# How Tough Is Location Anonymization? Re-identifying 100K Real-User Trajectories in Japan

Abhishek Kumar Mishra
Inria
Lyon, France
abhishek.mishra@inria.fr

Mathieu Cunche
INSA Lyon, CITI Laboratory, Inria
Lyon, France
mathieu.cunche@insa-lyon.fr

Héber H. Arcolezi
Inria
Grenoble, France
heber.hwang-arcolezi@inria.fr

## Abstract

Mobility traces are among the most revealing forms of personal data, yet trajectory releases are often protected only by ad hoc transformations. We stress-test such practices on recently-released `YJMob100K`, an anonymized dataset of 100,000 user trajectories in Japan. First, we show that the applied protection leaves enough spatial and temporal structure to recover both the real-world geographic frame and the actual calendar timeline by exploiting density signatures, urban correlations, and temporal activity profiles. On top of this reconstruction, we quantify privacy risks through trajectory-level metrics that capture spatio-temporal $k$-anonymity, $m$-point unicity, home–work and multi-anchor uniqueness, and exposure to secluded and sensitive locations. These metrics reveal extensive re-identification surfaces: a small number of observations, anchors, or sensitive venues often suffices to uniquely pinpoint users or their social neighborhoods. Finally, we evaluate representative sanitization strategies: geo-indistinguishability, local differential privacy, and aggressive spatial de-structuring; and observe a consistent pattern: strong privacy parameters destroy downstream utility, while utility-preserving settings leave structural leakage largely intact. Overall, our findings show that current sanitization techniques are insufficient for large-scale mobility data, and they highlight the urgent need for trajectory-aware privacy mechanisms and stronger publication standards.

## CCS Concepts

• **Security and privacy → Pseudonymity, anonymity and untraceability**; **Social aspects of security and privacy**.

## Keywords

Mobility traces, privacy attacks, anonymization, re-identification, trajectory data, location privacy.

**ACM Reference Format:**
Abhishek Kumar Mishra, Mathieu Cunche, and Héber H. Arcolezi. 2026. How Tough Is Location Anonymization? Re-identifying 100K Real-User Trajectories in Japan. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*Conference'17, Washington, DC, USA*
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Mobility data is central to modern urban analytics, transportation planning, and epidemiological modeling. At the same time, human movement is highly identifying: trajectories encode stable routines, anchor locations, and social structure [13]. Even after direct identifiers are removed, an adversary can often link pseudonymous traces to external information and re-identify individuals [16]. This tension becomes particularly acute for large-scale, high-resolution releases, where rich structure is what makes the data useful.

In this paper, we take a close look at `YJMob100K` [26], a recently-published public mobility dataset containing GPS trajectories for 100,000 users in Japan. The release replaces absolute timestamps and geographic coordinates with values defined relative to unknown spatial and temporal reference points, with the stated goal of breaking the linkage to real locations and dates. However, the transformed trajectories still preserve dense structural signals: spatial density gradients, road-aligned corridors, recurrent commuting pathways, and weekly temporal rhythms; that act as latent "privacy beacons" at the national scale.

We build an adversarial pipeline that operates directly on these structural cues. While prior work has hinted at the region of origin [18], we go significantly further by recovering meter-level coordinates, exact calendar dates, and conducting a full sensitivity analysis of potential defenses. Specifically, we first use population-density signatures and urban-form correlations to match `YJMob100K` to real cities and recover the hidden geographic frame. We then refine this alignment to the meter level, mapping anonymized grid cells back to latitude-longitude coordinates. In parallel, we exploit temporal activity profiles, weekday/weekend patterns, public holidays, and event-specific spikes to re-identify the true calendar timeline. Together, these steps convert the "anonymized" release into a time-stamped, geo-referenced dataset of real users moving through a real Japanese city.

On top of this re-identification, we systematically quantify privacy risks using state-of-the-art, trajectory-level metrics. Our framework covers spatio-temporal $k$-anonymity under realistic adversary queries, $m$-point unicity of partial trajectories, home–work and multi-anchor uniqueness, and exposure to secluded and sensitive locations. These metrics reveal extensive attack surfaces: small sets of points, stable anchors, or rare venues often suffice to single out users or small social neighborhoods.

Finally, we evaluate three representative sanitization strategies that span widely adopted and frequently advocated approaches in location privacy. First, we consider *geo-indistinguishability* [3], which is widely regarded as a state-of-the-art formal privacy guarantee for protecting individual locations in the absence of a trusted curator. Second, we study *Generalized Randomized Response (GRR)* [14], a

well-established primitive in the local differential privacy (LDP) literature that has been used for discrete location reporting and large-scale aggregate mobility analysis [22]. Third, we examine an *aggressive spatial de-structuring baseline* that deliberately destroys global spatial coherence while preserving marginal density distributions, serving as an upper bound on defenses that rely on structure removal rather than point-wise noise. Across all three mechanisms, we observe a consistent pattern: configurations that preserve analytical utility leave substantial structural leakage intact, whereas parameter regimes that provide strong privacy guarantees severely distort population distributions, degrade anchor inference, and compromise downstream analytics.

In summary, our main contributions are:

(1) We develop a scalable spatial re-identification methodology that leverages population-density structure to match anonymized trajectory data to public demographic maps, enabling recovery of the hidden urban area in YJMob100K and refinement to meter-level alignment between grid cells and real-world coordinates.
(2) We re-identify the dataset's true calendar by combining weekday inference, public holiday alignment, and event-level corroboration, thereby restoring real-world timestamps for 100,000 users.
(3) We propose a comprehensive set of trajectory-level privacy metrics quantifying identity & attribute leakage via partial traces, anchor locations, spatial seclusion, and visits to sensitive venues once structural re-identification is possible.
(4) We conduct an empirical evaluation of representative sanitization strategies on the re-identified dataset, revealing a pronounced privacy-utility trade-off and showing that current mechanisms fall short for large-scale mobility releases.

Our findings highlight a serious and underappreciated risk: once structural information in space and time is exposed, national-scale mobility releases can be pulled back into the real world with high fidelity, enabling detailed profiling of real individuals and communities. As mobility data increasingly informs AI systems, urban policy, and public services, the community needs trajectory-aware privacy mechanisms and publication standards that explicitly defend against structural, population-scale attacks, rather than relying on ad hoc coordinate warping and timestamp binning.

## 2 Related Work

Pinter et al. [18] analyze the YJMob100K dataset and demonstrate that its anonymization preserves sufficient spatial structure to enable identification of the underlying geographic region via density-based matching. Their analysis provides early evidence that relative-coordinate releases can leak large-scale geographic information. However, they do not recover fine-grained spatial coordinates, do not address temporal re-identification, and do not study how alternative sanitization strategies from the literature would behave on YJMob100K. As a result, their work stops short of quantifying downstream privacy risks or translating these findings into actionable guidance for dataset publication.

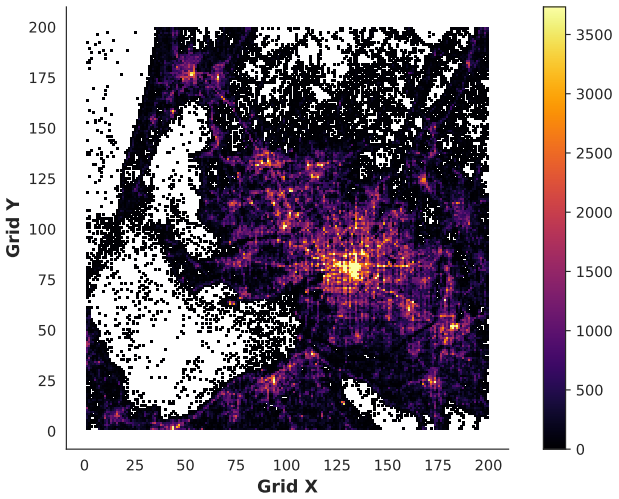



**Figure 1: Population densities of the released dataset. Grid-based trajectories in YJMob100K preserve structured mobility that aligns with urban infrastructure.**

A substantial body of prior work identifies location data as among the most privacy-sensitive forms of personal information. Location traces enable inference of sensitive personal attributes [7, 27] and frequently allow adversaries to single out individuals from large populations [16]. Crucially, this identifiability arises from structural properties of mobility data, such as regularity, sparsity, and anchor locations, rather than from explicit identifiers, making anonymization particularly challenging.

The literature on privacy-preserving data publishing proposes a wide range of mitigation strategies [10]. While general-purpose anonymization techniques are relatively well understood for tabular data, extending these approaches to spatial data [2, 22] and mobility traces remains difficult [9]. The sequential, highly correlated nature of trajectories complicates the application of both syntactic anonymization and formal privacy guarantees [9, 15, 19]. Empirical studies consistently show that anonymized mobility datasets remain vulnerable to a variety of re-identification and inference attacks, even after substantial perturbation [6, 11, 12, 20, 21, 25], highlighting a persistent privacy-utility trade-off.

This work extends prior findings by moving from isolated vulnerability demonstrations to an end-to-end, adversarial evaluation of a real-world mobility release. Whereas earlier studies primarily document identifiability or inference risks in abstract or controlled settings, we jointly reconstruct both spatial and temporal reference and use this reconstruction as a basis for systematic privacy measurement. By evaluating representative sanitization mechanisms under realistic utility constraints, our study connects structural re-identification directly to downstream privacy-utility trade-offs, offering an empirical perspective on how widely used defenses behave when confronted with population-scale mobility structure.

