Enhancing Data Privacy with Semantic Trajectories: A Raster-based Framework for GPS Stop/Move Management

Yang Wang and David McArthur

Abstract

Tracking facilities on smart phones generate enormous amount of GPS trajectories which provides new opportunities for study movement patterns and improve transportation planning. Converting GPS trajectories into semantically meaningful trips is attracting increasing research efforts with respect to the development of algorithms, frameworks and software tools. There are however few works focused on designing new semantic enrichment functionalities taking privacy into account. This paper presents a raster based framework which not only detects significant stop locations, segments GPS records into stop/move structure, brings semantic insights to trips but also provides possibilities to anonymize users' movements and sensitive stay/move locations into raster cells/regions so that a multi-level data sharing structure is achieved for a variety of data sharing purposes.

1. Introduction

The proliferation of smartphones has made it feasible to collect movement data, in the form of GPS trajectories, for a large number of people, generating new opportunities to study movement patterns and improve transportation planning (Liao 2007b, Hwang et al. 2013, Liu et al. 2012). In its raw form, such data are not particularly useful to transport planners. Planners have traditionally worked with data from travel diaries or link-based sensors. Travel diaries may give mode-specific origin/destination matrices according to trip purpose. Road sensors may give information about vehicle speed on a given link. Similar information can be extracted from GPS records. To do this ethically and legally, the privacy of data subjects must be considered.

Recent studies have considered the development of algorithms, frameworks and software tools to organize GPS records into semantically annotated trips (Alvares et al. 2007,

Zheng 2015, Yan et al. 2013). A semantic trajectory framework often processes GPS trajectories into a stop/move structure (Spaccapietra et al. 2008), where stop locations are used to infer the purpose of a trip while moves provide information about speed, direction, and mode of transport.

GPS trajectories are highly sensitive personal data, revealing locations such as homes and workplaces, as well as information about routes and schedules. Studies on large volumes of mobile phone data demonstrate that small segment of visit sequences reveals peoples' identity even the data is spatially and temporally coarsened (Zang & Bolot, 2011, De Montjoye et al, 2013). This poses a privacy threat as the complexity of location data makes it difficult to anonymize (Abul et al. 2008, De Montjoye et al. 2013). Trajectory mining frameworks tend not to consider privacy and data sharing, with Zheng (2015) being a rare exception.

This paper presents a raster-based semantic trajectory development and management framework which facilitates data sharing while protecting privacy. The framework employs a raster sampling method to detect significant stops and segment GPS trajectories into a stop/move structure. The same process also aggregates GPS records into raster cells and supports a variety of anonymization methods such as k-anonymity, generalization methods (grid masking) and 'stop/move' spatial cloaking. The framework allows easy measurement of information loss. The contribution of the paper is developing a flexible data structure and framework which can transform raw GPS data into a form which is useful for planners. Data owners may wish to share or sell the output. A key focus of our research is therefore how to include functionality to anonymise data within our proposed framework.

A review of previous work is presented in Section 2. Section 3 proposes a raster-based framework and describes how it processes raw GPS trajectories into semantically enriched datasets. The multi-level data sharing scheme with trajectory anonymization

supported by the framework is described in section 4 which leads to the discussion and conclusion.

2. Related Work

Processing raw GPS trajectories normally starts with detecting stops. Recently, several trajectory processing frameworks have been developed on top of the threshold (Li et al. 2008), feature-intersection (Bogorny et al. 2011), and density (Yan et al. 2013) stop detection methods. Other stop/stay-point detection methods such as density (Schoier & Borruso 2011, Hinneburg & Keim 1998, Ankerst et al. 1999, Campello et al. 2013) and threshold (Ashbrook & Starner 2002, Schuessler & Axhausen 2009, Srinivasan et al. 2009, Yan et al. 2013) adopt a 'bottom-up' strategy which scans and clusters GPS records into stop locations. As a result, pre-understanding of the GPS records is required to set suitable thresholds and parameters for specific cases e.g. different travel modes and study areas.

