PRIVIC in Practice: Evaluating Geolocation Privacy on Mecklenburg County Voter Data

Huong Tran
University of North Carolina at Charlotte
Charlotte, NC, USA
htran55@charlotte.edu

Liyue Fan
University of North Carolina at Charlotte
Charlotte, NC, USA
liyue.fan@charlotte.edu

ABSTRACT

The increasing availability of geolocation data in administrative records—such as voter rolls, utility usage logs, and public service requests—poses substantial privacy risks when these data are shared for research or policy analysis. In this study, we apply the existing PRIVIC framework—a mechanism that alternates between the Blahut-Arimoto rate distortion algorithm and Iterative Bayesian Update—to real-world dataset of Mecklenburg County, North Carolina voter addresses, discretized on a uniform spatial grid. We detail the end-to-end workflow for sampling partial observations, configuring the geo-indistinguishability parameter, and iteratively refining both the obfuscation channel and the estimated prior distribution. We discuss strategies for choosing distortion-privacy trade-off parameters, sampling fractions, and convergence criteria, and provide practical guidelines for deploying PRIVIC in similar administrative-data sharing scenarios. Our application demonstrate how PRIVIC can be improved and readily adopted to enforce formal privacy guarantees while retaining meaningful spatial patterns in sensitive datasets.

CCS Concepts

• Security and Privacy → Location Privacy; • Security and Privacy → Privacy preserving protocols; • Information systems → Data anonymization and sanitization; • Information systems → Geographic information systems.

Keywords

Location privacy, Geo-indistinguishability, PRIVIC, Blahut-Arimoto algorithm, Iterative Bayesian Update; voter registration data.

1. INTRODUCTION

The widespread collection and sharing of fine-grained geolocation data from administrative sources—such as voter registrations records, utility meter readings, and ride-hail logs—has provided new opportunities for researchers and public-policy analysis. Even though public datasets containing home addresses are often converted into latitude/longitude coordinates for analysis, this transformation does not guarantee anonymity. Adversaries may cross-reference GPS data with auxiliary sources (e.g., social media check-ins, property tax records) or apply machine-learning techniques to re-identify individuals, infer daily routines, or reveal sensitive attributes. Real-world incidents, including doxing and targeted harassment, underscore the urgent need for robust privacy protections in spatial data sharing. Ensuring strong privacy guarantees for spatial data is thus both scientifically and challenging and society crucial

Directly publishing raw latitude and longitude values thus fails to prevent inference and re-identification attacks. Even minimal coordinate disclosure can allow attackers to pinpoint a person's home, workplace, or their frequently visiting places. To address these vulnerabilities, geo-indistinguishability—a variant of differential privacy for location data—adds calibrated noise to each point so that any two locations within a given radius become statistically indistinguishable.

In this study, we leverage the existing PRIVIC framework to enforce geo-indistinguishability on a real-world administrative dataset: the Mecklenburg County, North Carolina voter registration roster. PRIVIC alternates between (i) the Blahut–Arimoto rate-distortion algorithm, which derives a channel that minimizes mutual information for a specified distortion bound, and (ii) an Iterative Bayesian Update, which refines the estimated location prior from the noisy outputs. By discretizing the county into a uniform spatial grid, sampling a fixed fraction of voter locations across sequential cycles, and exploring two privacy-utility trade-off settings ($\beta = 0.5$ and $\beta = 1.0$), we demonstrate a reproducible workflow for applying PRIVIC to large-scale administrative data.

The rest of this paper is organized as follows. Section 2 listed background works in location privacy and geo-indistinguishability. Section 3 describes the Mecklenburg County dataset, grid construction, and the algorithmic details of PRIVIC. Section 4 outlines our experimental design, parameter selection, and convergence criteria. Section 5 discusses practical insights, trade-offs between privacy and utility, and guidelines for deployment. Finally, Section 6 concludes with key lessons learned and directions for future applications of PRIVIC in administrative data sharing.

2. BACKGROUND

2.1 Differential Privacy

Differential privacy (DP) provides a formal guarantee that the inclusion or exclusion of any single individual's data has a limited effect on the output of a computation [1]. Formally, a randomized mechanism $\mathcal{M}$ over datasets $\mathcal{D}$ satisfies ε-DP if, for all adjacent datasets D, D' differing in one record, and for all measurable subsets S of outputs:

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

Cynthia Dwork and Aaron Roth's book [1] is the canonical reference on DP, offering both theoretical foundations and practical mechanisms, which is why we adopt their definitions and notation.