## 3 YJMob100K Dataset

The YJMob100K dataset comprises GPS trajectories from 100,000 anonymized users across Japan. Each record includes time-stamped

latitude-longitude coordinates, a user identifier, and additional metadata. The dataset also provides anonymized Points of Interest (PoIs), which represent frequently visited user locations.

Figure 1(a) provides a spatial overview of the grid-based movement trajectories. The patterns show structured mobility in an urban region: dense clusters, directional flows, and corridor-like traces that track transportation infrastructure and metropolitan dynamics. Even after anonymization, the retained spatial organization, including recurrent pathways and grid connectivity, mirrors real geographies such as transit axes and high-density cores. Prominent clusters suggest the presence of landmarks or PoIs that later act as strong anchors for re-identification.

## 4 Re-Identifying the Spatial Dimension

We re-identify the real-world location of `YJMob100K` using spatial structure and density. Our core insight is simple and powerful: an anonymization pipeline that preserves density and connectivity also preserves the city-level "fingerprint" of the underlying geography. We exploit that fingerprint with public population datasets.

### 4.1 Density-Based Re-identification Metric

We test the hypothesis that the anonymized dataset retains structural properties that match real population distributions. For `YJMob100K`, we compute daily spatial density as:

$$D(d) = \sum_{g \in G} P_g(d), \tag{1}$$

where $D(d)$ denotes the aggregated density on day $d$, $P_g(d)$ denotes the point count in grid cell $g$, and $G$ denotes the set of spatially selected grid cells. For comparison, we use public population density data [1] for Japan's ten largest cities. These distributions exhibit stable urban structure, including dense cores, corridor-like gradients, and tapering suburban regions. A successful anonymization pipeline must break the linkage between those structure markers and the released trajectory density.

### 4.2 Re-Identification of Location

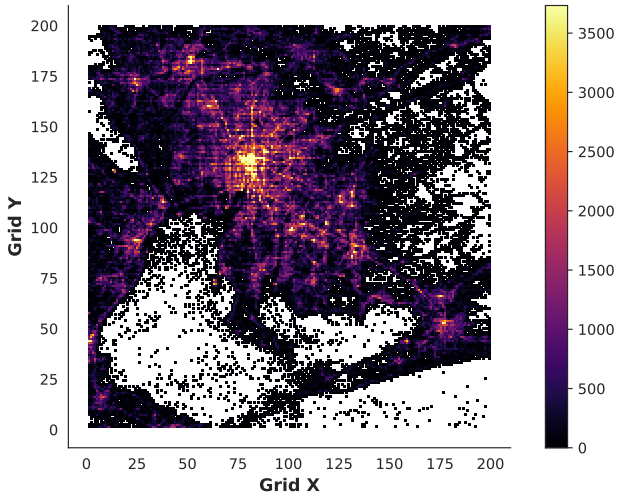
To infer the origin location of the dataset, we apply the re-identification procedure defined in Algorithm 1. The algorithm takes as input the anonymized dataset $D_{\text{raw}}$, a set of public population datasets $D_{\text{public}}$ corresponding to Japan's ten largest cities, and a transformation set $T$. The goal is to identify the most likely city of origin $C^*$ and the transformation $T^*$ that maximizes spatial similarity, with all intermediate correlation values stored in $S$.

We define a set of geometric transformations that account for potential obfuscation applied to $D_{\text{raw}}$: horizontal flip $F_x(D)$, vertical flip $F_y(D)$, 90° clockwise rotation $R_{90}(D)$, and 90° counterclockwise rotation $R_{-90}(D)$. Combining these operations produces eight transformed variants, which form the transformation set $T$.

For each transformed dataset $T_i \in T$, we compute a Spearman correlation $C_i(j)$ between $T_i(D_{\text{raw}})$ and each public dataset $D_{\text{public}}(j)$, where $j \in \{1, \ldots, 10\}$. To ensure robustness, we compute correlation over clustered spatial regions: we divide the full $200 \times 200$ grid into 25 non-overlapping clusters of size $40 \times 40$. We store the resulting scores in the dictionary $S$.




(a) Map of Nagoya




(b) Re-identified dataset

**Figure 2: Population densities of the re-identified dataset. (a) Map of Nagoya. (b) The same traces after our re-identification and alignment procedure (Algorithm 1), which reveals the underlying city and geography.**

We identify the optimal match by selecting $(C^*, T^*) = \arg\max_{(T_i, j)} C_i(j)$. The final output includes the re-identified city $C^*$ and the optimal transformation $T^*$.

Figure 2(b) visualizes the anonymized dataset after applying the optimal transformation ($R_{90}(F_y(D))$), and Figure 2(a) shows the corresponding match with the city of Nagoya. These transformations obscure orientation but preserve intrinsic movement structure, including cluster layout and displacement ranges. By inverting the transformations, we recover alignment with real geographies.

Quantitatively, Figure 3(a) and Figure 3(b) show Spearman correlations between the transformed dataset and public population data.

**Algorithm 1** Re-Identification Algorithm

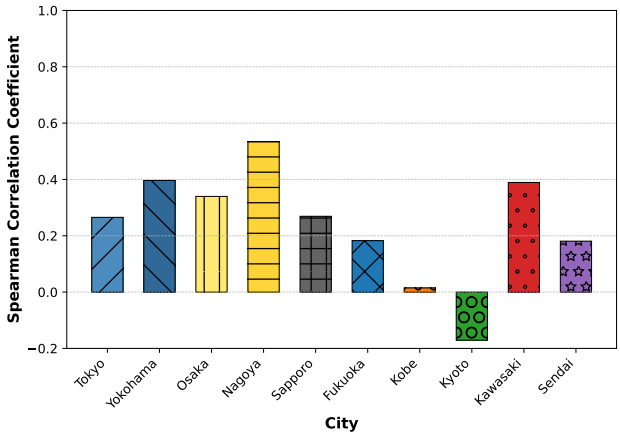
**Input:** $D_{\text{raw}}$, $D_{\text{public}}$
**Output:** $C^*$, $T^*$, $S$
$F_x(D) = \{(x, -y) \mid (x, y) \in D\}$ ▷ Flip x-axis
$F_y(D) = \{(-x, y) \mid (x, y) \in D\}$ ▷ Flip y-axis
$R_{90}(D) = \{(y, -x) \mid (x, y) \in D\}$ ▷ Rotate +90°
$R_{-90}(D) = \{(-y, x) \mid (x, y) \in D\}$ ▷ Rotate −90°
$T = \{D, F_x(D), F_y(D), F_x(F_y(D)), R_{90}(D), R_{-90}(D), R_{90}(F_x(D)), R_{90}(F_y(D))\}$
$S = \{\}$
**for** $T_i \in T$ **do**
  **for** $j = 1$ **to** 10 **do**
    $C_i(j) = C_{\text{corr}}(T_i(D_{\text{raw}}), D_{\text{public}}(j))$
    $S[(T_i, j)] = C_i(j)$
  **end for**
**end for**
$(C^*, T^*) = \arg\max_{(T_i, j) \in S} C_i(j)$
**Return** $(C^*, T^*, S)$

Nagoya consistently yields the strongest match, and the clustered strategy boosts robustness by suppressing cell-level noise. Under clustering, the correlation for Nagoya approaches unity, which indicates a near-perfect structural match.

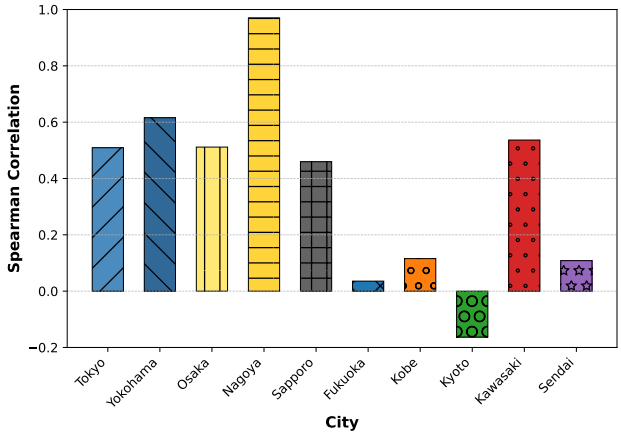
## 4.3 Fine-Grained Spatial Re-identification

After identifying Nagoya as the most probable city, we refine alignment with a fine-tuning step using hill climbing over a 100 km × 100 km area centered around Nagoya. This step targets exact latitude-longitude alignment that maximizes Spearman correlation between the public population grid and the anonymized dataset.

Starting from (35.05, 136.96), the algorithm iteratively perturbs the center by small steps (0.01°) and retains configurations that increase correlation. The process converges at (35.055019, 136.971202) once further perturbations yield no improvement. This procedure recovers the geocoordinates corresponding to anonymized grid cells in `YJMob100K` and completes spatial re-identification with meter-level precision. The recovered alignment, together with clustered correlation near 1.0 for Nagoya (Figure 3), confirms reliable spatial recovery. Prior work [18] hints at Nagoya as the underlying city; our method completes the attack surface by uncovering exact locations and validating the match through fine-grained optimization.




(a) Individual grid correlation




(b) Cluster-aggregated correlation

**Figure 3: Spearman correlations between the anonymized dataset and population data across 10 major Japanese cities. Nagoya emerges as the dominant match.**

# 5 Re-Identifying the Temporal Dimension

This section re-identifies the temporal dimension of `YJMob100K` by mapping its 75 consecutive days of mobility traces to exact calendar dates. Rather than treating time as anonymous indices, we ground the dataset in real chronology using population-scale behavioral regularities and distinctive temporal anchors.

