Breaking Anonymity at Scale: Re-identifying the Trajectories of 100K Real Users in Japan

Abhishek Kumar Mishra[1,*], **Mathieu Cunche**[1,*], **and Héber H. Arcolezi**[2,*]

[1]INSA-Lyon, Lab. CITI, Inria, Team Privatics, Lyon, France
[2]Inria, Team Privatics, Grenoble, France
[*]Email: {abhishek.mishra, mathieu.cunche, heber.hwang-arcolezi}@inria.fr

ABSTRACT

Mobility traces represent a critical class of personal data, often subjected to privacy-preserving transformations before public release. In this study, we analyze the anonymized `Yjmob100k` dataset, which captures the trajectories of 100,000 users in Japan, and demonstrates how existing anonymization techniques fail to protect their sensitive attributes. We leverage population density patterns, structural correlations, and temporal activity profiles to re-identify the dataset's real-world location and timing. Our results reveal that the anonymization process carried out for `Yjmob100k` is inefficient, and preserves enough spatial and temporal structure to enable re-identification. This work underscores the limitations of current trajectory anonymization methods and calls for more robust privacy mechanisms in the publication of mobility data.

1 Introduction

Mobility data drives advances in urban planning, transportation design, and epidemiological modeling. However, preserving privacy in such datasets remains challenging due to the intrinsic uniqueness of human movement patterns[1]. Even with pseudonymization, adversaries often re-identify individuals by linking spatio-temporal traces to external information[2].

This study analyzes the anonymization strategy behind `Yjmob100k`[3], a large-scale mobility dataset containing GPS trajectories from 100,000 users in Japan. The dataset replaces absolute timestamps and geographic coordinates with values relative to unknown reference points. Despite this transformation, it retains structural signals that reveal sensitive information.

Large-scale datasets like `Yjmob100k` aim to support innovation and research while preserving privacy. However, their anonymization techniques rarely undergo systematic evaluation at scale. Prior research[4] shows that even sparse or coarse-grained location traces enable re-identification when combined with auxiliary data. In `Yjmob100k`, anonymization shifts spatial and temporal references but preserves movement structure.

This work demonstrates how adversaries recover key attributes from the dataset. By aligning spatial densities, comparing urban structure, and profiling temporal behavior, we re-identify the hidden city of origin, infer actual timestamps, and link users to sensitive locations such as homes and workplaces.

These results expose fundamental weaknesses in widely used anonymization approaches. Mobility datasets often leak identity-related patterns, even when stripped of direct identifiers. As spatio-temporal data shapes AI systems and public health models, the need for effective privacy mechanisms becomes urgent. This study advocates for stronger anonymization standards and careful anonymization of mobility data releases to protect individual privacy without sacrificing data utility.

2 Related Work

Pinter et al.[5] analyze the `Yjmob100k` dataset and apply a similar spatial matching approach to infer its geographic origin, but do not reveal the fine-tuned spatial coordinates. Furthermore, their analysis does not explore temporal re-identification. In addition, they do not give insight into the capabilities of various protection strategies in the literature for obfuscating the `Yjmob100k` dataset.

Researchers consistently identify location data as one of the most privacy-sensitive data types. Location traces enable inference of personal attributes[6,7] and often allow adversaries to single out individuals[2]. This makes anonymization of mobility data particularly complex.

The field of privacy-preserving data publishing proposes several strategies to mitigate these risks[8]. While privacy-preserving data publishing methods exist for generic data, extending them to location data[9] and mobility traces[10] remains difficult. Studies demonstrate that even anonymized mobility datasets often remain vulnerable to re-identification attacks[4]. These results suggest that strong privacy guarantees typically require substantial sacrifices in utility (i.e. the practical value of the resulting data).

(a) Raw dataset (b) City of Nagoya (c) Transformed dataset, representing city Nagoya

Figure 1. Visual comparison of population densities in original vs. re-identified version. (a) Grid-based trajectories in the released `Yjmob100k` dataset exhibit structured mobility aligned with urban infrastructure. (b) Actual map of the city of Nagoya. (c) The same dataset after re-identification, revealing the city of Nagoya, as described in Algorithm 1

This work builds on that foundation by showing that naive anonymization of location data–such as geometric transformation and time shifting–does not provide sufficient protection. Our results reinforce the broader consensus that mobility data demands tailored, rigorous privacy mechanisms to prevent meaningful leakage.

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 further provides anonymized Points of Interest (PoIs), representing frequently visited locations inferred from user mobility patterns.

Figure 1(a) presents a spatial overview of the grid-based movement trajectories. The observed patterns reveal structured mobility within an urban context, with dense clusters and directional flows reflecting transportation infrastructure and metropolitan dynamics. Despite anonymization, the preserved spatial organization–manifested in recurrent pathways and grid connectivity–mirrors real-world geographies, including transit corridors and high-density regions. Prominent clusters in the distribution suggest the presence of key landmarks or PoIs, which can act as spatial anchors in subsequent re-identification analysis.

4 Re-identifying the spatial dimension

In this section, we leverage the spatial distribution of the records in the dataset in order to re-identify the real location of the data. We leverage the density of points in the dataset that should be similar to the population distribution in the real location.

4.1 Population density as a re-identification metric

We investigate the potential to re-identify the true geographic origin of the `Yjmob100k` dataset by leveraging the spatial distribution of mobility records. Our central hypothesis is that the anonymized dataset retains structural properties closely aligned with actual population distributions.

