UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

EDUARDO RODRIGUES DUARTE NETO

LONGITUDINAL GEOSPATIAL FREQUENCY ESTIMATION UNDER ADAPTIVE LOCAL DIFFERENTIALLY PRIVATE MODEL

FORTALEZA

2025

EDUARDO RODRIGUES DUARTE NETO

LONGITUDINAL GEOSPATIAL FREQUENCY ESTIMATION UNDER ADAPTIVE LOCAL DIFFERENTIALLY PRIVATE MODEL

Thesis submitted to the Programa de Pós-graduação em Ciência da Computação of the Centro de Ciências of the Universidade Federal do Ceará, as a partial requirement for obtaining the title of Doctor in Ciência da Computação. Concentration Area: Ciência da Computação

Advisor: Prof. Dr. Javam de Castro Machado

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

D8711 Duarte Neto, Eduardo Rodrigues.
Longitudinal Geospatial Frequency Estimation under Adaptive Local Differentially Private Model / Eduardo Rodrigues Duarte Neto. – 2025.
112 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação , Fortaleza, 2025.
Orientação: Prof. Dr. Javam de Castro Machado.

1. Privacidade de Dados Diferencial Local. 2. Privacidade. 3. Mobilidade. 4. Localização. I. Título.
CDD 005

EDUARDO RODRIGUES DUARTE NETO

LONGITUDINAL GEOSPATIAL FREQUENCY ESTIMATION UNDER ADAPTIVE LOCAL DIFFERENTIALLY PRIVATE MODEL

Thesis submitted to the Programa de Pós-graduação em Ciência da Computação of the Centro de Ciências of the Universidade Federal do Ceará, as a partial requirement for obtaining the title of Doctor in Ciência da Computação. Concentration Area: Ciência da Computação

Approved on: 18/08/2025

EXAMINATION BOARD

Prof. Dr. Javam de Castro Machado (Advisor)
Federal University of Ceará (UFC)

Prof. Dr. José Maria da Silva Monteiro Filho
Federal University of Ceará (UFC)

Prof. Dr. Rafael Castro de Andrade
Federal University of Ceará (UFC)

Prof. Dr. Felipe Timbó Brito
Laboratory of Systems and Databases
(LSBD/UFC)

Prof. Dra. Valéria Cesário Times
Federal University of Pernambuco (UFPE)

À minha família, por sempre ter acreditado em mim.

ACKNOWLEDGEMENTS

Agradeço primeiramente a Deus, por ser a fortaleza que me sustentou e a luz que guiou meus passos ao longo de toda esta jornada. Sua presença constante me deu a força e a sabedoria necessárias para superar cada obstáculo com fé e perseverança, tornando possível a concretização deste sonho.

À minha namorada, Ana Buosi, minha mais profunda e sincera gratidão. Seu apoio foi meu porto seguro, especialmente no momento mais difícil e turbulento que foi o encerramento deste doutorado. Quando o cansaço e a dúvida surgiram, foi a sua fé inabalável em mim que me deu a força necessária para continuar. Seu amor, paciência e incentivo foram o alicerce que me permitiu chegar até o fim. Esta conquista é tão sua quanto minha.

Agradeço com todo o meu coração aos meus pais, Eduardo Duarte e Samia Duarte, pela confiança, amor e pelo suporte incansável que sempre me ofereceram. Sua dedicação e seus sacrifícios foram a base fundamental para que eu pudesse chegar até aqui e realizar este sonho.

Sou imensamente grato aos meus irmãos, Gustavo Duarte e Raquel Duarte, e à minha cunhada, Talita Facó, pelo amor, amizade e por todo o apoio ao longo desta trajetória. O carinho e os momentos de descontração que compartilhamos foram essenciais para aliviar a pressão e me dar forças para continuar. A presença de vocês em minha vida é um presente inestimável.

Ao meu orientador, Professor Doutor Javam de Castro Machado, minha mais profunda gratidão. Sua orientação excepcional, sabedoria e dedicação foram cruciais para o meu desenvolvimento como pesquisador. Agradeço por cada conselho, pela paciência e por me guiar com maestria através dos desafios acadêmicos. Sua influência foi determinante para o sucesso deste trabalho.

Gostaria de expressar um agradecimento especial aos meus amigos de laboratório, Iago Castro, Malu, André Luis e Daniel Praciano. Vocês foram muito mais do que colegas; foram uma família. As longas horas de trabalho, as discussões produtivas, o apoio mútuo e a amizade sincera tornaram esta caminhada mais leve e significativa. Levarei comigo cada aprendizado e cada momento compartilhado.

Agradeço sinceramente a todos os meus colegas que, de alguma forma, contribuíram para esta tese, em especial a Paulo Amora, Serafim, Falcão, Felipe Timbó e Titi. Agradeço também a todos os membros do Laboratório de Sistemas de Banco de Dados (LSBD) pelo ambiente colaborativo e estimulante. A troca de ideias, a ajuda mútua e o companheirismo foram essenciais para o meu crescimento e para a conclusão desta pesquisa. Sou grato por ter tido a

oportunidade de aprender e evoluir ao lado de profissionais tão dedicados.

”Anyone who has never made a mistake has never tried anything new.”

(Albert Einstein)

ABSTRACT

The collection of geospatial data under Local Differential Privacy (LDP) enables valuable spatial analytics without compromising user privacy. However, existing LDP mechanisms rely on static spatial discretizations, such as uniform grids or fixed-depth quadtrees, that are ill-suited to the dynamic and non-uniform nature of real-world mobility data. These limitations are further amplified in longitudinal settings, where users report their locations repeatedly over time.

In this work, we propose **ALOQ** (Adaptive Longitudinal Quadtree), a novel data collection model for continuous, privacy-preserving location frequency estimation under LDP. ALOQ introduces a dynamic quadtree-based spatial representation that evolves in response to noisy user density distributions, improving estimation accuracy while preserving strong privacy guarantees. The model includes a Quadtree Adaptation Window (QAW) to detect significant temporal changes, a similarity-aware privacy budget allocation mechanism, and a bounded refinement strategy that ensures the cumulative privacy loss remains under control.

We provide a theoretical analysis of ALOQ's privacy guarantees and evaluate its performance on both synthetic and real-world datasets. Our results show that ALOQ consistently outperforms state-of-the-art LDP baselines in terms of utility and budget efficiency, particularly in scenarios with skewed spatial distributions and evolving mobility patterns.

Keywords: local differential privacy; differential privacy; location data; spatial partition.

RESUMO

A coleta de dados geoespaciais sob o modelo de Privacidade Diferencial Local (LDP) viabiliza análises espaciais valiosas sem comprometer a privacidade dos usuários. No entanto, os mecanismos LDP existentes baseiam-se em discretizações espaciais estáticas, como grades uniformes ou quadtrees de profundidade fixa, que são inadequadas para a natureza dinâmica e não uniforme dos dados de mobilidade do mundo real. Essas limitações se agravam em cenários longitudinais, nos quais os usuários reportam suas localizações repetidamente ao longo do tempo.

Neste trabalho, propomos o **ALOQ** (Adaptive Longitudinal Quadtree), um novo framework para estimativa contínua de frequência de localização com preservação de privacidade sob LDP. O ALOQ introduz uma representação espacial dinâmica baseada em quadtree que evolui em resposta a distribuições de densidade de usuários ruidosas, melhorando a acurácia das estimativas sem comprometer as garantias de privacidade. O framework inclui uma Janela de Adaptação da Quadtree (GAW) para detectar mudanças temporais significativas, um mecanismo de alocação de orçamento de privacidade baseado em similaridade e uma estratégia de refinamento com limites que assegura controle sobre o acúmulo de perda de privacidade ao longo do tempo.

Apresentamos uma análise teórica das garantias de privacidade do ALOQ e avaliamos seu desempenho em conjuntos de dados sintéticos e reais. Os resultados demonstram que o ALOQ supera consistentemente os principais métodos LDP da literatura em termos de utilidade e eficiência no uso do orçamento de privacidade, especialmente em cenários com distribuições espaciais assimétricas e padrões de mobilidade dinâmicos.

Palavras-chave: privacidade diferencial local; privacidade diferencial; dados de localização; particionamento espacial.

LIST OF FIGURES

LIST OF TABLES

LISTA DE ALGORITMOS

LIST OF SYMBOLS

S	Spatial domain (e.g., a city)
U	Set of users;
u_n	the *n-th* user;
T	Set of discrete timestamps;
$t_\tau\}$	The timestamp τ;
l	Location in the continuous plane
r_i^t	True report of user u_i at time t
$\tilde{r}_i^t = \Psi(r_i^t)$	Noisy (LDP) report produced by mechanism Ψ
Q_t	Spatial partition (quadtree) at timestamp t
$q_j \in Q_t$	A spatial cell in the partition at timestamp t
f_j^t	Frequency (count) of reports assigned to cell q_j at time t
k	Number of cells in the current partition
$tr_{\max}$	Upper report-count threshold for splitting a quadtree cell
$tr_{\min}$	Lower report-count threshold for merging quadtree cells
tr_s	Similarity threshold that triggers quadtree adaptation
sw	Size of the sliding window used by the Quadtree Adaptation Window (QAW)
$rec_{\max}$	Maximum recursion depth during the per-timestamp splitting phase
ϵ	Upper-bound privacy budget for a single timestamp
ϵ_t	Dynamically adjusted per-timestamp privacy budget
$\alpha_{\min}$	Minimum budget scaling factor used in QAW
F_t	Estimated frequency vector at time t
F_{avg}	Average of the last sw frequency vectors (QAW)

MSE	Mean Squared Error (utility metric)
MAE	Mean Absolute Error (utility metric)
B_u	Total privacy budget consumed by user u over T
$\overline{B}$	Average total privacy budget per user

CONTENTS

1 INTRODUCTION

The pervasive collection of geospatial data from mobile devices and location-based services presents an unprecedented opportunity for spatial analytics, yet simultaneously introduces profound privacy challenges (CHENG *et al.*, 2017; OUCHRA *et al.*, 2023). This surge in location data enables a wide range of applications, including intelligent transportation systems, emergency planning, and urban development (RIKALOVIC *et al.*, 2018; MIRANDA-PASCUAL *et al.*, 2023). A fundamental analytical primitive in these scenarios is frequency analysis, which quantifies how often users visit specific spatial regions (HONG *et al.*, 2021). Frequency patterns are commonly visualized via heatmaps, supporting decisions about infrastructure investment, traffic optimization, and public health interventions (HUANG *et al.*, 2024; PEIPEI *et al.*, 2020).

Despite the analytical value of geospatial frequency data, its collection inherently entails significant privacy risks. Location histories can reveal highly sensitive information, including home and work locations, health visits, religious practices, and behavioral routines (AHUJA *et al.*, 2023; LI *et al.*, 2017). This opens the door to user re-identification, behavioral profiling, and discriminatory practices (UNNIKRISHNAN; NAINI, 2013; MIRANDA-PASCUAL *et al.*, 2023; SOLYMOSI *et al.*, 2023). Therefore, robust privacy-preserving mechanisms are essential when designing systems that collect and analyze such data.

To contextualize the inherent privacy risks in location data collection, consider the representative scenario illustrated in Figure 1, commonly encountered in urban mobility systems: generating heatmaps to analyze user movements within a city. In this scenario, individuals (depicted in the figure at timestamps t_1, t_2, and t_3) continuously transmit their geographic locations, represented by latitude and longitude coordinates, along with persistent identifiers, to a centralized aggregator.

This type of collection process, where users report their data repeatedly over time, is known as a longitudinal setting. Rather than a one-time snapshot, longitudinal data captures user behavior as it evolves, enabling richer analysis but also amplifying privacy risks, since repeated observations can increase the likelihood of linking data back to individuals (YAO *et al.*, 2014).

For example, imagine a user who reports their location only a few times per day. If these reports consistently occur around 8:30 a.m. and 5:30 p.m. on weekdays at the same coordinates, an adversary could reasonably infer that this place corresponds to the user's workplace. Similarly, late-night or overnight reports at another fixed location could reveal their residence.

Even with limited and seemingly harmless disclosures, repeated observations over time enable adversaries to reconstruct highly sensitive aspects of an individual's life.

At each timestamp, the aggregator receives and compiles all location reports into an aggregated dataset. Using frequency analysis, which is defined as counting the number of occurrences of location reports within discrete spatial regions or cells, the aggregator identifies the areas with the highest concentration of reports. Subsequently, this frequency data can be visualized as a heatmap, highlighting the most frequently visited or densely populated regions, as depicted in Figure 1.

While frequency analysis is beneficial for urban planning and optimizing transportation services, this method introduces significant privacy risks. Specifically, if an adversary (which can be the aggregator itself or an external malicious actor) gains access to frequency analysis results, they could potentially link aggregated data back to individual users. For instance, if the adversary has external knowledge about the presence of very few individuals in a specific region (such as region A), observing a report from this region at a given timestamp could enable the adversary to infer the presence or movements of a specific individual.

Prior studies have demonstrated the practical feasibility of these re-identification attacks, where minimal auxiliary information combined with frequency analysis can lead to significant breaches of privacy (MONTJOYE *et al.*, 2013; SOLYMOSI *et al.*, 2023; PYRGELIS *et al.*, 2017). Such breaches may allow adversaries to reconstruct daily routines, infer sensitive locations like residences and workplaces, and expose behavioral patterns. Hence, the threat model addressed in this work seeks to enable meaningful spatial analyses via frequency estimation while rigorously preserving individual privacy under formally guaranteed protections.

To address these challenges, Differential Privacy (DP) has emerged as a prominent framework for enabling data analysis while preserving individual privacy (DWORK *et al.*, 2006; CUMMINGS *et al.*, 2024). DP provides a formal privacy guarantee, ensuring that any statistical query result is nearly indistinguishable regardless of the inclusion or exclusion of any individual's data. It protects individual privacy by perturbing the data or query responses through the addition of carefully calibrated random noise, introduced via a randomized mechanism. This mechanism ensures that the data is modified in a controlled manner, making it difficult for adversaries to confidently infer sensitive personal information about individuals from aggregate data analyses.

However, DP's traditional model depends on a trusted curator, which introduces a

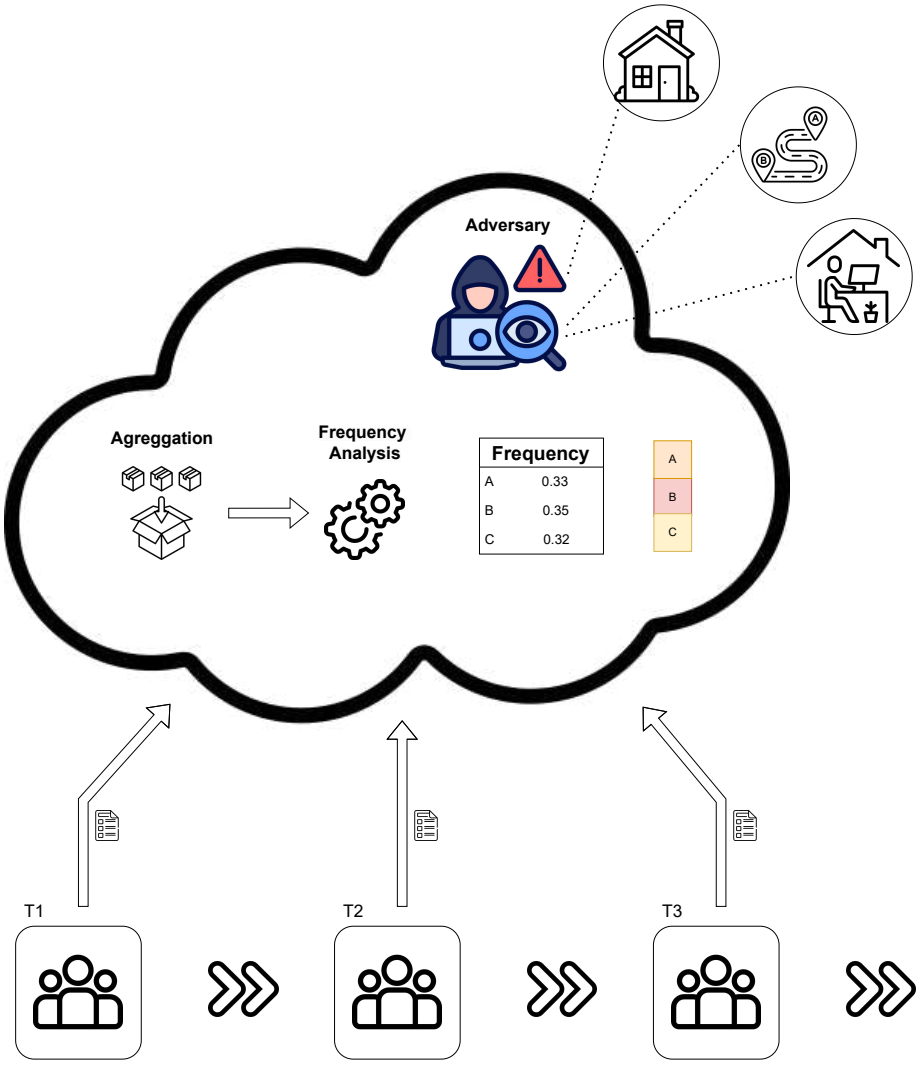


Figure 1 – Frequency analysis without privacy protection. Users continuously send location reports to an aggregator, enabling adversaries to infer sensitive individual information.

single point of failure (MARREIRAS NETO *et al.*, 2024). Several high-profile data breaches, including those involving Google, Facebook, Inter Bank, and Equifax, demonstrate the risks of relying on centralized data storage (Wakabayashi, Daisuke, 2018; DANIEL, 2024; UOL, 2017; CNN, 2017). Local Differential Privacy (LDP) mitigates this risk by shifting the responsibility for data anonymization to the user's device. In this model, each user perturbs its data locally before sharing it, using a randomized mechanism to add noise and ensure privacy, even if the server is compromised (DWORK; ROTH, 2014; KASIVISWANATHAN *et al.*, 2011).

A fundamental challenge in both DP and LDP frameworks is managing the inherent trade-off between privacy protection and data utility. We address this challenge through privacy budget allocation, which determines how much privacy is enforced at each step. A stronger privacy setting means injecting more randomness or noise into the data, often reducing

analytical accuracy. Conversely, achieving higher accuracy usually requires adding less noise, which weakens privacy guarantees. Striking the right balance is therefore essential to design privacy-preserving mechanisms that provide strong protection while still delivering meaningful analytical results.

Nevertheless, applying LDP to location data, especially in longitudinal settings where users submit repeated reports over time, introduces unique challenges. Existing LDP mechanisms often rely on static spatial discretizations, such as uniform grids or fixed-depth quadtrees (YE *et al.*, 2021; QARDAJI *et al.*, 2013), which fail to accommodate the dynamic and non-uniform nature of real-world mobility data. These methods suffer from non-uniformity error; coarse cells in dense regions limit resolution, while fine cells in the sparse areas incur high noise without meaningful signal. Moreover, static discretizations are misaligned with temporal shifts in user behavior, such as changes between peak and off-peak hours.

As part of this research, we initially developed a model named *ALOG* (Adaptive Longitudinal Grid) (DUARTE NETO *et al.*, 2026), which explored the potential of adaptive spatial discretization to improve longitudinal LDP utility. In this early design, the spatial domain was represented using a grid structure whose granularity evolved dynamically based on user report densities. A key contribution of ALOG was the introduction of the *Grid Adaptation Window* (GAW), a mechanism designed to limit structural updates to moments of significant distributional change. Experimental results confirmed that adjusting spatial granularity and controlling update frequency are essential for balancing privacy and utility in practice.

However, ALOG also revealed several architectural limitations. The grid structure lacked support for region coarsening (merging) in areas that became sparse over time, leading to inefficiencies in both utility and budget allocation. Furthermore, the use of fixed-interval adaptation was poorly aligned with the temporal variability typical of human mobility patterns.

These findings directly motivated the investigation of a more robust and flexible solution. We designed **ALOQ**, an Adaptive Longitudinal Quadtree, which generalizes and improves upon the ideas explored in ALOG. ALOQ introduces recursive merging, dynamic adaptation triggers, and hierarchical spatial encoding, providing a more expressive and privacy-efficient model for longitudinal geospatial data collection under Local Differential Privacy.

1.1 Problem Statement

We consider the problem of privately estimating the spatial frequency distribution of users over time under Local Differential Privacy (LDP). Let S be a bounded two-dimensional geographic region (e.g., a city), and let $U = \{u_1, u_2, \dots, u_n\}$ be a set of users moving within this region over a sequence of discrete timestamps $T = \{t_1, t_2, \dots, t_\tau\}$.

At each timestamp $t \in T$, a user $u_i \in U$ occupies a position $r_i^t \in S$. These positions are considered sensitive, and the system must ensure that no user is exposed to tracking or profiling. To protect privacy, we assume that each user perturbs their true location r_i^t using a randomized mechanism Ψ that satisfies local differential privacy, producing a privatized report $\tilde{r}_i^t = \Psi(r_i^t)$. The server collects these privatized reports from all users at each timestamp and aims to estimate the spatial frequency distribution over the region.

The spatial domain S is discretized into a collection of cells. Most existing methods adopt a fixed spatial partitioning structure, such as a uniform grid or static-depth quadtree. However, such static discretizations often misalign with the true distribution of user locations. Coarse partitions in dense areas result in a loss of spatial precision, while fine partitions in sparse regions introduce high noise due to underreporting. This phenomenon, known as *non-uniformity error*, leads to degraded utility under LDP.

This challenge becomes even more pronounced in *longitudinal settings*, where users submit location reports repeatedly over time. A fixed discretization cannot adapt to the temporal dynamics of mobility patterns, such as fluctuations caused by daily commuting cycles, traffic congestion, or special events. Figure 1 illustrates such a longitudinal collection scenario, where users report their location at multiple timestamps, and the server aggregates this data to perform frequency analysis.

To address these challenges, we aim to design a privacy-preserving model that adaptively adjusts the spatial partitioning structure over time, while ensuring that every individual report satisfies LDP. This mechanism must maintain high utility in the face of non-uniform and evolving data distributions, without relying on access to true location values.

Objective: Given a region S, a set of users U, and a privacy constraint, the goal is to estimate the frequency f_i^t of user visits to each spatial subregion (cell) $q_i \in Q^t$ at every timestamp $t \in T$, where Q^t is a (possibly adaptive) partition of S, using only privatized reports $\tilde{r}_i^t$. The model must:

- Satisfy local differential privacy for every user report;

- Operate over time without access to raw (non-perturbed) location values;
- Dynamically adjust the spatial granularity to reflect changing report densities;
- Balance the trade-off between utility and privacy under fixed privacy constraints.

In light of these challenges, we next present our research hypothesis, which motivates the adaptive solution proposed in this work.

1.2 Hypothesis

Given the challenges and limitations discussed above, we hypothesize that an adaptive spatial partitioning mechanism, driven by noisy frequency estimates under Local Differential Privacy (LDP), significantly improves the utility (i.e., reduces estimation error) of longitudinal geospatial frequency estimation compared to static partitioning methods, while maintaining strong privacy guarantees and efficient budget consumption.

1.3 Contributions

In this thesis, we propose **ALOQ** (Adaptive Longitudinal Quadtree), a novel model for privacy-preserving, continuous location data collection under LDP. ALOQ integrates four core innovations:

1. A **fully adaptive quadtree structure** that refines and merges spatial partitions over time, balancing resolution based on noisy report densities;
2. A **Quadtree Adaptation Window (GAW)** mechanism that detects when changes in report distributions justify structural updates;
3. A **dynamic privacy budget allocation strategy** that adjusts anonymization strength based on the similarity between current and past reports;
4. A **bounded refinement and one-time merge policy** that constrains privacy cost without sacrificing responsiveness.

All adaptations in ALOQ are driven exclusively by noisy, privatized data, never raw values, ensuring that privacy is maintained throughout the process. The model avoids the need for trusted curators and supports longitudinal deployments with minimal privacy loss due to redundant sanitization. ALOQ is designed to be modular and compatible with standard LDP protocols for frequency estimation, such as RAPPOR (ERLINGSSON *et al.*, 2014), L-OSUE (ARCOLEZI *et al.*, 2024), and LOLOHA (ARCOLEZI *et al.*, 2023). Users encode and perturb

their spatial positions using a provided quadtree structure, which is updated only when structural changes are justified by the GAW. Memoization is used to preserve efficiency across time, and quadtree depth is capped to avoid exponential privacy loss.

This section summarizes the main contributions of this work:

- We propose **ALOQ**, the first fully integrated model for longitudinal geospatial data collection under LDP, combining adaptive spatial partitioning, intelligent budget control, and privacy-preserving updates.
- We introduce a **quadtree adaptation algorithm** that performs recursive cell splits in high-density regions and a one-time merge for sparse regions, all under a bounded refinement depth.
- We design the **Quadtree Adaptation Window (GAW)**, a mechanism that triggers structural updates based on similarity between privatized frequency estimates, avoiding unnecessary recomputations and maintaining memoization efficiency.
- We implement a **dynamic privacy budget strategy**, which allocates more or less noise based on distributional drift, thus conserving privacy budget during stable periods and improving utility when significant changes occur.
- We provide a theoretical privacy guarantee for longitudinal reporting in ALOQ, showing that the cumulative privacy cost is bounded by the number of leaves generated across all time steps.
- We conduct extensive experiments on real-world and synthetic datasets (GeoLife, Porto, S1, S2) and show that ALOQ consistently outperforms leading LDP baselines in utility and budget efficiency across diverse spatial distributions and privacy constraints.

Published papers:

- DUARTE NETO, E.; COSTA FILHO, J. S.; MARREIRAS NETO, A. A.; MACHADO, J. C. ALOG: Adaptive Longitudinal Grids for Geospatial Data using Local Differential Privacy. *International Conference on Extending Database Technology, 2026*. (DUARTE NETO *et al.*, 2026).
- MARREIRAS NETO, A. A.; DUARTE NETO, E.; COSTA FILHO, J. S.; MACHADO, J. C. Locally differentially private and consistent frequency estimation of longitudinal data. *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados, 2024*. (MARREIRAS NETO *et al.*, 2024)
- SILVA, MARIA; MENDONÇA, A. L. C.; DUARTE NETO, E. R.; CHAVES, I. C.; CAM-

INHA, C.; BRITO, F. T.; FARIAS, V. A. E.; MACHADO, J. C. FACTO Dataset: A Dataset of User Reports for Faulty Computer Components. *Simpósio Brasileiro de Banco Dados, 2024, Florianópolis. 2024*. (LOURDES *et al.*, 2024)

- SENA, LUCAS ; PRACIANO, DANIEL ; CASTRO, IAGO ; DUARTE NETO, E. R.; MACHADO, JAVAM C.: AUDIO-MC: A General Framework for Multi-context Audio Classification. *24th International Conference on Enterprise Information System, 2022*. (SENA *et al.*, 2022)
- BENTO FILHO, M. E.; DUARTE NETO, E. R.; MACHADO, J. C. Privacy-preserving of patients with Differential Privacy: an experimental evaluation in COVID-19 dataset. *Journal of Information and Data Management, 2021*. (FILHO *et al.*, 2021)
- MACHADO, JAVAM C. ; DUARTE NETO, E.; BENTO FILHO, M. E. Técnicas de privacidade de dados de localização. *Anais do XXXIV Simpósio Brasileiro de Banco de Dados, 2019*. (MACHADO *et al.*, 2019)
- OLIVEIRA, DANIEL ; DUARTE NETO, E. R.: Um estudo comparativo de mecanismos de privacidade diferencial sobre um dataset de ocorrências do ZIKV no brasil. *Anais do XXXIV Simpósio Brasileiro de Banco de Dados, 2019*. (OLIVEIRA *et al.*, 2019)
- DUARTE NETO, E.; L. C. MENDONÇA, ANDRÉ ; C. MACHADO, JAVAM. PrivLBS: Preserving Privacy in Location-Based Services. *Journal of Information and Data Management - JIDM, 2019*. (DUARTE NETO *et al.*, 2019)

1.4 Organization

The remainder of this thesis is organized as follows:

Chapter 2 presents the foundations and essential definitions used throughout the work. In Chapter 3, we present the related work and discuss how the main existing approaches work and their limitations. We present our approach in Chapter 4. Experimental results are shown in Chapter 5. Finally, Chapter 6 concludes this work.

2 BACKGROUND AND DEFINITIONS

This section introduces the theoretical foundations and technical components used in our model, including differential privacy (DP), local differential privacy (LDP), and LDP-based frequency estimation protocols.

2.1 Differential Privacy

At its foundation, Differential Privacy (DP) relies on a *randomized algorithm*, often called a *mechanism*, to respond to queries. As illustrated in Figure 2, this mechanism interacts with a dataset and generates outputs from a carefully controlled *output distribution*. The goal is to ensure that the probability of any given output remains nearly unchanged regardless of whether any single individual's data is included or excluded from the dataset. This property ensures that even an adversary with extensive background knowledge cannot confidently infer the participation of any specific individual.

Differential privacy is a formal paradigm designed to provide rigorous privacy guarantees when releasing statistical analyses of sensitive data (DWORK; ROTH, 2014; DWORK *et al.*, 2006). The central principle is to make the result of an analysis substantially indistinguishable with respect to any single individual's presence. This robustness holds even if the adversary has access to all other records in the database.

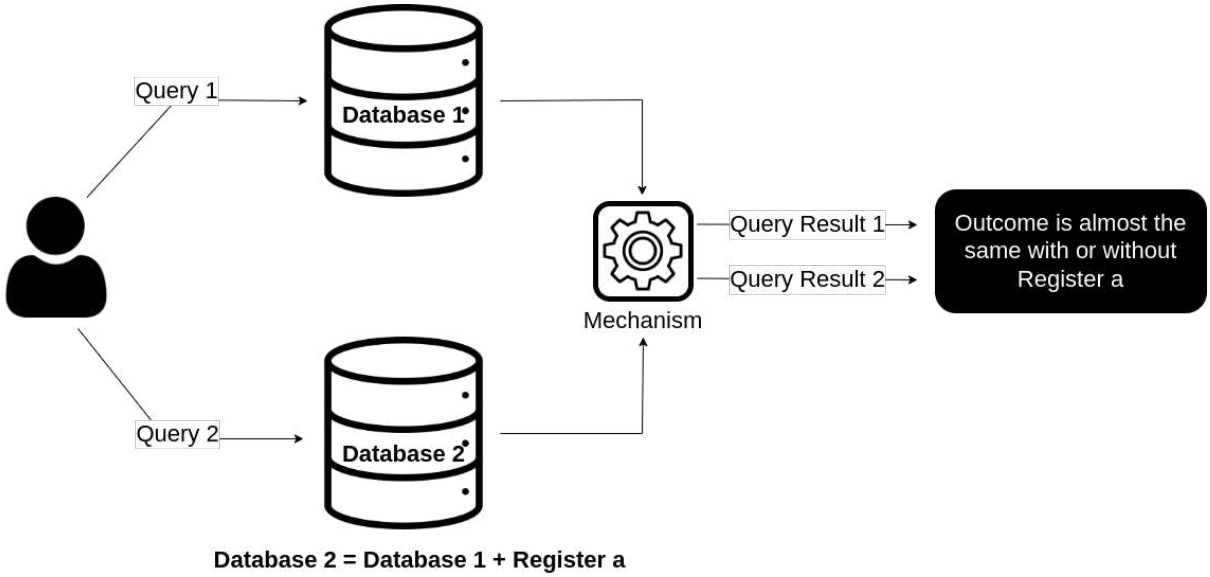


Figure 2 – Diagram illustrating how a randomized differentially private algorithm works.

Mechanisms designed for differential privacy carefully structure their output distri-

butions to balance analytical accuracy with the need to protect individuals. By doing so, they enable meaningful insights while safeguarding privacy.