Once stops are detected, information such as points of interest (POI), road networks, and land-use are used to add contextual meaning and infer trip purposes using, for example, complex probability models for automatic semantic annotation (Yan et al. 2013). These frameworks, however, place emphasis on the overall processing of GPS trajectories with little concern for protecting data subjects' privacy. Zheng (2015) summarizes a broad paradigm on mining insights from GPS records (Zheng et al. 2011, Yuan et al. 2011, Li et al. 2008, Zheng et al. 2008). Privacy concerns are also included but after stops/moves detection, hence, the trajectory processing and privacy protection are detached from the rest of the process. An interaction between managing GPS trajectories and anonymization is still unattended

On the other hand, GPS anonymization techniques have mostly developed in parallel. Among them, mixing an individual with k-1 others is one of the most popular. For anonymizing movements, 'Never Walk Alone' (NWA) (Abul et al. 2008), publishes the

mean locations of a co-localised k trajectories, within a given period. As a cluster method, the association with road links, the semantics of the moving, is lost. Better privacy protection may involve a large decline in data utility and a loss of trip semantics (Yin et al. 2015).

There are other anonymization methods which 'blend' stop points into larger stay zones (Huo et al 2012) or displace the GPS records (Armstrong et al. 1999, Kwan et al. 2004, Hampton et al. 2010, Zandbergen 2014). These techniques are less semantically-aware although, among them, location swapping (Zhang et al. 2017) tries to preserve semantics such as land cover and proximity to roads but there are increasing concerns about sensitive semantics being wrongly associated with the locations during anonymization (Seidl et al. 2017).

Generalisation-based methods, especially grid masking, are more relevant to our idea where trajectory points are aggregated or snapped to grid cells for publication (Leitner & Curtis 2006, Krumm 2007, Shi et al. 2009, Seidl et al. 2016). Seidl et al. (2016) and Sila-Nowicka & Thakuriah (2016) note a compromise of travel pattern with larger masking size for better privacy protection. This issue can be addressed by organizing and publishing GPS records under a multilevel framework where trip semantics are preserved, retrievable, and even released based on data usage agreements made at different levels.

3. A Raster based Framework for Developing and Managing GPS Trajectories

3.1. The Overall Semantic Trajectory Management Data Framework

The proposed framework uses raster cells as a unified data processing 'vehicle' which incorporates stop detection/annotation and trajectory anonymization/publication in a single framework. A raster-based stop detection algorithm (Section 3.2) is the core function that processes the GPS records into a stop/move structure while the raster cells are preserved throughout the processing and anonymization phase. As shown in Figure 1, there are three

layers in the framework: (a) the GPS records, (b) the unstructured layer where the segmented stop/move GPS records are denoted with 'rastervalues' from the stop detection and (c) the structured layer comprising the stop/move cells with aggregated trajectory attributes. Semantic annotation is conducted at the structured layer with contextual information such as land-use and road network data.

3.2. Detecting Stops

We firstly describe a top-down stop detection method. One of the major characteristics of the method is that it requires minimal parameter setting, with only the raster cell size and generic stop selection quantile breaks needing to be set. The method supports flexible post-processing functions which can improve the detection accuracy (see section 3.2.2) in some contexts. Our approach differs from existing raster approaches such as the kernel density approach (Thierry et al. 2013, Lei et al. 2011) as we are not sampling the density of GPS points but information such as total dwelling time.

3.2.1. Method

We calculate the estimated dwelling time inside a raster cell. If the duration between two consecutive GPS records is denoted as $dur_{(GPS_i,GPS_j)}$, to remove the impact of 'moving' time in-between cells or inside cells, the indicator is defined as

$$dur\_withoutTravel_{(GPS_i,GPS_j)} = dur_{(GPS_i,GPS_j)} - est\_travelTime_{(GPS_i,GPS_j)} \quad (1)$$

where $est\_travelTime_{(GPS_i,GPS_j)}$ is the estimated travel time between point i and j calculated from travel speed observed before and after a given GPS record within a five-record window (we do not use a specific temporal window as the temporal gaps between the GPS records are not evenly distributed),

$$est\_travelTime_{(GPS_i,GPS_j)} = {Distance_{(GPS_i,GPS_j)}}/{mean\left(speed_{window_{(i,j)}}\right)} \quad (2)$$

where $window_{(i,j)}$ is $< \{GPS_{i-5}, \dots, GPS_i\}, \{GPS_j, \dots, GPS_{j+5}\} >$.