2.2 Local Differential Privacy

Local differential privacy (LDP) strengthens DP by enforcing the privacy guarantee at the data-collection point, before any centralized aggregation [2]. Under ε-LDP, each user applies a randomized algorithm $\mathcal{R}$ to her own datum x, ensuring

$$\Pr[\mathcal{R}(x) = y] \leq e^{\varepsilon} \Pr[\mathcal{R}(x') = y]$$

for all x, x' and possible outputs y. Kasiviswanathan et al. [2] introduced the formal model of LDP, and Erlingsson et al.'s RAPPOR system [3] demonstrated its practical viability in large-scale telemetry collection, which motivates our focus on per-record privatization.

2.3 Geo-Indistinguishability

Standard DP does not account for the continuous, metric structure of location data. Geo-indistinguishability adapts DP to the two-dimensional plane by bounding privacy loss as a function of geographic distance d [4]. A mechanism $\mathcal{M}$ satisfies (ε, d)-geo-indistinguishability if for all locations x, x'

$$\Pr[\mathcal{M}(x) = y] \leq e^{\varepsilon * d(x,x')} \Pr[\mathcal{M}(x') = y].$$

Andrés et al.'s CCS 2013 paper [4] is the foundational work on geo-indistinguishability; we follow their formulation to enforce distance-based noise.

2.4 PRIVIC: Privacy-preserving Incremental Collection

Standard rate-distortion methods such as the Blahut–Arimoto (BA) algorithm compute an optimal channel given a known prior distribution [5]. However, in practice the true prior is unknown and must be learned from obfuscated data. PRIVIC ("PRIVacy-preserving Incremental Collection") alternates between:

1. **BA step:** compute the minimal-mutual-information channel $C^{(t)} = BA(\theta^{(t-1)}, c_0, \beta, \delta_{BA})$ for distortion parameter β and initial noise bound c_0.
2. **Data collection:** users sample new true locations x^t and report noisy outputs $y^{(t)}$ via $C^{(t)}$.
3. **IBU step:** perform an Iterative Bayesian Update $\mu^{(t)} = IBU(C^{(t)}, \theta^{(t-1)}, q^{(t)}, \delta_{IBU})$ to obtain a refined estimate of the prior.
4. **Aggregate priors:** combine $\mu^{(t)}$ with $\theta^{(t-1)}$, weighted by sample counts, to form $\theta^{(t)}$.

Shokri & Theodoridis's PETS 2019 paper [5] introduced the core BA + IBU cycle and proved convergence under mild sampling assumptions. The very recent PoPETs 2024 article "PRIVIC: A privacy-preserving method for incremental collection of location data" [6] extends this analysis, showing formally and empirically that each iteration brings the estimated prior closer to the true distribution and that the resulting mechanism satisfies geo-indistinguishability with an elastic metric.

3. RELATED WORKS

This section surveys prior efforts in applying (geo-)differential privacy to location data, developing incremental or adaptive privacy mechanisms, and using Bayesian or iterative techniques to estimate priors from noisy observations. We then contrast these with our empirical deployment of PRIVIC on Mecklenburg County voter data.

3.1 One-shot Geo Differential Privacy Mechanism

Early work on geo-differential privacy injects noise in a single step, assuming a known prior or uniform distribution over the domain. Andrés et al. [4] introduced geo-indistinguishability, a metric-based variant of differential privacy that scales privacy loss with Euclidean distance, and demonstrated the planar Laplace mechanism to guarantee (ε, d)-privacy under a fixed prior. Subsequent improvements, such as Bordenabe et al. [9], derived optimal geo-indistinguishable noise distributions tailored to arbitrary priors. However, these approaches require knowing (or assuming) the true prior distribution in advance and apply noise in a single "one-shot" release, without any mechanism to refine their prior estimate from observed data.

3.2 Bayesian and Iterative Prior-Estimation Techniques

Several privacy systems employ Bayesian inference or iterative estimation to recover distributions under privacy constraints. PrivBayes, introduced by Zhang et al. [7], constructs a differentially private Bayesian network over high-dimensional tabular data via an iterative posterior-update procedure. It first identifies a structure of correlated attributes, then repeatedly refines the network's conditional distributions by drawing noisy counts and applying posterior inference to approximate the true joint distribution. PrivBayes demonstrates strong utility on complex datasets while satisfying ε-DP by carefully calibrating noise to each update. Although PrivBayes excels at modeling arbitrarily high-dimensional data, it neither incorporates spatial distance metrics nor provides an incremental, cycle-based mechanism to refine the obfuscation channel alongside the prior. Instead, it targets a single, static data release, limiting its applicability to streaming or online location-collection scenarios.