The formal definition, first introduced by Dwork (DWORK, 2006), is as follows:

Definition 1 (ϵ-Differential Privacy)**.** A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if, for any two datasets D_1 and D_2 that differ in only one tuple (i.e., are neighboring), and for any possible output O of $\mathcal{A}$, we have:

$$\Pr[\mathcal{A}(D_1) = O] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(D_2) = O] \tag{2.1}$$

To make this concept more concrete, consider a scenario involving a database of user locations. Two **neighboring datasets**, D_1 and D_2, could be:

- **Dataset D_1 (without your data):** A database containing the locations of many users, where 15 people are currently at the *Fortaleza Public Library*.
- **Dataset D_2 (with your data):** The same as D_1, but with one additional record: your own, also indicating presence at the *Fortaleza Public Library*.

Consider a query posed to a randomized algorithm $\mathcal{A}$ that satisfies ϵ-differential privacy:

```
Query: "How many people are at the Fortaleza Public Library?"
```

The true answer for D_1 is 15, and for D_2 it is 16. Instead of returning these exact numbers, $\mathcal{A}$ perturbs the result with calibrated noise (e.g., drawn from a Laplace distribution) before releasing the answer:

- When run on D_1, the algorithm might return **15.3**.
- When run on D_2, it might return **15.8**.

The privacy guarantee arises from the fact that an adversary observing an output like 15.3 cannot confidently determine whether the true underlying count was 15 or 16. The outputs are statistically close enough that the presence or absence of any individual, including yourself, is effectively hidden.

Practically, differential privacy is typically implemented via *output perturbation*, wherein the true result of a query is modified by adding carefully calibrated random noise. The magnitude of this noise is determined in relation to a privacy budget parameter (discussed later), ensuring that the contribution of any single individual remains indistinguishable within the output.

2.2 Local Differential Privacy

Local Differential Privacy (LDP) represents a fundamental shift from the centralized model of Differential Privacy (DP) described previously. While centralized DP relies on a *trusted curator* to collect and anonymize raw user data, this creates a significant vulnerability: the centralized database becomes a single point of failure. If this central server is compromised, either by external attackers or a malicious insider, the privacy of all users is lost. This is not a theoretical concern; major data breaches at organizations like Google, Facebook, Equifax, and Inter Bank have repeatedly demonstrated the risks of entrusting sensitive data to centralized systems (Wakabayashi, Daisuke, 2018; DANIEL, 2024; UOL, 2017; CNN, 2017).

LDP eliminates this trust requirement entirely. In the LDP model, each user applies a randomization mechanism locally on their own device *before* any data is transmitted (KASIVISWANATHAN *et al.*, 2011). Only the perturbed, anonymized output is sent to the aggregator. Because the server never receives the true, sensitive data, user privacy is preserved even if the server is fully compromised. This makes the data "born private" and renders LDP the gold standard for settings where a trusted data curator cannot be assumed. This model is particularly well-suited for longitudinal data collection, where users repeatedly report information, as privacy is guaranteed at every single submission (ARCOLEZI *et al.*, 2023).

Definition 2 (ϵ - LDP (ERLINGSSON *et al.*, 2014))**.** An algorithm $\Psi(\cdot)$ satisfies ϵ-local differential privacy, where $\epsilon \geq 0$, if and only if for any pair of inputs (v, v'), and any set R of possible outputs of Ψ, we have

$$Pr[\Psi(v) \in R] \leq e^{\epsilon} Pr[\Psi(v') \in R] \tag{2.2}$$

Intuitively, *ϵ-LDP* ensures that given the perturbed output in R, it is not possible for the server or for an adversary to distinguish whether the original true value was v or v' beyond the probability ratio e^{ϵ}. Here, ϵ controls the privacy level. Lower ϵ yields stronger privacy.

Analogous to central DP, LDP also benefits from key foundational properties, including robustness to post-processing and composition (DWORK; ROTH, 2014).

Proposition 1 (Post-Processing)**.** If Ψ is ϵ-LDP, then $f(\Psi)$ is also ϵ-LDP for any function f.

The post-processing property, stated formally in Proposition 1, illustrated in Figure 4, ensures that once a user's data has been perturbed using a mechanism that satisfies ϵ-Local

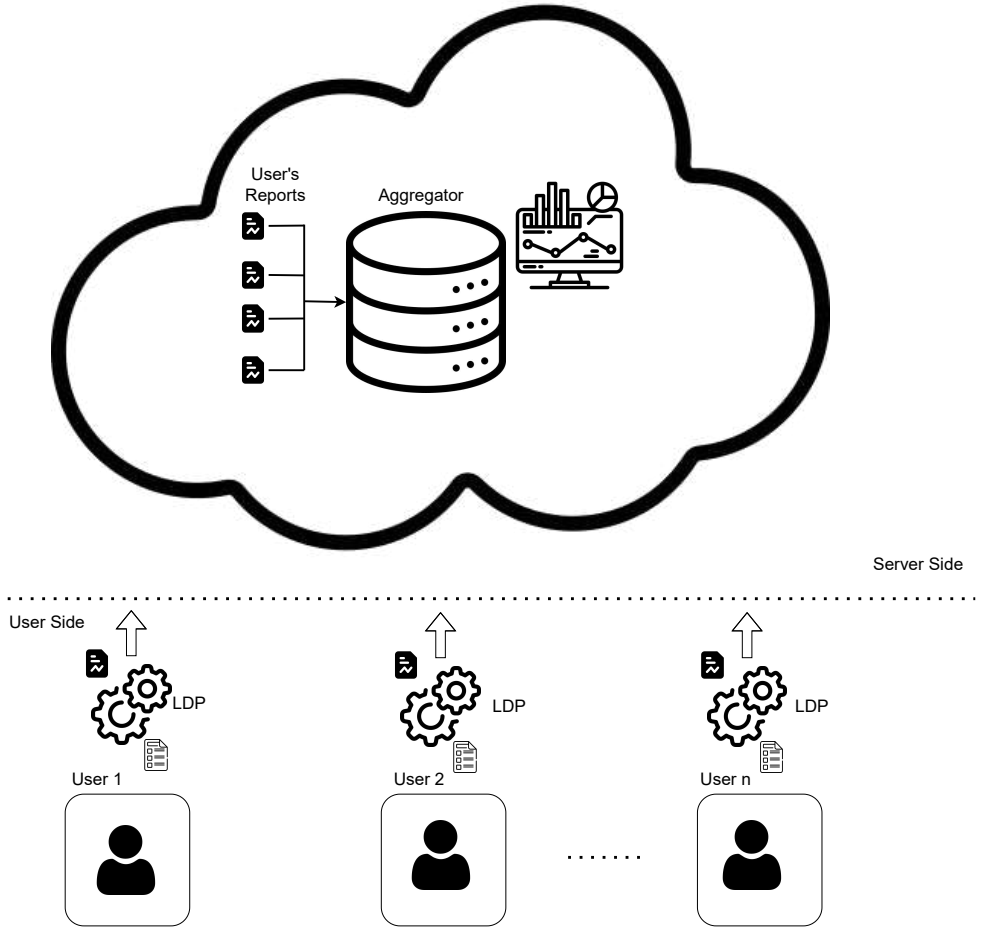


Figure 3 – LDP model: each user locally perturbs their true data before sending it to the aggregator. The server only receives noisy data and performs global analysis, preserving privacy even under compromise.

Differential Privacy (LDP), any subsequent computation applied to that output cannot degrade the privacy guarantee.

Proposition 2 (Sequential Composition)**.** Let each Ψ_i be an ϵ-LDP mechanism, and Ψ is the sequential composition Ψ_1,...,Ψ_m. Then, Ψ satisfies $\sum_{i=1}^{m} \epsilon$-LDP.

This property 2, illustrated in Figure 5, ensures that when multiple LDP mechanisms are applied independently to the same user's data, the total privacy loss accumulates linearly. That is, if a user provides m reports, each perturbed using an ϵ-LDP mechanism, the overall process guarantees $\sum_{i=1}^{m} \epsilon$-LDP.

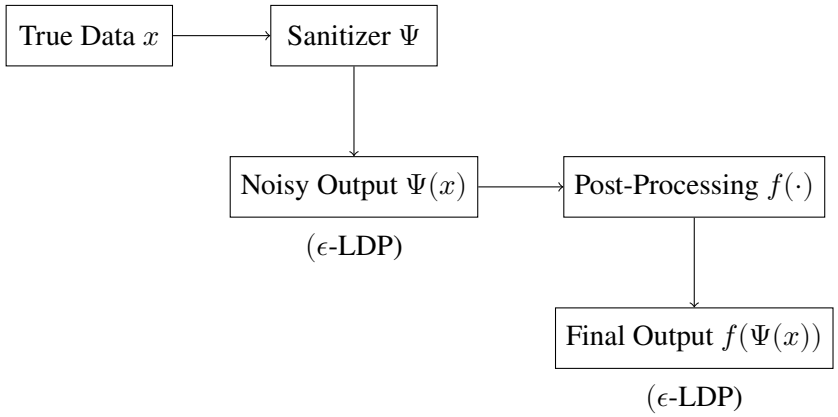


Figure 4 – Illustration of the Post-Processing Property: once a value has been sanitized under ϵ-LDP, any transformation applied thereafter cannot weaken the privacy guarantee.

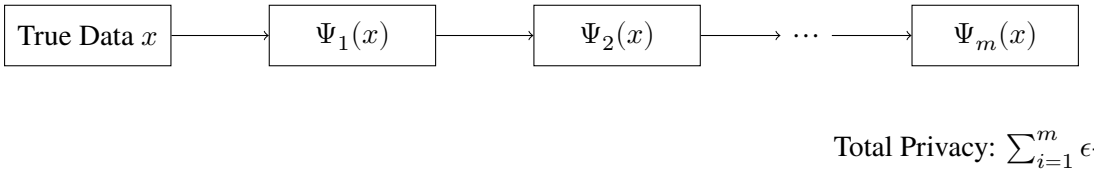


Figure 5 – Illustration of the Sequential Composition Property: applying m mechanisms, each satisfying ϵ-LDP, results in a total privacy cost of $\sum_{i=1}^{m} \epsilon$.

2.3 Longitudinal Data

Longitudinal data refers to information collected from the same subjects (e.g., users, devices) repeatedly over multiple distinct time points or periods (SINGER; WILLETT, 2003). This approach contrasts with data collected at a single instance, as it enables the analysis of changes, trends, and the evolution of phenomena over time. A key strength of longitudinal data is its ability to reveal within-individual transformations and to understand how initial states or prior events influence subsequent observations.

In the context of mobility data, longitudinal information captures sequences of user-reported locations over time. Rather than a snapshot of presence in a region, longitudinal mobility datasets allow us to reconstruct movement trajectories, commuting patterns, and temporal habits. These insights are crucial in applications such as traffic forecasting, urban planning, and epidemiological modeling.

Example. Consider a user who reports their location to a mobile application every 30 minutes during weekdays, from 7:00 to 19:00. Over several days, the sequence of coordinates might reveal a consistent pattern: departure from home in the morning, a stationary period near an office location during work hours, and return in the evening. The app collects the following (simplified) sequence of location reports for a given day:

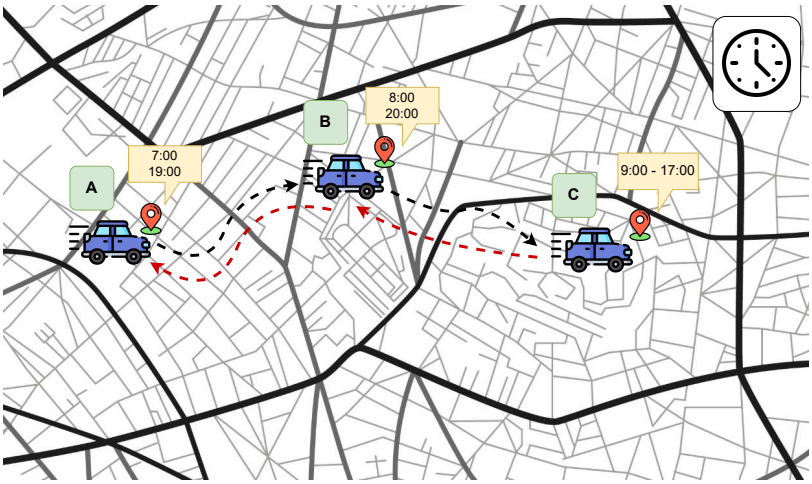


Figure 6 – Example of Longitudinal Mobility Data: User Trajectories Over Time.

- 07:00 – Location A (home)
- 08:00 – Location B (en route)
- 09:00 to 17:00 – Location C (workplace)
- 18:00 – Location B (en route)
- 19:00 – Location A (home)

When this reporting occurs over many days, analysis may identify regularities (e.g., Monday to Friday commute), deviations (e.g., detours or delays), and transitions (e.g., work-from-home changes or new job location).

Analytical Opportunities and Risks. Longitudinal mobility data enables fine-grained modeling of behavioral patterns, such as:

- *Predictive analytics:* Inferring next locations, travel times, or peak traffic periods.
- *Change detection:* Identifying anomalies or shifts in routines (e.g., holidays, illness, job change).
- *User profiling:* Modeling user-specific behavior for personalization or segmentation.

However, the same continuity that enables rich insights also amplifies privacy risks. Repeated reporting increases the probability of re-identification and allows for inferences about sensitive aspects of users' lives (e.g., health, habits, religious or political practices). The cumulative exposure across time points means that even if individual reports are obfuscated, patterns may still emerge when aggregated. This trade-off between temporal utility and longitudinal privacy forms a central motivation for privacy-preserving mechanisms like ALOQ.

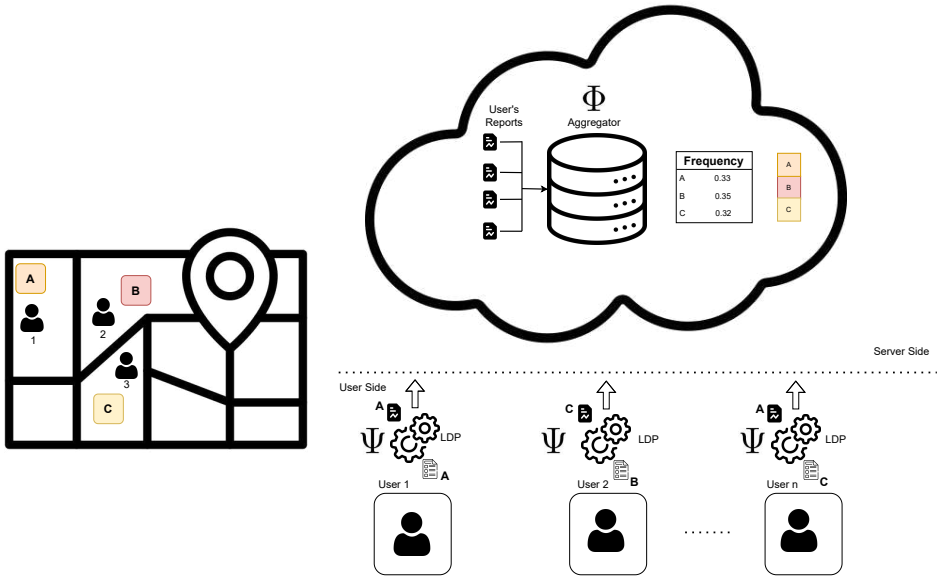


Figure 7 – Frequency Estimator: users sending their sanitized report to the aggregator to estimate the frequency of values for a given attribute.

2.4 Frequency Estimator

A frequency estimator is an LDP mechanism designed to estimate how often each value $r \in V$ appears in the dataset, where V is the set of all possible values of a given attribute (CORMODE *et al.*, 2021; BASSILY; SMITH, 2015). A frequency estimator consists of two components: A sanitizer and an aggregator. The sanitizer is employed by users to locally perturb their data. An aggregator receives the perturbed data and estimates the frequencies regarding the input. Frequency estimators have been employed for a variety of different tasks such as answering range queries (FILHO; MACHADO, 2023), identifying heavy hitters (ZHU *et al.*, 2024; QIN *et al.*, 2016), and the problem we attack in this thesis, namely private queries on geospatial data (HONG *et al.*, 2021; LAOUIR; IMINE, 2024; TIRE; GURSOY, 2024).

Figure 7 illustrates a typical frequency estimator. In this example, the domain consists of three possible spatial regions or cells: A, B, and C, which could represent specific urban areas or grid cells within a city. Each user applies a sanitization mechanism to their true location report (e.g., indicating which spatial region they are currently in) before submitting the sanitized version to the server. On the server side, the aggregator collects all sanitized reports and estimates the frequency of user presence within each defined spatial region in the domain.

In sequence, we present several state-of-the-art Local Differential Privacy (LDP) Frequency Oracle (FO) protocols, which are often used as building blocks for more complex

tasks (e.g., heavy hitter estimation (BASSILY *et al.*, 2020), machine learning (CHAMIKARA *et al.*, 2020), and private frequency monitoring (ARCOLEZI *et al.*, 2024; DING *et al.*, 2017))

2.5 Frequency Oracles

A *frequency oracle (FO)* protocol can be used to estimate the frequency of any value $v \in D$ under LDP, where D is the domain. A FO consists of two algorithms. The first one is Ψ, which users use locally to perturb their private data. The second one is Φ, and the aggregator uses it to estimate the frequencies based on the perturbed data received. In Figure 7 we can observe the algorithms.

2.5.1 Generalized Randomized Response

Randomized Response (RR), originally introduced for binary responses (WARNER, 1965), can be naturally extended to handle larger domains. This generalized version, known as the *Generalized Randomized Response (GRR)*, allows each user with a private value $v \in D$ to report their true value v with probability p. With the complementary probability $1 - p$, the user reports a randomly selected value $v' \in D \setminus \{v\}$.

Formally, the GRR mechanism perturbs a user's input according to the following probability distribution:

$$\forall_{x \in D} \quad \Pr\left[\Psi_{\text{GRR}_{(\epsilon)}}(v) = x\right] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + |D| - 1}, & \text{if } x = v \\ q = \frac{1-p}{|D|-1} = \frac{1}{e^\epsilon + |D| - 1}, & \text{if } x \neq v \end{cases}$$

This mechanism satisfies ϵ-local differential privacy (ϵ-LDP), since the ratio $p/q = e^\epsilon$ ensures indistinguishability between any two possible inputs.

From a population of n users, the aggregator collects a vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$, where each $x_i \in D$ is the reported (perturbed) value from the i-th user. To estimate the frequency of a value $v \in D$, let $C(v)$ be the number of times v appears in $\mathbf{x}$. The unbiased frequency estimator is given by:

$$\Phi_{\text{GRR}_{(\epsilon)}}(v) := \frac{C(v)/n - q}{p - q} = \frac{\frac{C(v)}{n} - \frac{1}{e^\epsilon + |D| - 1}}{\frac{e^\epsilon - 1}{e^\epsilon + |D| - 1}}. \tag{2.3}$$

This expression provides an unbiased estimate of the true frequency of v in the underlying population (WANG *et al.*, 2017a). The corresponding variance of the estimator is:

$$\text{Var}\left[\Phi_{\text{GRR}_{(\epsilon)}}(v)\right] = \frac{e^\epsilon + |D| - 2}{n(e^\epsilon - 1)^2}. \tag{2.4}$$

As evident from Equation (2.4), the variance grows linearly with the domain size $|D|$. Consequently, the accuracy of the GRR protocol deteriorates as the size of the input domain increases, making it less suitable for high-cardinality domains.

Example: Estimating Favorite Color with GRR

To illustrate the GRR mechanism, consider a simple survey where each user is asked to report their favorite color. The domain is defined as:

$$D = \{\text{Red}, \text{Green}, \text{Blue}\}, \quad \text{so } |D| = 3$$

Let the privacy budget be $\epsilon = \ln(3)$. Then:

$$e^{\epsilon} = 3, \quad p = \frac{3}{3+2} = \frac{3}{5} = 0.6, \quad q = \frac{0.4}{2} = 0.2$$

Each user reports their true favorite color with probability 0.6, and a randomly chosen incorrect color with probability 0.2 each.

Suppose the following 5 users respond:

User	True Color	Reported Color
1	Red	Red
2	Blue	Blue
3	Green	Red
4	Red	Green
5	Green	Green

The frequency of reported colors is:

$$C(\text{Red}) = 2, \quad C(\text{Green}) = 2, \quad C(\text{Blue}) = 1$$

We apply the GRR frequency estimator for each color, with $n = 5$, $p = 0.6$, and $q = 0.2$:

$$\Phi(v) = \frac{C(v)/n - q}{p - q}$$

- **Red:** $\Phi(\text{Red}) = \frac{0.4-0.2}{0.4} = 0.5$
- **Green:** $\Phi(\text{Green}) = \frac{0.4-0.2}{0.4} = 0.5$
- **Blue:** $\Phi(\text{Blue}) = \frac{0.2-0.2}{0.4} = 0.0$

These are unbiased estimates of the true frequency of each color in the underlying population, despite the perturbation introduced to protect user privacy.

2.5.2 Local Hashing (LH)

The Local Hashing (LH) protocol is a two-stage mechanism designed to address the challenge of frequency estimation under Local Differential Privacy (LDP) when the input domain is large (WANG *et al.*, 2017a). By combining domain reduction through hashing with a privatization mechanism over a smaller space, LH improves both efficiency and utility compared to applying a flat perturbation over the original domain.

The LH protocol operates as follows:

1. **Domain Reduction via Hashing**: Each user possesses a private value $v \in V$, where $|V| = k$ denotes the original domain size. The user randomly selects a hash function H from a universal family $\mathcal{H}$ and maps v to a smaller domain $[1..b]$ via $x = H(v)$. This step reduces the domain size from k to b, where $2 \leq b \ll k$, thereby limiting the impact of noise on utility. The universal property of $\mathcal{H}$ ensures that, for any distinct values $v_1, v_2 \in V$, the probability of a collision (i.e., $H(v_1) = H(v_2)$) is at most $1/b$, introducing controlled uncertainty regarding the original value.
2. **Perturbation via Generalized Randomized Response (GRR)**: Once the input has been hashed to a reduced domain, the Generalized Randomized Response (GRR) mechanism is applied to the hashed value $x \in [1..b]$. GRR is parameterized by the privacy budget $\epsilon > 0$ and operates by reporting the true hashed value x with probability p, and a different value $y \in [1..b] \setminus \{x\}$ with probability q. These probabilities are defined as:

$$p := \frac{e^{\epsilon}}{e^{\epsilon} + b - 1} \quad \text{and} \quad q := \frac{1}{e^{\epsilon} + b - 1},$$

The randomized output $x' \in [1..b]$, along with the chosen hash function H, is reported as the pair $\langle H, x' \rangle$.

Upon receiving a collection of reports $\langle H^u, x^u \rangle$ from n users, the aggregator estimates the normalized frequency of each original value $v \in V$ using the formula:

$$\hat{f}(v) := \frac{C(v) - nq}{n(p - q)},$$

where $C(v) = |\{u \in U \mid H^u(v) = x^u\}|$ is the count of users whose hash of v matches their reported value.

This approach enables efficient and accurate frequency estimation even when the original domain V is large, benefiting from reduced variance due to the smaller hashed space.

Example: Estimating Favorite Color via Local Hashing

To illustrate the LH mechanism in the same context as before, consider a survey where users report their favorite color from a large domain:

$$V = \{\text{Red}, \text{Green}, \text{Blue}, \text{Yellow}, \text{Purple}, \dots\}, \quad |V| = 100$$

To reduce noise, each user applies a hash function H that maps $V \to [1..5]$, i.e., $b = 5$. Assume a universal hash family $\mathcal{H}$ is shared between users and aggregator. Each user:

- Randomly chooses $H \in \mathcal{H}$
- Computes $x = H(v)$, where v is their true favorite color
- Applies GRR with privacy budget $\epsilon = \ln(3)$, yielding:

$$p = \frac{3}{3+4} = \frac{3}{7} \approx 0.4286, \quad q = \frac{1}{7} \approx 0.1429$$

- Sends $\langle H, x' \rangle$, where x' is the perturbed hash output

Example responses from 3 users:

User	**True Color**	**Hash Function H**	**Reported Bucket x'**
1	Red	$H_1(\text{Red}) = 3$	3
2	Blue	$H_1(\text{Blue}) = 1$	4
3	Green	$H_1(\text{Green}) = 3$	3

Goal: Estimate the frequency of Red.

Let $C(\text{Red})$ be the number of reports such that $H(\text{Red}) = x'$. Since $H_1(\text{Red}) = 3$, and users 1 and 3 reported 3, we have:

$$C(\text{Red}) = 2$$

Apply the LH estimator:

$$\hat{f}(\text{Red}) = \frac{2 - 3 \cdot 0.1429}{3(0.4286 - 0.1429)} \approx \frac{1.5713}{0.8571} \approx 1.833$$

Although the estimate is above 1 due to the small sample, this value would converge to a normalized frequency as n increases. This illustrates how LH leverages domain compression to enable accurate and private frequency estimation over large domains.

2.5.3 Unary Encoding (UE)

Unary Encoding (UE) protocols are a class of mechanisms for frequency estimation under Local Differential Privacy (LDP) that transform a categorical value into a binary vector representation before applying randomization (WANG *et al.*, 2017a). This approach enables efficient privatization while maintaining unbiased estimations for histogram computation.

The UE protocol operates as follows:

1. **One-Hot Unary Encoding**: Each user holds a private value $v \in V$, where $|V| = k$ is the domain size. The user encodes v as a one-hot (unary) vector $\mathbf{x} \in \{0,1\}^k$, where only the bit corresponding to the index of v is set to 1, and all other bits are set to 0. This step transforms categorical data into a binary vector suitable for bitwise randomization.
2. **Bitwise Perturbation**: Each bit of $\mathbf{x}$ is independently randomized to obtain a perturbed vector $\mathbf{x}'$. The randomization is governed by the following probabilities:

$$\Pr[x'_i = 1] = \begin{cases} p, & \text{if } x_i = 1 \\ q, & \text{if } x_i = 0 \end{cases} \quad \text{for all } i \in [1..k].$$

The vector $\mathbf{x}'$ is then sent to the server for aggregation. Different UE variants are defined by specific choices of p and q, which directly affect the privacy-utility trade-off.

Two well-known instantiations of the UE protocol include:

- **Symmetric Unary Encoding (SUE)** (ERLINGSSON *et al.*, 2014): Sets $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$, such that $p + q = 1$. This symmetric configuration simplifies the analysis and is adopted in Google's RAPPOR deployment.
- **Optimal Unary Encoding (OUE)** (WANG *et al.*, 2017a): Sets $p = \frac{1}{2}$ and $q = \frac{1}{e^{\epsilon}+1}$. This configuration minimizes variance and improves utility while maintaining ϵ-LDP.

After collecting the perturbed vectors from n users, the server estimates the normalized frequency of each $v \in V$ using the following unbiased estimator:

$$\hat{f}(v) := \frac{C(v) - nq}{n(p-q)},$$

where $C(v)$ is the number of times the bit corresponding to value v was reported as 1 across all users. This estimator remains unbiased under the defined perturbation model (WANG *et al.*, 2017a).

Both SUE and OUE protocols satisfy ϵ-LDP, with the privacy loss parameter ϵ de-

fined as:

$$\epsilon = \ln\left(\frac{p(1-q)}{(1-p)q}\right).$$

While UE protocols are efficient and statistically robust for moderate domain sizes, their communication cost grows linearly with k, making them less suitable for high-cardinality domains compared to hashing-based approaches.

Example: Estimating Favorite Color via Unary Encoding

Continuing with the same scenario, suppose users are asked to report their favorite color from a small domain:

$$V = \{\text{Red}, \text{Green}, \text{Blue}\}, \quad |V| = 3$$

Each value $v \in V$ is mapped to a one-hot vector of length 3:

$$\text{Red} \to [1,0,0], \quad \text{Green} \to [0,1,0], \quad \text{Blue} \to [0,0,1]$$

Assume the use of the **Optimal Unary Encoding (OUE)** variant with privacy budget $\epsilon = \ln(3)$. Then:

$$p = \frac{1}{2}, \quad q = \frac{1}{e^{\epsilon}+1} = \frac{1}{4} = 0.25$$

Each user perturbs their one-hot vector such that:

- The true bit (the 1) is kept as 1 with probability $p = 0.5$, and flipped to 0 otherwise.
- Each 0 bit is flipped to 1 with probability $q = 0.25$, and remains 0 otherwise.

3 users respond as follows:

User	True Color	Unary Vector	Perturbed Report
1	Red	[1, 0, 0]	[1, 0, 1]
2	Blue	[0, 0, 1]	[0, 1, 1]
3	Green	[0, 1, 0]	[0, 1, 0]

The aggregator sums the bits for each index (Red, Green, Blue):

$$C(\text{Red}) = 1, \quad C(\text{Green}) = 2, \quad C(\text{Blue}) = 2$$

Apply the OUE frequency estimator:

$$\hat{f}(v) = \frac{C(v) - nq}{n(p-q)} = \frac{C(v) - 3 \cdot 0.25}{3(0.5 - 0.25)} = \frac{C(v) - 0.75}{0.75}$$

- **Red:** $\widehat{f}(\text{Red}) = \frac{1-0.75}{0.75} = 0.33$
- **Green:** $\widehat{f}(\text{Green}) = \frac{2-0.75}{0.75} = 1.67$
- **Blue:** $\widehat{f}(\text{Blue}) = \frac{2-0.75}{0.75} = 1.67$

These results demonstrate how UE can recover frequency estimates from privatized binary vectors. As with previous methods, a larger number of users would reduce estimation variance and improve accuracy. Despite some overestimation (e.g., values > 1), the method provides unbiased estimators that converge with more samples.

2.6 Memoization in LDP Protocols

In longitudinal scenarios, where users submit data across multiple time steps, repeatedly applying randomized mechanisms to the same values may lead to significant privacy budget consumption. To address this issue, several Local Differential Privacy (LDP) protocols employ a strategy known as *memoization* (ERLINGSSON *et al.*, 2014; WANG *et al.*, 2017a).

The core idea behind memoization is that the user performs the randomization only once per observed value. The perturbed output is then stored locally and reused whenever the same input value reoccurs. As a result, even if the same true value is reported over time, the user consistently returns the same randomized response, thereby preventing multiple privacy losses for identical inputs.

For example, in protocols such as LOLOHA and L-OSUE (ARCOLEZI *et al.*, 2023), which are detailed in the Section 3.1. In these protocols, when a value $v \in D$ is encountered for the first time, it is passed through a perturbation mechanism (e.g., BitFlip or Unary Encoding) to generate a noisy output v'. The pair $\langle v, v' \rangle$ is memoized locally, and future reports of v simply return v'.

This process is illustrated in Figure 8, where the same input value maps to the same randomized output across multiple occurrences, highlighting how memoization helps mitigate cumulative privacy loss.

While memoization significantly improves privacy guarantees under fixed budgets, it introduces utility challenges, particularly when the representation of values changes over time due to domain partitioning. If the domain structure evolves (e.g., through quadtree splitting or merging), previously memoized responses may become inconsistent with the current representation. This limitation motivates several of the design choices in ALOQ, such as restricting arbitrary merging and splitting over time. These constraints help preserve structural consis-

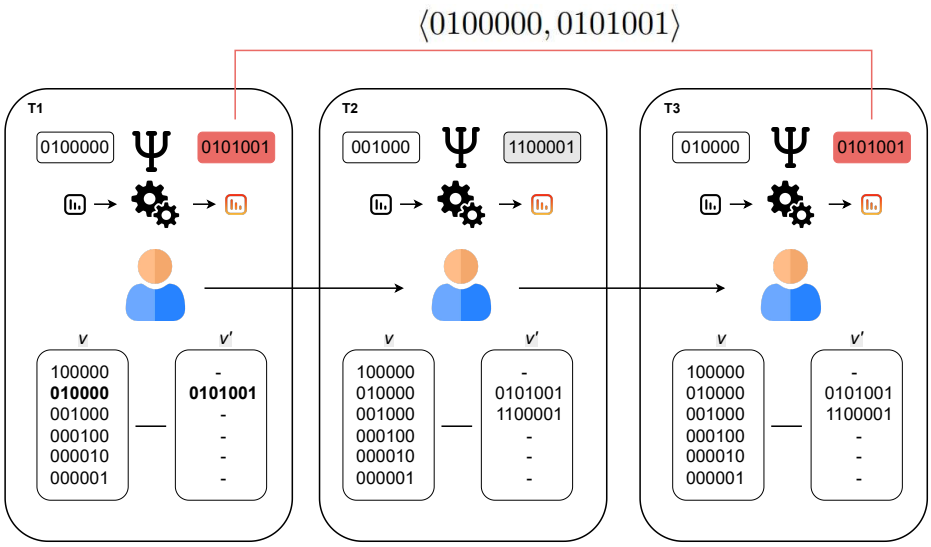


Figure 8 – Illustration of memoization in local differential privacy protocols. Each unique value v is randomized once into v' and reused on subsequent occurrences.

tency, ensuring that memoized responses remain valid across the reporting window, as further discussed in Chapter 4.