For the `Yjmob100k` dataset, we compute the spatial population density as:

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

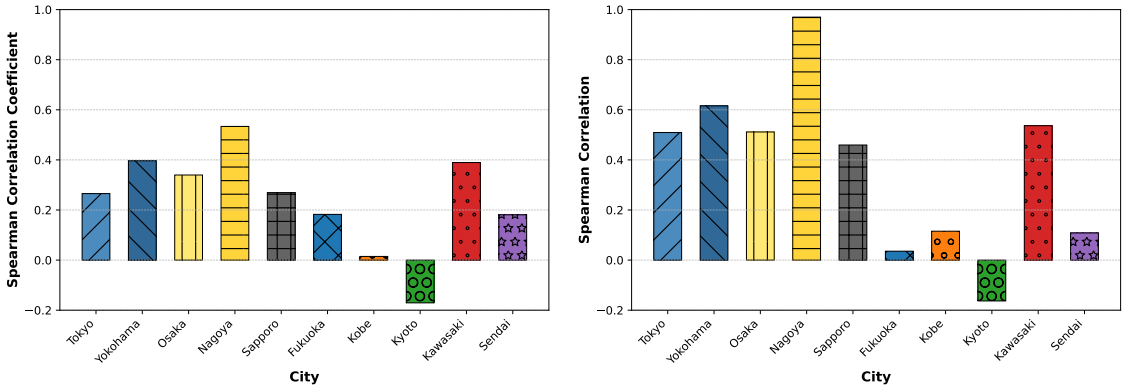
where $D(d)$ denotes the aggregated density on day d, $P_g(d)$ the point count in grid cell g, and G the set of spatially selected grid cells. For comparison, we utilize public population density data[11] of Japan's ten largest cities. We observe that these distributions exhibit well-defined urban structures, characterized by dense urban cores and tapering suburban areas.

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_{raw}, a set of public population datasets D_{public} corresponding to Japan's ten largest cities, and

Algorithm 1 Re-Identification Algorithm

Input: D_{raw}, D_{public}, T
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)



(a) Individual grid correlation (b) Cluster-aggregated correlation

Figure 2. Spearman correlation coefficients between the anonymized dataset and population data across 10 major Japanese cities. Nagoya clearly emerges as the most probable location.

a transformation set T. The goal is to identify the most likely city of origin C^* and the transformation T^* that maximize spatial similarity, with all intermediate correlation values stored in a dictionary S.

We define a set of geometric transformations that account for potential obfuscation applied to D_{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 calculate this correlation over clustered spatial regions, dividing the full 200×200 grid into 25 non-overlapping clusters of size 40×40. We store the resulting scores $C_i(j)$ in the dictionary S, indexed by the transformation T_i and city j.

We identify the best matching configuration by selecting the pair $(C^*, T^*) = \arg\max_{(T_i,j)} C_i(j)$ that yields the highest correlation. The final output includes the re-identified city C^* and the optimal transformation T^*.

Figure 2(a) illustrates the Spearman correlations between the transformed dataset and the public population data. The city of Nagoya consistently yields the highest score, suggesting a close match with the anonymized data. While minor discrepancies persist due to sampling noise, the overall structure remains visualy recognizable. The clustering-based strategy enhances correlation stability across cities. As shown in Figure 2(b), Nagoya's correlation approaches unity under clustering, indicating a near-perfect match.

Figure 1(c) visualizes the anonymized dataset post-transformation ($R_{90}(F_y(D))$), corresponding to the optimal match. These transformations obscure spatial orientation but preserve intrinsic movement patterns, such as cluster distributions and

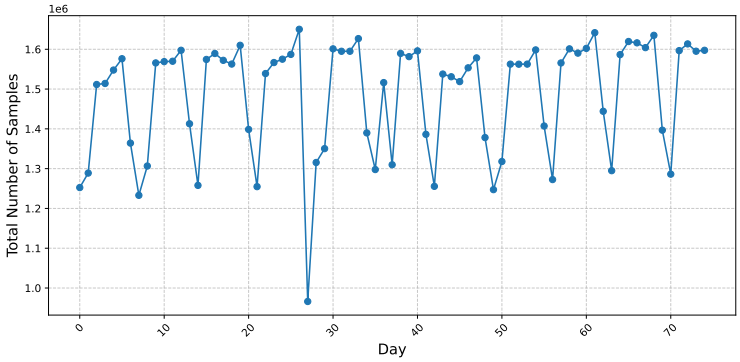


Figure 3. Overall activity.

displacement ranges. By inverting the transformations, we recover structural alignment with known geographies.

4.3 Fine-grained spatial re-identification

To refine the spatial alignment after identifying Nagoya as the most probable city, we perform a fine-tuning step using hill climbing over a 100 km × 100 km area centered around Nagoya's coordinates. The goal is to determine the optimal latitude-longitude alignment that maximizes Spearman correlation between the public population map and the anonymized dataset. Starting from the initial center $(35.05, 136.96)$, the algorithm iteratively perturbs the center by small steps $(0.01°)$ and retains the configuration yielding the highest correlation. The final correlation score stabilizes at the overall grid center $(35.055019, 136.971202)$ when no further improvement is observed. This procedure recovers the exact geocoordinates corresponding to the anonymized grid cells in `Yjmob100k`, completing the spatial re-identification process with meter-level precision. The recovered alignment, combined with the robust clustered correlation approaching 1.0 for Nagoya (Figure 2), confirms a highly reliable spatial re-identification. While Pinter et al.[5] previously hinted at Nagoya as the underlying city, our method completes the re-identification by uncovering exact locations and validating the match using a fine-grained optimization.

