Protecting participants or population? Comparison of k-anonymous Origin-Destination matrices

Protecting participants or population? Comparison of k-anonymous Origin-Destination matrices

PIETRO ARMENANTE[1], KAI HUANG[1], NIKHIL JHA[1], AND LUCA VASSIO[1,*]

[1] *Politecnico di Torino, Italy*
[*] *luca.vassio@polito.it*

1. INTRODUCTION

Origin-Destination (OD) matrices are a core component of research on users' mobility and summarize how individuals move between geographical regions. These regions should be small enough to be representative of user mobility, without incurring substantial privacy risks.

There are two added values of the NetMob2025 challenge dataset. Firstly, the data is extensive and contains a lot of socio-demographic information that can be used to create multiple OD matrices, based on the segments of the population. Secondly, a participant is not merely a record in the data, but a statistically weighted proxy for a segment of the real population. This opens the door to a fundamental shift in the anonymization paradigm. A population-based view of privacy is central to our contribution. By adjusting our anonymization framework to account for representativeness, we are also protecting the inferred identity of the actual population, rather than survey participants alone.

The challenge addressed in this work is to produce and compare OD matrices that are k-anonymous for survey participants and for the whole population. We compare several traditional methods of anonymization to k-anonymity by generalizing geographical areas. These include generalization over a hierarchy (ATG and OIGH) and the classical Mondrian. To this established toolkit, we add a novel method, i.e., ODkAnon, a greedy algorithm aiming at balancing speed and quality.

Unlike previous approaches, which primarily address the privacy aspects of the given datasets, we aim to contribute to the generation of privacy-preserving OD matrices enriched with socio-demographic segmentation that achieves k-anonymity on the actual population.

The full report for this work is available on arXiv[1], with open-source reproducible code available on GitHub[2].

2. METHODOLOGY

A. Geographical area hierarchy definition

We assume a fixed geographical generalization hierarchy of zones. A generalization hierarchy describes how smaller zones should be merged together. We can think of this hierarchy as a tree where the leaves are the initial small zones, and a parent node represents the union of the areas of its child nodes.

In this work, we use the well-known, Uber-developed H3 hierarchy.[3] In H3, the space is organized in hexagonal cells: every cell, up to the maximum resolution supported by H3, has seven child cells below it in this hierarchy—see Figure 1. Hexagons do not cleanly subdivide into finer ones; however, a subdivision into seven cells can be approximated.

B. Algorithms for k-anonymity

We want to find an aggregation of origins and destinations to ensure that each entry corresponds to a group of at least k indistinguishable individuals. We compare algorithms that perform an adaptive generalization such that, for every trip in our anonymized dataset, at least k-1 other trips exist that start in the same origin and finish in the same destination (possibly different from the origin). We consider the state-of-the-art ATG algorithm [1], OIGH uniform tree generalisation algorithm [2], and Mondrian [3]

We propose a new OD-matrix anonymization algorithm, ODkAnon, that is greedy, iterative, and aims at generalizing OD locations in a hierarchy to satisfy k-anonymity [4] constraints. The algorithm considers the fixed, H3-based generalization hierarchy, and iteratively selects and aggregates cells related to lower-density areas (considering both origins and destinations) until the k-anonymity condition is met. This algorithm (as OIGH) creates *homogeneous* geographical areas, i.e., origin areas choice does not depend on the given destination and vice-versa. This means that the final OD matrix will comprise non-overlapping areas. A detailed description of ODkAnon is available in our report, and its implementation is in our GitHub repository.

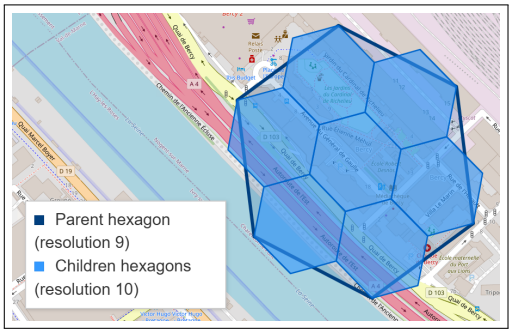

Fig. 1. Example of a H3 hierarchy of hexagons in Paris.

C. Protecting the participants vs. protecting the population

Common algorithms aim at obtaining a k-anonymous dataset, where every item—in this case, a trip—cannot be distinguished from at least k-1 other ones. However, each dataset only represents a subsample of the population, some segments of which may be over- or under-represented by the records in the dataset. The awareness of this data skewness could help re-identification attacks by an adversary. Therefore, risks are not only towards the participant from whom the data has been collected, but also towards the original population that has been sampled.

The weight attribute that the NetMob2025 challenge dataset offers hints at a solution in this direction, indicating the number of people in the Île-de-France area that each participant represents. By using this number to weight each entry in the dataset, we can tweak the algorithms to account for the population rather than the dataset, strengthening the privacy guarantees of the algorithms. Naively, the algorithms will see the same trip repeated as many times as the survey participant's population representativeness.

[1] https://arxiv.org/abs/2509.12950
[2] https://github.com/SmartData-Polito/ODkAnon
[3] https://h3geo.org/docs/

Table 1. Challenge dataset relevant characteristics.

Participants with GNSS records	3,320
Represented population	9,001,164
Trips with GNSS points	81,291
H3 hexagons at resolution 10	44,011
Relevant socio-demographic features	sex, age, profession

D. Performance metrics

The anonymization alters the values of the OD matrix, as well as the zone sizes, depending on the aggregation of the geographical areas in the hierarchy.