2.7 Space Partitioning

Spatial regions are commonly represented using map projections such as the Mercator projection, which transforms geographic coordinates (latitude and longitude) into a planar format suitable for computational processing (PETERSON, 2014). While effective for visualization, analytical tasks often require discretizing this continuous space into smaller, manageable subregions. Traditional approaches frequently employ uniform grids due to their simplicity; however, such fixed-size partitioning fails to account for variations in spatial data density (SHEKHAR *et al.*, 2011).

Quad-Tree

In contrast, the quad-tree is a hierarchical spatial index widely used for representing two-dimensional data (FINKEL; BENTLEY, 1974). The structure recursively partitions a region into four quadrants, or cells, until a stopping condition is met, such as reaching a minimum cell size or a threshold on the number of contained points. This recursive decomposition produces a tree structure where each internal node corresponds to a partitioned region and each leaf corresponds to a spatial cell. Quad-trees are attractive in data analysis because they provide a natural multi-resolution view of the space: dense areas are subdivided into fine-grained cells, while sparse areas remain aggregated in larger regions.

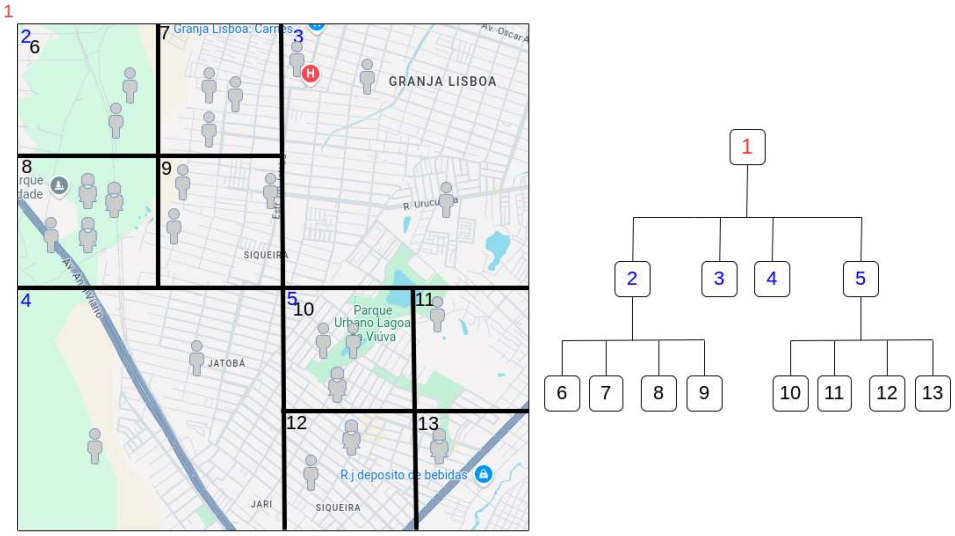


Figure 9 – Partitioning the space into quadtree

Several variations of quad-trees have been proposed to adapt the partitioning strategy to different application needs:

- Point quad-trees partition space based on the distribution of individual data points (FINKEL; BENTLEY, 1974).
- Region quad-trees subdivide the entire space uniformly, regardless of data distribution (SAMET, 1984).
- Adaptive quad-trees further refine dense regions while leaving sparse regions coarse, leading to more balanced trees and efficient storage (SAMET, 1990).
- Fuzzy quad-trees extend standard quad-trees by allowing cells to overlap through degrees of membership, so objects can belong partly to multiple regions. This soft partitioning makes them better suited for uncertain or imprecise spatial data, such as vague boundaries or noisy sensor inputs. (MINH *et al.*, 2024)

In summary, quadtree-based partitioning provides a flexible, hierarchical approach for spatial data representation. Its ability to adapt resolution to local density or uncertainty makes it especially well suited for skewed and heterogeneous distributions, offering both efficient storage and meaningful analytical granularity.

Location Encoding via Unary Encoding

Once a spatial quadtree has been constructed, each of its leaf nodes defines a distinct spatial cell. Let $Q = \{q_1, q_2, \dots, q_k\}$ be the current set of leaves, where $k = |Q|$ is the number of active leaf nodes. A user located in cell $q_i \in Q$ encodes their position as a one-hot binary

vector $\mathbf{v} \in \{0,1\}^k$, where:

$$v_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

This encoding is known as *Unary Encoding* (UE) and serves as the input to subsequent privacy-preserving mechanisms. Its main advantage lies in its compatibility with bitwise randomization techniques used in Local Differential Privacy (LDP).

Example: Consider a spatial quadtree with eight leaf nodes (i.e., $k = 8$). A user located in the region corresponding to leaf node q_3 encodes their location as:

$$\mathbf{v} = [0, 0, 1, 0, 0, 0, 0, 0]$$

This vector is then perturbed using a UE-based mechanism such as Optimal Unary Encoding (OUE), which introduces controlled noise to each bit according to parameters p and q. The randomized output is sent to the aggregator, who can estimate the spatial frequency distribution without directly observing any user's true location.

By discretizing continuous spatial data into an adaptive quadtree structure and encoding locations accordingly, it becomes possible to anonymize and aggregate geospatial information in a manner that preserves utility while ensuring rigorous privacy guarantees.

2.7.1 Non-Uniformity in Spatial Data

Real-world mobility and location datasets are rarely uniformly distributed. Instead, user reports tend to concentrate in specific areas, such as urban centers, workplaces, or transportation hubs, while being sparse or even absent in others, like residential zones or rural regions (ZANG; BOLOT, 2011; PAPPALARDO *et al.*, 2015).

This **non-uniformity** introduces a fundamental challenge when estimating spatial frequencies under Local Differential Privacy (LDP). Because LDP mechanisms (e.g., GRR, LH, UE) apply noise uniformly across the input domain, regions with different densities are affected differently:

- **High-density regions** (e.g., city centers) suffer from *noise saturation*: while there are many true signals, the added random noise can still dominate the signal if the domain is large or the privacy budget is tight.

- **Low-density regions** (e.g., suburban or rural areas) suffer from *over-noise amplification*: since few users report data in these areas, even modest noise can overwhelm the signal, causing frequency estimates to become unreliable or even negative (after debiasing).

Uniform discretizations, such as fixed grids or static hash buckets, tend to amplify this problem. Sparse regions accumulate noise without sufficient signal to counterbalance it, while dense regions are forced to aggregate many reports into coarse buckets, leading to a loss of resolution.

This phenomenon, often referred to as the **non-uniformity error**, has been documented in several studies (QARDAJI *et al.*, 2013; YE *et al.*, 2021), and is one of the main motivations for adaptive mechanisms. These adaptive approaches dynamically refine the spatial representation to match the data distribution, either by subdividing dense regions (e.g., via quadtrees) or clustering hotspots to optimize utility under LDP constraints.

Impact of Misconfigured Privacy Budgets

An inadequate choice of privacy budget ϵ can exacerbate the non-uniformity error. For instance, setting ϵ too high in an attempt to preserve utility in dense areas may inadvertently reduce protection in sparse regions, where individual presence becomes more distinguishable. Conversely, setting ϵ too low may add excessive noise across the board, suppressing meaningful signals even in high-density zones.

Example: Consider a sparse rural region in which only one or two users report their locations over a given time interval. Suppose the domain is discretized using a static grid, and the privacy mechanism assigns $\epsilon = 1.5$. In high-density areas, this noise level may suffice to mask individual contributions. However, in the rural region, if a single report is received in a grid cell, and the noise is not sufficient to obfuscate its origin, an attacker with auxiliary information (e.g., the user's home region or mobility history) could easily infer the individual's location, leading to a privacy breach.

This risk becomes particularly acute when privacy guarantees are evaluated globally but applied uniformly, neglecting local sparsity. Adaptive techniques, such as the ALOQ method introduced in this work, address this challenge by focusing refinement and noise allocation where they are most needed.

Designing privacy mechanisms that are robust to non-uniform data distributions and adaptive to local density remains an open research challenge, especially in high-dimensional or temporally evolving scenarios. Adaptive data structures like ALOQ are a promising direction to

mitigate this imbalance by coupling spatial partitioning with localized noise control and density-aware budget allocation.

2.8 Chapter Summary

This chapter presented the foundational concepts necessary to understand the privacy challenges and solutions in geospatial data collection and analysis. We introduced the principles of Differential Privacy (DP) and Local Differential Privacy (LDP), discussed key mechanisms for data perturbation, and explored frequency estimation techniques over discretized spatial domains. Additionally, we highlighted the limitations of static spatial partitioning methods when applied to dynamic and non-uniform mobility data.

These concepts provide the theoretical groundwork for the remainder of this thesis. In the next chapter, we review related work and examine how existing approaches address, or fail to address, the specific challenges posed by longitudinal scenarios and adaptive spatial partitioning under LDP.

3 RELATED WORK

Centralized models have been widely explored to address the problem of estimating the frequency of individuals in specific regions of a geospatial area, such as a city, while maintaining user privacy. These approaches typically assume a trusted curator who aggregates and anonymizes user data. A variety of privacy models have been employed in this context, including differential privacy and several of its relaxed or modified variants, each with distinct trade-offs between privacy guarantees and utility. Due to these differences in privacy definitions and corresponding parameter adjustments, direct comparisons across such models are not straightforward. A common element among centralized approaches is the use of grid-based partitioning, discretizing geographic space into cells to facilitate spatial density calculations (**??**WEI *et al.*, 2019; YAN *et al.*, 2022).

Geo-Indistinguishability (Geo-Ind) is a privacy notion specifically designed for location based systems, introduced by Andrés et al. (2013). This concept generalizes the standard definition of Differential Privacy (DP) by incorporating a spatial dimension, offering a formal privacy guarantee tailored for geographic data. Geo-Ind achieves privacy protection by adding controlled random noise, drawn from a planar Laplace distribution, to the user's precise geographic location before it is shared with a service provider. The fundamental idea behind Geo-Ind is that privacy protection weakens gradually with increasing distance: it provides strong protection within a small radius around the user's exact location, while allowing broader approximate information to be released at greater distances. Consequently, Geo-Ind ensures users remain indistinguishable within a small area (defined by a radius), but allows larger-scale location information, such as city-level location, to be revealed with higher probability as distances increase. The author in (KIM, 2024) proposes a novel approach that combines Geo-Indistinguishability (Geo-Ind) (ANDRÉS *et al.*, 2013) with adaptive grids, which dynamically adjust the granularity of spatial discretization according to local user density. This combination allows for finer granularity in densely populated regions and coarser granularity in sparse regions, thereby enhancing the utility of spatial queries relative to traditional static-grid methods. However, it is important to note that Geo-Ind represents a relaxed version of traditional differential privacy, making privacy adjustments based on the spatial context. As such, it is not straightforward to directly compare its guarantees with stricter, more standardized privacy models like Local Differential Privacy (LDP), which do not inherently adjust based on spatial proximity.

In contrast to Geo-Ind, several works have explored Local Differential Privacy (LDP),

a stricter privacy framework where data perturbation occurs locally on user devices, independently of any spatial proximity considerations. Unlike Geo-Ind and centralized approaches, LDP protocols typically provide uniform privacy guarantees that do not weaken based on spatial context. For instance, the authors in (WANG *et al.*, 2017b) focus on developing protocols that enable frequency estimation under LDP, ensuring user-level privacy without relying on a trusted curator. Their work introduces optimized protocols such as Optimized Unary Encoding (OUE) and Optimized Local Hashing (OLH), which significantly enhance the accuracy of frequency estimation tasks, thereby making LDP practical and robust for real-world privacy-preserving applications.

To apply LDP effectively in geospatial contexts, PrivTC (YANG *et al.*, 2022) introduces an adaptive spatial discretization method named PrivAG. PrivAG is designed specifically for constructing adaptive grids by dynamically adjusting cell granularity based on user data distribution. It employs a two-phase hierarchical partitioning strategy, initially dividing the geographic domain into a coarse-grained grid, which is broadcasted to a subset of users. Each user discretizes their entire trajectory based on this initial grid, mapping geographic points to corresponding cells. These discretized locations are then perturbed locally via OLH to ensure strict LDP constraints before being transmitted to an untrusted aggregator. Using these noisy frequency estimates, the aggregator calculates an average frequency for each cell and subsequently refines the grid granularity. Cells with higher frequencies (indicating dense data distributions) are subdivided further, while lower-frequency cells remain coarse-grained. This adaptive partitioning is executed once, resulting in an optimized grid utilized by all subsequent users, significantly enhancing the precision and utility of spatial analyses under LDP.

Beyond single-instance spatial analyses, LDP has also been extensively studied in longitudinal scenarios, where data is collected repeatedly over extended periods. Protocols such as RAPPOR (ERLINGSSON *et al.*, 2014), L-OSUE (ARCOLEZI *et al.*, 2024), and LOLOHA (ARCOLEZI *et al.*, 2023) have specifically addressed continuous and repeated data collection challenges. For example, RAPPOR facilitates longitudinal frequency estimation by locally perturbing users' data using Bloom filters, making it highly suitable for continuous, real-time location-based data collection scenarios.

Among these, the contributions in (ARCOLEZI *et al.*, 2024) stand out for enhancing the utility of longitudinal frequency estimations through refinements to existing LDP protocols. Their proposal, L-OSUE, integrates Optimized Unary Encoding (OUE) and Symmetric Unary Encoding (SUE) to improve accuracy across repeated collections, while LOLOHA builds

upon OLH by dynamically adjusting hash functions to accommodate evolving data distributions. These adaptations significantly enhance accuracy in real-time tracking scenarios. Further advancing this line of work, Cunningham et al. (CUNNINGHAM *et al.*, 2021) proposed an LDP-based mechanism tailored for trajectory data, using perturbation over hierarchically structured, overlapping n-grams. Their approach incorporates a multidimensional hierarchy of real-world points of interest to increase the semantic coherence and analytical value of perturbed trajectories, while preserving strong privacy guarantees in longitudinal and spatial-temporal data collection tasks.

While these LDP protocols offer important advances, they generally rely on static spatial discretization structures, such as uniform grids or fixed domain representations. In the context of real-time location data collection, however, such static structures may fail to adequately reflect the dynamic and uneven distribution of user mobility across space and time. This can lead to inefficiencies in both privacy protection and data utility. To more effectively support continuous geolocation reporting under LDP, it is crucial to adopt a spatial representation that can evolve in response to changing data distributions.

Our work addresses this gap by proposing a novel real-time data collection model under LDP, in which location data are discretized using a quadtree-based spatial partitioning structure. Unlike uniform grids, quadtrees offer hierarchical, non-uniform partitions that can adapt more flexibly to heterogeneous spatial densities. In our framework, the quadtree evolves continuously over time, guided by sanitized frequency estimates obtained under LDP, ensuring that the structure remains aligned with the actual data distribution without compromising user privacy. To the best of our knowledge, this is the first approach that integrates an evolving quadtree structure with LDP in a longitudinal, real-time setting.

To assess the effectiveness of our method, we conduct a comparative evaluation against leading LDP-based protocols, RAPPOR, L-OSUE, and LOLOHA, which rely on static, uniform spatial discretizations. These experiments will demonstrate the benefits of dynamic, data-driven partitioning in terms of both privacy guarantees and analytical accuracy, highlighting the potential of adaptive structures in the future of privacy-preserving location data collection. In addition to these baseline mechanisms, we also consider a recent and structurally more complex framework, PrivTC, which combines adaptive spatial partitioning with statistical modeling of user trajectories under LDP. Given its conceptual proximity to our approach, we dedicate a separate subsection to describing PrivTC's architecture in more detail, in order to support a fair

and focused comparison in later sections.

3.1 RAPPOR

RAPPOR (ERLINGSSON *et al.*, 2014) is an influential work in the development of LDP and frequency estimators. It makes use of two rounds of sanitization and *memoization* to provide LDP guarantees for several queries over time. The first round of sanitization is referred to as the PRR step (permanent randomized response) and the second as the IRR step (instantaneous randomized response), both a reference to the randomized response algorithm and their respective functions. PRR adds persistent noise to values still not memoized and then the sanitizer memoizes the noisy values, while IRR adds noise to every instance of a memoized value before sending a report to the aggregator. As the domain size gets larger, it becomes a challenge to apply noise in a utility-preserving way, thus RAPPOR implements Unary Encoding as a means of guaranteeing LDP while reducing data utility loss. In Unary Encoding, a value $v \in D$ with domain size k is encoded as a length-k binary vector $B = [0,\cdots,0,1,0,\cdots,0]$ where only the v-th position is 1.

The perturbation algorithm for RAPPOR and other UE-based LDP protocols is as follows:

Given the probabilities *p1, p2, q1, q2* to flip the i-th position of the binary vector B, for the first sanitization round, we have

$$Pr[\Psi_{UE_{(\epsilon)}} B'[i] = 1] = \begin{cases} p1, if B[i] = 1 \\ q1, if B[i] = 0 \end{cases}$$

and for the second

$$Pr[\Psi_{UE_{(\epsilon)}} B'[i] = 1] = \begin{cases} p2, if B[i] = 1 \\ q2, if B[i] = 0 \end{cases}$$

p1, p2, q1, q2 are computed as function of ϵ_∞ and ϵ_1 by symmetric unary encoding (SUE) to satisfy: $p1 = \frac{e^{\epsilon_\infty/2}}{e^{\epsilon_\infty/2}+1}$, $q1 = \frac{1}{e^{\epsilon_\infty/2}+1}$, $p1 + q1 = p2 + q2 = 1$ and

$$\epsilon_1 = ln(\frac{(p_1p_2 - q_2(p_1-1))(p_2q_1 - q_2(q_1-1)-1)}{(p_2q_1 - q_2(q_1-1))(p_1p_2 - q_2(p_1-1)-1)}) \tag{3.1}$$

ϵ_∞ is an upper bound for the privacy budget as the number of reports sampled by the frequency estimator tends to infinity, and ϵ_1 is a lower bound for a single report.

All frequency estimators presented in this research, including those based on the RAPPOR protocol discussed in this section and the L-OSUE and LOLOHA protocols covered in the following two sections, utilize the same unbiased function to estimate frequencies at the aggregator(ERLINGSSON *et al.*, 2014; ARCOLEZI *et al.*, 2024; ARCOLEZI *et al.*, 2023):

$$\Phi_{f_L}(v) := \frac{C(v) - nq_1(p_2 - q_2) - nq_2}{n(p_1 - q_1)(p_2 - q_2)} = \frac{\frac{C(v)/n - q_2}{p_2 - q_2} - q_1}{p_1 - q_1} \tag{3.2}$$

where *n* is the number of reports, *p1, p2 , q1, q2* are the probabilities for perturbing the data, and *C(v)* is the count of reports with a value *v*, given $v \in D$.

Example: Suppose a city is discretized into $k = 4$ zones: $\{\text{Downtown}, \text{North}, \text{East}, \text{South}\}$, indexed as $\{1, 2, 3, 4\}$. A user currently located in the East zone wants to report her location using the RAPPOR protocol with domain size $k = 4$.

First, the user's true location $v = 3$ is encoded via Unary Encoding as the binary vector:

$$B = [0, 0, 1, 0]$$

First Round (PRR): For each bit in B, noise is applied using the probabilities p_1 and q_1. Assume the protocol uses $\epsilon_\infty = 2$, so that:

$$p_1 = \frac{e^1}{e^1 + 1} \approx 0.73, \quad q_1 = \frac{1}{e^1 + 1} \approx 0.27$$

Each bit is perturbed:

- Bits 1, 2, and 4 (zeros): flipped to 1 with probability $q_1 = 0.27$
- Bit 3 (one): flipped to 1 with probability $p_1 = 0.73$

Suppose the perturbed vector becomes:

$$B' = [0, 0, 1, 1]$$

This result is memoized for the user.

Second Round (IRR): At each reporting instance, B' is further perturbed using probabilities p_2 and q_2, derived from ϵ_1. For simplicity, assume $p_2 = 0.7$ and $q_2 = 0.3$.

Then the final reported vector B'' is produced by flipping each bit in B':

- Bits 1 and 2: remain 0 with probability $1 - q_2 = 0.7$
- Bit 3: remains 1 with probability $p_2 = 0.7$
- Bit 4: remains 1 with probability $p_2 = 0.7$

Assume the reported vector is:

$$B'' = [0, 0, 1, 0]$$

This vector is sent to the aggregator. Over time, the aggregator collects many such reports and applies Equation (3.2) to estimate the frequency of each location zone.

3.2 L-OSUE

L-OSUE is a protocol proposed in (ARCOLEZI *et al.*, 2024) that enhances longitudinal data collection under Local Differential Privacy (LDP) by combining the advantages of two Unary Encoding (UE) protocols: Optimized Unary Encoding (OUE) for the first sanitization step and Symmetric Unary Encoding (SUE) for the second. This approach aims to improve the accuracy of frequency estimation while maintaining strong privacy guarantees across repeated data collections.

As with RAPPOR, L-OSUE relies on a two-stage sanitization process:

Step 1 – Permanent Randomized Response (PRR): The user first encodes the private value $v \in D$ as a one-hot binary vector B of length $k = |D|$. For the PRR step, L-OSUE applies OUE, where the bit corresponding to the true value is set to 1 and the rest to 0. The vector is then perturbed using fixed probabilities:

$$p_1 = \frac{1}{2}, \quad q_1 = \frac{1}{e^{\epsilon_\infty} + 1}$$

Here, ϵ_∞ denotes the total privacy budget consumed as the number of reports grows indefinitely. The perturbed vector is then memoized and reused as the input to the next step, ensuring longitudinal privacy over time.

Step 2 – Instantaneous Randomized Response (IRR): At each reporting time t, the user perturbs the memoized vector again using SUE. The probabilities for this second round are defined by:

$$p_2 = \frac{e^{\epsilon_{\text{IRR}}}}{e^{\epsilon_{\text{IRR}}} + 1}, \quad q_2 = \frac{1}{e^{\epsilon_{\text{IRR}}} + 1}$$

where ϵ_{IRR} is derived from both ϵ_∞ and the desired single-report privacy guarantee ϵ_1. Specifically:

$$\epsilon_{\text{IRR}} = \ln\left(\frac{e^{\epsilon_\infty+\epsilon_1}-1}{e^{\epsilon_\infty}-e^{\epsilon_1}}\right)$$

The output of this step, B'', is the sanitized report sent to the server.

Example: Consider a city discretized into $k = 4$ zones: $\{\text{Downtown}, \text{North}, \text{East}, \text{South}\}$, indexed as $\{1, 2, 3, 4\}$. A user currently located in the *East* zone wants to report her location using the L-OSUE protocol, with domain size $k = 4$.

First, the true location $v = 3$ is encoded using Unary Encoding as:

$$B = [0, 0, 1, 0]$$

Step 1 – PRR (OUE): The bit vector B is perturbed using the Optimized Unary Encoding mechanism. Let the total privacy budget be $\epsilon_\infty = 2$. Then:

$$p_1 = \frac{1}{2}, \quad q_1 = \frac{1}{e^2+1} \approx 0.119$$

Each bit is randomized independently:

- Bit 3 (value 1): kept as 1 with probability $p_1 = 0.5$
- Bits 1, 2, and 4 (value 0): flipped to 1 with probability $q_1 \approx 0.119$

Suppose the perturbed vector is:

$$B' = [0, 0, 1, 0]$$

This memoized vector B' is stored for future reporting.

Step 2 – IRR (SUE): At each report time, the memoized vector B' is perturbed again using SUE. Assume that a single-report privacy guarantee $\epsilon_1 = 1$ is desired.

Using:

$$\epsilon_{\text{IRR}} = \ln\left(\frac{e^{\epsilon_\infty+\epsilon_1}-1}{e^{\epsilon_\infty}-e^{\epsilon_1}}\right) = \ln\left(\frac{e^3-1}{e^2-e^1}\right) \approx 0.88$$

Then:

$$p_2 = \frac{e^{0.88}}{e^{0.88}+1} \approx 0.71, \quad q_2 \approx 0.29$$

Each bit in B' is randomized again:

- Bit 3 (value 1): kept as 1 with probability $p_2 \approx 0.71$
- Bits 1, 2, and 4 (value 0): flipped to 1 with probability $q_2 \approx 0.29$

Suppose this yields the final reported vector:

$$B'' = [0, 0, 1, 1]$$

This vector is sent to the aggregator. Over time, with multiple users and multiple reports, the server applies a debiasing estimator to recover the frequency of each location zone.

3.3 LOLOHA

Unlike RAPPOR and L-OSUE, LOLOHA (ARCOLEZI *et al.*, 2023) does not make use of unary encoding, it instead improves utility by shirking the domain size through local hashing. Like the previously cited estimators, LOLOHA implements two rounds of sanitization and memoization to provide LDP guarantees while maintaining utility for queries executed over time. In this thesis, we will use OLOLOHA, the optimal configuration of LOLOHA. It shrinks a domain size k to an optimal hashed size g, that is computed by

$$g = 1 + \max\left(1, \left\lfloor \frac{1 - a^2 + \sqrt{a^4 - 14a^2 + 12ab(1-ab) + 12a^3b + 1}}{6(a-b)} \right\rceil\right) \tag{3.3}$$

where $a = e^{\epsilon_\infty}$ and $b = e^{\alpha\epsilon_\infty}$ for $\alpha \in (0, 1)$.

In local hashing, given an original domain size k and input values $v \in V$, a randomly chosen universal hash function H maps the original domain to a range $[1...g]$ with $g \geq 2$. Approximately k/g values $v \in V$can be mapped to the same hashed value $H(v) \in [1...g]$ due to collision. The sanitization algorithm for LOLOHA is as follows:

$$\forall_{x \in [1...g]}\ Pr\left[\Psi_{LOLOHA_{(\epsilon_\infty)}}(v) = x\right] = \begin{cases} p_1 = \frac{e^\epsilon}{e^\epsilon + g - 1} & \text{if x = v} \\ q_2 = \frac{1}{e^\epsilon + g - 1} & \text{if x} \neq \text{v} \end{cases}$$

followed by a second round that outputs a report x':

$$\forall_{x' \in [1...g]}\ Pr\left[\Psi_{L-GRR_{(\epsilon_1)}}(x) = x'\right] = \begin{cases} p_2 & \text{if x' = x} \\ q_2 = \frac{1-p_2}{g-1} & \text{if x'} \neq \text{x} \end{cases}$$

where $p_2 = \frac{q_1 - e^{\epsilon_1} p_1}{(-p_1 e^{\epsilon_1}) + g q_1 e^{\epsilon_1} - q_1 e^{\epsilon_1} - p_1(g-1) + q_1}$ as $\epsilon_1 = ln(\frac{p_1 p_2 + q_1 q_2}{p_1 q_2 + q_1 p_2})$ for LOLOHA.

After the sanitization step, the reports, privacy budgets, and hash function seeds are sent to the aggregator, which then outputs estimates with the original domain size k.

Example: Suppose the original location domain contains $k = 8$ zones (e.g., neighborhoods), indexed from 1 to 8. A user is located in zone $v = 5$, and the protocol uses LOLOHA to report this value under LDP.

Step 1 – Hashing and First Round of Sanitization (L-GRR): The user applies a local hash function H from a universal hash family that maps the original domain $[1..8]$ to a reduced domain $[1..g]$, with $g = 4$.

Let the hash function be defined as:

$$H(v) = (v \mod g) + 1$$

In this case:

$$H(5) = (5 \mod 4) + 1 = 2$$

Thus, the true value $v = 5$ is mapped to hashed value $x = 2$.

LOLOHA then applies the first perturbation using L-GRR:

$$\Pr[\Psi_{\text{L-GRR}_{(\epsilon_\infty)}}(x) = x'] = \begin{cases} p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+g-1}, & \text{if } x' = x \\ q_1 = \frac{1}{e^{\epsilon_\infty}+g-1}, & \text{otherwise} \end{cases}$$

Assuming $\epsilon_\infty = 2$, we get:

$$p_1 \approx 0.71, \quad q_1 \approx 0.096$$

Suppose the randomized output of this first round is:

$$x' = 2$$

The value x' is memoized and used in future reports.

Step 2 – Second Round of Sanitization: Before each report, LOLOHA applies a second perturbation to x', again using L-GRR, but now parameterized by $\epsilon_1 = 1$.

The probabilities are calculated so that:

$$\Pr[\Psi_{\text{L-GRR}_{(\epsilon_1)}}(x') = x''] = \begin{cases} p_2, & \text{if } x'' = x' \\ q_2 = \frac{1-p_2}{g-1}, & \text{otherwise} \end{cases}$$

Suppose this yields:

$$p_2 = 0.6, \quad q_2 = 0.133$$

And the final reported value is:

$$x'' = 3$$

The user sends the perturbed hashed value $x'' = 3$ along with the seed used to generate H. The aggregator then collects these reports from multiple users, applies the inverse hash mapping to recover candidate original values, and estimates frequencies in the original domain $[1..8]$.

This example illustrates how LOLOHA reduces the domain via hashing and applies randomized response in two phases, balancing privacy and utility under longitudinal LDP constraints.

3.4 PrivTC

PrivTC (YANG *et al.*, 2022) is a recent and comprehensive solution for collecting individual mobility trajectories under local differential privacy (LDP). Its core idea is to synthesize a realistic trajectory dataset from perturbed user trajectories, preserving privacy through an adaptive spatial discretization and statistical modeling approach. The framework is composed of three key phases: adaptive grid construction, HMM learning, and synthetic data generation.

3.4.1 Adaptive Grid Construction (PrivAG)

To address the challenge of discretizing continuous space without losing critical spatial characteristics, PrivTC introduces PrivAG, an adaptive grid construction algorithm. Unlike static grids, PrivAG employs a two-level hierarchical partitioning strategy driven by noisy user data.

The process begins by dividing the user base into two distinct groups, U_1 and U_2. The first group, U_1, is dedicated solely to the grid construction phase. The aggregator first defines a coarse, uniform base grid, G_1, and broadcasts it to the users in U_1. These users discretize their trajectories according to G_1 and report the noisy frequency of each cell using the Optimized Local Hashing (OLH) mechanism.

Based on these noisy frequency reports, the aggregator then performs the second level of partitioning. Each cell in the coarse grid G_1 is further subdivided into a finer grid, G_2. Crucially, the granularity of this sub-grid is adaptive. The paper provides an analytical guideline for choosing the optimal granularities for both levels, aiming to balance the trade-off between sampling error, noise error, and non-uniformity error. Specifically, the granularity of the coarse grid, g_1, is determined by:

$$g_1 = \sqrt{2\alpha(e^\epsilon - 1)\sqrt{\frac{n}{e^\epsilon}}} \tag{3.4}$$

where α is a small constant, n is the total number of users, and ϵ is the privacy budget. Subsequently, the granularity of the sub-grid for each cell k, denoted as g_2^k, is adaptively set based on its noisy frequency f_k:

$$g_2^k = \sqrt{2\alpha f_k(e^\epsilon - 1)\sqrt{\frac{(1-\sigma)n}{e^\epsilon}}} \tag{3.5}$$

where σ represents the proportion of users in the first group, U_1. This method results in a more fine-grained partition for dense areas (higher f_k) and a coarser partition for sparse areas. The final, two-level adaptive grid is then used by the second user group, U_2, for the trajectory modeling phase.