The raster value for each grid cell is the sum of $dur\_withoutTravel_{(GPS_i,GPS_j)}$

$$< row_r\ , column_c\ , v >_{r \in rows, c \in columns} \quad (3)$$

Where v is $sum\left(dur_{withoutTravel_{(GPS_i,GPS_j)}}\right)$ if $geom(GPS_i) \in cell < row_r\ , column_c >$.

We use Natural Break ($Jenks$) to group raster values into classes (other clustering methods, such as k-means, could also be used). To avoid setting the number of classes, goodness of variance fit (over 0.8) is adopted. Taking the 'moves' being clusters into the lower value class, we select the raster cells with values higher than the 25% quantile of the clustering result as stops (50% quantile selection is tested but 25% produces better results. It is treated as a generic fixed setting in the framework.)

Post-processing functions are included, as illustrated in Figure 2. Firstly, segmenting GPS trajectories into a stop/move structure following people's sequence of visits to the detected distinct stops, secondly, merging neighbouring stops together if they share an edge. The former transforms the trajectories into trips separated by detected stops and further supports the generation of the structured layer. The latter reduces the number of detected redundant stops significantly and reflects the fact that the GPS records around stops are clustered without a fixed shape. Other steps such as cleaning brief visits to neighbouring cells, detecting round-trips and eliminating intermediate travel stops are optional but can be performed to improve accuracy based on specific data processing requirements.

3.2.2. Testing the Method

The method is illustrated using GPS records collected from the Catch! Smartphone app (http://www.travelai.info/catch.html). The app gathers GPS records with no user interventions and regularly synchronizes with a central server. As the app is developed for the general public, there is no travel survey facility. To obtain meaningful ground truth, we select three users with different travel behaviour/settings. The performance of the methods is verified with the ground truth worked out by manually selecting candidate stops confirmed with the users.

As shown in Figure 3, User A (using an iPhone 6+) is located in a suburban area with the majority of trips to/from Glasgow city center being made by car. Trips in Case B (using a OnePlus One), are mostly within Glasgow by subway and walking. Case C (iPhone6) collects activities of a London resident with combined walking/underground/bus travel. These three cases represent the differences between a person with a simple travel pattern and travelers with multimodal, short/long trips in cities of different sizes and complexities. The data, with sensed sleep/non-moving/moving status and 1-2min frequency for moving, are cleaned to keep non-duplicated records with a consecutive speed of less than 200 km/h to concentrate on surface transport (Table 1). The method is illustrated using daily GPS trajectories for days which have more than 50 valid records.

For comparison, we include a threshold-based method which selects stops from all the raster cells using both a long stay threshold (>5min) and a map matching method (Wang & McArthur, 2017). The threshold method represents one of the most commonly used stop detection methods linking with GPS segmentation. The map matching method represents a 'top-down' method where GPS records at least 10 meters from the transport network for at least 5 minutes are detected as stops.

A unified raster scale over the UK with cell size set to 0.00091 decimal degrees (roughly 60 meters) under WGS_1984 is adopted. To measure accuracy, we take the detected stops (centroid of the cell shape) and compare their spatial proximity to the ground truth stop. The total number of detected stops that fall into a 100, 200, and 300-meter radius of a given actual stop are collected. The precision/recall are defined in formula (4) and (5). For the baseline methods, we sample their detection output using the raster template to ensure the precision/recall measurements are comparable.

$$Recall = \frac{\sum(Accurate\ Stops)_{Distance\ Band}}{\sum Actual\ Stops} \quad (4)$$

$$Precesion = \frac{\sum(Accurate\ Stops)_{Distance\ Band}}{\sum Detected\ Stops} \quad (5)$$