Focusing on Nagoya, the aggregated user activity across all cells (Figure 4) exhibits clear weekly periodicity and sharp anomalies. These signals encode unique temporal signatures, including workdays, weekends, public holidays, and major events. By aligning these signatures with publicly known Japanese holidays and disruptions, we infer the dataset's time frame. The dataset publication timeline [26] bounds the latest possible end date, which we assume as no later than April 18, 2024, and we bound the earliest possible start date after January 1, 2015. Within this search interval, a single candidate sequence matches both the observed regularity and the exceptional deviations, which enables a precise temporal re-identification.

## 5.1 Identifying Weekdays

Figure 4 shows strong 7-day periodicity in global mobility, which signals a weekly cycle that alternates between working and non-working days. The sequence also contains deviations that indicate public holidays embedded in the data.

To strengthen day-type inference, we analyze activity in the top 10 residential areas. We identify the top grid cells by user count as follows: (82.0, 135.0), (77.0, 135.0), (81.0, 135.0), (82.0, 149.0), (77.0, 134.0), (87.0, 141.0), (80.0, 127.0), (52.0, 185.0), (101.0, 135.0) and (88.0, 124.0). We cluster daily activity profiles using K-Means (with $K = 2$ and a random initialization seed), after scaling the 48-bin

**Table 1: List of the 75 days in the dataset, along with their class (A – non-working day / B – working day) and the inferred weekday. Non-working days outside of weekends, $\mathcal{H}$, are highlighted in yellow.**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Day** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| **Class** | A | A | B | B | B | B | A | A | A | B | B | B | B | A | A |
| **Weekday** | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
| **Day** | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| **Class** | B | B | B | B | B | A | A | B | B | B | B | B | A | A | A |
| **Weekday** | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon |
| **Day** | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 |
| **Class** | B | B | B | B | A | A | B | A | B | B | B | A | A | B | B |
| **Weekday** | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue |
| **Day** | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 |
| **Class** | B | B | B | A | A | A | B | B | B | B | A | A | B | B | A |
| **Weekday** | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed |
| **Day** | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 |
| **Class** | B | B | A | A | B | B | B | B | B | A | A | B | B | B | B |
| **Weekday** | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon | Tue | Wed | Thu |

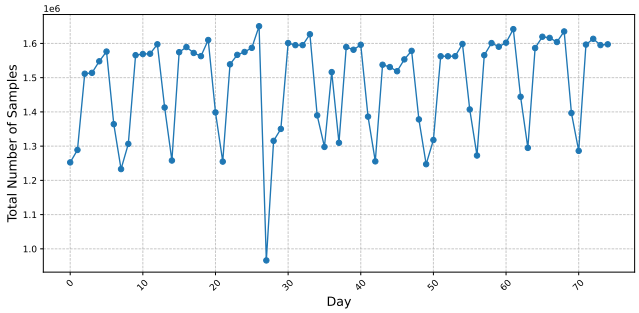



**Figure 4: Overall activity.**

time series using z-score normalization. Each day corresponds to a 48-dimensional vector of unique user counts per half-hour, and clustering uses Euclidean distance in this normalized space.

Figure 5 reveals two distinct behavioral regimes. Class A days show smoother morning ramps and gradual evening declines, consistent with relaxed or irregular routines that typically arise on weekends or holidays. Class B days show sharp morning and evening peaks that reflect commuting work schedules.

Day 27 stands out with unusually low activity and aligns with Class A, which signals a major disruption. The full day labeling (Table 1) follows a regular pattern of five consecutive Class B days followed by two Class A days, which confirms a standard Monday to Friday workweek with weekends every seventh day.

Based on this structure, we infer that the dataset begins on a Sunday (Day 0), which enables a deterministic mapping between day indices and weekdays. This weekday alignment enables the public-holiday matching step and supports event-level corroboration later in this section.

## 5.2 Identifying the Exact Date

Building on weekday inference, we now identify exact calendar dates by matching detected non-working days with Japan's public holiday calendar. In addition to regular weekends, we detect five non-working days outside the typical Saturday-Sunday pattern: Day 1 (Monday), Day 8 (Monday), Day 29 (Monday), Day 37 (Tuesday), and Day 50 (Monday). We denote this set of suspected public holidays as $\mathcal{H}$ (highlighted in Table 1).

To pinpoint the dataset timeline, we use Day 37, a Tuesday non-working day, as a strong anchor. We hypothesize that Day 37 aligns with a national holiday and we search for matching sequences in the 2015–2024 range using the official Japanese holiday calendar [23]. For each candidate Tuesday holiday, we test whether the remaining dates in $\mathcal{H}$ match known holidays under the weekday alignment implied by Table 1.

A single sequence satisfies all constraints: the dataset begins on 15 September 2019 (Day 0). Under this alignment, $\mathcal{H}$ matches Japan's national holidays exactly: *Respect for the Aged Day* (16/09/2019 - Day 1), *Autumn Equinox* (23/09/2019 - Day 8), *Health and Sports Day* (14/10/2019 - Day 29), *Enthronement Ceremony Day* (22/10/2019 - Day 37), and *Culture Day Observed* (04/11/2019 - Day 50). This unique match fixes Day 0 as 15/09/2019 and assigns calendar dates to the full 75-day sequence.

## 5.3 Additional Timeline Validation

We further validate the temporal reconstruction by matching real-world events in Nagoya to distinctive patterns in the dataset. These anchors, including disruptions and large gatherings, provide converging evidence that Day 0 corresponds to 15 September 2019.

*5.3.1 Activity anomaly on Day 27: Typhoon Hagibis.* Day 27 (Saturday) exhibits a pronounced activity collapse in Figure 4. The top 10 residential areas show an even sharper drop: Day 27 stands as the lowest-activity day in the 75-day sequence (Figure 5).

Under our alignment, Day 27 corresponds to 12 October 2019, the day Typhoon Hagibis makes landfall in Japan. The storm triggers nationwide disruption across transportation and public services, including impacts in the Nagoya region [4]. This tight temporal correspondence between a major disruption and a dataset-wide behavioral anomaly strongly corroborates our reconstructed timeline.

*5.3.2 Event correlation at Port Messe Nagoya.* Port Messe Nagoya, one of the city's largest exhibition venues, provides an additional temporal anchor. Activity at this location (Figure 6) shows multiple peaks during the dataset period. Each peak aligns with dates of major public events hosted at the venue. By cross-referencing public event calendars, we identify six events at Port Messe during 15/09/2019 to 28/11/2019, and each event coincides with a surge in user presence (see Table 2 in Section A).