3.4.2 Trajectory Modeling and Synthesis

Once the adaptive grid is established, users in the second group, U_2, discretize their trajectories accordingly. PrivTC then applies a privacy-preserving spectral learning method, PrivSL, to learn the parameters of an HMM that represents the collective mobility patterns. Finally, this learned HMM is used by the aggregator to generate a complete synthetic trajectory dataset that approximates the original data distribution while satisfying ϵ-LDP.

3.4.3 Discussion

PrivTC is particularly relevant to our work because it also combines adaptive spatial partitioning with local privacy mechanisms. However, its main objective is trajectory generation rather than direct spatial frequency estimation over time. Furthermore, PrivTC relies on a two-phase user grouping model and a single, upfront grid construction phase. In contrast, ALOQ

supports iterative temporal evolution of its spatial structure for all users and focuses on providing continuous frequency analysis. In our comparative evaluation, we adapt the grid construction phase of PrivTC to enable a direct comparison with ALOQ's adaptive quadtree partitioning in terms of utility, convergence, and privacy-budget alignment.

3.5 Summary Comparison of Related Works

To consolidate the discussion of existing work, we summarize in Table 1 the main characteristics of the most relevant LDP-based approaches for spatial data collection. The comparison highlights key design choices across the literature, such as the spatial partitioning strategy, support for user trajectories, encoding mechanisms, adaptivity to data distribution, and reporting model. This overview clarifies the methodological positioning of ALOQ relative to prior approaches, and sets the stage for the empirical evaluation presented in the next chapter.

Table 1 – Comparison of LDP-based spatial data collection protocols

Approach	Partitioning	Encode	Reporting	Post-processing
RAPPOR (2014)	Uniform Grid	Unary Enconing	Continuous	Frequency histogram
L-OSUE (2022)	Uniform Grid	Unary Encoding	Continuous	Frequency histogram
LOLOHA (2022)	Uniform Grid	Optimized Local Hashing	Continuous	Frequency histogram
PrivTC (2022)	Adaptive Grid	Optimized Local Hashing	Grouped	Trajectory Estimation
ALOQ (ours)	Adaptive Quadtree	Unary Encoding	Continuous	Frequency histogram

While Table 1 provides a structural and algorithmic comparison, Table 2 focuses on highlighting the key differentiators across these protocols. This view emphasizes which capabilities are fully supported by each method.

Table 2 – Key capabilities of spatial LDP protocols

Approach	Adaptive Partition	Adaptive Budget Allocation	Trajectory Protection	Iterative Reporting	Frequency Estimation
RAPPOR (2014)	✗	✗	✓	✓	✓
L-OSUE (2022)	✗	✗	✓	✓	✓
LOLOHA (2022)	✗	✗	✓	✓	✓
PrivTC (2022)	✓	✗	✓	✗	✗
ALOQ (ours)	✓	✓	✓	✓	✓

As shown in the tables, traditional methods such as RAPPOR, LOLOHA, and L-OSUE rely on static, uniform discretizations, which limits their ability to adapt to non-uniform data distributions. PrivTC and ALOQ represent a newer class of solutions that employ adaptive

spatial partitioning. However, they differ in purpose and design: PrivTC focuses on synthetic trajectory generation through spectral modeling, while ALOQ directly supports spatial frequency estimation over time using a dynamic quadtree and threshold-based refinement.

By combining adaptivity, frequency estimation, and iterative reporting, while protecting user trajectories, ALOQ uniquely addresses the needs of privacy-preserving, time-evolving location analytics.

4 THE ALOQ APPROACH

In this chapter, we introduce ALOQ, a novel approach for privacy-preserving location frequency estimation. ALOQ employs a frequency estimation model that leverages Local Differential Privacy to collect user location data in a privacy-preserving manner.

A key feature of ALOQ is its use of adaptive quadtree that evolve dynamically in response to longitudinal data, incorporating an adaptation window to improve resource usage, such as the privacy budget. By capturing correlations within users' data over time, ALOQ creates adaptive quadtrees that reduce non-uniformity errors, enhancing the accuracy and utility of spatial data analysis. This approach balances privacy, utility, and adaptability, making it an effective solution for geospatial data collection and analysis in privacy-sensitive contexts.

The architecture is driven by four design goals that mirror the contributions outlined in Chapter 1:

1. **Adaptive spatial granularity:** a quadtree that *splits* dense regions and *merges* sparse ones;
2. **Change detection (QAW):** trigger adaptations only when the distribution drifts beyond a similarity threshold t_{r_s};
3. **Similarity-aware budget scaling:** down-weight the LDP budget when successive distributions are near-identical;
4. **Bounded privacy cost:** a single merge phase and a controlled recursion ensure the worst-case privacy loss stated in Proposition 3.

To facilitate the understanding of our model, the specific parameters and notations used throughout this chapter are defined in Table 3.

4.1 From ALOG to ALOQ: Motivation and Evolution

Before designing ALOQ, we conducted a series of preliminary experiments with **ALOG** (Adaptive Longitudinal Grids for Geospatial Data under LDP). ALOG introduced an adaptive grid mechanism that adjusts spatial resolution based on user report density, aiming to reduce non-uniformity errors common in fixed grid-based LDP mechanisms.

A key component of ALOG was the *Grid Adaptation Window (GAW)*, which defined fixed-size intervals between grid updates. Instead of adapting the grid at every timestamp, ALOG performed updates only after w time steps. This strategy improved memoization sta-

Table 3 – ALOQ Model Parameters and Notations

Notation	Description
S	Spatial domain (e.g., a city)
$U = \{u_1, ..., u_n\}$	Set of users
$T = \{t_1, ..., t_\tau\}$	Set of discrete timestamps
l	Location in the continuous plane
r_i^t	True report of user u_i at time t
$\tilde{r}_i^t = \Psi(r_i^t)$	Noisy report under LDP
$q_i \in Q^t$	A spatial cell in the partition at timestamp t
f_i^t	Frequency of reports to cell q_i at time t
k	Number of cells in the current partition
tr_{max}	Upper report count threshold for splitting a quadtree cell.
tr_{min}	Lower report count threshold for merging quadtree cells.
tr_s	Similarity threshold to trigger quadtree adaptation.
s_w	The size of the sliding window for the Quadtree Adaptation Window (QAW).
rec_{max}	The maximum recursion depth for the cell splitting phase per timestamp.
ϵ	The upper-bound privacy budget for any single timestamp.
ϵ_t	The dynamically adjusted privacy budget for a single timestamp t.

bility and reduced cumulative privacy cost compared to continuous adaptation. However, it also introduced challenges. Shorter windows offered greater responsiveness at the cost of privacy, while longer windows risked misalignment with evolving user distributions. Experimental results (Figure 10) confirmed that the utility of ALOG degraded as the adaptation window increased, especially in dynamic datasets like Porto and GeoLife.

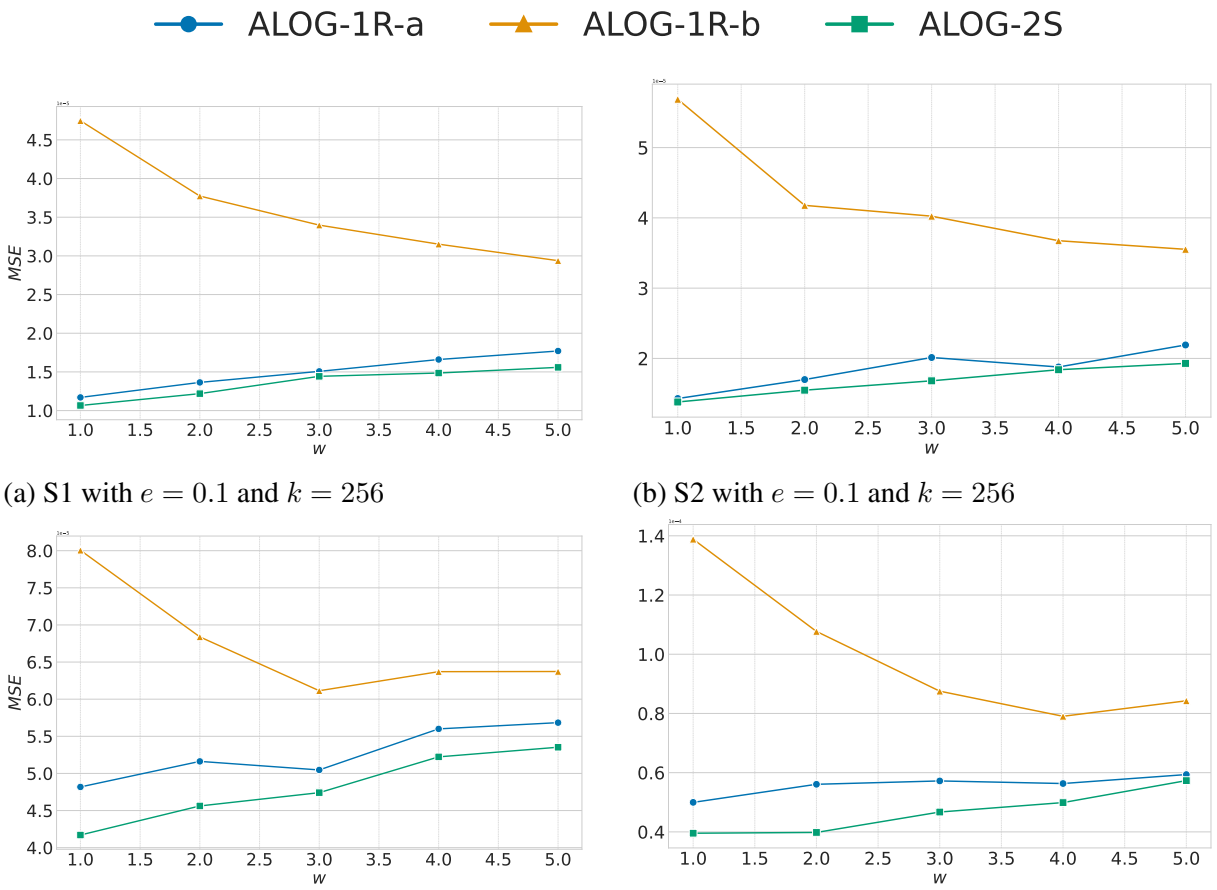
Moreover, ALOG exhibited several structural limitations:

- **Lack of merging:** It could refine grids but not coarsen them, causing over-fragmentation in sparse regions.
- **Uniform budget usage:** It applied the same privacy budget regardless of the temporal stability of the data.
- **Memoization inefficiency:** Frequent structural changes reset memoized responses, increasing privacy budget consumption.

These limitations motivated the development of ALOQ. Instead of a flat grid, ALOQ uses a hierarchical quadtree that allows both recursive refinement and controlled merging. ALOQ

replaces time-based triggers with a data-driven adaptation mechanism: the Quadtree Adaptation Window (QAW), which detects significant changes in the privatized frequency distribution to trigger structural updates. This approach improves responsiveness while maintaining memoization and privacy guarantees.

Figure 10 illustrates the performance sensitivity of ALOG to adaptation window size across multiple datasets.



(a) S1 with $e = 0.1$ and $k = 256$

(b) S2 with $e = 0.1$ and $k = 256$

(c) Geolife with $e = 0.1$ and $k = 256$

(d) Porto with $e = 0.1$ and $k = 256$

Figure 10 – MSE vs. adaptation window size for ALOG across S1, S2, GeoLife, and Porto datasets. Larger windows reduce update frequency but increase estimation error due to stale structures.

Another important observation was the trade-off between adaptation and privacy cost. As Table 4 shows, ALOG variants consumed significantly more privacy budget than static LDP baselines (e.g., LOLOHA, LOSUE, RAPPOR), especially in dynamic contexts. This was due to memoization resets caused by frequent structural changes.

In summary, ALOG introduced key ideas on adaptive discretization under LDP, but its limitations in merging, memoization efficiency, and adaptation triggering motivated the de-

Table 4 – Budget consumption comparison for S1, S2, Geolife, and Porto.

DataSet	ALOG-1R-a	ALOG-1R-b	ALOG-2R	LOLOHA	LOSUE	RAPPOR
S1	**1.78**	2.07	2.02	**0.20**	0.80	0.80
S2	**1.84**	2.11	2.05	**0.20**	0.84	0.86
GeoLife	**1.78**	2.24	2.10	**0.59**	0.99	0.99
Porto	**1.78**	2.07	2.02	**0.20**	0.80	0.80

sign of ALOQ. The following sections present the ALOQ model in detail, including its adaptive quadtree construction, similarity-aware budget strategy, and empirical performance compared to prior baselines.

4.2 Adaptive QuadTree

The adaptive quadtree is a non-uniform spatial data structure used to partition a geographic region, such as a city, into rectangular subregions. Each *leaf node* in the tree represents a cell covering a portion of this space, and all user-submitted location reports are mapped to their corresponding leaf cells based on geographic coordinates.

The quadtree adapts its structure according to the **density of reports** observed in each cell. At every timestamp, the aggregator collects the number of reports received in each leaf q_i, recording this value as c_i. This count serves as a proxy for the local density of user activity: regions with more reports are interpreted as denser, while regions with fewer reports are considered sparse.

By concentrating spatial detail in denser areas and reducing it elsewhere, the adaptive quadtree enables more accurate frequency estimation across heterogeneous data distributions. It also ensures a more efficient allocation of the privacy budget under Local Differential Privacy (LDP), reducing noise where precision is most critical.

The adaptation is governed by two privacy-aware thresholds. The **upper threshold** ($tr_{\max}$) limits the number of reports allowed in a leaf before that region becomes too coarse, while the **lower threshold** ($tr_{\min}$) determines when previously subdivided regions should be collapsed back into a single cell. These mechanisms enable the tree to adjust its resolution in line with the underlying report distribution, as illustrated in Figure 11.

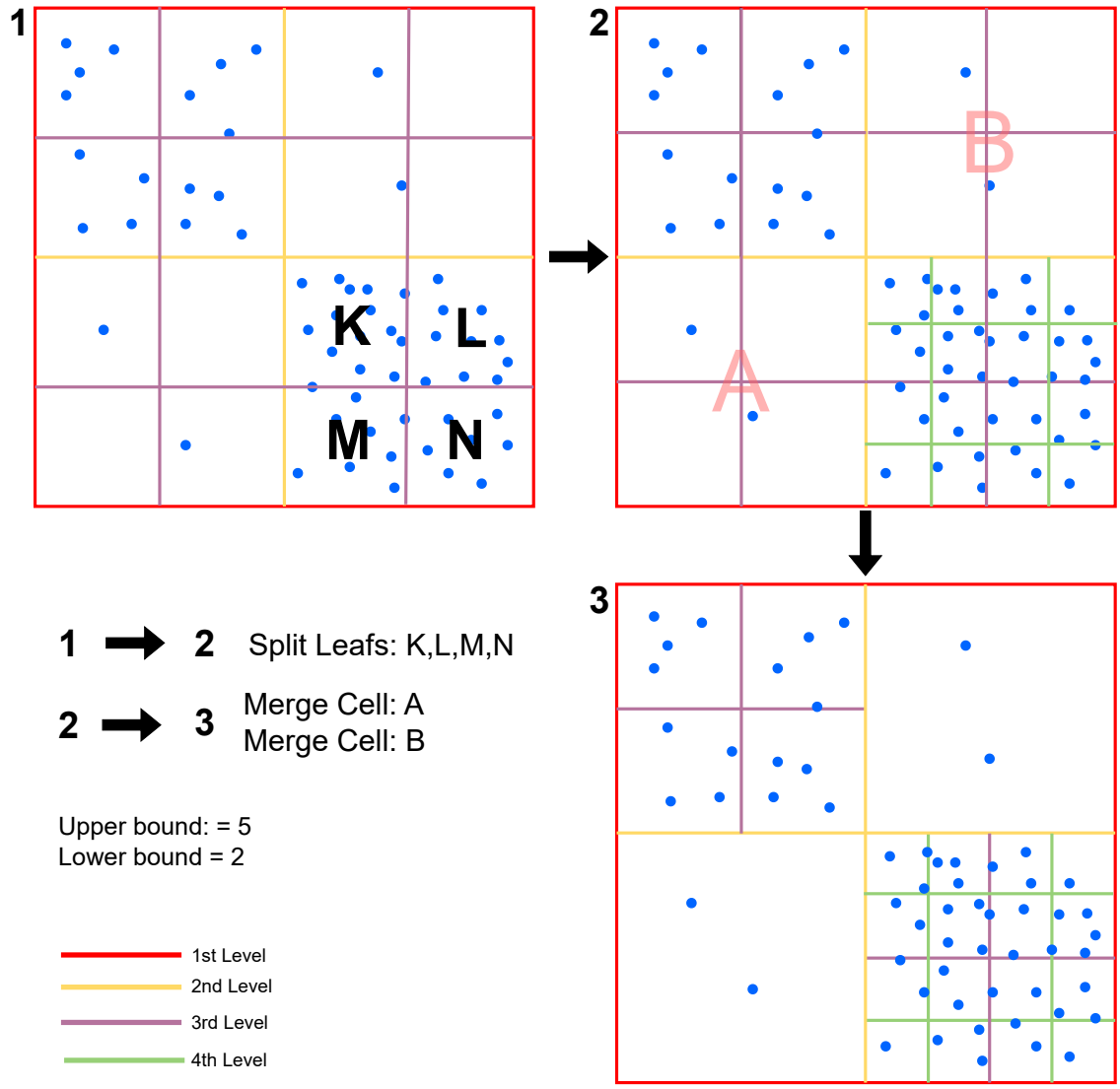


Figure 11 – Example of quadtree adaptation: cells K, L, M, and N are subdivided due to high report density; cells A and B are merged due to low activity.

4.2.1 Splitting Phase

To refine regions with high report concentration, the aggregator applies a recursive splitting procedure. Whenever a leaf node's count exceeds a predefined upper threshold $tr_{\max}$, the node is subdivided into four equal-area child cells, becoming an internal node. The count of the parent is evenly distributed among the children, with each child receiving $\lceil c_i/4 \rceil$ reports, ensuring the continuity of the frequency distribution.

To avoid unbounded quadtree expansion and premature consumption of the privacy budget, the recursion depth is bounded by a maximum value $rec_{\max}$ per timestamp. This depth limit ensures that each round of refinement respects both utility and privacy constraints. Im-

portantly, this limit applies locally to each timestamp; future timestamps may further refine the structure, allowing the quadtree to evolve gradually over time.

This recursive splitting mechanism strikes a balance between spatial granularity and privacy budget usage, adapting to local density while enforcing global guarantees. The process is formally defined in Algorithm 1.

Example: Figure 11 illustrates the splitting behavior across three stages. Initially, the space is uniformly partitioned into 16 cells. Suppose cells K, L, M, and N receive the following counts: 15, 10, 8, and 7, respectively. Given an upper threshold $tr_{\text{max}} = 5$, these four cells exceed the limit and trigger subdivision.

Each of the affected cells is recursively split into four equal-area subcells, resulting in 16 new children in the second stage. The original counts are evenly distributed: for example, cell K (with 15 reports) produces four subcells, each receiving $\lceil 15/4 \rceil = 4$ reports. The same logic applies to cells L, M, and N, whose children receive 3, 2, and 2 reports respectively.

If any of the new subcells still exceed the threshold, the recursion continues, provided the depth does not surpass rec_{max}. This adaptive strategy concentrates resolution in dense areas while avoiding over-refinement in sparse regions.

4.2.2 Merging Phase

Independently, the algorithm evaluates each internal node to determine whether it qualifies for *merging* based on sparsity. If the total report count across all of a node's children falls below a predefined lower threshold tr_{min}, those children are pruned: their individual counts are aggregated, and the parent node is converted into a new leaf holding this combined count. This merging operation simplifies the quadtree in regions with low data density, preventing over-partitioning where fine granularity is unnecessary.

Merging is applied *recursively*, allowing entire sparse subtrees to be coarsened efficiently. Once a node is merged, its parent may, in turn, become eligible for collapse, enabling cascaded simplifications. This recursive strategy reduces the structural complexity of the tree while preserving a faithful approximation of the spatial distribution.

To maintain stability and ensure strong privacy guarantees, merging is performed only once, during the initial timestamp ($t = 0$). This one-time application serves two key purposes. First, it corrects over-partitioning artifacts caused by the initial uniform quadtree, which is used to bootstrap the system under strict privacy constraints. Second, it prevents instability in

Algorithm 1: Adaptive Quadtree SplitNode

Input : Quadtree Q,
uper bound threshold tr_{max},
max recursion rec_{max}
Output : Adaptive quadtree with updated cell counts

```
Function TraverseAndSplit (Q):
    foreach node q in post-order traversal of Q do
        SplitNode(q, tr_max);
    end
Function SplitNode (q, tr_max):
    if g is leaf then
        c = q.count;
        if c > tr then
            Split node q into four sub-nodes of equal area;
            for each sub-node j of cell q do
                j.count = c/4;
                SplitNode(j, tr_max);
                // Recursive call (max rec_max times)
            end
        end
    end
    else
        for each child j of cell q do
            SplitNode(j, tr_max);
        end
    end
```

the tree structure, avoiding oscillations between splits and merges in subsequent rounds. After this initialization, only the splitting mechanism remains active, allowing the quadtree to evolve progressively as new reports are collected.

Example: Continuing the example from Figure 11, suppose that the leaf cells in regions A and B each received at most one report.

Now assume a lower merging threshold $tr_{\text{min}} = 2$. The total counts under each internal node in A and B sum to 0 or 1, falling below the threshold. As a result, these nodes are collapsed: their children are removed, and the parent becomes a new leaf node holding the aggregated count, 2 reports for each region.

This coarsening reduces the number of active leaf cells in sparse regions and limits the need for unnecessary privacy budget expenditure associated with fine-grained partitions.

Figure 11 illustrates this process: from Box 2 to Box 3, internal nodes in A and B are merged into single leaf nodes due to their children's low combined counts. This operation results in a more compact and balanced spatial representation that better reflects the underlying

distribution of user reports.

By applying *splitting* and *merging* as independent, recursive procedures, splitting allowed at each timestamp, merging restricted to $t = 0$, the adaptive quadtree evolves in response to the spatiotemporal structure of the data. High-density regions are incrementally refined, while low-density areas are compacted early on, yielding a data-sensitive representation that balances utility and privacy under Local Differential Privacy (LDP).

Budget Alignment through Frequency Convergence:

A critical consequence of enforcing frequency thresholds through tr_{min} and tr_{max} is the emergence of a bounded range of cell-level counts across the spatial domain. By splitting regions with excessive counts and merging those with insufficient signal, the quadtree adaptively reshapes the spatial structure to concentrate frequencies within a narrower, more uniform band.

This transformation has direct implications for privacy. In standard LDP settings, a single global privacy budget ϵ must be applied uniformly to all user reports, regardless of their location. However, when the frequency distribution is highly skewed, such uniform application can lead to privacy breaches in sparse regions or unusable noise levels in dense ones.

ALOQ addresses this tension not by adjusting ϵ, but by equalizing the scale of frequency estimation across the space. By ensuring that most leaf nodes fall within a controlled frequency range, the mechanism enables a more effective and secure application of LDP noise. In essence, ALOQ prepares the spatial domain so that a globally fixed ϵ becomes statistically meaningful and privacy-compliant across all regions.

This property is particularly important in scenarios where budget adaptation is infeasible or undesirable. Through spatial adaptation alone, ALOQ achieves implicit budget calibration by modulating where and how privacy noise interacts with true signal.

4.3 ALOQ - Frequency Estimator Model

Running Example: ALOQ in Action

To make the ALOQ model more concrete, we will follow a running example throughout this chapter. Let's consider a scenario with a single user, Alice, and an aggregator. The aggregator's goal is to estimate user frequency in a city, which it initially divides into a 4x4 uniform quadtree (16 cells).

At timestamp t=0:

Algorithm 2: Adaptive Quadtree Merge

Input : Quadtree Q,
Lower bound threshold tr_{min}
Output Adaptive quadtree after merging sparse branches
:

Function `TraverseAndMerge` *(Q)*:
  foreach *internal node q in post-order traversal of Q* **do**
    MergeNode(q, tr_{min});
  end
Function `MergeNode` *(q, tr_{min})*:
  if *q is internal* **then**
    $c = \sum_{j=1}^{4} q.child_j.count$;
    if $c < tr_{\text{min}}$ **then**
      Remove the 4 children of q;
      Set q as a leaf node;
      Set $q.count = c$;
      `// Collapse the branch into the parent node`
    end
  end

1. **Aggregator Broadcast:** The aggregator broadcasts the initial 4x4 uniform quadtree and the initial privacy budget, ϵ_0, to all users.
2. **User-Side (Alice):** Alice is located in the area corresponding to cell 10. Her device encodes this as a one-hot vector of length 16: $[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$.
3. **Sanitization & Transmission:** Her device applies the L-OSUE protocol to sanitize this vector, producing a perturbed report: $[0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0]$. This noisy report is then transmitted to the aggregator.
4. **Aggregation & Adaptation:** The aggregator collects this (and all other users') perturbed reports to estimate the frequency of visits for each of the 16 cells. Based on these noisy estimates, it will decide if and how to adapt the quadtree for the next timestamp, as we will detail in the following sections.

The remainder of this section describes the reporting pipeline that operates on top of the adaptive quadtree just presented.

ALOQ operates through two core entities: the *users* and the *aggregator* (Figure 12). Users are responsible for locally encoding their private location information based on a shared spatial structure and applying local differential privacy (LDP) mechanisms before submitting their obfuscated reports. The aggregator, in turn, collects these privatized reports to estimate the frequency of visits to different spatial regions, each represented by a leaf node in an adaptive quadtree.

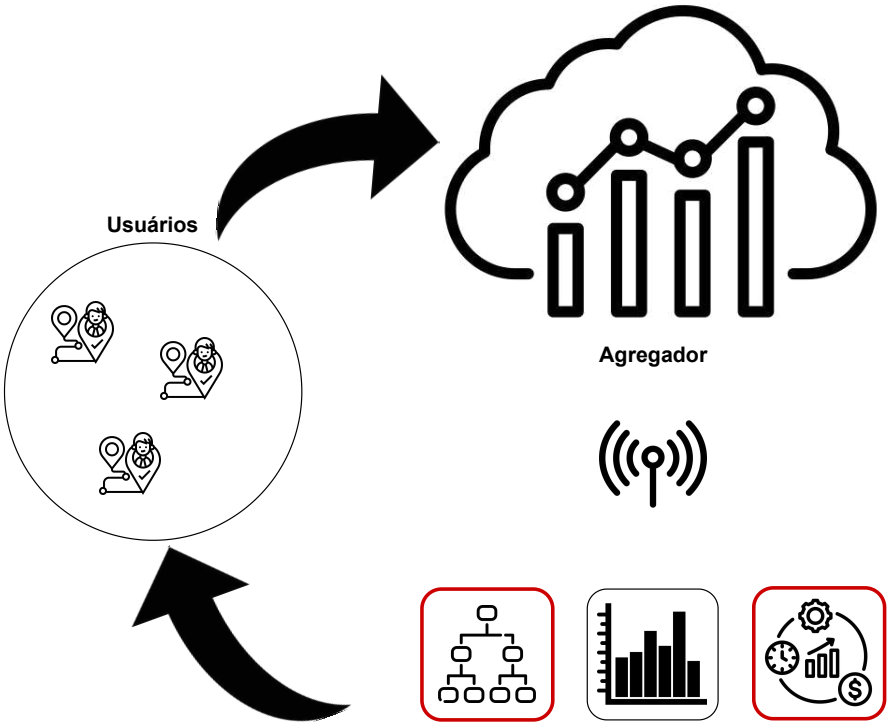


Figure 12 – ALOQ's frequency estimation protocol

Beyond frequency estimation, the aggregator plays a central coordination role. It is responsible for constructing and updating the spatial partitioning used throughout the protocol. Initially, the aggregator creates a uniform quadtree that divides the reference area, such as a city, into equally sized square cells. This structure is distributed to all users and serves as the common reference for location encoding.

On the user side, each individual maps their current location to the appropriate leaf in the received quadtree, as detailed in Section 2.7. To protect their privacy, users apply an LDP mechanism before transmitting their encoded report. ALOQ adopts the L-OSUE protocol (see Section 3.2) for this task, given its strong empirical performance in longitudinal settings. However, the model remains agnostic to the specific LDP mechanism and can accommodate alternative protocols. The complete user-side process is presented in Algorithm 3.

As user reports are received, the aggregator updates the frequency estimates associated with each leaf node. It also refines the quadtree over time, adapting its granularity in response to spatial density variations. This adaptive behavior improves the accuracy of frequency estimation by aligning spatial resolution with the underlying population distribution, achieving

Algorithm 3: ALOQ Algorithm - User Side

Input : Timestamp t, Quadtree Q_i, Privacy budget ϵ
Output Report $\hat{r}_i$
:

for *each timestamp* t **do**
    $l_i \leftarrow$ Capture current location;
    $Q_i \leftarrow$ Receive the Quadtree from Server;
    $\epsilon_t \leftarrow$ Receive the privacy budget from Server;
    $r_i \leftarrow$ Encode(l_i, Q_i);
    if $t = 0 \vee Q_i \neq Q_{i-1}$ **then**
        $M \leftarrow$ Initialize the Memoization Vector(Q_i);
    end
    $\hat{r}_i \leftarrow$ LDP Protocol(r_i, G_i, M,ϵ);
end

a better trade-off between utility and privacy.

4.4 Quadtree Adaptation Window (QAW)

To decide when the quadtree should evolve, we introduce the Quadtree Adaptation Window (QAW).

Refining the quadtree is essential for reducing non-uniformity error and improving the utility of frequency estimates. However, evolving the tree structure too frequently can undermine memoization efficiency, leading to increased privacy budget consumption. This occurs because each new quadtree introduces previously unseen leaf nodes, invalidating the canonical mapping used for memoization. As a result, the system must reinitialize the memoization cache, which reduces the benefits of repeated reporting under local differential privacy.

To balance this trade-off between spatial adaptivity and budget efficiency, ALOQ introduces the *Quadtree Adaptation Window (QAW)* mechanism. Instead of adapting the quadtree at every timestamp, QAW determines when an update is necessary based on observed changes in the data distribution.

Specifically, QAW maintains a *similarity window* of size s_w, which stores the frequency estimates from the most recent s_w timestamps. At each new timestamp, the current frequency estimate is compared to the aggregated estimate over the window, typically using an average. If the similarity between them falls below a predefined threshold tr_s, indicating a significant shift in the spatial distribution, the quadtree is re-adapted and redistributed to users.

This change-driven approach ensures that spatial refinements are only triggered by

meaningful distributional drift. In doing so, QAW reduces unnecessary structure updates, preserves memoization longevity, and limits privacy budget consumption without sacrificing responsiveness to dynamic spatial patterns.

Revisiting our example with Alice, let's see how the QAW works in practice. At timestamp t=1, the aggregator collects a new round of reports. It then compares the new frequency distribution with the one from t=0 using the cosine similarity metric described below. If the resulting value indicates a significant change (i.e., it crosses the similarity threshold tr_s), the QAW triggers a quadtree adaptation. If the distributions are similar and the threshold is not met, the quadtree from t=0 is reused to save privacy budget and maintain memoization consistency.

To measure similarity between frequency distributions, ALOQ employs *cosine similarity*. This metric captures the directional alignment between two vectors, making it invariant to scale differences. In the context of spatial frequency estimation, this property is particularly valuable: it enables the system to detect shifts in the *structure* of the spatial distribution while remaining robust to variations in absolute counts, such as those caused by fluctuations in user participation or privacy-induced noise.