5 Re-identifying the temporal dimension

This section re-identifies the temporal dimension of the `Yjmob100k` dataset by uncovering the exact calendar dates corresponding to its 75 consecutive days of mobility traces. Rather than treating time as abstract indices, this analysis grounds the data in real-world chronology using large-scale behavioral patterns.

Focusing on the urban area of Nagoya, the temporal dynamics of aggregated user activity across all cells, shown in Figure 3, exhibit prominent anomalies and regularities. These signals have unique temporal signatures, including workdays, weekends, and major events. By aligning these patterns with publicly known holidays and disruptions in Japan, we infer the dataset's time frame. The dataset likely ends before the official publication date[3], assumed here to be no later than April 18, 2024. To narrow the search space, the analysis bounds the starting date after January 1, 2015. Within this interval, only one candidate sequence matches both statistical regularity and exceptional deviations observed in the data, confirming the temporal re-identification.

5.1 Identifying weekdays from activity patterns in residential areas

Figure 3 displays a strong 7-day periodicity in global mobility, characteristic of a weekly cycle alternating between workdays and non-working days. Several deviations from this regular pattern suggest public holidays embedded within the sequence.

To strengthen the identification of day types, we examine user activity in the top 10 residential areas (cell indices listed in the Appendix). We cluster daily activity profiles using K-Means (with $K = 2$ and a random initialization seed), after scaling the 48-bin time series using z-score normalization. Each day is represented by a 48-dimensional vector of unique user counts per half-hour, and clustering is based on Euclidean distance in this normalized space. Figure 4 reveals two distinct behavioral patterns. Class A days show smooth morning increases and evening declines, consistent with relaxed or irregular routines typically observed on weekends or holidays. In contrast, Class B days present sharp morning and evening peaks, aligning with daily commutes and work schedules.

Day 27 stands out with an unusually low activity level and aligns clearly with Class A, indicating a potential non-working day. The full sequence of labeled days, shown in Table 1, follows a regular pattern of five consecutive Class B days followed by two Class A days. This structure confirms a standard Monday–Friday workweek, with weekends occurring every seventh day.

Based on this analysis, we determine that the dataset begins on a Sunday (Day 0), enabling precise alignment between each day index and its corresponding weekday. This temporal structure is critical for downstream analyses, including event

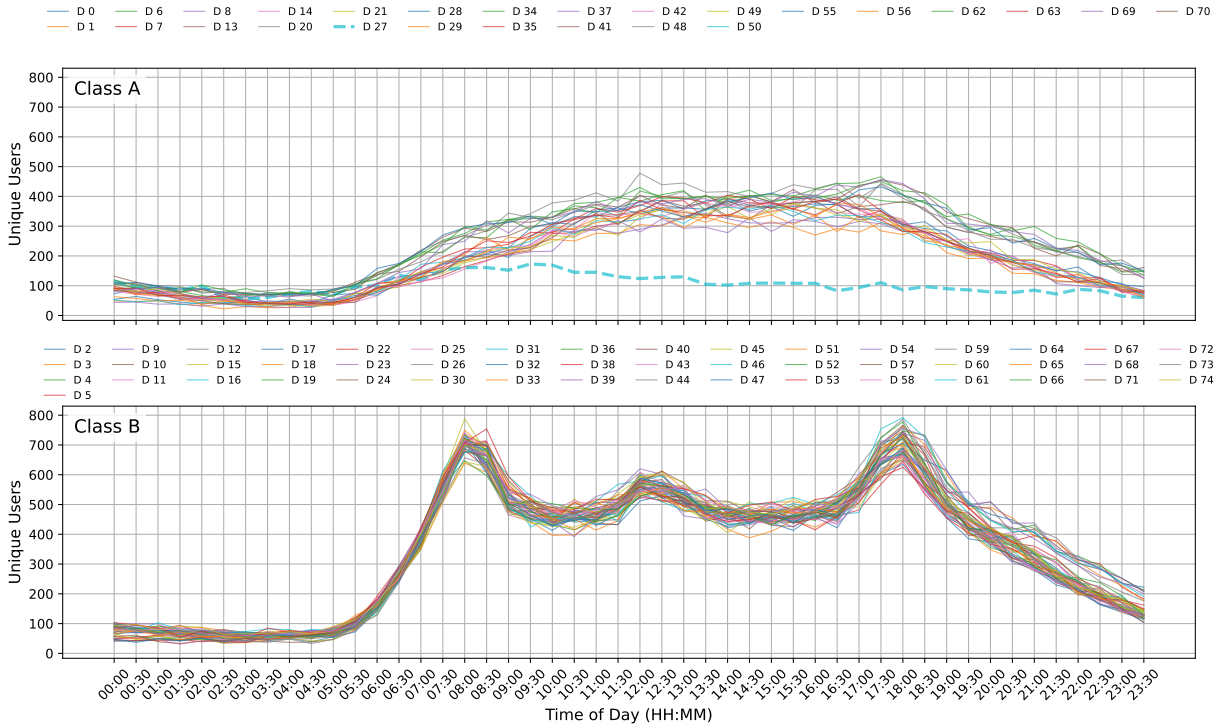


Figure 4. Daily temporal activity in top 10 residential areas organized in two classes obtained from a 2-class clusterisation.