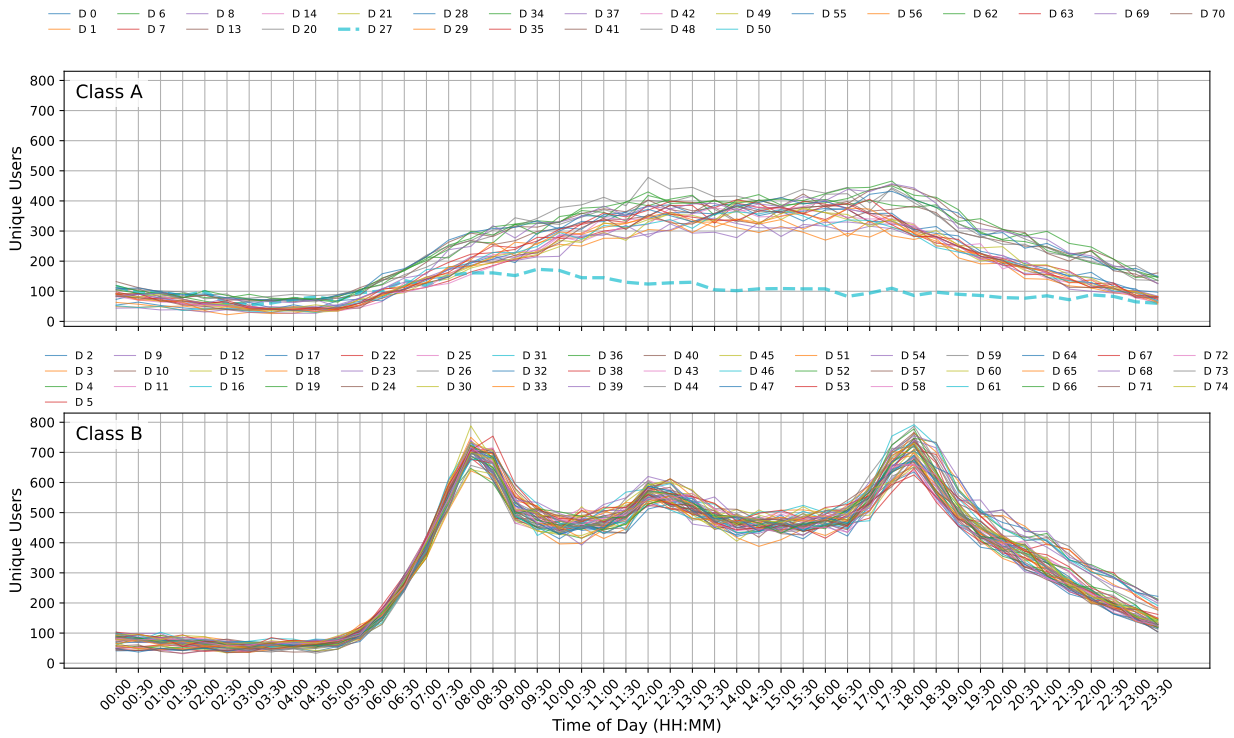



**Figure 5: Daily temporal activity in top 10 residential areas organized into two classes from a 2-class clustering.**

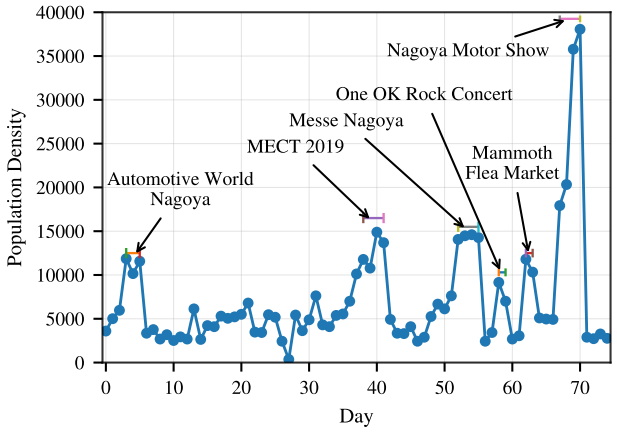



**Figure 6: Activity at Port Messe Nagoya (grid 69,110; GPS 35.05016252379402, 137.39573944062153) with hosted events.**

Furthermore, the magnitude of each peak correlates with the event's expected attendance. For example, the largest peak aligns with the *Nagoya Motor Show*, which attracts approximately 205,900 visitors. This relationship connects the dataset's spatio-temporal signals to real footfall at a known venue.

Taken together, these anchors, including public holidays, a large-scale disaster disruption, and location-specific event spikes, provide compelling corroboration. We conclude that the dataset begins on 15 September 2019, and our approach re-identifies the exact dates embedded in the anonymized trajectories.

## 6 Privacy Risks for Users in YJMob100K

After establishing that we can reverse the protection strategy for YJMob100K, we quantify the resulting privacy risks under the recovered location and timeline. To ensure ethical compliance, we refrain from deeper individualized tracing or identity linkage, even though the reconstructed spatio-temporal frame enables such steps. Prior work [16] highlights how these analyses risk serious harm when they connect mobility traces to identifiable individuals.

We introduce state-of-the-art privacy risk metrics that directly quantify leakage once an adversary recovers real coordinates and dates. We design each metric to (i) reflect a concrete adversary capability, (ii) admit rigorous mathematical definition, and (iii) support reproducible measurement and visualization.

### 6.1 Threat Model and Notation

We consider an adversary who obtains the released Yjmob100k trajectories and applies our spatial and temporal re-identification pipeline. The adversary therefore operates in the real-world reference frame: each record maps to a real grid cell and a real timestamp. The adversary also holds auxiliary knowledge about a target user, such as a home neighborhood, a work area, a small set of spatio-temporal check-ins, or attendance at a public event. This

threat model matches the standard linkage setting for mobility privacy [6, 16].

*Discrete representation.* We represent the mobility trace of user $u$ as a sequence

$$\mathcal{T}_u = \{(\ell_{u,i}, \tau_{u,i})\}_{i=1}^{n_u}, \tag{2}$$

where $\ell_{u,i} \in \mathcal{L}$ denotes a spatial cell (after re-identification) and $\tau_{u,i} \in \mathcal{H}$ denotes a time bin. We use 48 half-hour bins per day when we study daily temporal profiles, and we use day indices $d \in \{0, \ldots, 74\}$ when we study calendar-level signals.

*Visits and unique visitors.* For any cell $\ell \in \mathcal{L}$, let

$$\mathcal{U}(\ell) = \{u \mid \exists i \text{ s.t. } \ell_{u,i} = \ell\} \quad \text{and} \quad s(\ell) = |\mathcal{U}(\ell)| \tag{3}$$

denote the set and the count of unique users who have ever visited $\ell$. We use $s(\ell)$ as a principled measure of spatial seclusion.

## 6.2 Spatio-Temporal k-Anonymity

A dataset leaks identity when small fragments of a trace isolate a single user. We capture this effect through a spatio-temporal k-anonymity metric that counts how many users remain plausible matches under an adversary query.

*Adversary query model.* Let a query $Q$ consist of $m$ spatio-temporal constraints:

$$Q = \{(\ell_j, \Delta\tau_j)\}_{j=1}^{m}, \tag{4}$$

where each constraint says: "the user appears in spatial cell $\ell_j$ during time window $\Delta\tau_j$." Time windows let us model realistic uncertainty (for example, "around 9am" or "during lunch").

We define the *candidate set* of users consistent with $Q$ as

$$\text{Cand}(Q) = \left\{u \,\middle|\, \forall(\ell_j, \Delta\tau_j) \in Q,\ \exists i :\ \ell_{u,i} = \ell_j \wedge \tau_{u,i} \in \Delta\tau_j\right\}. \tag{5}$$

We define the *k-anonymity of query* $Q$ as

$$k(Q) = |\text{Cand}(Q)|. \tag{6}$$

*Population risk summary.* We report the fraction of queries that yield small candidate sets:

$$\text{Risk}_k(m, \Delta) = \Pr_{Q \sim \mathcal{D}(m,\Delta)} [k(Q) \le k], \tag{7}$$

where $\mathcal{D}(m, \Delta)$ samples $m$ constraints from real user traces and uses a time-window size $\Delta$ (for example, $\Delta = 30$ minutes or $\Delta = 2$ hours). This metric answers a critical security question: *how many auxiliary check-ins does an attacker need to isolate a user?*

Figure 7 shows that only a few spatio-temporal constraints are enough to isolate users even with coarse temporal tolerance $\Delta$. In Fig. 7(a), the uniqueness probability $\Pr[k(Q) = 1]$ is small for $m$=1 (only a few percent, depending on $\Delta$), but it jumps dramatically at $m$=2 (already around 0.75–0.90 as $\Delta$ tightens) and exceeds $\approx 0.94$ by $m$=3 for all shown $\Delta$, after which it rapidly saturates close to 1.

## 6.3 Trajectory Unicity Under Partial Knowledge

The mobility privacy literature repeatedly shows that a small number of points uniquely identifies individuals [16]. After re-identification, `Yjmob100k` supports the same style of attack at city scale and with real dates. We therefore quantify *unicity* directly, and we do so in a way that matches our discrete representation.

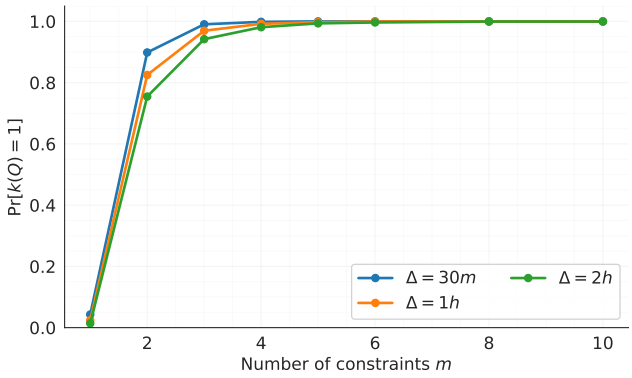



**Figure 7: Spatio-temporal $k$-anonymity risk under realistic adversary queries. Uniqueness probability rapidly increases with the number of spatio-temporal constraints $m$, even with coarse temporal tolerance $\Delta$.**

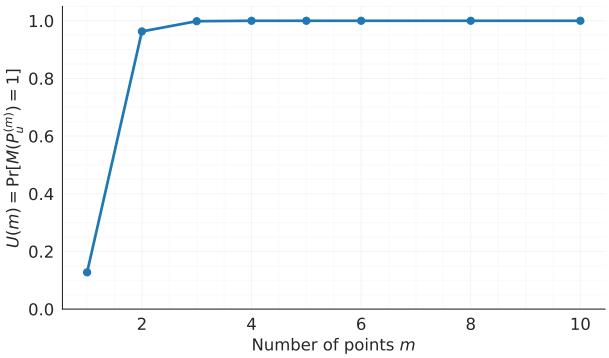



**Figure 8: Trajectory unicity under as a function of the number of known spatio-temporal points $m$.**

*Point-set representation.* Let $\text{Set}(\mathcal{T}_u)$ denote the set of distinct spatio-temporal points visited by $u$:

$$\text{Set}(\mathcal{T}_u) = \{(\ell, \tau) \mid \exists i :\ (\ell_{u,i}, \tau_{u,i}) = (\ell, \tau)\}. \tag{8}$$

*m-point unicity of a user.* For a fixed $m$, sample a subset $P_u^{(m)} \subset \text{Set}(\mathcal{T}_u)$ with $|P_u^{(m)}| = m$. We define the match count

$$M(P_u^{(m)}) = \left|\left\{v \mid P_u^{(m)} \subseteq \text{Set}(\mathcal{T}_v)\right\}\right|. \tag{9}$$

We call $u$ *unique under m points* when $M(P_u^{(m)}) = 1$.

*Population unicity curve.* We define the dataset-level unicity as

$$U(m) = \Pr_{u \sim \mathcal{U},\ P_u^{(m)}} \left[M(P_u^{(m)}) = 1\right], \tag{10}$$

where $\mathcal{U}$ denotes the uniform distribution over users and $P_u^{(m)}$ denotes uniform sampling over $m$-subsets from $\text{Set}(\mathcal{T}_u)$.