Cosine similarity is especially well-suited for high-dimensional, sparse frequency vectors common in geospatial settings, where most regions have zero or low counts. Unlike metrics such as L_1 or L_2 distances, cosine similarity downplays changes in magnitude and focuses on the relative allocation across regions, which is precisely the signal needed to decide whether structural adaptation is warranted.

In our implementation, we use a *cosine distance* variant that inverts the standard cosine similarity scale. Specifically, it returns values in the range $[0, 1]$, where $\mathbf{0}$ denotes perfect similarity and $\mathbf{1}$ indicates maximum divergence. This inversion is consistently taken into account when comparing against the similarity threshold tr_s to decide whether to trigger quadtree adaptation.

Let $\mathbf{F_t}$ be the frequency estimate at the current timestamp, and let $\overline{\mathbf{F}}_{\mathbf{avg}}$ be the aggregated frequency estimate over the similarity window of size s_w. The cosine similarity is defined as:

$$\text{CosSim}(\mathbf{F}_t, \mathbf{F}_{\text{avg}}) = 1 - \frac{\mathbf{F}_t \cdot \overline{\mathbf{F}}_{\text{avg}}}{\|\mathbf{F}_t\| \, \|\mathbf{F}_{\text{avg}}\|} \tag{4.1}$$

Here, $\cdot$ denotes the dot product, and $\|\cdot\|$ denotes the Euclidean norm. The subtraction

from 1 ensures that the result ranges from $[0, 1]$, where 0 indicates maximum similarity and 1 indicates complete divergence, following the convention of the cosine similarity implementation used.

This drift signal not only decides when the quadtree is refined, but also dictates how much privacy budget we should spend at each step, as we now detail.

4.4.1 Similarity-guided privacy-budget scaling

A second pillar of ALOQ is the dynamic allocation of the privacy budget ϵ_t for each timestamp t. Because the privacy budget directly controls the noise magnitude, and hence the trade-off between privacy loss and estimation accuracy. Its allocation must evolve in concert with the quadtree refinement policy.

Let $\text{CosSim}(\mathbf{F}_t, \overline{\mathbf{F}}_{avg}) \in [0, 1]$ be the cosine distance between the current frequency vector $\mathbf{F}_t$ and the similarity-window average $\overline{\mathbf{F}}_{avg}$, using Equation 4.1. ALOQ scales the per-timestamp budget by a reduction factor

$$\alpha_t \;=\; \max\big(\alpha_{\min},\, \text{CosSim}(\mathbf{F}_t, \overline{\mathbf{F}}_{avg})\big),$$

and sets $\epsilon_t = \alpha_t\, \epsilon_{\max}$, with $0 < \alpha_{\min} < 1$ preventing the budget from vanishing entirely.

- **High similarity** ($\text{sim} \approx 0$). The current distribution is almost unchanged; additional reports convey little new information. A small ϵ_t suffices, preserving cumulative privacy loss while keeping the quadtree stable.
- **Low similarity** ($\text{sim} \approx 1$). The distribution has shifted markedly; accurate estimation now requires less noise. A larger ϵ_t improves utility and enables the quadtree to adapt where needed.

Allocating the budget in this similarity-aware manner achieves two goals simultaneously: it *delays* unnecessary expenditure when the data are stable, and it *invests* budget precisely when structural change demands higher accuracy. This adaptive strategy was crucial for the performance gains reported in Section 5, where ALOQ outperformed fixed-budget baselines under identical global privacy constraints.

Figure 13 illustrates the operation of the *Quadtree Adaptation Window*.

Step 1:Reference vector construction. A sliding window of size s_w stores the most recent frequency estimates. At each new timestamp, their element-wise mean $\overline{\mathbf{F}}_{avg}$ is computed and used as the reference distribution.

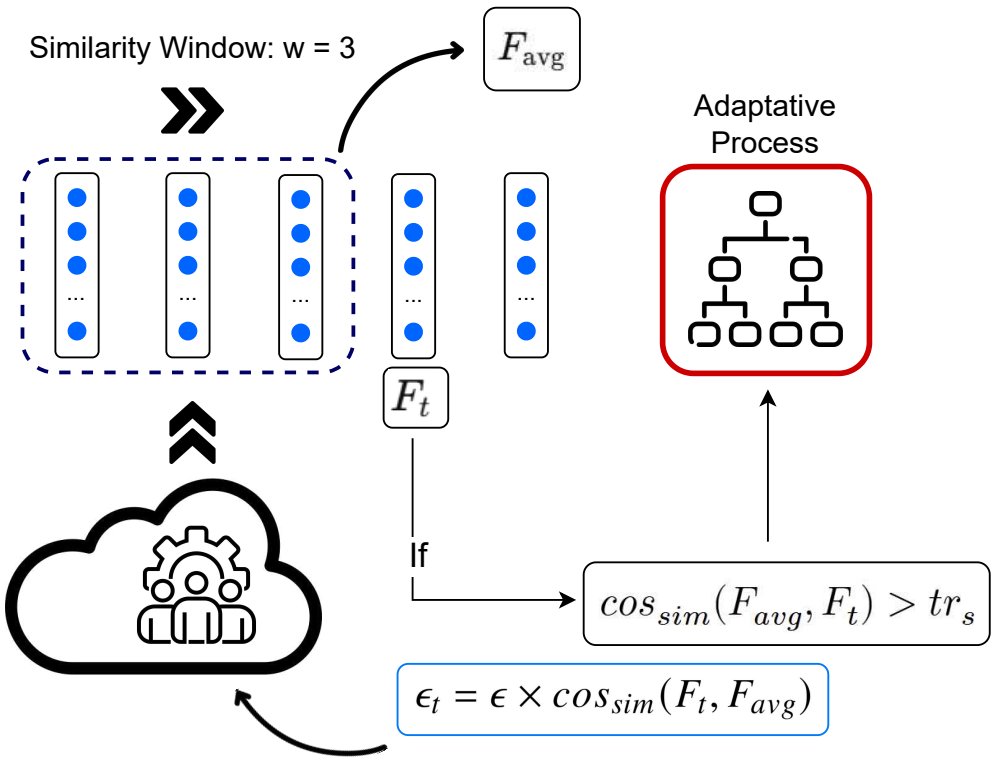


Figure 13 – QAW: Quadtree Sliding Adaptation Window

Step 2:Similarity test. The cosine distance between the current estimate $\mathbf{F}_t$ and $\overline{\mathbf{F}}_{avg}$ is evaluated, yielding a score in $[0, 1]$ where 0 denotes perfect similarity and 1 denotes maximal divergence. If this score exceeds the threshold tr_s, the quadtree is re-adapted and the updated structure is broadcast to users.

Step 3:Budget scaling. Independently of whether adaptation is triggered, the same similarity score acts as a reduction factor for the per-timestamp privacy budget ϵ_t. High similarity results in a smaller ϵ_t, conserving the global privacy budget; low similarity increases ϵ_t, improving utility when the distribution has shifted.

The complete algorithm for the server side of ALOQ can be seen on 4.

The similarity test and adaptive budget form the temporal counterpart to spatial splitting/merging, completing the core ALOQ flow described next.

4.5 Variants of ALOQ

While ALOQ fundamentally presents as an excellent solution, some strategies can be explored to achieve greater data utility without compromising privacy. To this end, we will

Algorithm 4: ALOQ Algorithm - Server Side

Input : tr_{max} (Upper-bound Cell Threshold), tr_{min} (Lower-bound Cell Threshold), tr_s (Similarity Threshold), s_w (Similarity Windows), ϵ, r_{max} (Max Recursion),α_{min}

Output : Adaptive Quadtree Q, Frequency Estimations F_i, Privacy Budget ϵ_t

$Q \leftarrow$ Initialize the initial Quadtree;
$Est_{Freq} \leftarrow$ Initialize an empty estimation vector;
for *each timestamp* t **do**
    $R \leftarrow$ Receive reports from all users;
    $F_i \leftarrow$ Aggregate reports to estimate frequencies;
    Compute counts with (Q, F_i);
    Capture the last s_w frequency vectors in a queue F;
    if $|Ext_{freq}| < 2$ **then**
        Append F_i to Ext$_{freq}$;
        $\epsilon_t \leftarrow \epsilon$;
    end
    else
        $F_{avg} \leftarrow$ Mean of frequency queue F;
        $cos_{sim} \leftarrow$ Cosine similarity between F_{avg} and F_i;
        $\epsilon_t \leftarrow \epsilon \times \max(\alpha_{min}, cos_{sim})$;
    end
    if $t = 0 \vee cos_{sim} > tr_s$ **then**
        $Q \leftarrow$ Adaptive Quadtree SplitNode$(Q, F_i, tr_{max}, r_{max})$;
        if $t = 0$ **then**
            $Q \leftarrow$ Adaptive Quadtree Merge(Q, F_i, tr_{min});
        end
        $Est_{Freq} \leftarrow$ Reset estimation vector;
    end
    Publish frequency estimator F_i;
    Broadcast privacy budget ϵ_t;
    Broadcast adaptive quadtree Q;
end

present three variations of ALOQ, summarized in Table 5.

Table 5 – ALOQ Variations

Approach	Stages of Sanitization	Base Quadtree for Adaptation
ALOQ-2S	Two stages	Previous Quadtree
ALOQ-1S-Adap	Single round	Previous Quadtree
ALOQ-1S-Base	Single stage	Uniform Quadtree

The first variation, **ALOQ-2S**, implements a two-stage sanitization process designed

to improve both quadtree refinement and the utility of the data. In the first stage, each user i anonymizes their location report using the initially provided uniform quadtree, allocating only a fraction of the available privacy budget to this step. For example, suppose a user i has a true location at (x_i, y_i), and the initially assigned leaf cell is g_0. The user first encodes their location into g_0 (e.g., by rounding their location to the nearest cell in the quadtree). The user then applies a Local Differential Privacy (LDP) protocol, utilizing a fraction of the total privacy budget to anonymize the encoded location. The resulting anonymized report, denoted as $\hat{r}_i$, is sent to the aggregator. This initial stage aims to provide a preliminary anonymized data sample to the aggregator.

Once the aggregator has collected the anonymized reports from all users, it proceeds to refine the spatial quadtree based on these reports. Specifically, the aggregator applies the SplitNode (1) and Merge (2) algorithms to construct a refined, adaptive quadtree. This quadtree is tailored to capture variations in the density of user locations, allowing for more precise spatial partitioning that better reflects the underlying data distribution.

In the second stage, the user i now anonymizes their true location using the refined quadtree. For instance, if the user's true location (x_i, y_i) falls in a newly created, smaller leaf cell g_1, the user will now encode their location into this smaller, more precise cell g_1. The user then applies the remaining portion of their privacy budget to anonymize the location, using the same LDP protocol as in the first stage. This refined anonymization ensures that the user's location is represented with higher precision according to the new adaptive quadtree. The anonymized report is then sent to the aggregator. After all users have submitted their second-stage reports, the aggregator collects these updated anonymized reports and performs frequency estimation for the given timestamp. With quadtree refinement in the first stage and a more precise anonymization in the second stage, this two-stage process results in improved utility. By incorporating the adaptive quadtree refinement early on, ALOQ-2S ensures that the data is processed with an optimized spatial partitioning, leading to more accurate frequency estimations while maintaining a strong privacy guarantee.

Alternatively, **ALOQ-1S-Adap** (One-Stage Sanitization — Adaptive) and **ALOQ-1S-Base** (One-Stage Sanitization — Baseline) utilize a single-round sanitization process. In both variations, users initially report perturbed data using a uniform quadtree, and the adaptive quadtree is applied only in subsequent reports. The key distinction between ALOQ-1S-Adap and ALOQ-1S-Base lies in the choice of the base quadtree used during the adaptation process.

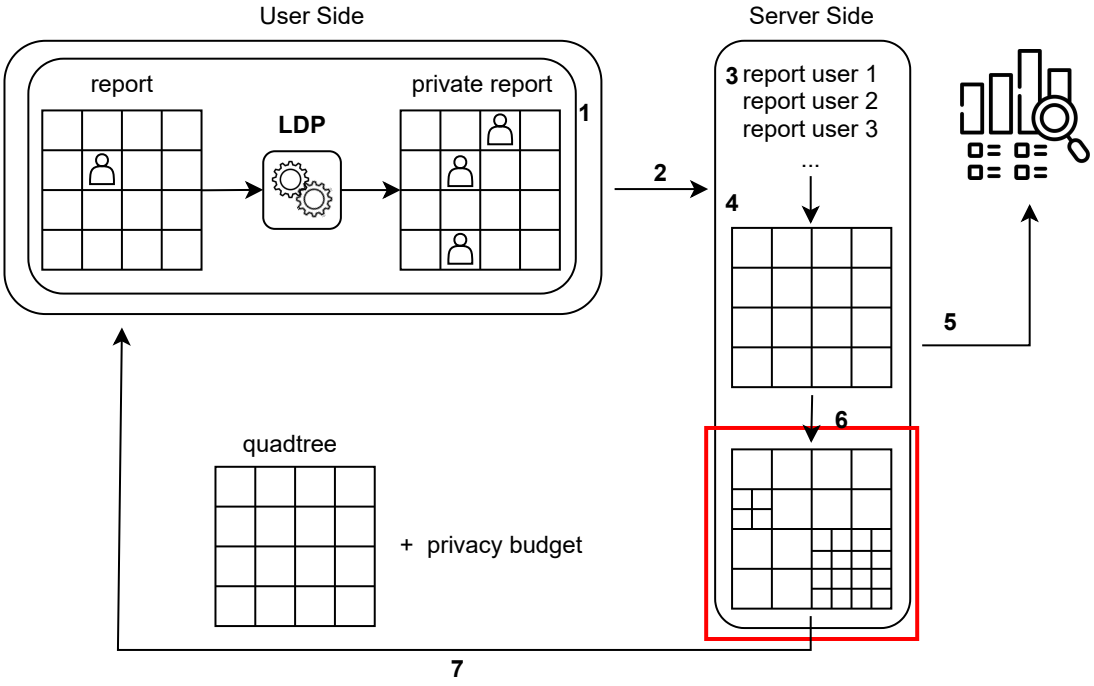


Figure 14 – ALOQ Workflow

In ALOQ-1S-Adap, the base quadtree used for adaptation is not static. Instead, the quadtree from the previous adaptation step serves as the input for the Adaptive Quadtree algorithms. This allows for incremental refinements over time, where the quadtree is continuously adjusted to reflect the evolving spatial patterns of user locations. For instance, suppose in the first timestamp, a user i has a true location at (x_i, y_i), which initially maps to the cell g_0. After anonymizing the location using LDP and sending the anonymized data to the aggregator, the adaptive quadtree is refined based on the collected reports. When the user submits their report in the next timestamp, the adaptive quadtree generated from the prior stage is used as the base for further refinement. If the user's true location (x_i, y_i) now falls into a more precise, smaller cell g_1 in this updated version, the user will anonymize their location based on this refined quadtree, improving the accuracy of the spatial partitioning with each adaptation step. This dynamic approach allows the system to adapt to changing patterns in user data, improving the overall utility while maintaining privacy.

Conversely, in ALOQ-1S-Base, the algorithm takes a different approach. Here, the adaptive quadtree is reset at every adaptation step, and the algorithm consistently uses the default uniform quadtree as a base for every iteration of the refinement process. In each adaptation step, the algorithm ignores any previous refinements and starts from the uniform quadtree, applying

the Adaptive Quadtree algorithms as if it were the first round of processing. For example, a user i's true location (x_i, y_i) is first mapped to the cell g_0. After anonymizing their location using LDP and sending the anonymized report to the aggregator, the quadtree is adapted based on the collected data. This new adaptive quadtree will be used in the next timestamp for users to anonymize their location, but the quadtree is discarded for the next adaptation process. During this new adaptation, again the users will send their reports based on the initial uniform quadtree, and the adaptation process repeat. This results in a stable reference framework for the quadtree adaptation process, which may be more predictable but does not account for changes in user spatial distributions over time.

The workflow of ALOQ is illustrated in Figure 14. The process begins with Step 1, performed on the user side, where each user encodes their location using the current base quadtree provided by the aggregator. This quadtree could either be the initial uniform quadtree (for ALOQ-1S-Base) or the refined one from the previous moment (for ALOQ-1S-Adap and ALOQ-2S). The user applies a Local Differential Privacy (LDP) protocol to anonymize the location report in all approaches. Once anonymized, the report is transmitted securely to the aggregator, completing Step 2.

In Step 3, the aggregator collects all received reports. Step 4 follows, where the aggregator counts the occurrences within each cell to estimate the spatial distribution of user locations. Step 5 is executed at every timestamp, where the estimated spatial distribution for each quadtree leaf cell is published. If the distribution similarity, calculated as explained in Figure 13, exceeds the threshold tr_s, indicating a significant difference, the quadtree adaptation process is triggered in Step 6. In this step, the quadtree cells are adjusted based on report density to improve spatial resolution and precision. This adjustment will occur as follows for each approach:

- ALOQ-2S and ALOQ-1S-Adap: Quadtree refinement occurs using the current quadtree used in step 1, allowing for dynamic, incremental adaptations.
- ALOQ-1S-Base: In this approach, the quadtree used as base in the adaptation process is always the same initial uniform quadtree, discarding any previous refinements. This provides a stable reference framework but does not consider previous adaptations.

Finally, in Step 7, the aggregator broadcasts the updated quadtree to the users, ensuring an adaptive and continuously updated spatial representation that aligns with real-time report patterns. Along with the quadtree, the privacy budget, calculated using the previous similarity measurement, is also broadcast. This entire workflow is designed to repeat iteratively,

maintaining a dynamic balance between spatial precision and privacy guarantees.

4.5.1 Illustrative Examples of ALOQ Variants

To make the differences between the three variants perfectly clear, we will now walk through our running example for each one. In all scenarios, we assume the QAW triggers an adaptation at the end of timestamp t=0 after detecting a high density in cell 10, causing it to be split.

Example 1: ALOQ-1S-Adap (Incremental Adaptation) This variant uses the most recently adapted quadtree as the starting point for the next round.

- **At t=0:** Users report on the initial uniform quadtree, Q_0. Based on the noisy counts, the aggregator creates a new adapted quadtree, Q_1 (where cell 10 is split).
- **At t=1:** The aggregator broadcasts Q_1. Users now report their locations based on this **new, refined** Q_1. The adaptation process for the end of t=1 will use Q_1 as its input, making the changes incremental.

Example 2: ALOQ-1S-Base (Stateless Adaptation) This variant always uses the original uniform quadtree as the foundation for refinement, regardless of intermediate changes.

- **At t=0:** Users report on the initial uniform quadtree, Q_0. The aggregator creates an adapted quadtree, Q_1 (where cell 10 is split).
- **At t=1:** The aggregator broadcasts Q_1, and users report using this adapted quadtree. However, when the aggregator runs its adaptation algorithms for the *next* round, it does not use Q_1 as the input. Instead, it reverts to the **original uniform quadtree** Q_0 and applies the splitting/merging logic to it based on the new frequency counts from t=1.

Example 3: ALOQ-2S (Two-Stage Sanitization) This variant performs a two-step report-and-refine process within each timestamp.

- **Within t=0:**
  1. **Stage 1:** Alice sends a first report using the uniform quadtree Q_0 and a *fraction* of her privacy budget.
  2. **Intermediate Adaptation:** The aggregator immediately uses these initial reports to create the adapted quadtree Q_1 (with cell 10 split). It broadcasts Q_1 back to the users.
  3. **Stage 2:** Alice sends a *second report for t=0*, this time using the more precise

quadtree Q_1 and the remainder of her privacy budget. The final frequency estimation for t=0 is based on these more accurate second-stage reports.

- **At t=1:** This entire two-stage process would repeat.

4.6 ALOQ - Algorithmic Complexity

We now study the time and space complexity of ALOQ using Big-O notation.

Let n be the number of users per epoch, $T = \{t_1, \dots, t_{|T|}\}$ the set of timestamps, q_t the number of active cells at epoch t (and $q = \max_t q_t$), k the domain size of S, d the maximum quadtree depth (so $q_t \leq 4^d$), r the per-epoch recursion cap for splitting, and sw the QAW window length.

Core routines and costs.

- **Local randomization (per user, epoch t).** First, locate the cell containing the user's (lat, lon) in the quadtree: tree descent costs $O(d) = O(\log q_t)$. With unary encoding (UE) over the domain of size k_t, the naive randomized response flips each bit independently, giving $O(\log q_t + k_t)$ per user.
- **Server aggregation per epoch** Counting k bits for each of n users gives $O(n\,k)$ updates.
- **Adaptation (split phase).** Scan all cells to test thresholds (e.g., $tr_{\max}$) in $O(q_t)$; each accepted split creates up to 4 children and may recurse up to r levels. Worst-case new cells per epoch: $O(4^r)$. Hence $O(q_t + 4^r)$ (worst case per epoch).
  Amortized costs are $O(q_t)$ if adaptation fires infrequently and r is small.
- **(If merging were enabled).** Symmetric to splitting: $O(q_t)$ scan; merges $\leq O(q_t) \Rightarrow O(q_t)$ amortized.
  Total overall by epoch using $q_t \leq q$ is:

$$O\big(|T|\,(nk + n\log q + q + 4^r)\big).$$

4.7 Privacy and Utility Analysis

We now establish that all three ALOQ variants meet a bounded longitudinal privacy guarantee **in the worst case**, and we discuss the resulting impact on utility. The worst–case scenario is driven by the variants that carry their quadtree forward across timestamps, namely **ALOQ-1S-Adap** and **ALOQ-2S**; for **ALOQ-1S-Base**, which re-initialises the quadtree at each

adaptation step, the cumulative privacy loss is strictly lower.

In our approach, we employ Local Differential Privacy (LDP) to protect user privacy while estimating the frequency of each quadtree cell in a spatial domain. Each user perturbs their location data locally before transmitting it to the aggregator, using the LOSUE protocol as described in Section 3.2.

At each timestamp, ALOQ enforces an upper bound on privacy leakage through a total budget of ϵ per report. In the case of ALOQ-2S, this budget is split between two stages, with each round consuming a fraction of ϵ. Thus, similar to other LDP data collection models, ALOQ preserves strong per-report privacy guarantees.

However, the effectiveness of the chosen privacy budget is inherently linked to the structure of spatial partitioning. In static discretizations (e.g., uniform grids or fixed quadtrees), the distribution of reports across cells is often highly imbalanced: some cells are densely populated, receiving many reports, while others remain sparse. Under LDP, the noise added to each report is independent of the true value, and the estimation error in each cell is inversely proportional to the number of reports it receives. Consequently, the utility in sparse cells deteriorates more severely, and a single privacy budget must be selected to protect the worst-case (i.e., most vulnerable) scenario.

This leads to a trade-off: to ensure privacy in low-density areas, one must reduce the privacy budget (i.e., increase the noise), but this imposes unnecessarily high noise in dense regions, degrading overall utility.

Formally, the expected variance of the frequency estimate in a cell $q_i \in Q$ under unary encoding mechanisms is:

$$\text{Var}[\hat{f}_i] \propto \frac{1}{n_i(\epsilon)^2}$$

where n_i is the number of reports mapped to cell q_i. For fixed ϵ, $\text{Var}[\hat{f}_i]$ increases rapidly as n_i decreases.

In contrast, ALOQ's adaptive quadtree structure mitigates this imbalance by equalizing the report distribution across cells. By recursively splitting high-density cells and merging low-density regions, ALOQ ensures that each active cell receives a more balanced number of reports. As a result:

- The overall distribution of n_i becomes more uniform;
- The chosen ϵ is better matched to the actual density of all cells;

- The utility across the space is significantly improved without increasing the privacy risk in any single cell.

This can be interpreted as an implicit normalization of per-cell noise variance, where the adaptive structure reshapes the domain to optimize the privacy-utility trade-off under a fixed global privacy constraint.

This insight suggests a new perspective: adaptive spatial representations enable more effective use of a given privacy budget, shifting the problem from a pessimistic "minimum protection" setting (driven by sparsity) to a more equitable distribution of privacy risk. Our experimental results in Section 5.4 validates this intuition—ALOQ consistently achieves lower MAE and MSE even at tighter budgets compared to static counterparts.

In summary, beyond reducing non-uniformity error, ALOQ's adaptation mechanism aligns the granularity of spatial representation with data distribution, enabling a better calibration of the privacy budget to heterogeneous real-world conditions and ultimately improving utility without weakening privacy guarantees.

To formally assess the privacy guarantees in this adaptive setting, we now establish a worst-case bound for longitudinal privacy consumption under the ALOQ model. This bound accounts for the dynamic quadtree refinements but assumes no direct access to raw data by the aggregator.

Importantly, ALOQ's splitting procedure is constrained by a maximum recursion depth r_{max}, which ensures that the tree growth remains bounded across timestamps. Additionally, the merging operation is applied only once, during the initial timestamp $t = 0$, as a preprocessing step to collapse low-density branches. Because no further merges are allowed during runtime, the structure evolves solely through bounded refinement.

These design constraints ensure that the quadtree evolves in a privacy-preserving manner while maintaining a bounded number of leaves. Since the adaptation logic only operates on previously sanitized frequency estimates, it does not consume additional privacy budget.

Proposition 3 provides a pessimistic worst-case privacy guarantee for our approach. Nevertheless, our experiments demonstrate that ALOQ achieves Pareto efficiency, a state where no further improvement can be made to one objective (e.g., utility) without degrading another (e.g., privacy), even with a slightly higher theoretical privacy budget than competing methods.

Proposition 3 (ALOQ's Privacy Guarantee)**.** The ALOQ algorithm guarantees, in the worst case, $\eta\epsilon$-local differential privacy, where:

- $\eta = \left(\sum_{r=0}^{R}(k-\bar{k})\cdot 4^r + \bar{k}\right)$,
- $\bar{k}$ is the number of initial leaf cells satisfying the threshold condition,
- $R = \left\lceil \log_4 \frac{|U|}{tr} \right\rceil$ is the overall maximum recursion depth of the Algorithm 1,
- ϵ is the per-report privacy budget under Ψ, the underlying LDP mechanism.

Proof. At initialization, the quadtree contains k leaf cells, of which $\bar{k}$ satisfy the threshold tr. The remaining $k - \bar{k}$ cells are recursively split (at most R times) until all subcells meet tr, as formalized in Algorithm 1.

For recursion level $r \in [0, R]$, the quadtree contains $(k-\bar{k})\cdot 4^r + \bar{k}$ cells. Summing over all levels, the *total cumulative cell count* is:

$$\sum_{r=0}^{R}\left((k-\bar{k})\cdot 4^r + \bar{k}\right).$$

Since each user's location report is privatized via Ψ (an ϵ-LDP protocol), the *longitudinal privacy cost* of ALOQ, accounting for adaptive quadtree construction and repeated reporting, is bounded by the worst-case total cell count multiplied by ϵ. This holds because:

1. *User-side perturbation*: Each report is independently ϵ-LDP.
2. *Aggregator-side adaptation*: The quadtree's dynamic refinement does not access raw data, relying only on privatized counts.

The total cumulative cell count reflects the worst-case scenario where Algorithm 1 exhausts all possible recursions. Each cell is sanitized only once during the longitudinal process by employing memoization in the LDP protocol. Thus, since the upper bound on cell count is η, the maximum number of sanitizations is precisely the cumulative cell count. As each report is independently ϵ-LDP and there are at most η leaf cells in the quadtree, our proposed algorithm is $\eta\epsilon$-local differentially private. □

4.8 Discussion on Limitations and Design Choices

Like any modeling approach, ALOQ incorporates a set of design choices that imply trade-offs between adaptability, efficiency, and rigor in its privacy guarantees. It is therefore important to discuss the limitations inherent in these choices.

- **Single Merge Strategy:** One of the main design decisions in ALOQ is the restriction of the merge operation to be executed only once, at the initial timestamp ($t = 0$). The resulting limitation is that the spatial structure cannot become coarser in regions where

user density decreases over time. This was a deliberate choice to ensure the stability of the quadtree, avoiding oscillations (cells being repeatedly merged and split), and, more importantly, to allow for a simpler and more rigorous privacy analysis, ensuring that the cumulative privacy cost remains formally bounded, as demonstrated in Proposition 3. Allowing a fully dynamic merge mechanism is a promising direction for future work, as discussed in Chapter 7.

- **Empirical Selection of Adaptation Thresholds:** ALOQ's performance is influenced by the choice of adaptation thresholds (tr_{max}, tr_{min}, and tr_s). The definition of these parameters is not automated by the model and depends on an empirical choice, based on the characteristics of the dataset and the application's objectives. Although this represents a limitation, the sensitivity analysis presented in Chapter 6 demonstrates that ALOQ's superiority over static approaches is robust, holding across a wide range of values for these thresholds. This indicates that while fine-tuning can optimize the results, the fundamental benefit of the adaptive structure is not fragile and prevails even with suboptimal parameter choices.
- **Definition of the Global Privacy Budget:** The choice of the maximum privacy budget (ϵ) is a fundamental challenge in the practical application of Differential Privacy and, in this work, it is treated as an input parameter defined by a privacy specialist. The scope of ALOQ is not to determine the ideal value of ϵ for a given application, but rather to use any provided budget as efficiently as possible. The contribution of our framework lies in its ability to dynamically allocate this budget intelligently, both temporally (through the QAW) and spatially (through quadtree adaptation), to maximize data utility without exceeding the pre-established privacy ceiling. The formal derivation of an optimal budget based on ALOQ's structural properties is, in itself, a complex and relevant line of research for future work.

5 EXPERIMENTAL EVALUATION

In this chapter, we present a comprehensive empirical evaluation of the ALOQ framework. The experiments are designed to validate our central hypothesis: that an adaptive spatial partitioning mechanism can significantly improve the utility of longitudinal frequency estimation under Local Differential Privacy (LDP). We demonstrate that ALOQ achieves a superior privacy-utility trade-off compared to both traditional static LDP protocols, namely RAPPOR (Section 3.1), L-OSUE (Section 3.2), and LOLOHA (Section 3.3), and the state-of-the-art adaptive grid model, PrivTC.

This chapter is organized as follows. First, we detail the experimental setup, including the synthetic and real-world datasets, the utility metrics, and the baseline configurations. Next, we present and analyze the results, comparing ALOQ's performance across multiple dimensions, such as initial granularity, privacy budget, and longitudinal depth. Finally, we conduct a sensitivity analysis on ALOQ's internal parameters to assess the model's robustness. All experiments were implemented in Python.

5.1 Experimental Setup

We evaluate our approach using two synthetic datasets created for this study and two real-world mobility datasets. The synthetic datasets simulate user movement within a 100 km^2 area and allow us to assess behavior under contrasting spatial distributions.

- **S1**: Users are distributed uniformly across the region.
- **S2**: Users follow a normal distribution centered in the region.

Each synthetic dataset simulates 10,000 users moving at a constant speed of 40 km/h, generating one location report per minute over 40 timestamps. Locations are constrained by the speed to ensure realistic trajectories.

The **GeoLife** dataset[1] contains GPS trajectories from 183 users in Beijing. We extracted 8,442 trajectories, each composed of 120 timestamps. This dataset covers a wide geographic area of approximately 478 km^2, with a highly skewed spatial distribution, some regions exhibit high concentration of points, while others are sparsely populated. The **Porto** dataset[2] comprises data collected from 442 taxis operating in Porto, Portugal. We selected 10,000 trajectories, each also containing 120 timestamps. In contrast to GeoLife, the Porto dataset spans a

[1] <[https://www.microsoft.com/en-us/download/details.aspx?id=52367>

[2] <[https://www.kaggle.com/datasets/crailtap/taxi-trajectory?select=train.csv>

smaller area of about 73 km^2, resulting in a more evenly distributed set of trajectory data. Table 6 summarizes the characteristics of both datasets.

We acknowledge the use of only two real-world datasets, a limitation imposed by the scarce availability of longitudinal datasets with precise coordinate data, a key requirement for our analysis. Nevertheless, the selected datasets offer meaningful and representative insights.

