- $R_y = 10$: $\varepsilon \approx 0.69$
- $R_y = 100$: $\varepsilon \approx 7.8$

Robustness (Goal 6):

- 8,000 simulations with estimated σ_v^2
- Bound validity rate >98%
- Safe to use plug-in estimates

Weight impact (Goal 5):

- $\varepsilon \propto w_{\max}$
- Equal-probability designs: best privacy
- For $\varepsilon \leq 1$: need $w_{\max} \lesssim 0.15$

Full simulation results available in the paper