$U(m)$ quantifies the re-identification surface directly: a high $U(m)$ means an attacker who learns $m$ check-ins (for example, from posts, receipts, or observed commutes) isolates a user with high probability.

Figure 8 reports the unicity curve, i.e., the probability that a randomly chosen user $u$ becomes uniquely identifiable when an

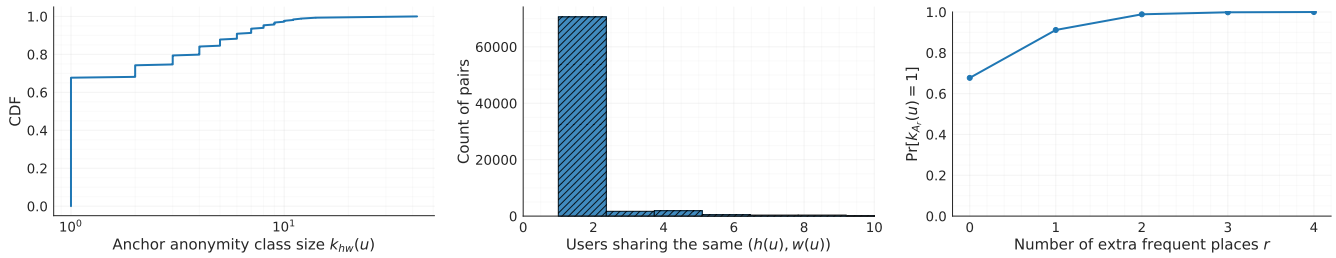



**Figure 9: Anchor-based privacy leakage. (a) Distribution of home–work anonymity class sizes $k_{hw}(u)$, where smaller values indicate users whose home and work pair is shared with few others. (b) Histogram of the number of users sharing each $(h(u), w(u))$ pair, highlighting how many home–work combinations are effectively unique or very sparsely shared. (c) Probability that the full anchor signature $\mathbf{A}_r(u) = (h(u), w(u), p_1(u), \ldots, p_r(u))$ uniquely identifies a user as the attacker learns additional frequent non-home, non-work places.**

adversary knows $m$ sampled spatio-temporal points $P_u^{(m)} \subset \mathrm{Set}(\mathcal{T}_u)$. The curve rises sharply: with only one point, unicity is low (about 0.13), but with two points it jumps to near-certain re-identification (about 0.97), and it essentially saturates to $U(m) \approx 1$ from $m \geq 3$. This indicates that, in this dataset and discretization, very small partial knowledge (only a couple of check-ins) is sufficient to isolate almost any individual trajectory, confirming the strong linkage surface captured by $M(P_u^{(m)})$ and the definition of $U(m)$.

## 6.4 Anchor-based Uniqueness

Home and work anchors drive linkage attacks because they provide stable, repeated signals. After re-identification, these anchors correspond to real neighborhoods and real routines. We therefore define metrics that quantify how strongly anchors isolate users.

*Home and work operators.* Let $\mathcal{T}_u$ contain visits across days. Define two time masks:

$$\mathcal{H}_{\text{home}} = \{\tau \mid \tau \in [22{:}00, 06{:}00)\} \tag{11}$$

$$\mathcal{H}_{\text{work}} = \{\tau \mid \tau \in [09{:}00, 17{:}00)\} \tag{12}$$

Define the home and work cells as

$$h(u) = \arg\max_{\ell \in \mathcal{L}} \#\{i \mid \ell_{u,i} = \ell \wedge \tau_{u,i} \in \mathcal{H}_{\text{home}}\} \tag{13}$$

$$w(u) = \arg\max_{\ell \in \mathcal{L}} \#\{i \mid \ell_{u,i} = \ell \wedge \tau_{u,i} \in \mathcal{H}_{\text{work}}\} \tag{14}$$

*Home-work uniqueness.* We define the equivalence class of a user under anchors as

$$\mathcal{E}_{hw}(u) = \{v \mid h(v) = h(u) \wedge w(v) = w(u)\}, \quad k_{hw}(u) = |\mathcal{E}_{hw}(u)|. \tag{15}$$

We report the *unique-anchor rate*

$$\mathrm{UA}_{hw} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbf{1}[k_{hw}(u) = 1], \tag{16}$$

and we also report the distribution of $k_{hw}(u)$ to show how quickly anonymity collapses as an attacker learns anchors.

*Multi-anchor extension.* Real attackers often know more than home and work. Let $p_1(u), \ldots, p_r(u)$ denote the top-$r$ most visited non-home, non-work cells for user $u$ over the full horizon. We define the anchor signature

$$\mathbf{A}_r(u) = \big(h(u), w(u), p_1(u), \ldots, p_r(u)\big). \tag{17}$$

We define

$$k_{A_r}(u) = |\{v \mid \mathbf{A}_r(v) = \mathbf{A}_r(u)\}|. \tag{18}$$

We report $\Pr[k_{A_r}(u) = 1]$ as a function of $r$. This metric captures a realistic escalation: each additional frequent place acts as an extra key in a linkage attack.

Figure 9 shows that home and work already act as strong quasi-identifiers, and that a few extra anchors almost completely collapse anonymity. From panel 9(a), we find that about $\approx 68\%$ of users have $k_{hw}(u) = 1$ (unique home–work pair), and roughly $\approx 88\%$ satisfy $k_{hw}(u) \leq 5$, so most users share their $(h(u), w(u))$ combination with at most a handful of others. Panel 9(b) confirms that many anchor pairs are highly specific: around $\approx 68\%$ of distinct $(h(u), w(u))$ pairs are used by a single user, and over $\approx 87\%$ by at most three users. The multi-anchor curve in panel 9(c) then shows how this risk escalates as we include more frequent locations: the uniqueness probability jumps from $\Pr[k_{A_0}(u) = 1] \approx 0.68$ (home+work only) to $\approx 0.91$ with one extra place ($r = 1$), exceeds $\approx 0.98$ for $r = 2$, and reaches essentially full uniqueness ($> 0.99$) by $r = 3$–$4$, indicating that knowing just a few stable locations beyond home and work is enough to uniquely pinpoint almost every user.

## 6.5 Seclusion Exposure

Seclusion creates high-risk moments: sparse cells allow easy isolation, and visits to sparse cells often correspond to sensitive places and increase individual re-identification of users. We therefore quantify *how much* of each trajectory occurs in low-support regions.

*Per-user seclusion exposure.* For a threshold $\kappa$ (for example, 1, 3, 10 users per cell), we define the per-user seclusion exposure ratio:

$$\mathrm{SE}_\kappa(u) = \frac{\sum_{i=1}^{n_u} \mathbf{1}\big[s(\ell_{u,i}) \leq \kappa\big]}{n_u}. \tag{19}$$

This metric measures the fraction of a user's samples that fall in cells with at most $\kappa$ unique visitors.

Figure 10 summarizes how much of each trajectory runs through low-support regions. It shows that a considerable number of users visit truly isolated cells: about 2,000 users have $\mathrm{SE} * 1(u) > 0$, rising to roughly 4,000 for $\kappa = 3$ and around 11,000 for $\kappa = 10$, signifying a high individual re-identification threat.

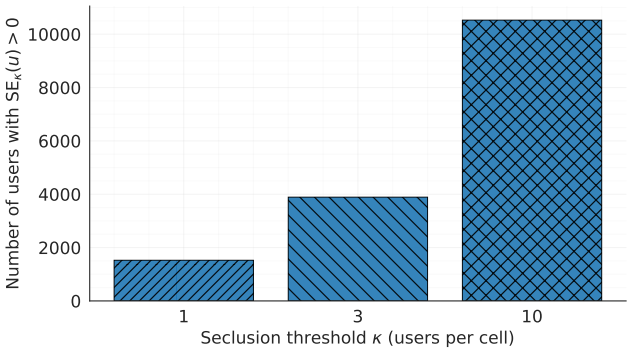



**Figure 10: Users in cells with seclusion thresholds $\kappa$.**

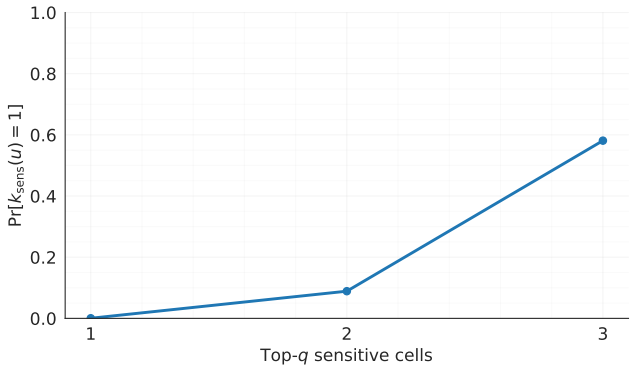



**Figure 11: Sensitive-POI uniqueness. Probability that the top-$q$ sensitive cells $\mathcal{S}_q(u)$ uniquely identify a user.**

## 6.6 Sensitive-POI Disclosure

After re-identification, a cell maps to a physical region. This mapping enables semantic inference because public POI catalogs label categories such as hospitals, places of worship, nightlife, political venues, and support groups (listed in Section B). We quantify risk without performing individual identification by measuring exposure and uniqueness of sensitive visits.