3.3 Our Approaches

Unlike PrivBayes and other one-shot Bayesian methods, our work applies PRIVIC to Mecklenburg County voter data in an incremental fashion—jointly learning the location prior and deriving a geo-indistinguishable channel across different collection cycles. We explore practical choices of sampling fractions, distortion parameters (β), and convergence criteria, delivering deployment guidelines that PrivBayes and similar frameworks do not address.

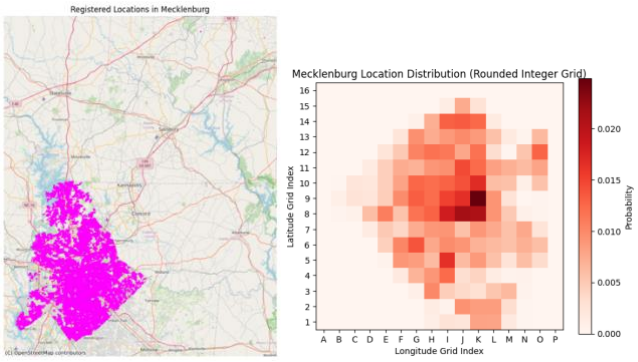
4. METHOD

4.1 Data Preprocessing

We started with parsing addresses from the raw voter-registration file (933,532 addresses) and converted them into GPS using OpenStreetMap Nominatim API. There were request delay set up to 30s per call to respect rate limits. We also trimmed unit numbers or other special characters of the addresses to avoid the None and improve geocoding accuracy. However, due to the time limit and other technical issues, not all addresses were geocoded. We just processed data around 200,000~ 300,000 addresses and then randomly sample 100,000 addresses out of that to ensure a clean and valuable set of geocoded points suitable for downstream privacy evaluation.

4.2 Spatial Grid & Prior Initialization

To apply PRIVIC's discrete-state algorithms, we first mapped each geocoded voter location into one of $G \times G = 16 \times 16$ grid cells over Mecklenburg County. By using a square grid tuned to our county's bounding box, we ensured uniform cell areas and simplified indexing, while avoiding the larger computational cost of non-square grids. We used latitude limits $[-35.0°, 35.0°]$ N and longitude limits $[-81.0°, 81.0°]$ W, chosen to cover the full extent of the county. To achieve uniform partitioning, each axis was divided into 16 equal intervals (step size $0.5/16 = 0.03125°$), yielding 256 rectangular cells. We stored the latitude and longitude edge values in descending order so that grid index (0,0) corresponds to the northwest corner. For each of the 100,000 sampled points, we located the unique pair (i, j) satisfying $Lat_{i+1} < lat < Lat_i$, $Lon_{j+1} < lon < Lon_j$, and recorded that cell index in a CSV file. Finally, we normalized the 256 cell counts by the total sample size to obtain the initial empirical prior $\theta^{(0)}$ over the grid. This prior served as the starting point for the Blahut–Arimoto and IBU steps in PRIVIC (Figure 1).



(a) The original locations from 100,000 sampled locations from Mecklenburg County.

(b) Density of the original locations from Mecklenburg County.

Figure 1: (a) visualizes the original locations from 100,000 sampled locations from Mecklenburg County. (b) illustrates a heatmap representation of the locations in Mecklenburg County.

4.3 PRIVIC Implementation

We follow the PRIVIC cycle (Algorithm 1 in [6]) with the following configuration:

- **Distortion-privacy parameter** $V \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ (analogous to β).
- **Number of cycles** $N \in \{10,15\}$.
- **BA precision**: run Blahut–Arimoto for 10 iterations per cycle.
- **IBU precision**: run Iterative Bayesian Update for 10 iterations per cycle.

At each cycle t = 1,… ,N:

1. **BA step**: compute obfuscation channel

   $$C^{(t)} = BA(\theta^{(t-1)}, c_0, \beta, \delta_{BA}),$$

   where c_0 is the uniform channel and δ_{BA} the fixed iteration count.

2. **Data collection & obfuscation**: sample 10 % of the original points, apply $C^{(t)}$ to obtain noisy reports $y^{(t)}$
3. **IBU step**: estimate posterior

   $$\mu^{(t)} = IBU(C^{(t)}, \theta^{(t-1)}, q^{(t)}, \delta_{IBU})$$

   where $q^{(t)}$ is the empirical PMF of $y^{(t)}$.

4. **Prior update**: combine $\mu^{(t)}$ and $\theta^{(t-1)}$ by weighting according to sample counts to form $\theta^{(t)}$.

We implemented these steps in Python, re-using our geocoded CSV and grid definitions from Section 4.1–4.2.