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

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, $\mathscr{H}$, are highlighted in yellow.

localization and public holiday matching.

5.2 Identifying the exact date from public holidays

Building on the weekday classification, we now identify the exact calendar dates corresponding to the dataset by matching detected non-working days with Japan's public holiday calendar. In addition to regular weekends, we observe five non-working days that fall outside the typical Saturday–Sunday pattern: Day 0 (Monday), Day 8 (Monday), Day 29 (Monday), Day 37 (Tuesday), and Day 50 (Monday). We denote this set of suspected public holidays as $\mathscr{H}$ (highlighted in Table 1).

To pinpoint the dataset's timeline, we use the unique occurrence of Day 37—a Tuesday non-working day—as an anchor. We hypothesize that this day aligns with a known national holiday and search for matching sequences within the 2015–2024 range using the official Japanese holiday calendar[12]. Starting from each potential Tuesday holiday, we iteratively test whether

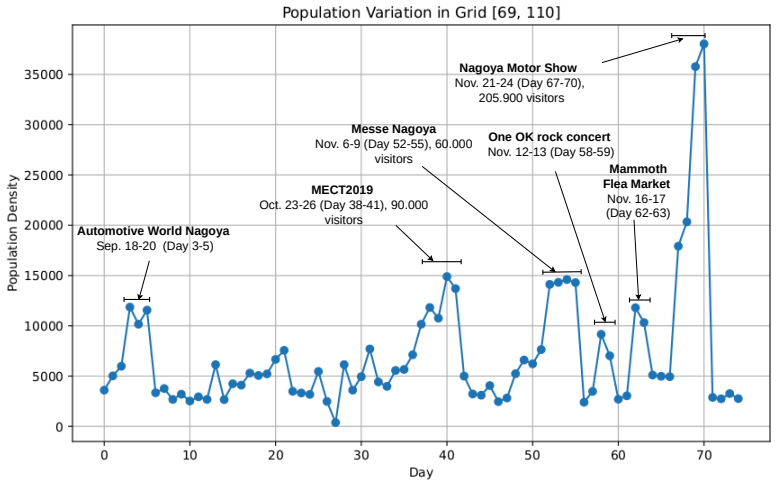


Figure 5. Activity at Port Messe Nagoya (grid 69,110; GPS 35.05016252379402, 137.39573944062153) with hosted events.

the four remaining dates in $\mathscr{H}$ align with known holidays.

Only one sequence satisfies all constraints: the dataset begins on 15 September 2019 (Day 0). Under this alignment, the five holidays in $\mathscr{H}$ match exactly with Japan's national holidays: *Respect for the Aged Day* (16/09/2019 - Day 1), *Autumn Equinox* (23/09/2019 - Day 9), *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 matching provides strong evidence that Day 0 corresponds to 15/09/2019, allowing us to assign calendar dates to the entire 75-day sequence in the `Yjmob100k` dataset.

5.3 Further confirmation of the identified timeline

We further validate our temporal re-identification hypothesis by presenting real-world events in Nagoya that align with distinct patterns observed in the dataset. These events—ranging from anomalous disruptions to major gatherings—provide strong corroborative evidence that Day 0 in the dataset corresponds to 15 September 2019.

5.3.1 Activity anomaly on Day 27: Typhoon Hagibis

A striking anomaly is observed on Day 27 (a Saturday), where overall activity across the entire region sharply declines, as shown in Figure 3. This decrease is even more pronounced in the top 10 residential areas, where the activity on Day 27 is the lowest in the 75-day sequence (Figure 4).

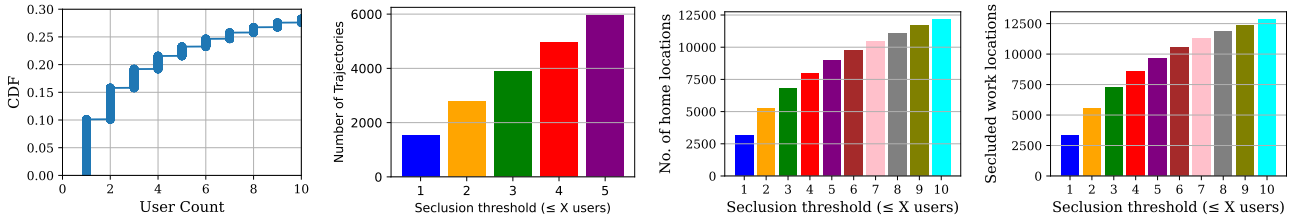
According to our proposed alignment, Day 27 corresponds to 12 October 2019—the day Typhoon Hagibis made landfall in Japan. This typhoon led to widespread disruptions in transportation and public services across the country, including the Nagoya region[13]. The temporal and geographic correspondence between this significant real-world event and the observed data anomaly strongly reinforces the correctness of our identified timeline.