*Sensitive categories.* Let $\mathcal{S} \subseteq \mathcal{L}$ denote the set of cells that intersect sensitive PoIs under a public taxonomy (for example, medical, religious, sexual health, addiction support, or political organizations). Let $C$ denote categories, and let $c(\ell) \in C$ denote the category label of a cell when available.

*Sensitive signature uniqueness.* Attackers often use a small set of rare venues as quasi-identifiers. We define the top-$q$ sensitive cells of user $u$:

$$\mathcal{S}_q(u) = \text{Top-}q \text{ cells in } \{\ell \in \mathcal{S} \mid \ell \text{ appears in } \mathcal{T}_u\} \text{ by visit count.} \tag{20}$$

We define the sensitive uniqueness class size:

$$k_{\text{sens}}(u) = \left|\left\{v \mid \mathcal{S}_q(v) = \mathcal{S}_q(u)\right\}\right|. \tag{21}$$

We report $\Pr[k_{\text{sens}}(u) = 1]$ for small $q$ (for example, $q \in \{1, 2, 3\}$). This metric captures a concrete leakage channel: a small set of sensitive visits uniquely pins down a user even without home and work.

Figure 11 quantifies how often users appear near sensitive venues and how quickly such visits become identifying. It focuses on users who ever visit such cells and tracks the sensitive signature uniqueness $\Pr[k_{\text{sens}}(u) = 1]$ as we reveal the top-$q$ sensitive cells $\mathcal{S}_q(u)$. A single sensitive cell is almost never uniquely identifying (uniqueness well below 1%), but adding one more already makes around 9% of these users unique, and by $q = 3$, more than half of them are uniquely pinned down by their top three sensitive locations alone. Thus, a small set of frequently visited sensitive cells quickly acts as a powerful quasi-identifier.

# 7 Protecting Mobility Datasets like YJMob100K

Having established that the protection measures applied to the dataset are ineffective, we investigate if other sanitization schemes can provide satisfactory protection, specifically in the spatial dimension. We start by applying state-of-the-art schemes that are working at the local level by modifying the individual traces. Then, we consider a more global and destructive approach by removing the spatial structure.

## 7.1 Possible Sanitization Techniques

We evaluate three representative privacy-preserving mechanisms that span different points in the design space for trajectory anonymization: `Geo-Ind` [5] for point-wise differential privacy, Generalized Randomized Response (GRR) [14] for location-level local differential privacy, and spatial de-structuring to eliminate global spatial coherence. We select these techniques for their theoretical foundations, empirical relevance, and practical scalability to large-scale mobility datasets. `Geo-Ind` targets quality-preserving perturbation for local services. GRR targets aggregate analyses such as flow estimation under local privacy. Spatial de-structuring provides an aggressive baseline: it preserves density profiles while discarding spatial continuity.

Recent methods such as PrivTrace [24] aim for trajectory-level privacy using Markov modeling and convex optimization. However, these approaches often impose computational complexity that challenges large-scale reproducibility. In particular, the overall time complexity scales as $O(m|D| + m^3)$, where $m$ denotes the number of spatial cells and $|D|$ denotes the number of trajectories. This complexity arises from discretization, shortest-path estimation, and higher-order transition modeling, and it becomes difficult to scale to datasets like `YJMob100K` with high spatial granularity and very large user counts. In contrast, our selected techniques balance theoretical footing, interpretability, and feasibility for large-scale evaluation.

## 7.2 Geo-Indistinguishability

In this study, we employed Geo-indistinguishability (`Geo-Ind`) [3] to anonymize individual location points within user trajectories. `Geo-Ind` offers strong theoretical guarantees and is widely regarded as the *de facto* standard for location privacy in the absence of a trusted data curator. Its design extends Differential Privacy (DP) [8]

to the spatial domain, ensuring that nearby locations yield statistically indistinguishable outputs, thereby limiting the adversary's ability to infer precise user positions.

Our choice to use Geo-Ind is motivated by its established role in the literature and by recent evidence that re-identification risks remain high in the absence of formal privacy guarantees. However, while Geo-Ind effectively obfuscates individual locations, protecting complete mobility trajectories remains an open challenge. As emphasized in the recent survey by Miranda-Pascual et al. [15], achieving DP for trajectory data is particularly difficult due to the sequential and highly correlated nature of spatiotemporal movements. The authors highlight that existing techniques often struggle to simultaneously ensure privacy and maintain utility and that designing privacy-preserving mechanisms capable of protecting full trajectories is still an unresolved research problem.

More formally, a mechanism satisfies $\epsilon$-Geo-Ind if for any two locations $x_1, x_2 \in \mathbb{R}^2$ within a radius $r$, and any possible output $y \in \mathbb{R}^2$, we have:

$$\frac{\Pr(y|x_1)}{\Pr(y|x_2)} \leq e^{\epsilon r}, \forall r > 0, \forall y, \forall x_1, x_2 : d(x_1, x_2) \leq r.$$

Intuitively, for any point $x_2$ within distance $r$ of $x_1$, Geo-Ind ensures that their output distributions are close-bounded by $l = \epsilon r$. For example, an adversary might distinguish that a user is in Paris rather than London, but may struggle to identify a specific street in Paris, depending on $\epsilon$. Although Geo-Ind and classical DP use $\epsilon$ to denote a privacy budget, they are not directly comparable: in Geo-Ind, $\epsilon$ is scale-dependent and incorporates spatial units (e.g., meters$^{-1}$).

In the continuous plane setting considered in this paper, Geo-Ind can be instantiated using the planar Laplace mechanism [3]. Rather than reporting the true location $x \in \mathbb{R}^2$, the mechanism samples a perturbed location $y \in \mathbb{R}^2$ using the probability density function:

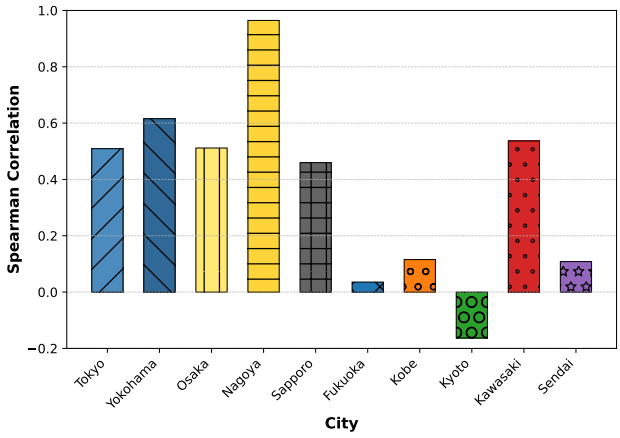
$$D_\epsilon(y) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d(x,y)}.$$

where $d(x, y)$ is the Euclidean distance. The algorithmic implementation of this mechanism is shown in Algorithm 2.

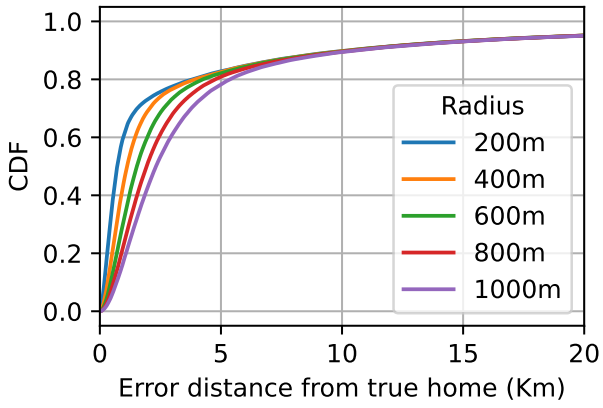
**Algorithm 2** Polar Laplace mechanism in continuous plane [3]

1: **Input :** $\epsilon > 0$, real location $x \in \mathbb{R}^2$.
2: **Output :** sanitized location $y \in \mathbb{R}^2$.
3: Draw $\theta$ uniformly in $[0, 2\pi)$
4: Draw $p$ uniformly in $[0, 1)$
5: Set $r = C_\epsilon^{-1}(p) = -\frac{1}{\epsilon}\left(W_{-1}\left(\frac{p-1}{e}\right) + 1\right)$
6: **Return :** $y = x + \langle r\cos(\theta), r\sin(\theta)\rangle$

**Observations:** We applied Geo-Ind at the level of individual location points to assess whether such point-wise perturbations could meaningfully mitigate re-identification risks in the YJMob100K large-scale mobility dataset. To instantiate the mechanism, we used a privacy parameter $\epsilon$ corresponding to a radius of 1 km, which is considered a reasonably strong privacy guarantee in the context of location-based services [3]. This setting introduces a significant amount of spatial uncertainty, yet, as our experiments reveal, it does not fully obscure high-level structural information.