The choice of configuration was based on the experiment of PRIVIC experiment for Paris and San Francisco check-in locations. However, we changed some of them based on their advantages. The numbers of cycles was chose within the convergence range of PRIVIC experiments, which is 8-15. Ten inner iterations per cycle provide near-optimal channel computation and posterior estimation without excessive runtime.

5. RESULTS

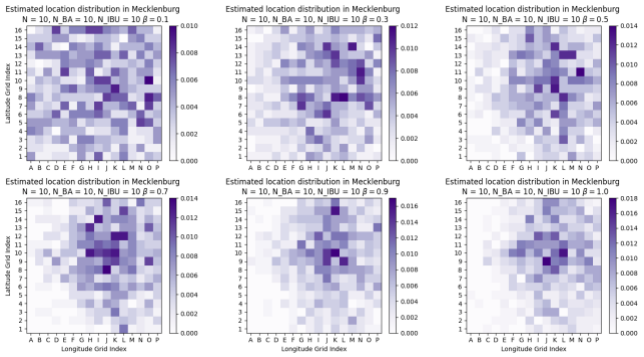
5.1 Obfuscation Simulation

This experiment measures the immediate distortion caused by a single application of a geo-indistinguishable Laplace mechanism:

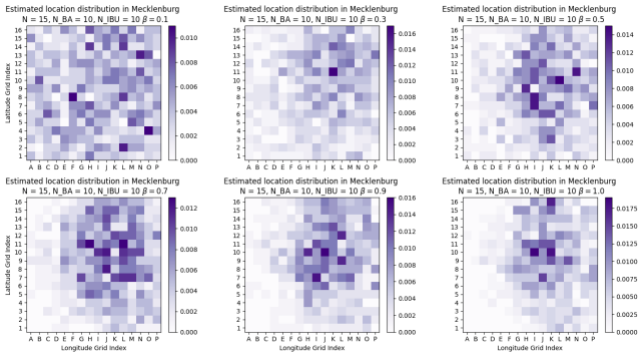
1. **Parameters:** privacy levels $V \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$, 10 independent runs per V, sample size 1000.
2. **Workflow per run:**
   - Randomly sample 1 000 points, compute the true cell-histogram p_{true} .
   - Build the Laplace channel C for parameter V.

- Obfuscate the sample to obtain noisy reports and compute the obfuscated histogram p_{obf}.
- Compute $EMD_{(p_{true},\ p_{obf})}$.
- Save each posterior and EMD for later analysis.

This one-shot evaluation establishes a baseline: how much utility is lost without any iterative prior refinement (Figure 2).



a) 10 runs



b) 15 runs

Figure 2: Visualization of the estimated true distribution of the locations in Mecklenburg with 10 cycles (a) and (b) 15 cycles run

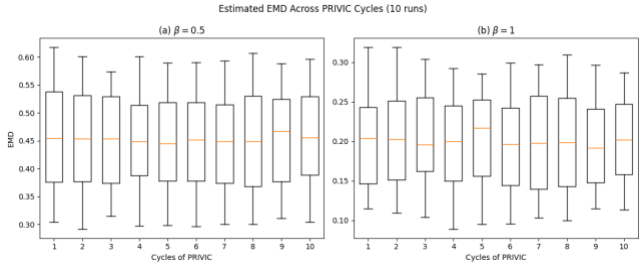
5.2 Iterative PRIVIC Convergence Analysis

To observe how PRIVIC's BA + IBU cycles progressively recover utility, we ran the full algorithm for N={10, 15} cycles under two distortion parameters (β=0.5 and β=1.0), repeating each setting over 5 independent trials. At the end of each cycle t, we recorded the estimated prior $\theta^{(t)}$ and computed $EMD_{(\theta_{true},\ \theta^{(t)})}$.

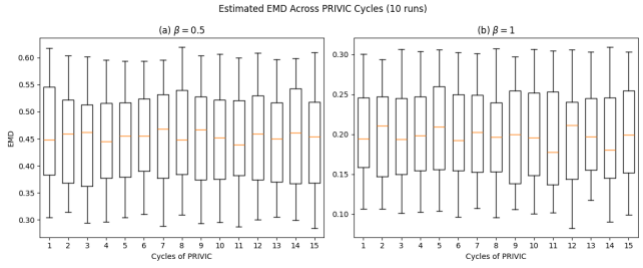
By aggregating results over runs, we generated a boxplot of EMD versus cycle number for each β. This visualization summarizes the distribution of EMD—median, interquartile range, and outliers—at each iteration, allowing us to:

- **Quantify convergence speed:** observe how many cycles are needed for EMD to stabilize.
- **Compare privacy settings:** higher β (weaker noise) converges faster and to lower EMD, while lower β (stronger privacy) requires more cycles.