5.3.2 Event correlation at Port Messe Nagoya

Port Messe Nagoya, one of the city's largest exhibition venues, provides another temporal anchor. Activity levels at this specific location (see Figure 5) exhibit multiple peaks during the dataset period. Each peak aligns precisely with the date of a major public event hosted at the venue. Cross-referencing with publicly available event calendars, we identify six events held at Port Messe during the proposed time frame, all of which coincide with surges in user presence at the site (Table 2).

Furthermore, the magnitude of each peak correlates with the event's expected attendance. For example, the highest peak in the dataset corresponds to the *Nagoya Motor Show*, which attracted approximately 205,900 visitors, confirming a robust association between the dataset signals and actual footfall.

Taken together, these converging lines of evidence—including public holidays, anomalous low-activity days due to natural disasters, and location-specific activity tied to scheduled events—offer compelling validation of the reconstructed timeline. We confidently conclude that the dataset begins on 15 September 2019 and that our approach successfully re-identifies the exact dates embedded in the anonymized spatiotemporal traces.



(a) Presence in secluded grids (b) Trajectories (c) Home locations (d) Work locations

Figure 6. Sensitive user attributes with varying seclusion thresholds.

6 What Are the Privacy Risks for Users in This Dataset?

Having assessed that the implemented protection measures for `Yjmob100k` can be reversed, we want to study the potential threat to the user privacy of the dataset with actual date and location. We assess the privacy impact of our spatio-temporal re-identification by quantifying individual-level risks using standard metrics. In particular, we evaluate user exposure in low-density areas and the uniqueness of behavioral patterns such as home-work commutes.

Figure 6(a) shows the distribution of users across secluded grids–defined as grid cells with very few unique individuals. Approximately 20% of grids contain three or fewer users, and 10% are visited by only one user, signaling high re-identification potential. In Figure 6(b), we plot the number of distinct user trajectories that pass through such areas. Even at a threshold of just three users per grid, over 4,000 trajectories intersect with these secluded regions, indicating frequent exposure to privacy-sensitive zones.

To examine anchor point leakage, we infer home and work locations using a classical approach[14]: we identify each user's most frequently visited grid during nighttime (10:00 PM–6:00 AM) and working hours (9:00 AM–5:00 PM), respectively. Among the 100,000 users, 67,342 have unique home-work location pairs, making them vulnerable to linkage attacks. Figures 6(c) and (d) display the distribution of these anchor points relative to grid seclusion levels. Approximately 3,000 users have completely unique home and work locations. When we relax the threshold to 10 users per grid, this number rises to about 12,000, showing that a significant portion of users anchor their mobility in sparsely populated regions.

To ensure ethical compliance, we refrain from deeper re-identification or individual tracing. Prior work[2] highlights how such analyses risk crossing ethical boundaries, even when technically feasible.

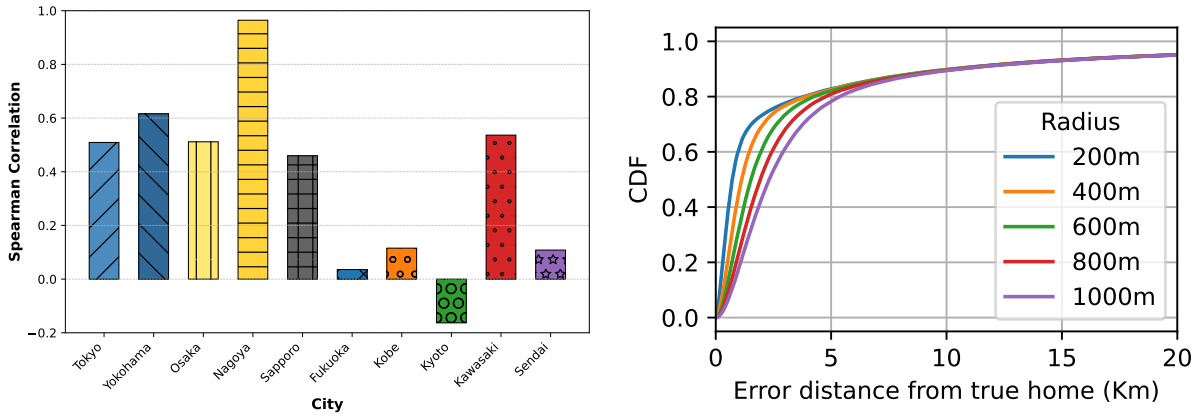
7 How to protect mobility datasets like `Yjmob100k`?

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

In this study, we evaluate three representative privacy-preserving mechanisms that span different points in the trajectory anonymization design space: `Geo-Ind`[15] for point-wise differential privacy, Generalized Randomized Response (GRR)[16] for location-level local differential privacy, and spatial de-structuring to eliminate global spatial coherence. These techniques were selected for their theoretical foundations, empirical relevance in the literature, and practical scalability to large-scale mobility datasets. `Geo-Ind` enables quality-preserving location perturbation, suitable for applications such as local recommendations and proximity-based services. GRR offers a balance between privacy and statistical utility, making it appropriate for aggregate analyses like population flow estimation. Spatial de-structuring, while a more destructive transformation, serves as an upper bound on anonymization by preserving density profiles while discarding spatial continuity.