(a) Clustered correlation



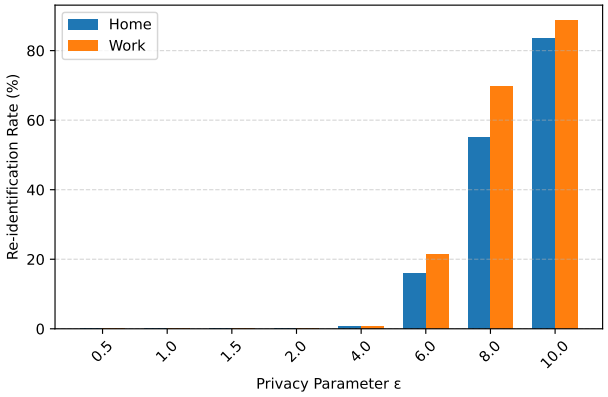

(b) Error in home inference

**Figure 12: (a) Clustered correlation after applying Geo-Ind with a 1 km radius; Nagoya remains the dominant match, which supports location re-identification. (b) Error in home inference when comparing pre- and post-perturbation anchors under Geo-Ind.**

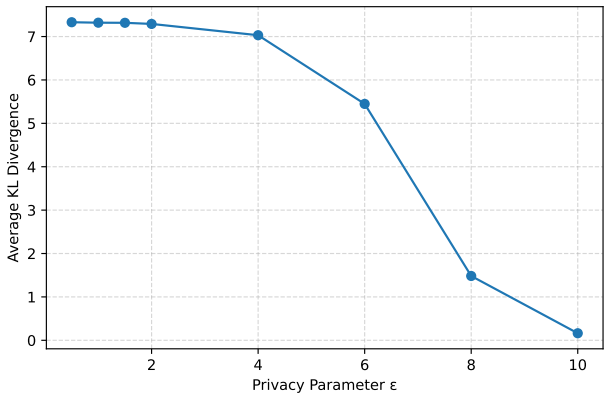
Figure 12(a) presents the results of our clustered correlation analysis after applying Geo-Ind noise. Despite the perturbations, the transformed dataset remains highly correlated with the population distribution and still reveals Nagoya as the recorded city, suggesting that large-scale patterns in user mobility persist and can still be exploited by adversaries.

In Figure 12(b), we further examine the effectiveness of Geo-Ind in obfuscating sensitive points such as home locations. We observe a considerable yet limited increment in the error when varying the privacy parameter by increasing the radius.

This highlights a critical limitation: while Geo-Ind ensures privacy at the level of individual points, the aggregation of these points, especially over repeated visits to the same locations, can leak identifying information. Taken together, these results emphasize that point-wise mechanisms, although theoretically robust, may fall short in practical scenarios involving rich trajectory data.




(a) Re-identification Rate




(b) KL Divergence

**Figure 13: Effect of GRR-based LDP on home–work inference. (a) Home–work re-identification rate vs. $\epsilon$. (b) KL divergence between true and perturbed population distributions.**

## 7.3 Generalized Randomized Response (GRR)

To evaluate the privacy–utility trade-off under a canonical local mechanism, we apply the Generalized Randomized Response (GRR) [14] to protect users' location trajectories. GRR is one of the most widely studied mechanisms in the local differential privacy (LDP) literature: it offers an analytically simple and optimally private channel for categorical data, and it is routinely used as a baseline when comparing LDP mechanisms. In our setting, the city is discretized into spatial grid cells; once each location is mapped to a discrete symbol, GRR becomes a natural choice to stress-test what can be achieved with point-wise LDP on trajectories without relying on a trusted curator.

*Discrete representation and notation.* Let $\mathcal{L}$ denote the set of spatial cells obtained from tessellating `YJMob100K`, and let $|\mathcal{L}| = k$. We index cells by integers and write

$$\phi : \mathcal{L} \to \{1, \dots, k\}, \quad \ell_{u,i} \in \mathcal{L}, \quad z_{u,i} = \phi(\ell_{u,i}) \in \{1, \dots, k\},$$

where $\ell_{u,i}$ is the $i$-th true location of user $u$ and $z_{u,i}$ its discrete index. GRR acts on this discrete domain $\mathcal{X} = \{1, \dots, k\}$: for a true symbol $x \in \mathcal{X}$, the mechanism outputs a sanitized symbol $\tilde{x} \in \mathcal{X}$.

*GRR mechanism and privacy guarantee.* Fix a privacy parameter $\epsilon > 0$. GRR specifies, for each true value $x \in \{1, \dots, k\}$, the distribution of the reported value $\tilde{x}$:

$$\Pr[\tilde{x} = v \mid x] = \begin{cases} p = \dfrac{e^{\epsilon}}{e^{\epsilon} + k - 1}, & \text{if } v = x, \\ q = \dfrac{1}{e^{\epsilon} + k - 1}, & \text{if } v \neq x. \end{cases}$$

In words, the mechanism preserves the true cell with probability $p$, and with the remaining probability $1 - p$ it outputs a value drawn uniformly from the $k-1$ other cells. This channel is $\epsilon$-locally differentially private: for any two inputs $x, x' \in \{1, \dots, k\}$ and any output $v$,

$$\frac{\Pr[\tilde{x} = v \mid x]}{\Pr[\tilde{x} = v \mid x']} \leq e^{\epsilon}.$$

Indeed, the largest ratio occurs between *in-channel* and *out-of-channel* probabilities, and we have $p/q = e^{\epsilon}$ by construction. Thus, each reported cell $\tilde{x}$ hides its true origin among all $k$ possibilities, with indistinguishability controlled by $\epsilon$.

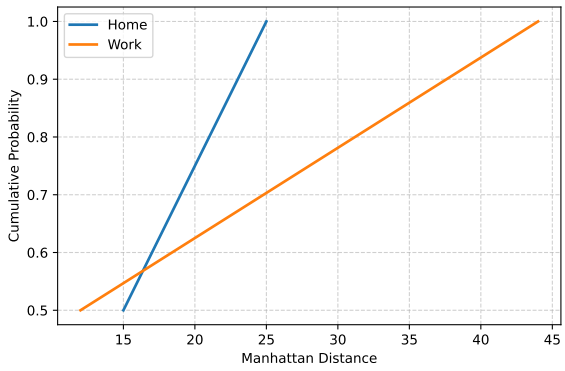
*Application to trajectories.* We apply GRR independently to every point in each user's trajectory. For each sample $(\ell_{u,i}, t_{u,i})$, we compute its discrete index $z_{u,i} = \phi(\ell_{u,i})$, run GRR with input $z_{u,i}$, and obtain a sanitized index $\tilde{z}_{u,i}$. Mapping $\tilde{z}_{u,i}$ back through $\phi^{-1}$ yields a perturbed cell $\tilde{\ell}_{u,i}$. This step produces a synthetic trajectory

$$\widetilde{\mathcal{T}}_u = \{(\tilde{\ell}_{u,i}, t_{u,i})\}_i$$

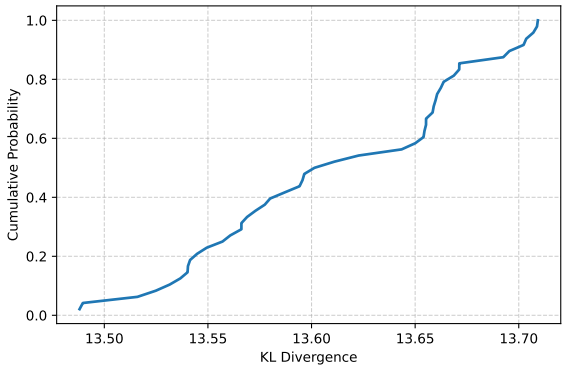
for each user $u$, in which every reported location is protected by $\epsilon$-LDP at the point level.

*Evaluation methodology.* We generate GRR-perturbed datasets for a range of privacy budgets $\epsilon$ and assess both privacy and utility. On the privacy side, we re-run our home/work inference pipeline directly on $\widetilde{\mathcal{T}}_u$ and measure the re-identification rate of anchors ($h(u)$ and $w(u)$) as a function of $\epsilon$, yielding the curve in Figure 13(a). On the utility side, we estimate population-level spatial distributions at each time slot. Let $f_\tau \in \Delta^{k-1}$ denote the true frequency vector over cells at time $\tau$ and $\tilde{f}_\tau$ the empirical distribution obtained from GRR outputs. Since GRR is a linear channel with known parameters $(p, q)$, we apply the standard debiasing estimator for GRR [14] to recover an estimate $\widehat{f}_\tau$ of $f_\tau$, and we quantify distortion via the Kullback–Leibler divergence $\mathrm{KL}(f_\tau \,\|\, \widehat{f}_\tau)$. Figure 13(b) reports the average KL divergence across time.

**Observations:** Figure 13 summarizes the trade-off induced by GRR. For small privacy budgets (e.g., $\epsilon \leq 2$), the home–work re-identification rate in Figure 13(a) decreases substantially, indicating that anchor linkage becomes difficult for the attacker; however, Figure 13(b) shows that the average KL divergence between true and reconstructed population distributions increases sharply, signaling severe loss of utility for aggregate analyses. As $\epsilon$ grows (e.g., $\epsilon \geq 6$), GRR preserves most of the population structure (low KL divergence), but the re-identification rate climbs back toward the original level, meaning that anchor privacy is largely lost. Overall, GRR illustrates a fundamental limitation of point-wise LDP on rich mobility traces: either we choose $\epsilon$ small enough to have meaningful protection but destroy most analytic value, or we choose $\epsilon$ large enough to preserve utility but provide little effective defense against trajectory-based re-identification attacks.




(a) Home and Work Location Error




(b) Utility Loss

**Figure 14: Spatial de-structuring effects. (a) Home/work inference error. (b) KL divergence after permutation.**

## 7.4 Spatial De-Structuring

We finally consider an extreme sanitization mechanism designed to probe how much re-identification power remains when spatial structure is deliberately destroyed. Prior work, including Section 4 and [18], shows that the *shape* of spatial density: how activity concentrates across space, often uniquely characterizes a city. To counter this effect, we apply a *spatial de-structuring* transformation in which we randomly permute the grid-cell indices for each user. This permutation preserves, for every cell, its marginal visit frequency and the temporal usage pattern, but breaks the geographic layout: neighbor relationships and large-scale spatial geometry are no longer meaningful.