Table 6 – Summary of Datasets Used in the Experiments

Dataset	**Type**	**Distribution**	**# Trajectories**
S1	Synthetic	Uniform	10,000
S2	Synthetic	Normal	10,000
GeoLife	Real-world	Real-world	8,442
Porto	Real-world	Real-world	10,000

For each dataset, we simulate users submitting anonymized location reports using longitudinal LDP protocols. These reports are used to estimate cell frequencies over a uniform quadtree. We compare these results with those obtained using ALOQ's three variations (ALOQ-2S, ALOQ-1S-Adap, and ALOQ-1S-base) under identical conditions.

In addition to the experiments already described, we conducted a comparative analysis of our approach, specifically the ALOQ-2S model, against PrivTC, an adaptive spatial partitioning method previously introduced in the literature. We adapted PrivTC to our longitudinal context and compared its performance to ALOQ across key evaluation dimensions.

Data Utility: To assess utility, we compute two metrics: the **Mean Squared Error (MSE)** and the **Mean Absolute Error (MAE)** between the true and estimated frequency distributions over time. Let $F_i = \{f_0, \dots, f_{|G_i|-1}\}$ and $\hat{F}_i = \{\hat{f}_0, \dots, \hat{f}_{|G_i|-1}\}$ denote the true and estimated frequencies at timestamp t_i, where G_i is the set of cells at time t_i, and $|T|$ is the number of timestamps. The metrics are defined as:

$$\text{MSE} = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{|G_i|} \sum_{j=0}^{|G_i|-1} (f_j - \hat{f}_j)^2 \tag{5.1}$$

$$\text{MAE} = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{|G_i|} \sum_{j=0}^{|G_i|-1} |f_j - \hat{f}_j| \tag{5.2}$$

These metrics provide complementary insights: MSE penalizes larger errors more heavily, while MAE offers a more interpretable average deviation.

Budget consumption. Let $T = \{t_1, \dots, t_{|T|}\}$ be the reporting epochs and U the users. The per-epoch budget cap is ϵ_t. Due to memoization, user u is charged either 0 or some

$\epsilon_{u,i} \in [0, \epsilon_t]$ (only when a fresh randomized report is emitted). The per-user cumulative privacy loss is

$$B_u = \sum_{i=1}^{|T|} \epsilon_{u,i}.$$

We report the *average total budget per user* to capture the cumulative privacy cost:

$$\overline{B} = \frac{1}{|U|} \sum_{u \in U} B_u = \frac{1}{|U|} \sum_{u \in U} \sum_{i=1}^{|T|} \epsilon_{u,i}.$$

5.2 Adaptation of the PrivTC Framework

To ensure a direct and equitable comparison between spatial partitioning mechanisms, it was necessary to adapt the PrivTC framework to our longitudinal and iterative collection context. Our evaluation is interested exclusively in the performance of its adaptive grid construction component, PrivAG, as a method for spatial discretization.

The original PrivTC framework involves multiple phases, including dividing users into separate groups for grid construction and trajectory modeling, followed by HMM learning and synthetic trajectory generation. Our adaptation simplifies this process significantly to isolate its core grid mechanism. Specifically, we do not divide the user base into groups; all users participate in frequency reporting at every timestamp. Furthermore, we omit the subsequent trajectory modeling and synthesis phases, as our goal is frequency estimation, not trajectory generation.

In our adapted workflow, the PrivAG algorithm is executed by the aggregator at each timestamp. It uses the noisy frequency estimates collected from all users in that instant to generate a new adaptive grid. This resulting grid is then broadcast to all users and employed as the reference structure for encoding and anonymizing their locations in the subsequent timestamp. This modification allows us to directly evaluate the effectiveness of PrivAG's grid adaptation strategy against ALOQ's under identical, continuous data collection conditions.

5.3 Revisiting ALOG: Motivation for ALOQ's Design

The evaluation results presented thus far highlight the strengths of ALOQ in addressing the limitations observed in our earlier model, ALOG. While ALOG introduced the concept of adaptive spatial discretization under Local Differential Privacy (LDP), it exhibited several

structural and practical weaknesses that constrained its utility and efficiency in longitudinal scenarios.

First, ALOG lacked support for merging sparse regions. Its grid refinement process was unidirectional, leading to over-fragmentation in low-density areas and resulting in excessive noise injection. Second, the model used fixed-size adaptation windows, which failed to reflect the real dynamics of spatiotemporal data. This caused unnecessary or delayed updates, negatively affecting accuracy. Lastly, frequent grid updates invalidated memoized responses, forcing repeated re-randomization and increasing cumulative privacy cost.

These limitations directly motivated the design of ALOQ. In contrast to ALOG's flat grid structure, ALOQ employs a hierarchical quadtree with recursive refinement and controlled merging, allowing it to dynamically adapt to data density while preserving structure stability. The Quadtree Adaptation Window (QAW) mechanism replaces static time-based triggers with a similarity-aware criterion, enabling more effective adaptation based on actual data changes. Additionally, ALOQ introduces a similarity-guided budget scaling strategy to reduce unnecessary noise when report distributions remain stable.

The following section consolidates the experimental comparison across multiple datasets and budget regimes, demonstrating how these design changes result in consistent gains in both utility and privacy cost across diverse conditions.

5.4 Results

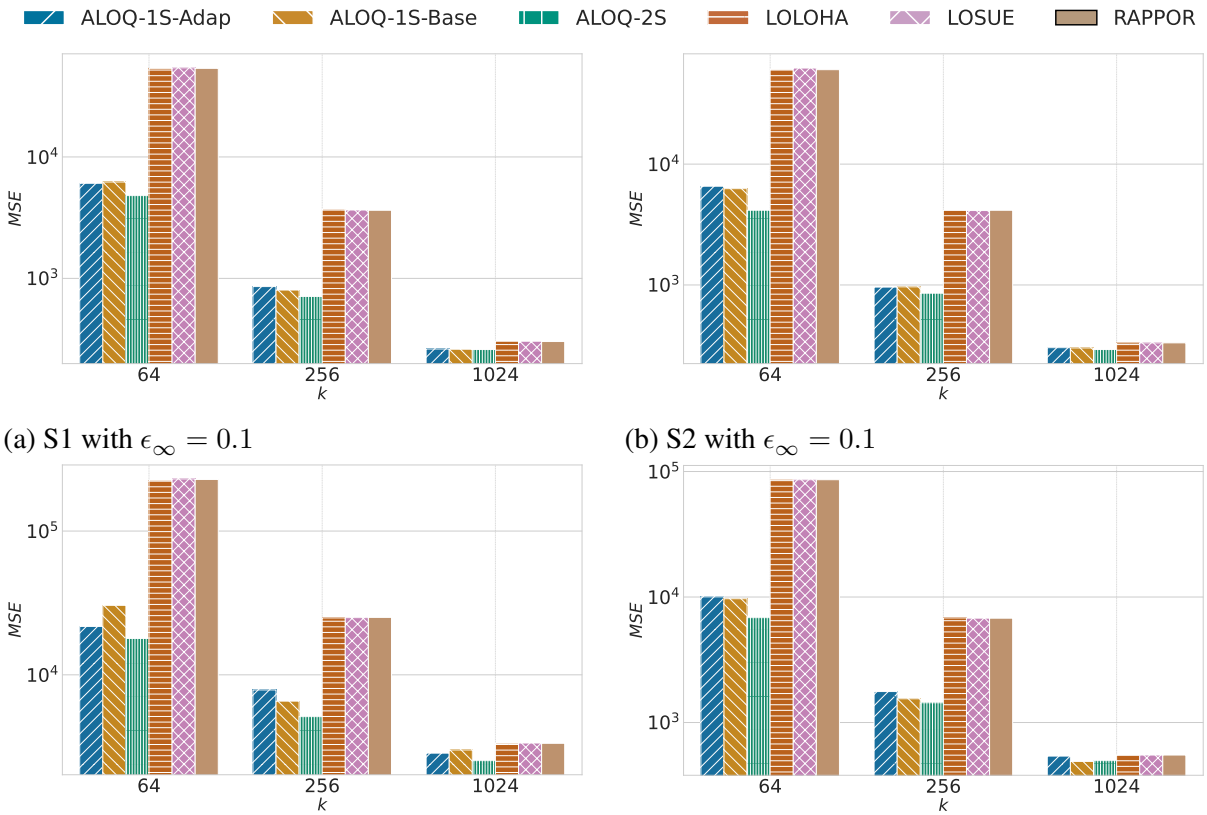
This section presents the experimental results comparing ALOQ to baseline methods, including LOLOHA, LOSUE, RAPPOR, and PrivTC, across multiple datasets and privacy budgets. We evaluate the models using utility metrics such as Mean Squared Error (MSE), budget consumption, and adaptability to spatiotemporal dynamics. The results are organized to highlight the advantages of ALOQ's adaptive quadtree, similarity-aware budget scaling, and structural stability under longitudinal reporting. Through these comparisons, we demonstrate how ALOQ consistently achieves superior accuracy and privacy efficiency in a variety of real-world and synthetic settings.

5.4.1 Comparison with static approaches

We first compare ALOQ to traditional static methods, such as LOLOHA, LOSUE, and RAPPOR. These methods rely on fixed partitioning structures and uniform privacy budgets, making them less responsive to dynamic changes in user mobility patterns. Our goal is to highlight the benefits of ALOQ's adaptivity and privacy efficiency under varying spatial and temporal conditions.

5.4.1.1 Quadtree Granularity vs. Utility

In this experiment, we analyze how varying the initial quadtree granularity affects estimation utility. ALOQ is evaluated alongside static methods using multiple grid resolutions, allowing us to assess how well each model copes with non-uniform spatial distributions.



(a) S1 with $\epsilon_\infty = 0.1$

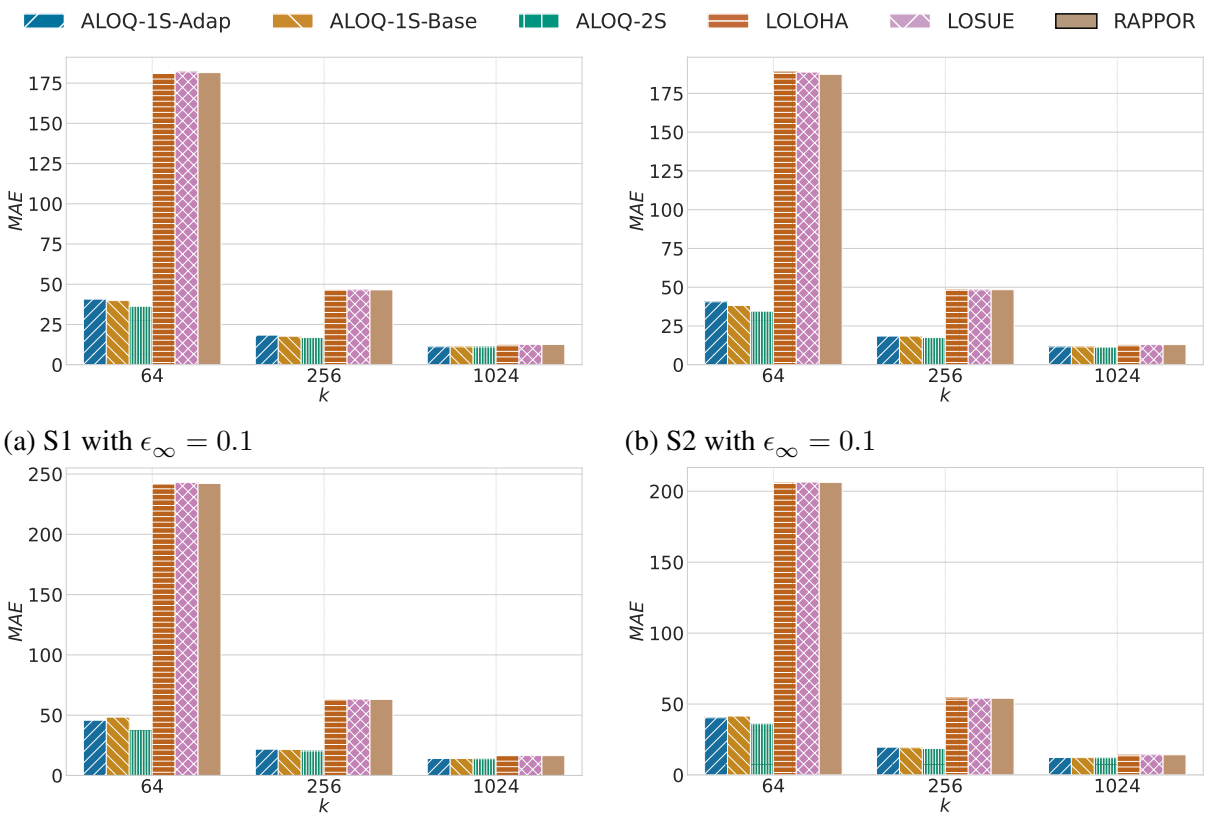
(b) S2 with $\epsilon_\infty = 0.1$

(c) Geolife with $\epsilon_\infty = 0.3$

(d) Porto with $\epsilon_\infty = 0.3$

Figure 15 – MSE vs. initial grid size for S1, S2, GeoLife, and Porto (log scale).

Theoretically, the utility of each ALOQ variant is driven by the trade-off between the noise introduced by local differential privacy (LDP) perturbation and the granularity of the quadtree cells. In frequency estimation protocols such as LOSUE, the variance is inversely pro-



(a) S1 with $\epsilon_\infty = 0.1$

(b) S2 with $\epsilon_\infty = 0.1$

(c) Geolife with $\epsilon_\infty = 0.3$

(d) Porto with $\epsilon_\infty = 0.3$

Figure 16 – MAE vs. initial grid size for S1, S2, GeoLife, and Porto.

portional to the number of reports per cell. Consequently, uniform partitioning approaches (e.g., RAPPOR) are more susceptible to high estimation error in regions with skewed data distributions due to non-uniformity error.

Figures 16 and 15 illustrate the effectiveness of our adaptive approaches in improving utility compared to uniform baselines. Across all datasets, the ALOQ variants consistently outperform traditional methods, with the most notable gains observed in the Porto, Uniform, and Normal datasets for both MAE and MSE metrics. Among them, ALOQ-2S achieves the lowest error in most cases, leveraging its two-stages reporting mechanism and refined spatial adaptation. Even in the challenging GeoLife dataset, where overall errors are higher due to spatial sparsity and skewness, ALOQ approaches maintain a clear advantage over RAPPOR and LOSUE.

5.4.1.2 Privacy Budget vs. Utility

Here, we investigate the trade-off between privacy and utility across different values of the privacy parameter ϵ. The results demonstrate ALOQ's ability to maintain lower estimation error under tighter privacy constraints compared to static baselines.

Table 7 – Mean Squared Error (MSE, $\times 10^{-3}$) for $k = 256$.

DataSet	Budget	ALOQ-1S Adap	ALOQ-1S Base	ALOQ-2S	LOLOHA	LOSUE	RAPPOR
S1	$\epsilon = 0.05$	0.797	0.743	**0.702**	3.665	3.678	3.636
	$\epsilon = 0.1$	0.763	0.797	**0.684**	3.661	3.690	3.648
	$\epsilon = 0.3$	0.815	0.796	**0.726**	3.638	3.686	3.660
	$\epsilon = 0.6$	0.819	0.789	**0.701**	3.672	3.634	3.660
S2	$\epsilon = 0.05$	0.915	0.867	**0.846**	4.075	4.089	4.070
	$\epsilon = 0.1$	0.910	0.942	**0.815**	4.065	4.072	4.096
	$\epsilon = 0.3$	0.890	0.899	**0.830**	4.084	4.087	4.089
	$\epsilon = 0.6$	0.894	0.898	**0.864**	4.066	4.100	4.069
Geolife	$\epsilon = 0.05$	6.433	6.702	**5.698**	24.261	24.361	24.280
	$\epsilon = 0.1$	**5.534**	6.380	5.545	24.222	24.240	24.261
	$\epsilon = 0.3$	**4.020**	11.137	4.515	24.212	24.229	24.284
	$\epsilon = 0.6$	2.406	11.401	**2.356**	24.076	24.139	24.135
Porto	$\epsilon = 0.05$	1.747	1.784	**1.422**	6.806	6.829	6.837
	$\epsilon = 0.1$	1.698	1.632	**1.509**	6.863	6.865	6.838
	$\epsilon = 0.3$	**1.738**	1.637	1.755	6.794	6.792	6.786
	$\epsilon = 0.6$	1.459	3.128	**1.201**	6.760	6.799	6.792

Table 8 – Mean Absolute Error (MAE) for $k = 256$.

DataSet	Budget	ALOQ-1S Adap	ALOQ-1S Base	ALOQ-2S	LOLOHA	LOSUE	RAPPOR
S1	$\epsilon = 0.05$	18.03	17.78	**17.36**	48.07	48.09	47.95
	$\epsilon = 0.1$	18.06	18.38	**17.18**	48.01	48.03	48.11
	$\epsilon = 0.3$	17.98	17.87	**17.20**	48.03	48.10	48.05
	$\epsilon = 0.6$	18.12	17.99	**17.34**	47.99	48.19	48.00
S2	$\epsilon = 0.05$	17.88	17.41	**17.04**	46.63	46.73	46.46
	$\epsilon = 0.1$	17.53	17.87	**16.62**	46.62	46.71	46.52
	$\epsilon = 0.3$	17.88	17.88	**17.07**	46.53	46.65	46.57
	$\epsilon = 0.6$	17.79	17.77	**16.91**	46.58	46.45	46.59
Geolife	$\epsilon = 0.05$	21.48	21.99	**20.63**	62.99	63.21	63.18
	$\epsilon = 0.1$	21.13	21.71	**20.41**	63.09	63.02	63.14
	$\epsilon = 0.3$	14.90	32.53	**13.76**	63.13	63.19	63.17
	$\epsilon = 0.6$	10.25	33.29	**8.69**	62.77	62.94	62.85
Porto	$\epsilon = 0.05$	19.18	19.60	**18.69**	54.22	54.18	54.29
	$\epsilon = 0.1$	19.48	19.33	**18.70**	54.40	54.50	54.38
	$\epsilon = 0.3$	19.66	19.55	**18.49**	54.25	54.16	54.18
	$\epsilon = 0.6$	17.76	30.52	**16.06**	54.03	54.13	54.19

In our experiments we varied the privacy budget, ϵ, from 0.05 to 0.6. The quadtree was initialized with $k = 256$ leaf cells, ensuring a uniform spatial resolution across all methods and datasets.

Utility is reported using Mean Absolute Error (MAE) and Mean Squared Error

(MSE), shown in Tables 8 and 7, respectively. Table 7 lists MSE values scaled by 10^{-3}; multiply the displayed numbers by 10^3 to obtain the actual error.

Across the synthetic datasets (**S1** and **S2**), the **ALOQ-2S** approach consistently yields the best performance, achieving the lowest MAE and MSE in all configurations, showing that two-round refinement is effective even in lower-density synthetic data.

In the highly skewed **Geolife** dataset, ALOQ-2S again performs best across nearly all privacy budgets. It achieves the lowest MAE and MSE at $\epsilon = 0.6$ (8.69 and 2.356×10^3, respectively), showing significant gains over ALOQ-1S-Base, whose MAE exceeds 33 and MSE exceeds 11×10^3. In particular, ALOQ-1S-Adap performs slightly better than ALOQ-2S at $\epsilon = 0.1$ and $\epsilon = 0.3$ in MSE, suggesting that single-round adaptation is more robust under tight budget conditions in extremely skewed data. However, the two-round strategy quickly surpasses it as the privacy constraints are relaxed.

In the more balanced but real-world **Porto** dataset, ALOQ-2S delivers the best MAE and MSE at three out of four budget levels. At $\epsilon = 0.6$, it achieves an MAE of 16.06 and MSE of 1.201×10^3, while ALOQ-1S-Base degrades significantly, with an MAE of 30.52 and MSE of 3.128×10^3. ALOQ-1S-Adap remains competitive across all budgets but is consistently slightly behind ALOQ-2S. The performance gap between adaptive and baseline methods (LOLOHA, LOSUE, RAPPOR) remains substantial, with baselines showing little responsiveness to increased budget and high errors throughout.

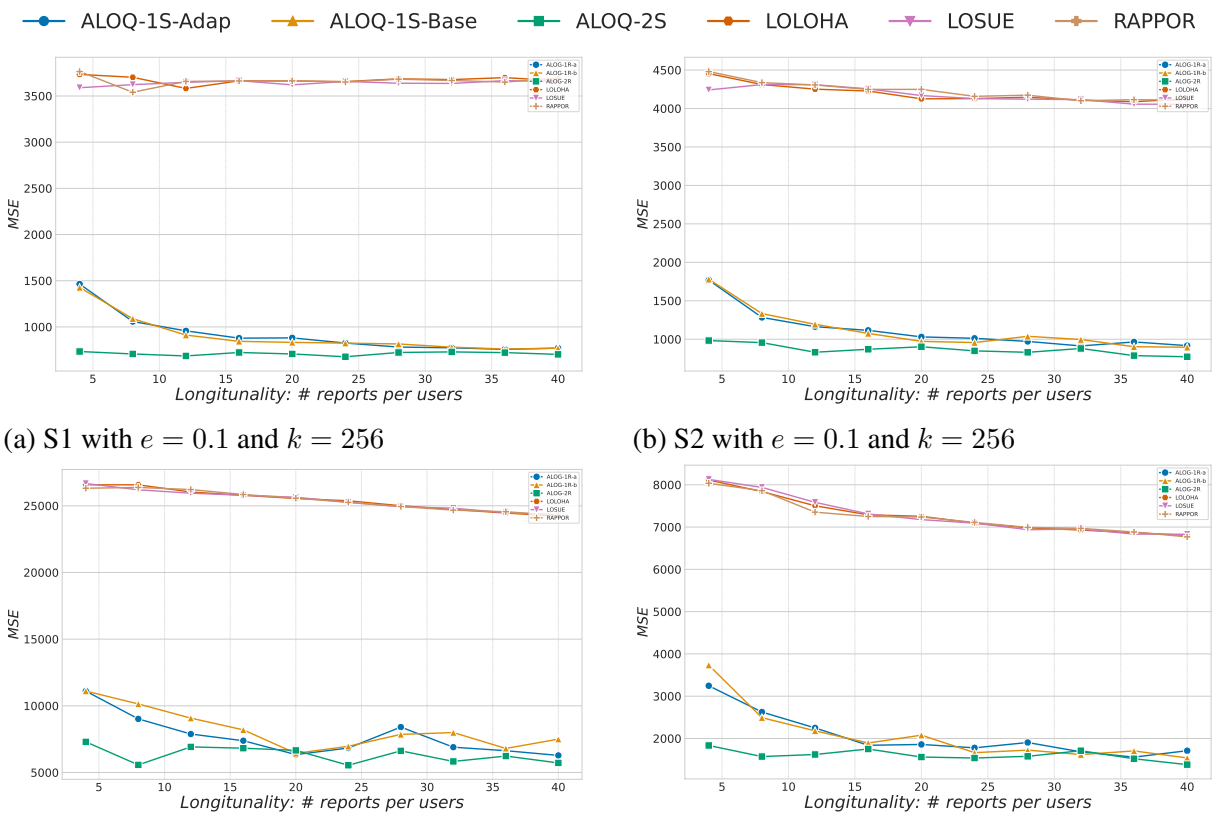
In summary, the results show that **ALOQ-2S** is the most robust and accurate method overall, providing the best trade-off between privacy and utility in most scenarios. Although ALOQ-1S-Adap may briefly outperform at lower budgets in highly skewed settings, ALOQ-2S dominates as the budget increases.

5.4.1.3 Longitudinal Setting vs. Utility

This subsection examines the impact of repeated user reporting over time. We assess how each method performs under longitudinal data collection, where maintaining utility while preserving privacy becomes increasingly challenging.

Both MAE and MSE plots reveal a consistent pattern: as the number of user reports increases, the utility improves across all methods and datasets. However, the magnitude and stability of this improvement differ significantly between adaptive and non-adaptive approaches.

Across all datasets, **ALOQ-2S** consistently delivers the lowest error in both MAE



(a) S1 with $e = 0.1$ and $k = 256$

(b) S2 with $e = 0.1$ and $k = 256$

(c) Geolife with $e = 0.1$ and $k = 256$

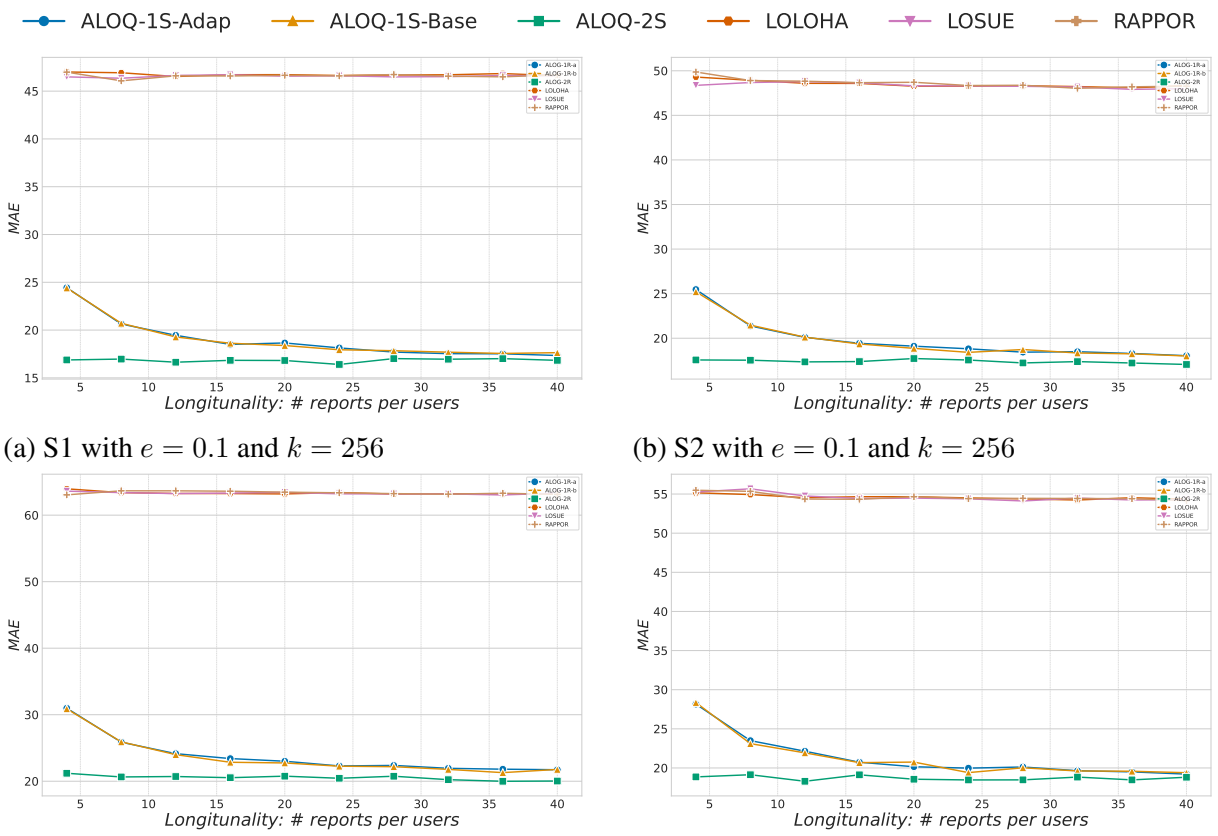
(d) Porto with $e = 0.1$ and $k = 256$

Figure 17 – MSE variation with the number of reports per user for the four datasets: S1, S2, GeoLife, and Porto.

and MSE, especially at higher longitudinal depths. **ALOQ-1S-Adap** often performs closely, sometimes slightly outperforming ALOQ-2S at lower report counts. This is evident in datasets like S1 and S2, where the performance gap between ALOQ-1S-Adap and ALOQ-2S narrows as the number of reports increases. These methods leverage their spatial refinement mechanisms to effectively absorb temporal fluctuations in user behavior.

In contrast, baseline methods such as LOLOHA, LOSUE, and RAPPOR exhibit higher and more stable error curves, showing limited sensitivity to longitudinal variation. Particularly in the **Geolife** and **Porto** datasets, both characterized by skewed or real-world movement patterns, these baselines fail to capitalize on the growing amount of user data, maintaining consistently high MAE and MSE.

The consistent superiority of ALOQ-2S and ALOQ-1S-Adap across all longitudinal depths and datasets reinforces their effectiveness in dynamic environments. Their ability to adapt to evolving user behavior and spatial patterns ensures significantly lower error, making them highly suitable for longitudinal privacy-preserving systems. These findings confirm that advanced spatially adaptive methods are essential for achieving high utility in real-world



(a) S1 with $e = 0.1$ and $k = 256$

(b) S2 with $e = 0.1$ and $k = 256$

(c) Geolife with $e = 0.1$ and $k = 256$

(d) Porto with $e = 0.1$ and $k = 256$

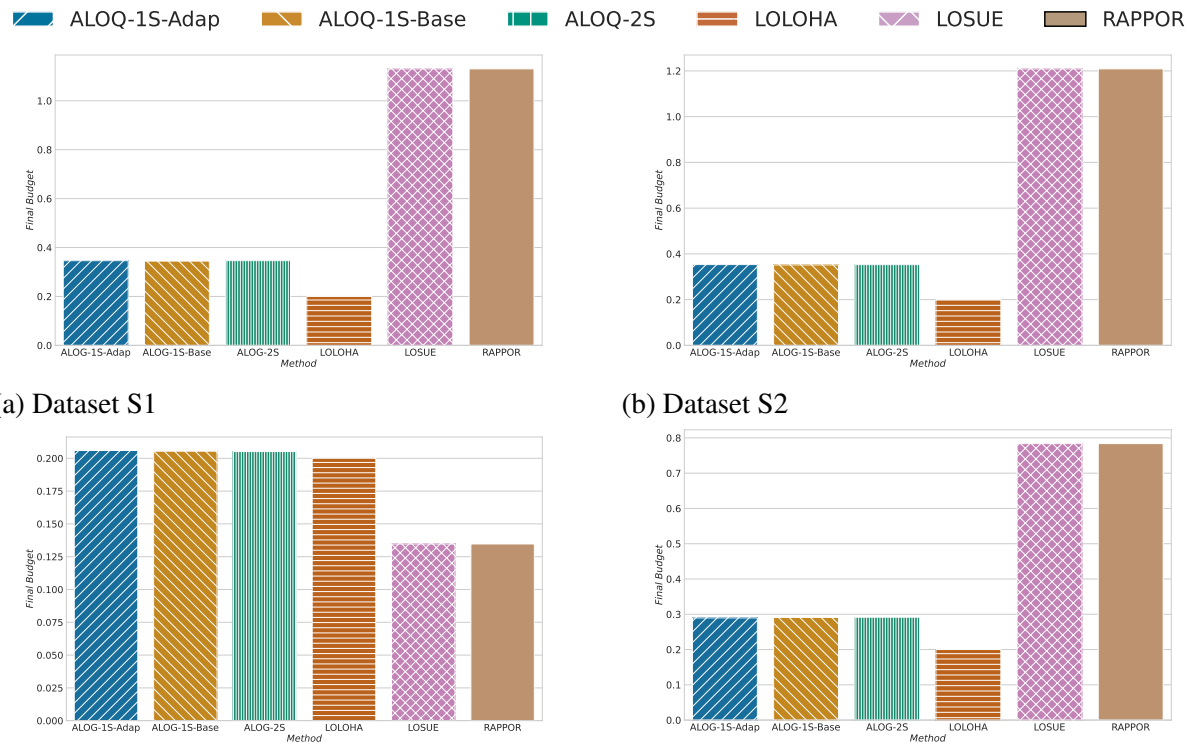
Figure 18 – MAE variation with the number of reports per user for the four datasets: S1, S2, GeoLife, and Porto.

deployments where users continuously contribute data.