Recent methods, such as PrivTrace[17], propose trajectory-level privacy mechanisms based on Markov modeling and convex optimization. However, these approaches exhibit computational complexity that may be prohibitive in large-scale settings. Specifically, the overall time complexity is $\mathcal{O}(m|D| + m^3)$, where m is the number of spatial cells and $|D|$ the number of trajectories. This complexity, arising from discretization, shortest-path estimation, and second-order transition modeling, poses challenges when scaling to datasets like `Yjmob100k`, which contain hundreds of thousands of users and high spatial granularity. In contrast, our selected techniques strike a balance between theoretical guarantees, interpretability, and computational feasibility, making them more suitable for reproducible research on privacy in real-world mobility traces.



(a) Clustered correlation

(b) Error in home inference

Figure 7. a) Clustered correlation after applying `Geo-Ind` using radius as 1 Km. The results still indicate a significantly higher correlation for Nagoya, allowing for successful location re-identification. b) Error in home location inference when comparing before and after applying perturbations using `Geo-Ind`.

7.2 Geo-indistinguishability

We apply Geo-indistinguishability (`Geo-Ind`)[18] to perturb individual location points in the `Yjmob100k` dataset. `Geo-Ind` extends differential privacy to the spatial domain by ensuring that any two locations $x_1, x_2 \in \mathbb{R}^2$ within a radius r produce statistically similar outputs. Formally, for any output y, the mechanism guarantees:

$$\frac{\Pr(y|x_1)}{\Pr(y|x_2)} \leq e^{\varepsilon r},$$

where ε defines the privacy budget in spatial units. This bound limits fine-grained inference while tolerating coarse spatial reasoning. We implement `Geo-Ind` using the planar Laplace mechanism, which perturbs the true location x by sampling y from a distribution centered at x with density:

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

where $d(x,y)$ is the Euclidean distance. The mechanism samples a random angle θ and radial distance r from an inverse CDF derived from the Laplace distribution[18].

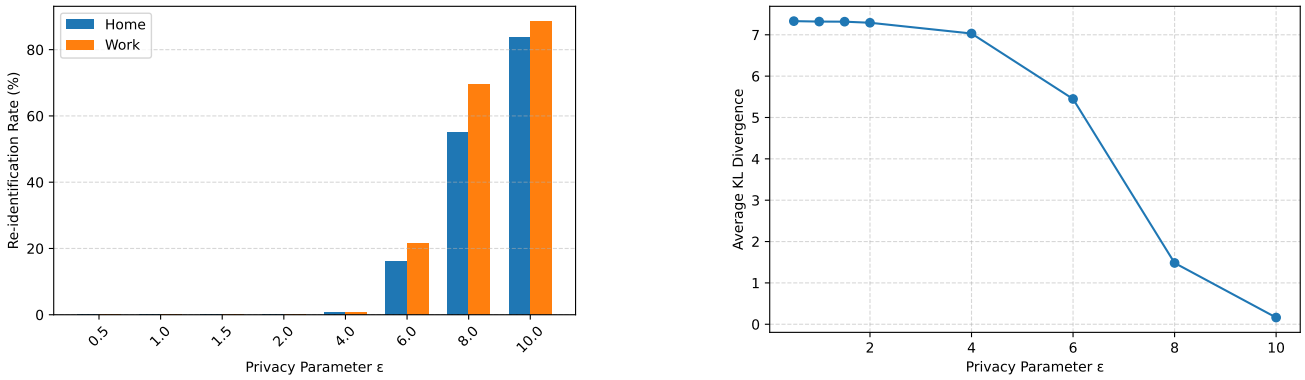
We perturb each user location using `Geo-Ind` with ε calibrated to a 1 km radius–considered a strong privacy baseline[18]. Figure 7(a) shows that even after perturbation, the clustered correlation with public population data remains high, particularly for Nagoya, enabling successful city-level re-identification. In Figure 7(b), we analyze home location inference error. Although `Geo-Ind` introduces spatial noise, the resulting error distribution remains narrow, revealing residual anchor points.

These results expose a fundamental limitation: while `Geo-Ind` effectively protects isolated location queries, it struggles to anonymize full mobility trajectories. Repeated visits and correlated movements enable aggregation attacks that undermine point-wise protection. In practice, adversaries still exploit large-scale structural patterns, suggesting that trajectory-level privacy remains an open challenge even under formally private mechanisms like `Geo-Ind`.

7.3 Generalized Randomized Response (GRR)

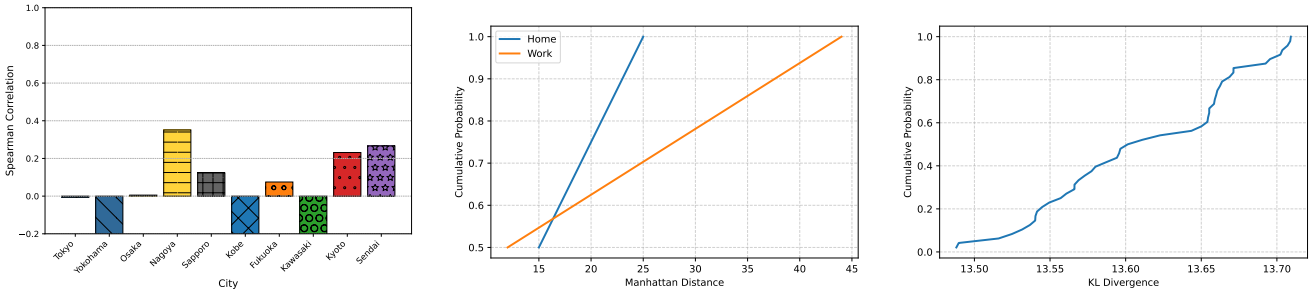
To evaluate the privacy-utility trade-off in mobility anonymization, we apply the Generalized Randomized Response (GRR)[16] mechanism to protect users' location trajectories. GRR operates over a discrete domain of spatial grid cells derived from the `Yjmob100k` dataset. Each location (x,y) is first mapped to a unique integer in a domain of size k, and then perturbed using the GRR mechanism with privacy parameter ε. At each step, GRR preserves the true location with probability $p = \frac{e^\varepsilon}{e^\varepsilon + k - 1}$, and otherwise replaces it with a uniformly sampled alternative from the remaining grid cells. This design ensures ε-local differential privacy for each reported location while enabling controlled utility loss. We generate GRR-perturbed versions of the dataset for multiple values of ε and evaluate the impact on both privacy and data utility.