**Observations:** After de-structuring, the Spearman correlation between clustered user density and census data drops substantially to 40%, confirming that spatial coherence has largely collapsed. However, coarse density signals remain: Nagoya still emerges as the top candidate city, with a correlation margin exceeding 10% over any other city under consideration. Figure 14(a) shows the CDF of home/work inference error after permutation, revealing a strong degradation in anchor accuracy, while Figure 14(b) reports very large KL divergence between true and estimated population distributions, indicating extreme utility loss. Thus, even when we erase spatial order to an unrealistic degree, residual density patterns still support meaningful city-level inference, but only at the cost of rendering the data nearly unusable for most practical analyses.

## 7.5 Takeaways

Our experiments with various sanitization methodologies collectively paint a stark picture: current mechanisms are fundamentally ill-equipped to protect rich mobility trajectories without inflicting severe damage on utility. `Geo-Ind` offers rigorous guarantees for isolated point queries, yet our trajectory-level re-identification results show that correlated movements, repeated visits, and stable anchors readily survive point-wise noise, allowing an attacker to recover city structure and infer home and work locations. GRR, in turn, provides formal local differential privacy over discrete grid cells, but our evaluation demonstrates that suppressing home–work re-identification requires very small $\epsilon$ values, under which KL divergence explodes, and population-level statistics become unusable. At the other extreme, spatial de-structuring aggressively destroys geographic coherence: while it severely harms anchor inference and downstream analytics, residual density patterns still suffice to single out Nagoya at the city scale.

These findings underscore that trajectory re-identification is not a corner case but a structural vulnerability of mobility data: even after applying state-of-the-art point-wise mechanisms or extreme spatial transformations, our attacks continue to extract meaningful information. In other words, our trajectory re-identification study raises a clear alarm that existing anonymization techniques are misaligned with the complexity of real-world traces. Designing mechanisms that suppress structural patterns: recurrent routes, anchor constellations, and group co-location; while preserving high-resolution spatio-temporal utility remains an open and pressing research problem. Developing such trajectory-aware privacy protections is, in our view, a necessary next step for any realistic deployment of large-scale mobility datasets.

## 8 Conclusion

This study demonstrates that current anonymization practices for mobility data are dangerously inadequate. Using `YJMob100K` [26] as a case study, we show that a structure-aware adversary can recover both the underlying geography and real-world timeline, enabling direct re-identification of sensitive patterns such as presence in sparse regions and large-scale uniqueness of home–work anchors. Despite coordinate shifting and timestamp normalization, users remain exposed to location inference, anchor profiling, and downstream linkage through public auxiliary data. These results are a strong warning: as long as structural information in space and time is left intact, "anonymized" trajectories can still be traced back to individuals. There is an urgent need for new privacy mechanisms that explicitly target this structural leakage, and trajectory re-identification should now be treated as a central threat model for the design of future spatio-temporal data releases.

## 9 Responsible Disclosure

The vulnerabilities uncovered in this paper show that users included in the public `YJMob100K` dataset face concrete risks of privacy compromise. Prior to conducting our analysis, we obtained

approval from the ethical and legal committee of our institution. Upon confirming our findings, we notified the authors responsible for publishing the YJMob100K dataset and shared a technical summary of the identified vulnerabilities. We also disclosed the issues to the relevant Data Protection Authorities, including the Japanese *Personal Information Protection Commission* [17], to inform them of their potential impact.

## References


[1] 2018. Japan: High Resolution Population Density Maps + Demographic Estimates. Available online: https://data.humdata.org/dataset/japan-high-resolution-population-density-maps-demographic-estimates (accessed on 3 June 2025).

[2] Gergely Acs and Claude Castelluccia. 2014. A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris. In *ACM SIGKDD (KDD '14)*. ACM, New York, NY, USA, 1679–1688. doi:10.1145/2623330.2623361

[3] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *ACM CCS* (Berlin, Germany) *(CCS '13)*. 901–914. doi:10.1145/2508859.2516735

[4] BBC News. 2019. Typhoon Hagibis: Japan suffers deadly floods and landslides from storm. https://www.bbc.com/news/world-asia-50020108. Accessed: 2025-06-03.

[5] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *PETS*. Springer, 82–102. doi:10.1007/978-3-642-39077-7_5

[6] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T. Silva. 2016. Anonymizing NYC Taxi Data: Does It Matter?. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, 140–148. http://ieeexplore.ieee.org/abstract/document/7796899/

[7] Kostas Drakonakis, Panagiotis Ilia, Sotiris Ioannidis, and Jason Polakis. 2019. Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta)Data. doi:10.48550/arXiv.1901.00897 arXiv:1901.00897 [cs].

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.

[9] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. 2019. Privacy in trajectory micro-data publishing: a survey. *arXiv preprint arXiv:1903.12211* (2019).

[10] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)* 42, 4 (2010), 1–53.

[11] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2014. De-anonymization attack on geolocated data. *J. Comput. System Sci.* 80, 8 (2014), 1597–1614.

[12] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS* (San Jose, California) *(SPRINGL '10)*. Association for Computing Machinery, New York, NY, USA, 34–41. doi:10.1145/1868470.1868479

[13] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779–782.

[14] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. 2016. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*. PMLR, 2436–2444.

[15] Alex Miranda-Pascual, Patricia Guerra-Balboa, Javier Parra-Arnau, Jordi Forné, and Thorsten Strufe. 2023. SoK: Differentially Private Publication of Trajectory Data. *Proceedings on Privacy Enhancing Technologies* 2023, 2 (April 2023), 496–516. doi:10.56553/popets-2023-0065

[16] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3 (March 2013), 1376. doi:10.1038/srep01376

[17] Personal Information Protection Commission Japan. 2025. Personal Information Protection Commission (PPC). https://www.ppc.go.jp/en/. Accessed: 2025-06-03.

[18] Gergő Pintér. 2024. Revealing urban area from mobile positioning data. *Scientific Reports* 14, 1 (Dec. 2024). doi:10.1038/s41598-024-82006-5

[19] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. 2019. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2772–2793. doi:10.1109/COMST.2018.2873950

[20] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. What Does The Crowd Say About You? Evaluating Aggregation-based Location Privacy. *Proceedings on Privacy Enhancing Technologies* 2017 (2017), 156 – 176.

[21] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. Knock knock, who's there? Membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145* (2017).

[22] Daniele Quercia, Ilias Leontiadis, Liam McNamara, Cecilia Mascolo, and Jon Crowcroft. 2011. SpotME If You Can: Randomized Responses for Location Obfuscation on Mobile Phones. In *2011 31st International Conference on Distributed Computing Systems*. 363–372. doi:10.1109/ICDCS.2011.79

[23] Time and Date AS. 2021. Japanese Holidays Calendar. https://www.timeanddate.com/calendar/custom.html?year=2021&country=26&cols=3&hol=9&df=1. Accessed: 2025-06-03.

[24] Haiming Wang, Zhikun Zhang, Tianhao Wang, Shibo He, Michael Backes, Jiming Chen, and Yang Zhang. 2023. PrivTrace: Differentially Private Trajectory Synthesis by Adaptive Markov Models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 1649–1666.

[25] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1241–1250.

[26] Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data* 11, 1 (2024), 397.

[27] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. 2015. You Are Where You Go: Inferring Demographic Attributes from Location Checkins. In *ACM WSDM*. ACM, Shanghai China, 295–304. doi:10.1145/2684822.2685287


# A Events Hosted at the Exhibition Center

**Table 2: Events hosted at Port Messe exhibition center between 15/9/2019 and 28/11/2019.**

| Event Name | Dates | Attendance |
|---|---|---|
| *Automotive World Nagoya* 2019[a] | September 18–20 (Day 3–5) | unknown |
| *MECT2019*[b] | October 23–26 (Day 38–41) | 90,000 |
| *Messe Nagoya 2019*[c] | November 6–9 (Day 52–55) | 60,000 |
| *One OK Rock Concert*[d] | November 12–13 (Day 58–59) | unknown |
| Mammoth Flea Market Z vol. 61[e] | November 16–17 (Day 62–63) | unknown |
| *Nagoya Motor Show 2019*[f] | November 21–24 (Day 67–70) | 205,900 |

[a] https://www.automotiveworld-nagoya.jp/en-gb/visit/visitorsguide-dl.html
[b] https://mect-japan.com/2019/en/about/
[c] https://www.nic-nagoya.or.jp/en/nagoya-calendar/assets_c/nagoya_calendar_2019_10.pdf
[d] https://www.japanconcerttickets.com/one-ok-rock-2019-2020/
[e] https://www.nic-nagoya.or.jp/en/nagoya-calendar/assets_c/nagoya_calendar_2019_11.pdf
[f] https://www.jetro.go.jp/en/database/j-messe/tradefair/detail/104782

## B Sensitive POIs

**Table 3: Sensitive POI keyword used to flag cells as sensitive.**

| Keyword substring | Keyword substring |
|---|---|
| Hospital | Addiction |
| Clinic | Rehab |
| Medical | Counsel |
| Pharmacy | Therapy |
| Relig | Politic |
| Church | Party |
| Mosque | Campaign |
| Temple | Adult |
| Shrine | Strip |
| Nightclub | |