5.4.1.4 Budget Consumption vs. Utility

We compare the cumulative privacy cost incurred by each method, focusing on the relationship between budget consumption and utility. The results show that ALOQ conserves privacy budget more effectively by leveraging adaptation and memoization strategies.

In the synthetic datasets, **Uniform** and **Normal**, Figures 19a and 19b, the ALOQ approaches (**ALOQ-1S-Adap**, **ALOQ-1S-Base**, and **ALOQ-2S**) exhibit remarkably similar behavior. All three consume a low and nearly identical portion of the privacy budget, closely matching the most efficient method, **LOLOHA**. This demonstrates that the adaptive mechanisms in ALOQ, while flexible, can maintain efficiency in uniformly distributed scenarios. LOLOHA's local hashing strategy proves effective in reducing domain size and, consequently, limiting budget expenditure. In contrast, the static baselines **RAPPOR** and **LOSUE** display poor performance, with significantly higher budget usage in these settings.



(a) Dataset S1

(b) Dataset S2

(c) Geolife

(d) Porto

Figure 19 – Budget Consumption for Datasets S1, S2, Geolife and Porto

In the **Porto** dataset (Figure 19d), the same trend holds: ALOQ methods remain budget-efficient, and LOLOHA continues to perform competitively. RAPPOR and LOSUE again lag in efficiency, consuming more of the allocated privacy budget.

A shift occurs in the **Geolife** dataset (Figure 19c), which features a highly skewed spatial distribution. Here, **ALOQ** approaches and LOLOHA maintain their low budget consumption, showing that their adaptivity does not lead to unnecessary refinement in concentrated data. However, in contrast to the previous datasets, **RAPPOR** and **LOSUE** show a significant improvement, consuming substantially less budget than they did in uniform and semi-structured contexts. This improvement is likely due to the effectiveness of memoization when user reports are spatially concentrated, allowing static methods to reuse responses more frequently and reduce budget expenditure. This phenomenon is confirmed by the results in Figure 20, which displays the number of memoization hits for each method and reveals a marked increase in hits for RAPPOR and LOSUE in the Geolife dataset. The higher count of memoization hits directly correlates with more frequent reuse of privatized responses, substantiating the observed reduction in budget consumption.

In summary, ALOQ methods and LOLOHA are budget-efficient across all scenarios. Yet, in skewed datasets like Geolife, RAPPOR and LOSUE can close the efficiency gap,

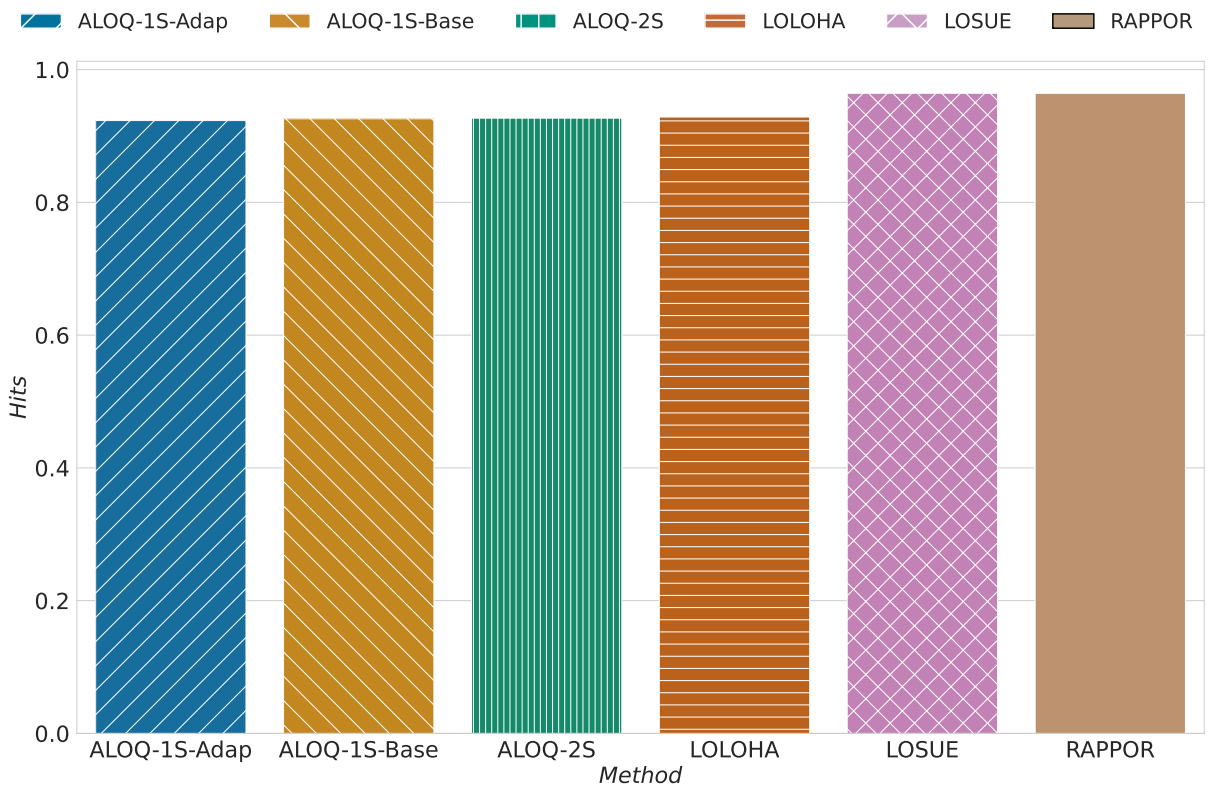


Figure 20 – Memoization hits for each method in the Geolife dataset, showing the frequency of repeated response reuse due to spatial concentration.

highlighting the influence of spatial distribution on privacy budget consumption.

To deepen this investigation, we extend our analysis to explicitly evaluate privacy budget consumption using Pareto frontier analysis (Figure 21). This approach enables us to simultaneously assess utility and efficiency by plotting the trade-off between MSE and the final budget used. Methods positioned closer to the Pareto frontier exhibit superior performance, as they achieve lower error with minimal privacy budget consumption. The resulting frontiers reveal how adaptive and static methods compare not only in isolation but also in terms of their overall cost-benefit effectiveness under local differential privacy constraints.

In the synthetic datasets, LOLOHA remains close to the frontier only under conditions of very low privacy budget consumption. In contrast, adaptive methods such as ALOQ-2S and ALOQ-1S-Adap dominate all other approaches at higher budget levels, consistently achieving lower MSE values. For the GeoLife dataset, ALOQs outperform other methods starting from privacy budget consumptions above approximately 0.1 with ALOQ-2S and ALOQ-1S-Base dominating with higher privacy budget. For the Porto dataset, ALOQ-2S dominates at budget consumptions exceeding 0.05.

In the **GeoLife** dataset, which exhibits high spatial skewness, the adaptive approaches

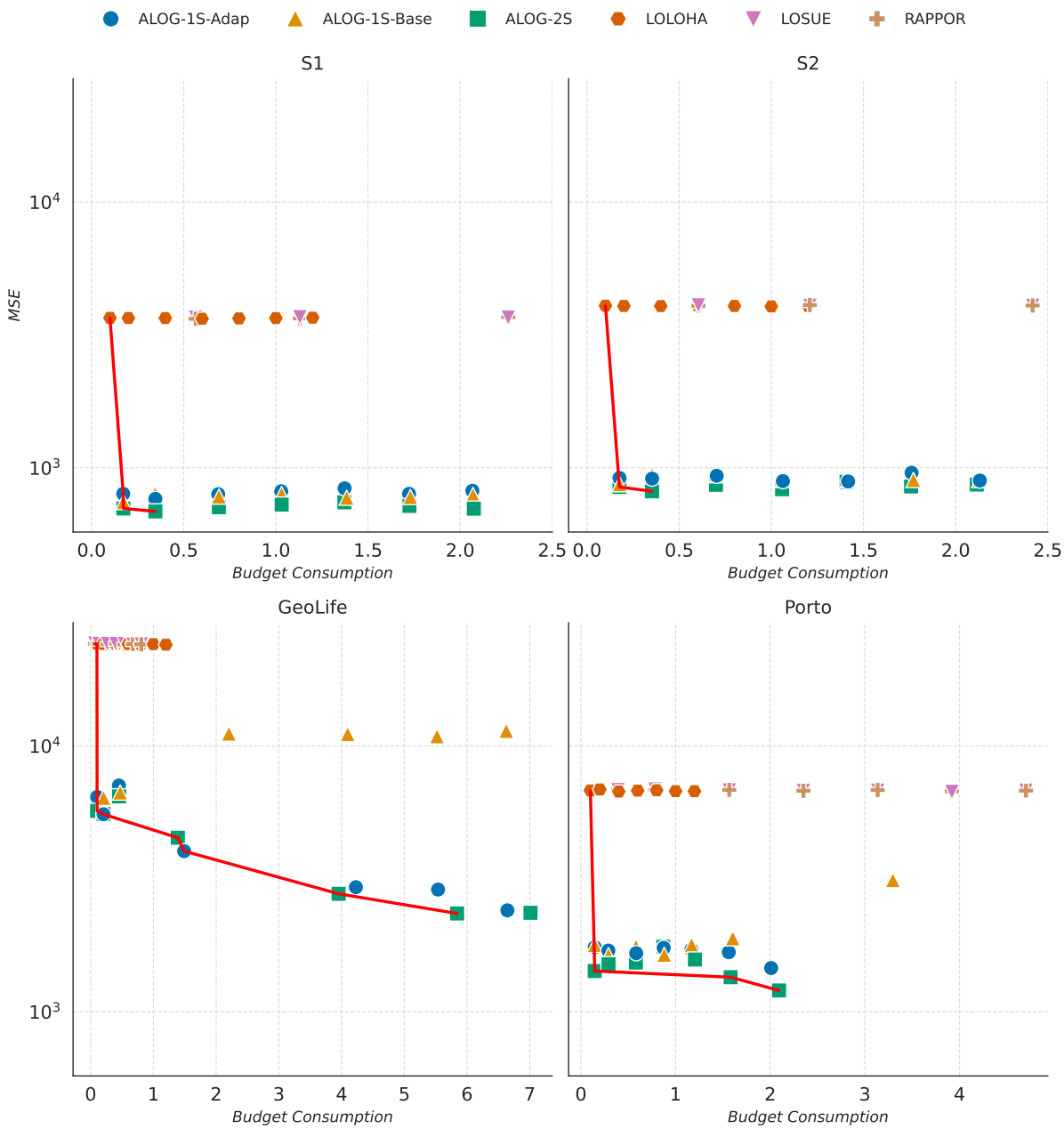


Figure 21 – ALOQ dominates the Pareto frontier, achieving the best trade-off between MSE and budget in most scenarios. Only under very low budgets does LOLOHA slightly outperform it

begin to outperform others once the privacy budget exceeds approximately $\epsilon = 0.1$. In this scenario, both ALOQ-2S and ALOQ-1S-Base progressively reduce estimation error, with ALOQ-2S achieving the lowest MSE at higher budgets. Similarly, in the **Porto** dataset, ALOQ-2S dominates the trade-off space once the budget exceeds $\epsilon = 0.05$, offering significantly better utility than static baselines.

LOLOHA consistently demonstrates lower budget consumption across all scenarios, owing to its local hashing mechanism, which reduces the effective domain size and improves memoization efficiency. In contrast, RAPPOR and LOSUE exhibit the highest budget consumption in most datasets, with the sole exception of the GeoLife dataset. In this highly skewed setting, frequent repetition of spatial cells allows memoization to mitigate budget usage more effectively. Nevertheless, these static methods remain significantly distant from the Pareto frontier in terms of utility, highlighting their limited adaptability and lower estimation accuracy when compared to the proposed adaptive quadtree-based approaches.

These results highlight that although ALOQ approaches may incur slightly higher budget consumption, particularly when compared to LOLOHA, they consistently offer superior utility. This makes them especially suitable for applications where accuracy, adaptability, and responsiveness to temporal and spatial dynamics are critical, including longitudinal settings.

5.4.2 Comparison with PrivTC

In this section, we evaluate the performance of ALOQ-2S, the variant that achieved the best overall results, in comparison with PrivTC (Section 3.4), a recent adaptive method for location privacy under Local Differential Privacy (LDP). This analysis focuses on spatial granularity, privacy budget efficiency, and utility in longitudinal settings, providing a comprehensive assessment of ALOQ's advantages over state-of-the-art adaptive partitioning strategies.

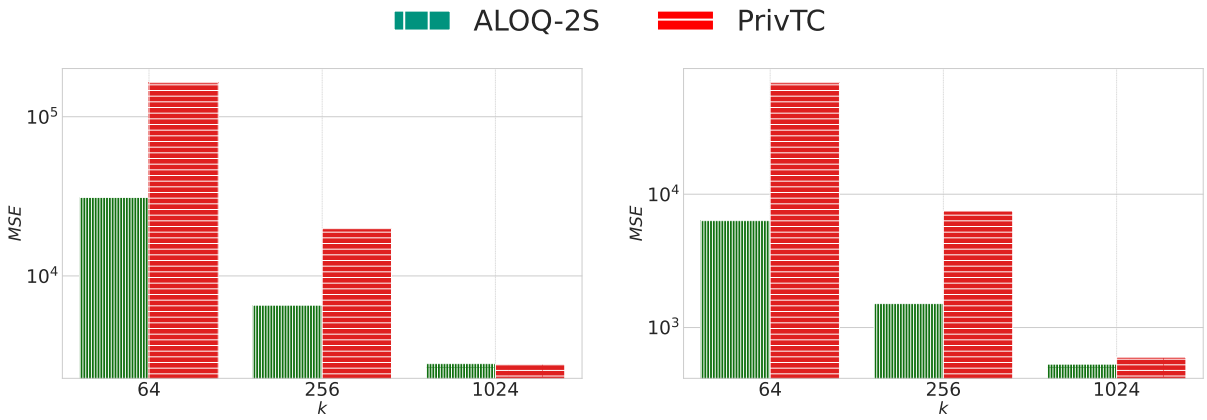
5.4.2.1 Quadtree Granularity vs. Utility

We begin by analyzing how the initial quadtree granularity affects the utility of ALOQ-2S and PrivTC. By varying the starting spatial resolution, we evaluate the sensitivity of both methods to partitioning choices. The results show that ALOQ-2S maintains superior estimation accuracy across most grid sizes, particularly in datasets with highly skewed spatial distributions.

Interestingly, as the initial partition becomes more fine-grained (e.g., with 1024

cells), the behavior diverges between datasets. In the highly skewed GeoLife dataset, the difference between the two methods narrows, with PrivTC exhibiting a slightly lower MSE at the finest partitioning level. However, in the Porto dataset, where the spatial distribution is more concentrated and less skewed, ALOQ-2S maintains its advantage across all granularities, including when the initial partition is very fine. This suggests that the recursive, data-driven spatial adaptation of ALOQ-2S remains beneficial even under high-resolution initial configurations, especially in more homogeneous or moderately skewed scenarios.

Overall, for practical partition sizes and across varying data distributions, ALOQ-2S provides robust and generally superior utility due to its ability to adapt recursively throughout the longitudinal process.

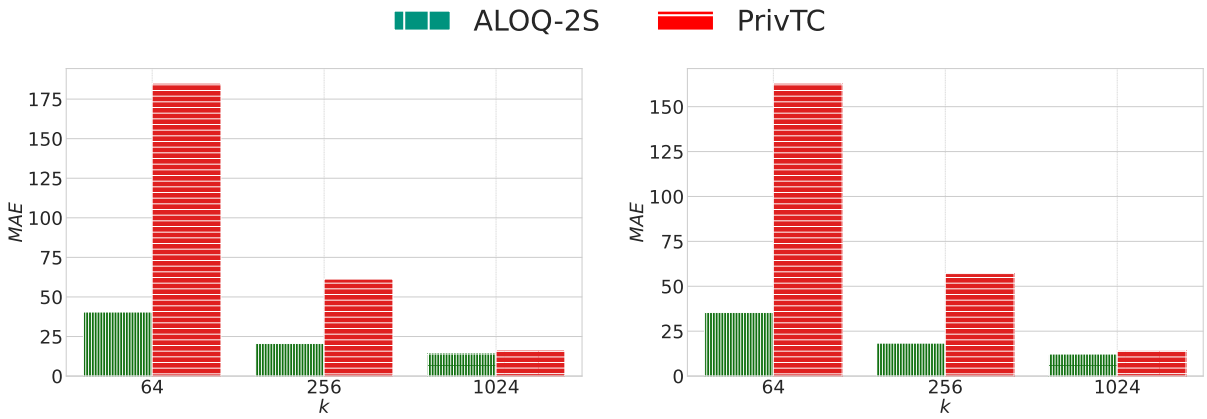


(a) Geolife with $\epsilon_\infty = 0.1$ (b) Porto with $\epsilon_\infty = 0.1$

Figure 22 – MSE vs. initial grid size for GeoLife and Porto datasets (log scale).

With respect to Mean Absolute Error (MAE), ALOQ-2S also demonstrates superior performance, achieving lower MAE than PrivTC on both real-world datasets. This trend is particularly evident in scenarios where spatial distributions are skewed or dynamically changing, further highlighting the benefits of quadtree-based, recursive refinement in maintaining accuracy over time.

In summary, these results confirm that while PrivTC can approach ALOQ's performance under highly granular initial partitions, ALOQ-2S generally provides more robust utility, both in terms of MSE and MAE, across a broad range of partition sizes and real-world data distributions.

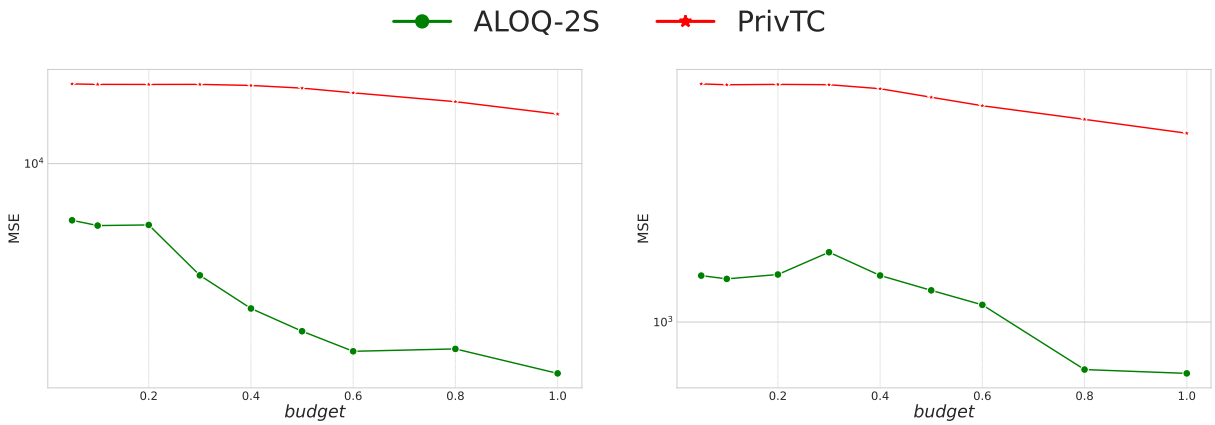


(a) Geolife with $\epsilon_\infty = 0.1$ (b) S2 with $\epsilon_\infty = 0.1$

Figure 23 – MAE vs. initial grid size for GeoLife and Porto datasets.

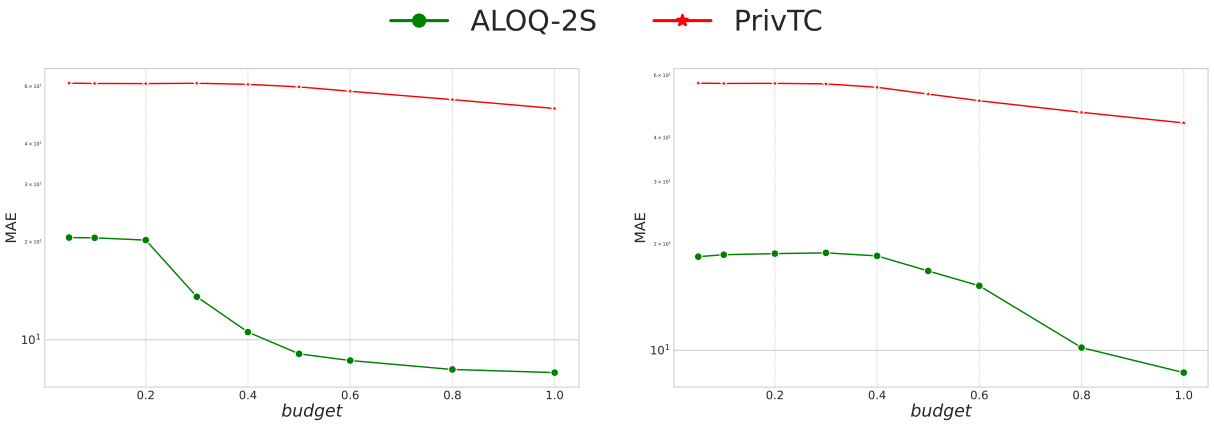
5.4.2.2 Privacy Budget vs. Utility

Next, we compare ALOQ-2S and PrivTC under varying privacy budgets. The MSE and MAE are reported for multiple values of ϵ to assess how each method balances privacy protection and accuracy. As illustrated in Figures 24 and 25, which report the MSE and MAE respectively, ALOQ-2S achieves lower estimation errors throughout the entire range of evaluated privacy budgets. This consistent advantage demonstrates the effectiveness of ALOQ's dynamic and recursive spatial adaptation, ensuring robust utility without compromising privacy, regardless of the budget allocation.



(a) Geolife with $k = 256$ (b) Porto with $k = 256$

Figure 24 – MSE variation with the privacy budget for the datasets: GeoLife, and Porto.

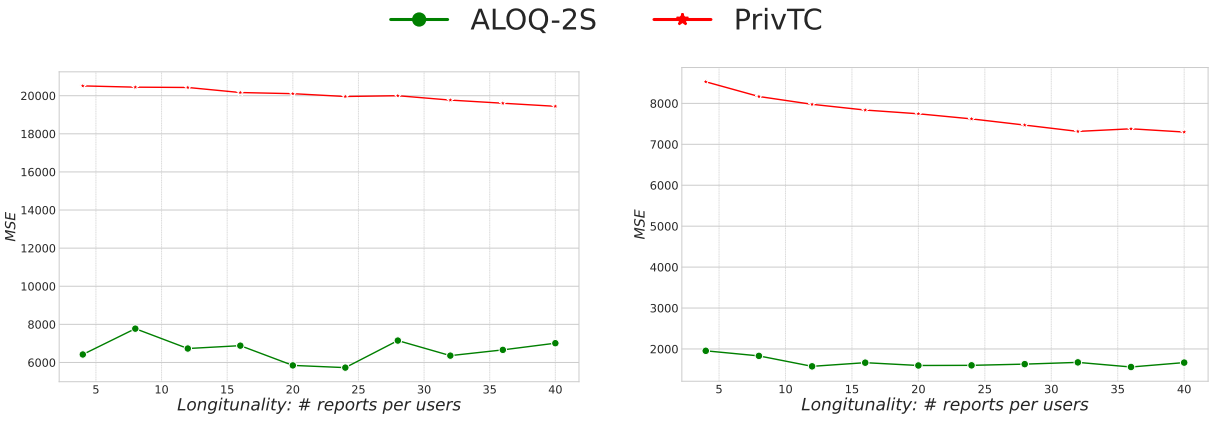


(a) Geolife with $k = 256$ (b) Porto with $k = 256$

Figure 25 – MAE variation with the privacy budget for the datasets: GeoLife, and Porto.

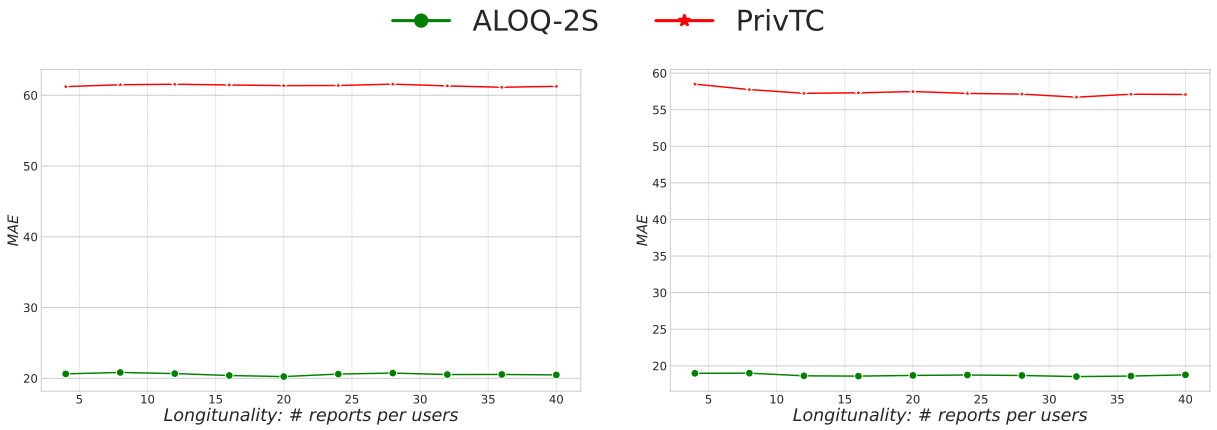
5.4.2.3 Longitudinal Setting vs. Utility

Here, we examine how both approaches perform over time in a longitudinal setting, where users report their location repeatedly. We analyze the utility degradation or improvement across timestamps. ALOQ-2S and PrivTC exhibit relative insensitivity to the range of values, particularly in terms of MAE. The MAE remains stable for both methods across all evaluated longitudinal depths. In the case of MSE, we observe a slightly greater variance for ALOQ-2S; however, it is important to note that ALOQ-2S still achieves better MSE and MAE than PrivTC across the entire range of values, as shown in Figures 26 and 27. These results indicate that, despite minor fluctuations, ALOQ-2S maintains superior utility in longitudinal scenarios.



(a) Geolife with $e = 0.1$ and $k = 256$ (b) Porto with $e = 0.1$ and $k = 256$

Figure 26 – MSE variation with the number of reports per user for the datasets: GeoLife, and Porto.

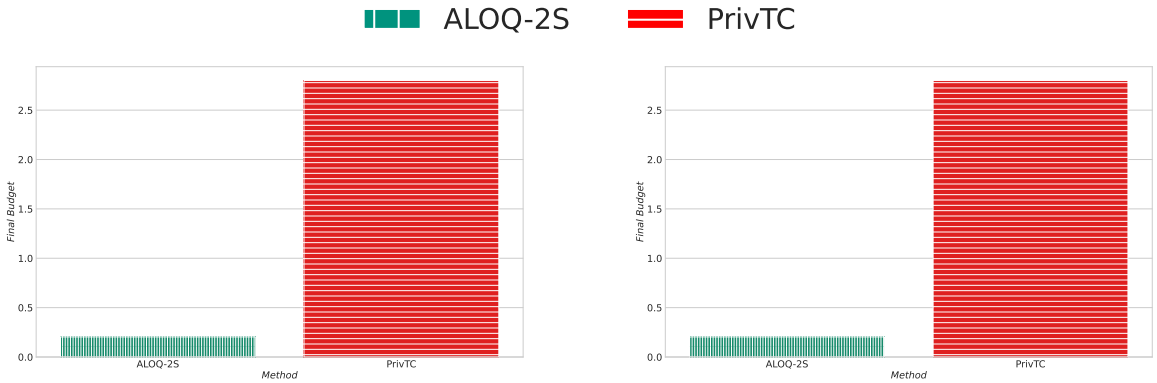


(a) Geolife with $e = 0.1$ and $k = 256$ (b) Porto with $e = 0.1$ and $k = 256$

Figure 27 – MAE variation with the number of reports per user for the datasets: GeoLife, and Porto.

5.4.2.4 *Budget Consumption vs. Utility*

Finally, we compare the utility achieved relative to the total privacy budget consumed. We observe in Figure 28 that **ALOQ-2S achieves significantly lower privacy budget consumption compared to PrivTC**, while simultaneously maintaining high utility. This demonstrates that ALOQ-2S is not only more accurate in its estimations, but also more efficient in privacy budget management, highlighting its practical advantages for longitudinal privacy-preserving applications.



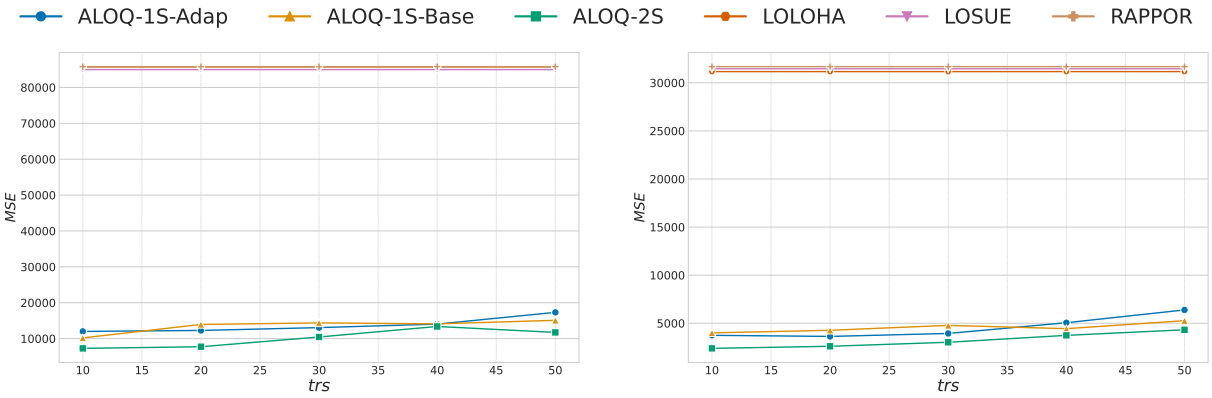
(a) Dataset Geolife (b) Dataset Porto

Figure 28 – Budget Consumption for Datasets Geolife and Porto

5.4.3 Sensitivity to Adaptation Thresholds

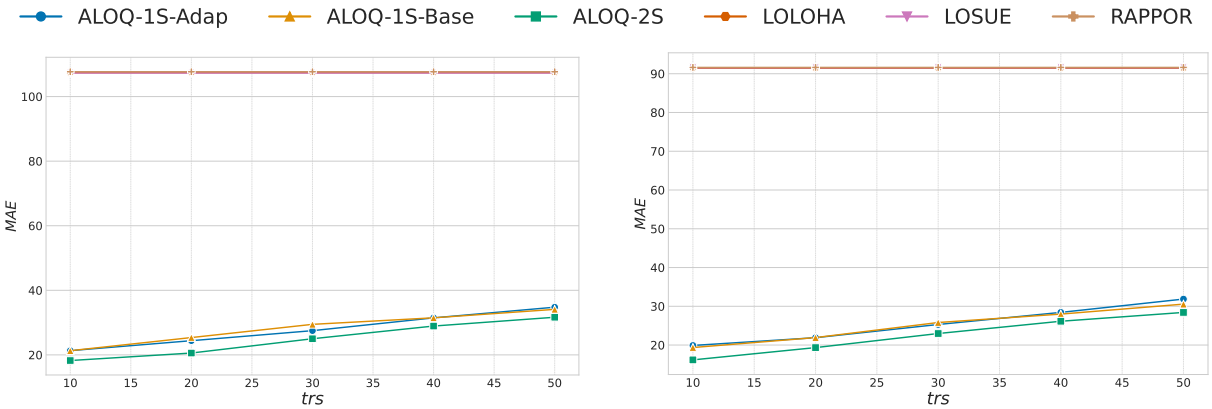
In this section, we analyze the impact of the quadtree adaptation thresholds, tr_{max} and tr_{min}, on the utility of the ALOQ framework. Understanding the sensitivity to these parameters is crucial for assessing the model's robustness and providing guidance for practical deployment.