Figure 8 illustrates this trade-off across two axes: (a) shows the re-identification rate of home and work locations as a function of ε, while (b) presents the average KL divergence between the true and estimated population distributions at each



(a) Re-identification Rate (b) KL Divergence

Figure 8. Effect of GRR-based LDP on home and work inference. (a) Re-identification rate of users' home and work locations under varying ε. (b) KL divergence between true and estimated population distributions, reflecting the utility loss.



(a) Clustered Correlation (b) Home and Work Location Error (c) KL Divergence

Figure 9. Evaluation of spatial de-structuring effects. (a) Spearman correlation between spatially-permuted user distributions and official census data for the top 10 Japanese cities. (b) CDF of Manhattan distance errors between inferred and true home/work locations. (c) CDF of KL divergence between temporal-spatial distributions before and after permutation.

timestamp. For low ε values (e.g., $\varepsilon \leq 2$), re-identification rates drop significantly, indicating stronger privacy guarantees; however, this comes at the cost of a sharp rise in KL divergence, which signals substantial utility degradation. Conversely, for high ε values (e.g., $\varepsilon \geq 6$), KL divergence decreases and the perturbed distributions converge toward the original data, but the protection becomes negligible–leaving users exposed to exact home-work inference. GRR fails to safeguard user privacy unless we either accept severe utility loss or assume near-perfect knowledge of the underlying population distributions.

7.4 Spatial De-structuring

We now evaluate a sanitization mechanism that intentionally removes the spatial structure of mobility datasets to assess whether cell-level density alone suffices for spatial re-identification. As previously demonstrated in Section 4 and supported by findings in[5], the shape of the spatial density is often sufficient to uniquely characterize cities.

To counteract this, we apply a spatial de-structuring strategy whereby each user's grid cells are randomly permuted, preserving the density of each cell but destroying its spatial arrangement. This maintains the temporal usage patterns and marginal frequencies of cell visits while eliminating geographic coherence.

Figure 9 presents three metrics for evaluating this transformation. Figure 9(a) shows the Spearman correlation of clustered user density with census data across major cities. While the spatial structure is reduced, density indicators persist, and Nagoya still emerges to an attacker as the most probable guess for the anonymized location. Figure 9(b) shows the CDF of home/work location error after permutation, which confirms immense degradation in precision. Furthermore, Figure 9(c) shows the very high KL divergence between the true and estimated population distributions, reflecting the extreme loss of utility. Even after eliminating all spatial order, the remaining density structure still encodes sufficient information to enable city re-identification.

7.5 Discussion

As opposed to the previous schemes, spatial de-structuring does not perturb individual location points but rather alters the dataset at a macroscopic level by scrambling spatial coherence. The persistence of partial re-identification underscores the difficulty of obfuscating population-scale structures, even under seemingly destructive transformations.

When considered alongside `Geo-Ind` and GRR, a broader pattern emerges. `Geo-Ind` provides strong theoretical guarantees under the differential privacy framework by perturbing each location with calibrated noise. However, our experiments demonstrate that repeated visits and spatio-temporal correlations in user trajectories allow adversaries to recover salient structures, such as home or city of origin. GRR, on the other hand, applies privacy at the granularity of discrete grid cells and achieves formal local differential privacy. Yet, our results show that its protection is only effective under very strong privacy budgets (i.e., low ε), which in turn leads to significant degradation in analytical utility, such as population density estimation or home/work inference. Spatial de-structuring avoids formal guarantees but aggressively disrupts layout semantics. Despite this, residual statistical signals persist, revealing the limitations of even non-parametric transformations.

Taken together, these findings reinforce that trajectory anonymization remains an open and unresolved challenge. Simple point-wise noise mechanisms fail to hide users in long, structured trajectories, while aggressive approaches severely degrade the utility of the data. There is currently no general-purpose mechanism that can simultaneously offer strong privacy guarantees and preserve high-resolution utility for spatio-temporal datasets. Addressing this challenge requires effective structural pattern suppression–an important direction for future research in mobility privacy.

8 Responsible disclosure

The issue reported in this paper demonstrates that users included in the now public dataset `Yjmob100k` can have their privacy compromised. We obtained approval from the ethical and legal committee of our research institution (Inria COERLE, Decision 2025-18). We contacted both the French and the Japanese Data Protection Authorities (respectively the CNIL[19] and the *Personal Information Protection Commission*[20]) to inform them of our findings.