Privacy metrics validate the strength of the anonymization. We measure the minimum k-anonymity obtained for participants when protecting the population, and the other way around.

Data utility metrics measure how well the anonymized OD matrix preserves the original data characteristics, comparing the pre- and post-anonymization versions. Among the ones we used, we report here: a generalisation error ($\bar{G}$) that counts the number of merges in the hierarchy [1], a discernibility metric (C_{DM}) that accounts for how many trips are aggregated [3], and a reconstruction loss (E) [1].

3. DATASET CHARACTERISTICS

The NetMob2025 challenge dataset offers a multi-dimensional view of mobility behavior in the Île-de-France region over a continuous 7-day period. We select only participants with complete GNSS records. Then, we consider a trip completed after the absence of GNSS records for 3 minutes. Finally, we extract the origin and destination coordinates of each trip. These are anonymized, generalized to the center of the closest H3 hexagon of resolution 10.

Together with origins and destinations, we employ the weight feature assigned to every trip, describing how many people the record is representative of, and three socio-demographic features (sex, age, and profession) to build segment-specific OD matrices. The main characteristics relevant to our work are reported in Table 1.

4. MAIN RESULTS AND DISCUSSION

The complete results are available in our report. We run the OD-kAnon algorithm on the trips within the Île-de-France, obtaining a single OD matrix aggregating all data. We first consider protecting the participants in the survey, and we set $k = 10$ for obtaining a k-anonymous OD matrix. We allow for suppression of up to 10% of the trips. Merging over the hierarchy, the original thousands of resolution 10 hexagons, we obtain the same 29 zones both for the origins and destinations. Figure 2 reports a zoom over Paris of the obtained generalized areas (blue hexagons).

When protecting the population, k should be adapted, accounting for the representativeness of each participant. Given that a participant on average accounts for 2,674 people, we use $k = 10 \times 2,674$ in order to keep a fair comparison of the two approaches. We now obtain the same 29 destination zones, but more fine-grained 35 origin zones. We show the origin zones in Figure 2 as red hexagons. Protecting the population produces a different anonymization. In particular, 7 smaller hexagons are now generalized to their parent node (the large blue hexagon). Likely, these zones contain few trips from the participants, but are related to a large portion of the population. Even more interesting, when the protection is applied to the population, the participants' OD matrix loses k-anonymity: 13 hexagons fall below the threshold (k = 10), with the smallest value equal to 4.

Table 2. Data utility metrics, after anonymization with OD-kAnon. Metrics are computed considering participant trips. A lower value means that the OD matrix retains more utility.

Protection	Data	OD dimension	$\bar{G}$	C_{DM}	E
Participants	All	29x29	601	5.4×10^7	1.91
Participants	Men	5x2	4,388	4.6×10^7	1.43
Participants	Women	29x29	447	1.8×10^7	1.51
Population	All	35x29	539	5.3×10^7	1.91
Population	Men	2x2	6,727	1.3×10^8	1.86
Population	Women	29x29	450	1.8×10^7	1.87

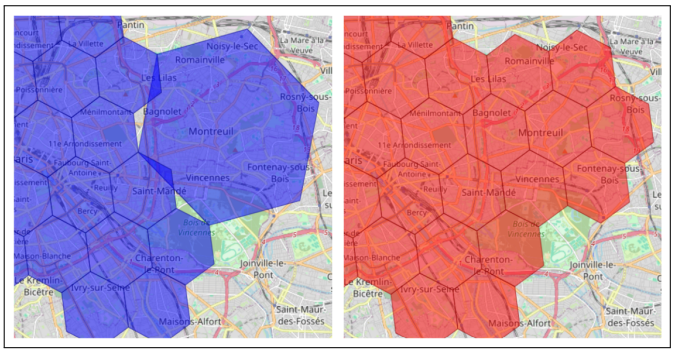

Fig. 2. Detail over Paris of the origin generalization hexagons for the participant-protecting (left, blue hexagons) and the population-protecting (right, red hexagons) definitions.

Moreover, we observed amplified differences when segmenting the population. Consider the example of sex (men and women), with the same k thresholds. The two segments have a similar total number of trips. When protecting the participants, anonymizing the men dataset produces a coarse 5x2 OD matrix, whereas anonymizing the women dataset results in a fine-grained 29×29 matrix. This indicates that protecting men is more challenging, as it requires very coarse hexagons, while for women the resulting hexagons remain much finer. Furthermore, when applying protection to the population, the difference becomes even more pronounced: the anonymized men dataset reduces to a 2×2 matrix.

We report the data utility metrics for the obtained anonymized datasets in Table 2. We computed these metrics considering the survey's participant trips. All these metrics indicate more utility when lower. Notice how the men segment loses utility when anonymized. Full comparisons, including (i) different anonymization algorithms, (ii) multiple socio-demographic segmentations, and (iii) metrics computation for population, are available in our arXiv report.

Our results showed that significant differences may exist when considering population-protecting for OD matrices anonymization, rather than survey participant-protecting ones. Moreover, we showed how these differences can even be amplified across socio-demographic segments.

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. B. Matet, A. Furno, M. Fiore, *et al.*, Transp. Res. Part C: Emerg. Technol. **154**, 104236 (2023).
2. W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, World Wide Web **24**, 1551 (2021).
3. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *ICDE'06*, (IEEE, 2006), pp. 25–25.
4. P. Samarati, IEEE Trans. on Knowl. Data Eng. **13**, 1010 (2002).