A side-by-side boxplot for β =0.5 and β =1.0 clearly illustrates these trade-offs across the 10 and 15 cycles (Figure 3).



a) 10 cycles



b) 15 cycles

Figure 3: shows the EMD between the true PMF of Mecklenburg locations and its estimation by PRIVIC in each of its cycles for $\boldsymbol{\beta}$ = 0.5 and $\boldsymbol{\beta}$ = 1.0.

5.3 Summary of Metrics and Visualizations

While heatmaps visualized final post-cycle priors $\theta^{(N)}$ as 16×16 heatmaps to inspect spatial fidelity, boxplots displayed EMD distributions over runs and cycles to capture variability and convergence behavior. Together, these evaluations provide a comprehensive picture of how privacy parameters and iterative refinement interact to balance geo-indistinguishability.

6. CONCLUSION

In our experiments, we found that the privacy parameter β exerts a strong influence on how sharply the estimated location posteriors concentrate around true hotspots in Mecklenburg County. When β was very low (0.1 or 0.3), the added noise dominated, and the heatmaps remained essentially diffuse and uniform. As β increased to 0.5 and above, however, the posteriors began to sharpen markedly, with the β = 1.0 estimates closely matching the underlying "ground-truth" distribution. Quantitatively, the Earth Mover's Distance (EMD) measured over ten independent runs dropped steeply during the first three to five PRIVIC cycles—falling to a median of approximately 0.45 for β = 0.5 and 0.20 for β = 1.0—and then plateaued, indicating that most of the utility gain is realized early and that further iterations yield diminishing returns.

The development and testing of our PRIVIC pipeline taught us several valuable lessons. By automating ten rounds each of Bayesian Aggregation (BA) and Iterative Bayesian Updates (IBU) on a fixed 16×16 grid, we were able to characterize the privacy–utility trade-off systematically across a range of β values. This process highlighted the importance of grid resolution: had we explored coarser or finer spatial discretizations, we might have observed different convergence behavior or EMD profiles. We also recognized that a rigid schedule (10 BA and 10 IBU iterations per cycle) could be suboptimal; an adaptive stopping criterion—halting early once posteriors stabilize—would reduce computation without sacrificing utility. Finally, while our six β values offered a clear view of broad trends, a denser sampling could uncover subtler transitions in privacy–utility performance.

Looking ahead, several avenues could extend and strengthen this work. First, applying PRIVIC to the full Mecklenburg County voter-registration dataset—comprising hundreds of thousands of addresses—would test the pipeline's scalability and robustness under realistic data volumes. Second, incorporating additional metrics, such as formal differential-privacy guarantees (ε, δ) or adversarial inference risk assessments, would provide a more holistic evaluation of privacy. Third, adapting the method to dynamic, time-varying data (for example, mobility traces) would enable us to study how posteriors evolve as new observations arrive. Algorithmic refinements—such as variational-Bayes updates or moment-matching techniques—might accelerate convergence under stringent privacy settings (very low β). Sampling is another aspects should be improved since during experiments, there were some cases the algorithm crashed due to not enough data for sampling. Finally, we envision an interactive tool that recommends β and cycle counts based on a user's target utility threshold, making PRIVIC more accessible to practitioners and enabling on-the-fly, data-driven parameter tuning.

7. REFERENCES

[1] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends® in Theoretical Computer Science, 2014.

[2] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. "What Can We Learn Privately?" *FOCS* 2011

[3] Ú. Erlingsson, V. Pihur, and A. Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." *CCS* 2014.

[4] M. E. Andrés, N. E. Fouque, G. Rechich, and C. Troncoso-Pastoriza. "Geo-Indistinguishability: Differential Privacy for Location-Based Systems." *CCS* 2013

[5] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. "Optimal Geo-Indistinguishable Mechanisms for Location Privacy." In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS)*, pp. 251–262, 2014.

[6] N. Zhang, M. Zhang, G. Cormode, and X. Xiao. "PrivBayes: Private Data Release via Bayesian Networks." In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 1427–1442, 2017.

[7] C. Dwork, M. Hardt, and K. Nissim. "Differentially Private Histogram Publication." *Journal of Privacy and Confidentiality*, 4(1):37–76, 2012.

[8] R. Shokri and S. Theodoridis. "PRIVIC: Private Information Collection via Iterative Bayesian Estimation." *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2019(3):101–120, 2019.

[9] S. Biswas and C. Palamidessi. "PRIVIC: A Privacy-Preserving Method for Incremental Collection of Location Data." *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2024(1):582–596, 2024.
DOI: 10.56553/popets-2024-0033.