9 Conclusion

This study shows that re-identification techniques remain highly effective against anonymized mobility datasets. Using the `Yjmob100k` dataset as a case study, we apply publicly available auxiliary data and structured analytical methods to recover both spatial and temporal information. We re-identify the original geographic coordinates embedded in user trajectories and infer the underlying timeline (dates) of the dataset. Our analysis highlights the substantial privacy risks associated with such re-identification. The results confirm that commonly applied anonymization strategies–such as coordinate shifting and timestamp normalization–fail to provide meaningful protection. Individuals remain vulnerable to location inference, home-work profiling, and potential linkage to real-world identities. These findings underscore the urgent need for more robust privacy-preserving frameworks tailored to spatio-temporal datasets. Future work must balance utility with individual privacy by advancing anonymization methods that address structural leakage at both spatial and temporal levels.

Data Availability

The `Yjmob100k` dataset is publicly available[21], and the code used in this study is available on request.

Acknowledgments

The authors acknowledge support from Inria, PEPR IPOP, and ANR GTTP projects.

References

1. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *nature* **453**, 779–782 (2008).
2. Montjoye, Y.-A. d., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Reports* **3**, 1376, DOI: 10.1038/srep01376 (2013).
3. Yabe, T. *et al.* Yjmob100k: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Sci. Data* **11**, 397 (2024).
4. Douriez, M., Doraiswamy, H., Freire, J. & Silva, C. T. Anonymizing NYC Taxi Data: Does It Matter? In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, 140–148 (IEEE, 2016).
5. Pintér, G. Revealing urban area from mobile positioning data. *Sci. Reports* **14**, DOI: 10.1038/s41598-024-82006-5 (2024).

6. Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F. & Xie, X. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. In *ACM WSDM*, 295–304, DOI: 10.1145/2684822.2685287 (ACM, Shanghai China, 2015).
7. Drakonakis, K., Ilia, P., Ioannidis, S. & Polakis, J. Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta)Data, DOI: 10.48550/arXiv.1901.00897 (2019). ArXiv:1901.00897 [cs].
8. Fung, B. C., Wang, K., Chen, R. & Yu, P. S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv. (Csur)* **42**, 1–53 (2010).
9. Acs, G. & Castelluccia, C. A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris. In *ACM SIGKDD*, KDD '14, 1679–1688, DOI: 10.1145/2623330.2623361 (ACM, New York, NY, USA, 2014).
10. Fiore, M. *et al.* Privacy in trajectory micro-data publishing : a survey. *arXiv:1903.12211 [cs]* (2020). ArXiv: 1903.12211.
11. Japan: High resolution population density maps + demographic estimates (2018). Available online: https://data.humdata.org/dataset/japan-high-resolution-population-density-maps-demographic-estimates (accessed on 3 June 2025).
12. Time and Date AS. Japanese holidays calendar. https://www.timeanddate.com/calendar/custom.html?year=2021&country=26&cols=3&hol=9&df=1 (2021). Accessed: 2025-06-03.
13. BBC News. Typhoon hagibis: Japan suffers deadly floods and landslides from storm. https://www.bbc.com/news/world-asia-50020108 (2019). Accessed: 2025-06-03.
14. Gramaglia, M., Fiore, M., Furno, A. & Stanica, R. Glove: Towards privacy-preserving publishing of record-level-truthful mobile phone trajectories. *ACM/IMS Trans. Data Sci.* **2**, DOI: 10.1145/3451178 (2021).
15. Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E. & Palamidessi, C. Broadening the scope of differential privacy using metrics. In *PETS*, 82–102, DOI: 10.1007/978-3-642-39077-7_5 (Springer, 2013).
16. Kairouz, P., Bonawitz, K. & Ramage, D. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, 2436–2444 (PMLR, 2016).
17. Wang, H. *et al.* PrivTrace: Differentially private trajectory synthesis by adaptive markov models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1649–1666 (2023).
18. Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K. & Palamidessi, C. Geo-indistinguishability: differential privacy for location-based systems. In *ACM CCS*, CCS '13, 901–914, DOI: 10.1145/2508859.2516735 (2013).
19. Commission Nationale de l'Informatique et des Libertés. CNIL – French Data Protection Authority. https://www.cnil.fr (2025). Accessed: 2025-06-03.
20. Personal Information Protection Commission Japan. Personal Information Protection Commission (PPC). https://www.ppc.go.jp/en/ (2025). Accessed: 2025-06-03.
21. Yabe, T. *et al.* YJMob100K: City-Scale and Longitudinal Dataset of Anonymized Human Mobility Trajectories, DOI: 10.5281/zenodo.10836269 (2024).

A Additional data

A.1 Top 10 residential grids

We identify the top 10 residential 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).

A.2 Port Messe exhibition center

We identify six events hosted at Port Messe exhibition center over the period from 15/9/2019 to 28/11/2019 (see Table 2).

Event name	Dates	Attendance
Automotive World Nagoya 2019[1]	September 18-20 (Day 3-5)	unknown
MECT2019[2]	October 23-26 (Day 38-41)	90.000
Messe Nagoya 2019[3]	November 6-9 (Day 52-55)	60.000
One OK rock concert[4]	November 12-13 (Day 58-59)	unknown
Mammoth Flea Market Z vol.61 [5]	November 16-17 (Day 62-63)	unknown
Nagoya Motor Show 2019[6]	November 21-24 (Day 67-70)	205.900

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