For this experiment, we varied the upper threshold tr_{max} across a range of values, while the lower threshold tr_{min} was consistently set to 25% of the corresponding tr_{max} value. We measured the resulting estimation error (MSE and MAE) for the ALOQ variants and plotted these against the performance of the static baselines (RAPPOR, LOSUE, LOLOHA), whose performance is independent of these thresholds. The results for the GeoLife and Porto datasets are presented in Figures 29 and 30.



(a) Dataset Geolife (b) Dataset Porto

Figure 29 – Impact of the adaptation threshold tr_{max} on Mean Squared Error (MSE).



(a) Dataset Geolife (b) Dataset Porto

Figure 30 – Impact of the adaptation threshold tr_{max} on Mean Absolute Error (MAE).

The results clearly demonstrate the robustness of the ALOQ framework. While the choice of tr_{max} does influence the exact error magnitude, the ALOQ variants consistently and significantly outperform all static baselines across a wide range of threshold values. For both MSE and MAE, even a suboptimal choice of thresholds yields substantially better utility than the best-performing static method. This indicates that the fundamental benefit of the adaptive structure is not contingent on exhaustive parameter tuning. An optimal performance range can be identified, but the key takeaway is that the adaptive approach is inherently superior, providing a strong practical advantage for real-world deployments where fine-grained tuning may not be feasible.

5.5 Discussion & Key Findings

The comprehensive experimental evaluation confirms our central hypothesis: an adaptive spatial partitioning mechanism, driven exclusively by noisy frequency estimates, can significantly outperform both static and state-of-the-art adaptive LDP baselines in terms of data utility and privacy budget efficiency. The key findings from our experiments are summarized below:

- **Superior Utility Over Static Baselines:** Across all four datasets, the ALOQ variants consistently and substantially lowered estimation error compared to static methods (RAPPOR, LOSUE, LOLOHA). As shown in Figures 15 and 16, ALOQ-2S reduced the Mean Squared Error (MSE) by as much as 92% and the Mean Absolute Error (MAE) by up to 84% compared to the strongest static baseline. The performance gains were most pronounced in the highly skewed GeoLife dataset, where ALOQ's adaptive splitting effectively mitigated the non-uniformity error that plagues fixed-grid approaches.
- **Dominance Over State-of-the-Art Adaptive Methods:** Crucially, ALOQ's advantages extend beyond static baselines. The direct comparison with PrivTC, a contemporary adaptive grid framework, establishes ALOQ as the superior adaptive approach. In nearly all tested configurations of granularity, privacy budget, and longitudinal depth, ALOQ-2S achieved significantly lower MSE and MAE. This demonstrates that ALOQ's fully recursive quadtree refinement and continuous, similarity-driven adaptation provide greater utility and responsiveness than PrivTC's adaptation grid model.
- **Efficient and Intelligent Budget Utilization:** ALOQ's architecture is not only more accurate but also more efficient. The similarity-guided budget scaling mechanism effectively

conserves the privacy budget by adapting ϵ_t to the level of distributional drift. This is further validated by the direct comparison with PrivTC, which shows ALOQ-2S consuming a fraction of the privacy budget while delivering higher utility. The Pareto frontier analysis synthesizes this finding, illustrating that ALOQ variants consistently dominate the frontier, offering the best trade-off between error and budget consumption in nearly all practical scenarios.

- **Longitudinal Robustness and Variant Ranking:** The ALOQ model proves to be highly robust in long-term deployments. As the number of reports per user increases, all ALOQ variants maintain their performance advantage over the baselines. Among the variants, **ALOQ-2S** emerges as the best all-around performer, providing the most consistent and lowest error across the majority of datasets and privacy regimes. **ALOQ-1S-Adap** remains a strong competitor, particularly under tight budget constraints in skewed data, while **ALOQ-1S-Base** offers a simpler, nearly stateless design at the cost of a slight reduction in utility.
- **Robustness to Internal Parameters:** The sensitivity analysis on the adaptation thresholds (tr_{max}, tr_{min}) confirms the robustness of our framework. As shown in Figures 29 and 30, ALOQ maintains its significant performance advantage over static methods across a wide spectrum of threshold values. This finding underscores the practical viability of ALOQ, as it does not depend on fragile, fine-grained parameter tuning to deliver its core benefits.

5.6 Threats to Validity

Dataset bias. Only two longitudinal datasets containing full trajectory coordinates are currently available; both stem from urban transport and may not generalize to rural mobility. **Parameter tuning.** Thresholds $(tr_{\text{max}}, tr_{\text{min}}, tr_s)$ were chosen via empiric evaluation; other tunes might alter the frontier.

5.7 Chapter Summary

Chapter 5 presented a comprehensive empirical evaluation of the ALOQ framework, demonstrating that its core mechanisms, adaptive quadtree partitioning, QAW-based drift detection, and similarity-guided budget scaling, yield a superior privacy-utility trade-off. The results

confirmed that ALOQ variants significantly outperform not only traditional static LDP baselines but also the state-of-the-art adaptive method, PrivTC. In particular, ALOQ-2S was identified as the most robust variant, achieving up to 84% lower MAE than the strongest static baseline and demonstrating consistently lower error and more efficient budget use than PrivTC. These empirical findings validate the theoretical claims of Chapter 4 and motivate the deployment-oriented future work outlined in the final chapter.

6 CONCLUSIONS

The research presented in this thesis set out to test the hypothesis that an adaptive spatial partitioning mechanism, driven by noisy frequency estimates, could significantly improve the utility of longitudinal geospatial frequency estimation compared to static methods. The comprehensive experimental evaluation provides strong evidence to support this hypothesis.

We introduced ALOQ, a novel model for continuous and privacy-preserving geospatial data collection under the model of Local Differential Privacy (LDP). To the best of our knowledge, ALOQ is the first to combine dynamic spatial representation, longitudinal reporting, and similarity-aware budget allocation into a single, adaptive architecture that does not rely on a trusted curator or any assumptions about the underlying data.

At its core, ALOQ lies an adaptive quadtree that evolves in response to changing spatial patterns across time. By relying solely on privatized reports to refine spatial granularity, ALOQ maintains rigorous privacy guarantees while improving accuracy. The adaptation is governed by the Quadtree Adaptation Window (QAW), which triggers structural changes only when significant shifts in estimated distributions are detected. Additionally, a similarity-driven budget allocation strategy allows ALOQ to modulate privacy budget usage according to temporal stability, conserving resources during stable periods and reinforcing utility during transitions.

We formally analyzed the privacy guarantees of our model and conducted extensive experiments on synthetic and real-world mobility datasets. The results confirmed that ALOQ consistently outperforms traditional LDP protocols with static discretizations, like RAPPOR, LOSUE, and LOLOHA, and the state-of-the-art adaptive grid framework, PrivTC, in both utility and budget efficiency. Among its variations, ALOQ-2S emerged as the most robust strategy across privacy regimes, achieving up to 92% lower MSE than static baselines and demonstrating consistently lower error and more efficient budget use than PrivTC.

Key insights from this research include:

- Adaptive spatial partitioning under LDP significantly improves estimation accuracy by reducing non-uniformity error, even when compared to other adaptive techniques.
- Dynamic, similarity-guided privacy budget allocation ensures efficient use of resources without compromising privacy guarantees.
- Even under continuous longitudinal reporting, cumulative privacy loss remains bounded due to memoization and a controlled, privacy-preserving tree expansion mechanism.
- In skewed, real-world datasets, ALOQ adapts to shifting densities more effectively than

both static and other adaptive structures, yielding superior accuracy with more efficient budget consumption.

A current limitation of ALOQ is that merging operations are restricted to the initialization phase. This prevents the structure from reducing granularity in regions where density decreases over time. This design choice maintains stability but could be extended in future work.

Beyond methodological advances, ALOQ enables concrete, privacy-preserving public health action. Frequency patterns derived from privatized mobility data are a proven decision primitive for situational awareness (e.g., hotspot heatmaps) and can directly support interventions such as resource allocation and distancing policies. At the same time, raw location histories pose acute risks of re-identification and profiling, which historically impede responsible data use in health crises.

Consider Fortaleza facing a fast-spreading respiratory virus. With citizen opt-in, smartphones locally perturb GPS reports under LDP before transmission, data are "born private," and repeated reporting across days remains protected, allowing continuous, city-scale monitoring without a trusted curator. ALOQ's adaptive quadtree aligns spatial granularity to noisy demand: it refines dense outbreak zones and merges sparse regions, mitigating the non-uniformity errors of static grids/quadtrees while conserving the privacy budget.

In practice, health authorities can (i) prioritize testing teams, ICU beds, and supplies toward cells with persistent frequency surges; (ii) implement targeted, time-bounded distancing in specific neighborhoods or hours instead of blanket restrictions; and (iii) evaluate policy effectiveness by tracking pre/post shifts in privatized visit frequencies, all as post-processing that never weakens privacy guarantees.

Empirically, ALOQ consistently achieves a superior privacy–utility trade-off versus widely used LDP baselines, strengthening its suitability for operational epidemiology where timely, accurate signals must coexist with strict privacy budgets. This impact perspective underscores ALOQ not only as a technical contribution, but as a practical foundation for ethics-aligned epidemic management in Fortaleza and similar urban contexts.

In conclusion, ALOQ provides a foundational contribution toward adaptive, privacy-preserving geospatial analytics. It opens new research opportunities for studying the balance between privacy and utility in dynamic environments where users continuously contribute data. It also lays the groundwork for future advances in real-time privacy-aware spatial modeling.

6.1 Future Work

Building on the contributions of this thesis, several promising directions for future research emerge:

- **Fully Dynamic Split-and-Merge Strategy:** A primary research topic for extending our results is to investigate a fully dynamic refinement strategy that allows the quadtree to both split and merge cells throughout the entire longitudinal process. This would enable the structure to coarsen in regions where user density decreases over time, potentially leading to greater efficiency and accuracy in long-term deployments with highly variable mobility patterns.
- **Advanced Adaptation Triggers:** The current QAW mechanism uses cosine similarity to detect distributional drift. Future work could explore alternative or hybrid triggers. For instance, directly using the non-uniformity error as a criterion for splitting or merging could provide a more targeted adaptation mechanism that is explicitly tied to the primary source of utility loss in LDP.
- **Automated Parameter Tuning:** ALOQ's performance depends on several parameters, including the adaptation thresholds (tr_{max}, tr_{min}, tr_s) and recursion bounds. While these were set empirically in our evaluation, future research could investigate learning-based approaches to automate parameter tuning, allowing the model to self-optimize for different datasets and application requirements.
- **Formal Derivation of the Optimal Privacy Budget:** The selection of the upper-bound privacy budget for each timestamp (ϵ) is a critical task in LDP, and is typically handled empirically by a privacy specialist. This ad-hoc process can lead to suboptimal privacy-utility trade-offs. A significant avenue for future work is to leverage ALOQ's unique structural properties to address this challenge. Since our model enforces an expected report density per cell (bounded by tr_{min} and tr_{max}) and the analytical task is known (frequency estimation), it may be possible to formally derive an optimal or near-optimal value for ϵ. This would replace empirical tuning with a principled, data-driven approach, providing stronger and more justifiable privacy guarantees for a given utility target.
- **Personalized and context-aware privacy budgeting:** Currently, LDP deployments, including ALOQ, apply a single privacy budget to all users and locations, yet people value privacy differently and certain places (e.g., hospitals, schools, residential areas) warrant stronger protection. As future work, we propose an ALOQ variant that lets users select a

simple privacy tier (e.g., high/medium/low) and automatically enforces stricter protection in designated sensitive zones and time windows. A lightweight *budget manager* would reconcile these choices under a global cap, increasing signal where patterns are changing and tightening protection where risk is higher, while maintaining formal privacy accounting and clear, revocable consent. Equity audits would ensure that accuracy does not degrade disproportionately across neighborhoods, and we would publish transparent summaries of protections. This extension preserves ALOQ's adaptive nature while aligning it with human preferences and contextual risk, improving utility without compromising trust.

REFERENCES

AHUJA, R.; ZEIGHAMI, S.; GHINITA, G.; SHAHABI, C. A neural approach to spatio-temporal data release with user-level differential privacy. **Proceedings of the ACM on Management of Data**, ACM New York, NY, USA, v. 1, n. 1, p. 1–25, 2023.

ANDRÉS, M. E.; BORDENABE, N. E.; CHATZIKOKOLAKIS, K.; PALAMIDESSI, C. Geo-indistinguishability: differential privacy for location-based systems. *In*: SADEGHI, A.; GLIGOR, V. D.; YUNG, M. (Ed.). **2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013**. ACM, 2013. p. 901–914. DOI 10.1145/2508859.2516735. Available at: https://doi.org/10.1145/2508859.2516735.

ARCOLEZI, H. H.; COUCHOT, J.; BOUNA, B. al; XIAO, X. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. **Digit. Commun. Networks**, v. 10, n. 2, p. 369–379, 2024. DOI 10.1016/J.DCAN.2022.07.003. Available at: https://doi.org/10.1016/j.dcan.2022.07.003.

ARCOLEZI, H. H.; PINZÓN, C.; PALAMIDESSI, C.; GAMBS, S. Frequency estimation of evolving data under local differential privacy. *In*: **Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023**. [*S.l.*]: OpenProceedings.org, 2023. p. 512–525. DOI 10.48786/EDBT.2023.44.

BASSILY, R.; NISSIM, K.; STEMMER, U.; THAKURTA, A. Practical locally private heavy hitters. **J. Mach. Learn. Res.**, v. 21, p. 16:1–16:42, 2020. Available at: https://jmlr.org/papers/v21/18-786.html.

BASSILY, R.; SMITH, A. D. Local, private, efficient protocols for succinct histograms. *In*: SERVEDIO, R. A.; RUBINFELD, R. (Ed.). **Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015**. ACM, 2015. p. 127–135. DOI 10.1145/2746539.2746632. Available at: https://doi.org/10.1145/2746539.2746632.

CHAMIKARA, M. A. P.; BERTÓK, P.; KHALIL, I.; LIU, D.; CAMTEPE, S.; ATIQUZZAMAN, M. Local differential privacy for deep learning. **IEEE Internet Things J.**, v. 7, n. 7, p. 5827–5842, 2020. DOI 10.1109/JIOT.2019.2952146. Available at: https://doi.org/10.1109/JIOT.2019.2952146.

CHENG, X.; FANG, L.; YANG, L.; CUI, S. Mobile big data: The fuel for data-driven wireless. **IEEE Internet Things J.**, v. 4, n. 5, p. 1489–1516, 2017. DOI 10.1109/JIOT.2017.2714189. Available at: https://doi.org/10.1109/JIOT.2017.2714189.

CNN. **Equifax data breach: 143 million people could be affected**. 2017. Available at: https://money.cnn.com/2017/09/07/technology/business/equifax-data-breach/index.html. Accessed on: 2024-06-23.

CORMODE, G.; MADDOCK, S.; MAPLE, C. Frequency estimation under local differential privacy [experiments, analysis and benchmarks]. **CoRR**, abs/2103.16640, 2021. Available at: https://arxiv.org/abs/2103.16640.

CUMMINGS, R.; DESFONTAINES, D.; EVANS, D.; GEAMBASU, R.; HUANG, Y.; JAGIELSKI, M.; KAIROUZ, P.; KAMATH, G.; OH, S.; OHRIMENKO, O.; PAPERNOT, N.; ROGERS, R.; SHEN, M.; SONG, S.; SU, W.; TERZIS, A.;

THAKURTA, A.; VASSILVITSKII, S.; WANG, Y.-X.; XIONG, L.; YEKHANIN, S.; YU, D.; ZHANG, H.; ZHANG, W. **Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment**. 2024. Available at: https://arxiv.org/abs/2304.06929.

CUNNINGHAM, T.; CORMODE, G.; FERHATOSMANOGLU, H.; SRIVASTAVA, D. Real-world trajectory sharing with local differential privacy. **Proc. VLDB Endow.**, v. 14, n. 11, p. 2283–2295, 2021. DOI 10.14778/3476249.3476280. Available at: http://www.vldb.org/pvldb/vol14/p2283-cunningham.pdf.

DANIEL, L. **Facebook Data Breach Fallout: Millions May Receive Compensation**. 2024. Available at: https://www.forbes.com/sites/larsdaniel/2024/11/18/facebook-data-breach-fallout-millions-may-receive-compensation/. Accessed on: 2025-06-23.

DING, B.; KULKARNI, J.; YEKHANIN, S. Collecting telemetry data privately. *In*: GUYON, I.; LUXBURG, U. von; BENGIO, S.; WALLACH, H. M.; FERGUS, R.; VISHWANATHAN, S. V. N.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA**. [s.n.], 2017. p. 3571–3580. Available at: https://proceedings.neurips.cc/paper/2017/hash/253614bbac999b38b5b60cae531c4969-Abstract.html.

DUARTE NETO, E. R.; FILHO, J. S. C.; NETO, A. A. M.; MACHADO, J. C. ALOG: adaptive longitudinal grids for geospatial data using local differential privacy. *In*: **Proceedings 29th International Conference on Extending Database Technology, EDBT 2026, Tampere, Finland, March 24-27, 2026**. OpenProceedings.org, 2026. p. 1–14. DOI 10.48786/EDBT.2026.01. Available at: https://doi.org/10.48786/edbt.2026.01.

DUARTE NETO, E. R.; MACHADO, J. C.; MENDONÇA, A. L. Privlbs: Preserving privacy in location based services. **J. Inf. Data Manag.**, v. 10, n. 2, p. 81–96, 2019. Available at: https://sol.sbc.org.br/journals/index.php/jidm/article/view/2038.

DWORK, C. Differential privacy. *In*: BUGLIESI, M.; PRENEEL, B.; SASSONE, V.; WEGENER, I. (Ed.). **Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II**. Springer, 2006. (Lecture Notes in Computer Science, v. 4052), p. 1–12. DOI $10.1007/11787006_1.Availableat : https : //doi.org/10.1007/11787006\_1$.

DWORK, C.; MCSHERRY, F.; NISSIM, K.; SMITH, A. D. Calibrating Noise to Sensitivity in Private Data Analysis. *In*: **Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings**. Springer, 2006. (Lecture Notes in Computer Science, v. 3876), p. 265–284. DOI $10.1007/11681878_14.Availableat : https : //doi.org/10.1007/11681878\_14$.

DWORK, C.; ROTH, A. The algorithmic foundations of differential privacy. **Found. Trends Theor. Comput. Sci.**, v. 9, n. 3-4, p. 211–407, 2014. DOI 10.1561/0400000042. Available at: https://doi.org/10.1561/0400000042.

ERLINGSSON, Ú.; PIHUR, V.; KOROLOVA, A. RAPPOR: randomized aggregatable privacy-preserving ordinal response. *In*: AHN, G.; YUNG, M.; LI, N. (Ed.). **Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014**. ACM, 2014. p. 1054–1067. DOI 10.1145/2660267.2660348. Available at: https://doi.org/10.1145/2660267.2660348.

FILHO, J. S. da C.; MACHADO, J. C. FELIP: A local differentially private approach to frequency estimation on multidimensional datasets. *In*: STOYANOVICH, J.; TEUBNER, J.; MAMOULIS, N.; PITOURA, E.; MÜHLIG, J.; HOSE, K.; BHOWMICK, S. S.; LISSANDRINI, M. (Ed.). **Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023**. OpenProceedings.org, 2023. p. 671–683. DOI 10.48786/EDBT.2023.56. Available at: https://doi.org/10.48786/edbt.2023.56.

FILHO, M. E. B.; NETO, E. R. D.; MACHADO, J. C. Privacy-preserving of patients with differential privacy: an experimental evaluation in COVID-19 dataset. **J. Inf. Data Manag.**, v. 12, n. 5, 2021. Available at: https://sol.sbc.org.br/journals/index.php/jidm/article/view/1947.

FINKEL, R. A.; BENTLEY, J. L. Quad trees: A data structure for retrieval on composite keys. **Acta Informatica**, v. 4, p. 1–9, 1974. DOI 10.1007/BF00288933. Available at: https://doi.org/10.1007/BF00288933.

HONG, D.; JUNG, W.; SHIM, K. Collecting geospatial data with local differential privacy for personalized services. *In*: **37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021**. IEEE, 2021. p. 2237–2242. DOI 10.1109/ICDE51399.2021.00230. Available at: https://doi.org/10.1109/ICDE51399.2021.00230.

HUANG, H.; CHEN, Z.; LI, B.; MA, Q.; HE, H. Festgcn: A frequency-enhanced spatio-temporal graph convolutional network for traffic flow prediction under adaptive signal timing. **Appl. Intell.**, v. 54, n. 6, p. 4848–4864, 2024. DOI 10.1007/S10489-024-05401-5. Available at: https://doi.org/10.1007/s10489-024-05401-5.

KASIVISWANATHAN, S. P.; LEE, H. K.; NISSIM, K.; RASKHODNIKOVA, S.; SMITH, A. D. What can we learn privately? **SIAM J. Comput.**, v. 40, n. 3, p. 793–826, 2011. DOI 10.1137/090756090. Available at: https://doi.org/10.1137/090756090.

KIM, J. Improving data utility in privacy-preserving location data collection via adaptive grid partitioning. **Electronics**, MDPI, v. 13, n. 15, p. 3073, 2024.

LAOUIR, A. E.; IMINE, A. Diapprox: Differential privacy-based online range queries approximation for multidimensional data. *In*: HONG, J.; PARK, J. W. (Ed.). **Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC 2024, Avila, Spain, April 8-12, 2024**. ACM, 2024. p. 337–344. DOI 10.1145/3605098.3636070. Available at: https://doi.org/10.1145/3605098.3636070.

LI, M.; ZHU, L.; ZHANG, Z.; XU, R. Achieving differential privacy of trajectory data publishing in participatory sensing. **Inf. Sci.**, v. 400, p. 1–13, 2017. DOI 10.1016/J.INS.2017.03.015. Available at: https://doi.org/10.1016/j.ins.2017.03.015.

LOURDES, M. S. Maria de; MENDONÇA, A. L.; NETO, E. R.; CHAVES, I. C.; CAMINHA, C.; BRITO, F. T.; FARIAS, V. A.; MACHADO, J. C. Facto dataset: A dataset of user reports for faulty computer components. *In*: SBC. **Dataset Showcase Workshop (DSW)**. [*S.l.*], 2024. p. 91–102.

MACHADO, J.; NETO, E. D.; FILHO, M. B.; CAVALCANTI, M.; TRAINA, A. Técnicas de privacidade de dados de localização. *In*: **Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2019**. [*S.l.*]: SBC, 2019. p. 8–37.

MARREIRAS NETO, A. A.; NETO, E. D.; FILHO, J. C.; MACHADO, J. Locally differentially private and consistent frequency estimation of longitudinal data. *In*: **Anais do XXXIX Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2024. p. 367–380. ISSN 2763-8979. DOI 10.5753/sbbd.2024.240837. Available at: https://sol.sbc.org.br/index.php/sbbd/article/view/30706.

MINH, H. N.; AN, H. T.; CAM, G. T. T.; HA, L. D. T.; QUOC, H. D. Fuzzy logic and quadtree-based control for mobile robots in dynamic environments. *In*: SPRINGER. **International Conference on Intelligent Robotics and Applications**. [*S.l.*], 2024. p. 429–445.

MIRANDA-PASCUAL, À.; GUERRA-BALBOA, P.; PARRA-ARNAU, J.; FORNÉ, J.; STRUFE, T. Sok: Differentially private publication of trajectory data. **Proc. Priv. Enhancing Technol.**, v. 2023, n. 2, p. 496–516, 2023. DOI 10.56553/POPETS-2023-0065. Available at: https://doi.org/10.56553/popets-2023-0065.

MONTJOYE, Y.-A. de; HIDALGO, C. A.; VERLEYSEN, M.; BLONDEL, V. D. Unique in the crowd: The privacy bounds of human mobility. **Scientific reports**, Nature Publishing Group, v. 3, p. 1376, 2013.

OLIVEIRA, D. de; RODRIGUES, E.; COSTA, S.; AMORA, P.; CALDAS, A.; OCAÑA, K.; HORTA, M.; FILIPPIS, A. M. de; VIDAL, V.; MACHADO, J. Um estudo comparativo de mecanismos de privacidade diferencial sobre um dataset de ocorrências do zikv no brasil. *In*: SBC. **Simpósio Brasileiro de Banco de Dados (SBBD)**. [*S.l.*], 2019. p. 253–258.

OUCHRA, H.; BELANGOUR, A.; ERRAISSI, A. An overview of geospatial artificial intelligence technologies for city planning and development. *In*: IEEE. **2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)**. [*S.l.*], 2023. p. 1–7.

PAPPALARDO, L.; SIMINI, F.; RINZIVILLO, S.; PEDRESCHI, D.; GIANNOTTI, F.; BARABáSI, A.-L. Returners and explorers dichotomy in human mobility. **Nature Communications**, v. 6, n. 1, p. 8166, 2015.

PEIPEI, S.; YINGHUI, X.; QINGMING, Z. A study on spatial and temporal aggregation patterns of urban population in wuhan city based on baidu heat map and poi data. **Sustainable Development**, v. 8, n. 3, 2020.

PETERSON, M. P. **Mapping in the Cloud**. New York: Guilford Publications, 2014.

PYRGELIS, A.; TRONCOSO, C.; CRISTOFARO, E. D. Knock knock, who's there? membership inference on aggregate location data. **CoRR**, abs/1708.06145, 2017. Available at: http://arxiv.org/abs/1708.06145.

QARDAJI, W. H.; YANG, W.; LI, N. Differentially private grids for geospatial data. *In*: JENSEN, C. S.; JERMAINE, C. M.; ZHOU, X. (Ed.). **29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013**. IEEE Computer Society, 2013. p. 757–768. DOI 10.1109/ICDE.2013.6544872. Available at: https://doi.org/10.1109/ICDE.2013.6544872.

QIN, Z.; YANG, Y.; YU, T.; KHALIL, I.; XIAO, X.; REN, K. Heavy hitter estimation over set-valued data with local differential privacy. *In*: WEIPPL, E. R.; KATZENBEISSER, S.; KRUEGEL, C.; MYERS, A. C.; HALEVI, S. (Ed.). **Proceedings of the 2016 ACM**

SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016. ACM, 2016. p. 192–203. DOI 10.1145/2976749.2978409. Available at: https://doi.org/10.1145/2976749.2978409.

RIKALOVIC, A.; SOARES, G. A.; IGNJATIĆ, J. Spatial analysis of logistics center location: A comprehensive approach. **Decision Making: Applications in Management and Engineering**, v. 1, n. 1, p. 38–50, 2018.

SAMET, H. The quadtree and related hierarchical data structures. **ACM Comput. Surv.**, v. 16, n. 2, p. 187–260, 1984. DOI 10.1145/356924.356930. Available at: https://doi.org/10.1145/356924.356930.

SAMET, H. **Applications of spatial data structures - computer graphics, image processing, and GIS**. [*S.l.*]: Addison-Wesley, 1990. ISBN 978-0-201-50300-5.

SENA, L. B.; PRACIANO, F. D. B. S.; CHAVES, I. C.; BRITO, F. T.; NETO, E. R. D.; MONTEIRO, J. M.; MACHADO, J. C. AUDIO-MC: A general framework for multi-context audio classification. *In*: **Proceedings of the 24th International Conference on Enterprise Information Systems, ICEIS 2022, Online Streaming, April 25-27, 2022, Volume 1**. SCITEPRESS, 2022. p. 374–383. DOI 10.5220/0011071500003179. Available at: https://doi.org/10.5220/0011071500003179.

SHEKHAR, S.; EVANS, M. R.; KANG, J. M.; MOHAN, P. Identifying patterns in spatial information: A survey of methods. **WIREs Data Mining Knowl. Discov.**, v. 1, n. 3, p. 193–214, 2011. DOI 10.1002/WIDM.25. Available at: https://doi.org/10.1002/widm.25.

SINGER, J. D.; WILLETT, J. B. **Applied longitudinal data analysis: Modeling change and event occurrence**. Oxford: Oxford university press, 2003.

SOLYMOSI, R.; BUIL-GIL, D.; CECCATO, V.; KIM, E.; JANSSON, U. Privacy challenges in geodata and open data. **Area**, Wiley Online Library, v. 55, n. 4, p. 456–464, 2023.

TIRE, E.; GURSOY, M. E. Answering spatial density queries under local differential privacy. **IEEE Internet Things J.**, v. 11, n. 10, p. 17419–17436, 2024. DOI 10.1109/JIOT.2024.3357570. Available at: https://doi.org/10.1109/JIOT.2024.3357570.

UNNIKRISHNAN, J.; NAINI, F. M. De-anonymizing private data by matching statistics. *In*: **51st Annual Allerton Conference on Communication, Control, and Computing, Allerton 2013, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013**. IEEE, 2013. p. 1616–1623. DOI 10.1109/ALLERTON.2013.6736722. Available at: https://doi.org/10.1109/Allerton.2013.6736722.

UOL. **Banco Inter expõe dados de milhares de clientes**. 2017. Available at: https://www.uol.com.br/tilt/noticias/redacao/2018/08/17/banco-inter-confirma-vazamento-de-dados-apos-ataque-hacker.html. Accessed on: 2025-06-23.

Wakabayashi, Daisuke. **Google+ to shut down after security glitch exposes users' data**. 2018. Available at: https://www.nytimes.com/2018/10/08/technology/google-plus-security-disclosure.html. Accessed on: 2024-06-23.

WANG, T.; BLOCKI, J.; LI, N.; JHA, S. Locally differentially private protocols for frequency estimation. *In*: KIRDA, E.; RISTENPART, T. (Ed.). **26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017**. USENIX Association, 2017. p. 729–745.

Available at: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

WANG, T.; BLOCKI, J.; LI, N.; JHA, S. Locally differentially private protocols for frequency estimation. *In*: KIRDA, E.; RISTENPART, T. (Ed.). **26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017**. USENIX Association, 2017. p. 729–745. Available at: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

WARNER, S. L. Randomized response: A survey technique for eliminating evasive answer bias. **Journal of the American statistical association**, Taylor & Francis, v. 60, n. 309, p. 63–69, 1965.

WEI, J.; LIN, Y.; YAO, X.; ZHANG, J. Differential privacy-based location protection in spatial crowdsourcing. **IEEE Transactions on Services Computing**, IEEE, v. 15, n. 1, p. 45–58, 2019.

YAN, Y.; SUN, Z.; MAHMOOD, A.; XU, F.; DONG, Z.; SHENG, Q. Z. Achieving differential privacy publishing of location-based statistical data using grid clustering. **ISPRS International Journal of Geo-Information**, MDPI, v. 11, n. 7, p. 404, 2022.

YANG, J.; CHENG, X.; SU, S.; SUN, H.; CHEN, C. Collecting individual trajectories under local differential privacy. *In*: IEEE. **2022 23rd IEEE International Conference on Mobile Data Management (MDM)**. [*S.l.*], 2022. p. 99–108.

YAO, Q.; LUO, Z.; YUAN, X.; YU, Y.; ZHANG, C.; XIE, J.; LEE, J. Y. Assembly of nanoions via electrostatic interactions: ion-like behavior of charged noble metal nanoclusters. **Scientific Reports**, Nature Publishing Group UK London, v. 4, n. 1, p. 3848, 2014.

YE, Y.; ZHANG, M.; FENG, D. Collecting spatial data under local differential privacy. *In*: IEEE. **2021 17th International Conference on Mobility, Sensing and Networking (MSN)**. [*S.l.*], 2021. p. 120–127.

ZANG, H.; BOLOT, J. Anonymization of location data does not work: A large-scale measurement study. **Proceedings of the 17th annual international conference on Mobile computing and networking**, p. 145–156, 2011.

ZHU, Y.; CAO, Y.; XUE, Q.; WU, Q.; ZHANG, Y. Heavy hitter identification over large-domain set-valued data with local differential privacy. **IEEE Transactions on Information Forensics and Security**, v. 19, p. 414–426, 2024. DOI 10.1109/TIFS.2023.3324726.