Figure 4 illustrates the precision/recall rates for the three methods and distance bands. The raster method is in red stars while the two baselines are in 'green' and 'blue' arrows. The darker the color the larger the distance band. The first row in Figure 4 illustrates the precision/recall rates of the raster sampling result. The bottom row incorporates the post-possessing functions which improve both precisions and recalls to around 0.8 in three cases. Focusing on the bottom row, in Case A, where the travel patterns are simple, the three methods produce high precision/recall. The precision/recall of the raster method is relatively stable across cases, around 0.8, compared to the two baselines, which show dramatic differences in B and C where different travel modes and complex travel environments are analyzed. The map matching method performs less well with a complex network by producing a higher number of incorrectly detected stops, while the threshold method is sensitive to different scenarios.

Figure 5 presents a sensitivity analysis of cell size choice. Cell sizes are increased by a factor of 1.5 and 2 (cell sizes are roughly 60, 85, 110 meters) and their impact on

precision/recall rates for the raster method (orange) and threshold method (green). The arrow in between shows the sequence of the change from small to large. For simplification, we include the 100m distance band result. The patterns are preserved for the other distances considered. The recall rates, around 0.8, for the raster method settle higher than the threshold method, particularly in B and C. The raster method detects a higher number of stops around the ground truths. The precisions of the raster method are around 0.7-0.8 in the three cases and more stable than the threshold method when increasing the cell sizes. There is a drop in both precision and recall rates in Case A with 85m cells, perhaps due to changes in shapes when merging neighbouring stops.

3.3. Structured Trajectory Stop/Move and Annotation

To semantically annotate the detected 'move' cells for each trip, we perform a map matching process using Barefoot (https://github.com/bmwcarit/barefoot). As recognized in previous studies (Liao et al. 2007a, Kang et al. 2004), time is an important parameter in distinguishing significant locations. Similar to Siła-Nowicka et al. (2016), we annotate stop cells with the longest dwelling at night as 'home' and stops with the most visits during Monday-Friday as 'work' and other stops as 'others' The semantics of a trip, in addition to GPS enter/leave time and stay durations, is attached to the stop/move cells, illustrated in Figure 1, where the resulting 'move' table has a 'road' field which contains the road's osm_id. The 'stop' cells are denoted with 'home', 'work' and 'others' as trip purposes[1].

3.4. Querying Semantic Trip Information

The framework supports semantic queries based on the annotated stops/moves in cells using the foreign keys, 'startTripid' and 'endTripid', in the stop table referencing to the 'tripid' in

[1] The Catch! app, developed by TravelAI, has an online travel mode detection method which adds the travel mode as another part of semantics but is not reported here.

the move table (Figure 1) without performing spatial joins. A strength of using raster cells to index stops/moves is the convenient calculation of similarities between users if their trips spatially or semantically related. Given that all the 'rasterids' are initially referencing a pre-defined raster template without a vector shape, a 'rasterindex' file (Figure 1) is created to convert all the stop/move raster cells into polygons for spatial queries such as the travel volume (not necessarily stops) around a museum at weekends.

4. Data Anonymization for Multi-Level Data Sharing

After describing the construction of the framework, we illustrate the flexibility of the framework in supporting generalizing, blurring and cloaking of the stops/moves. The three use cases are further used in this section to illustrate different data sharing strategies. Although user groups are small, they help to demonstrate our idea which is scalable if data are processed and managed in the same way. Our focus is not on discussing specific parameter but to show that the proposed framework supports multiple strategies and effective assessments of different settings. We report examples on the structured layer although some operations can be performed on the unstructured layer for simplification.

4.1 Sharing Aggregated 'Moves' Information

The framework splits the GPS trajectories into stop/move segments with derived trip purposes which offers the possibility of aggregating trips according to different spatiotemporal scales and trip purposes. Data in this format is a common input into transport planning models. Figure 6 shows the number of times each cell was visited on a home-based trip in Glasgow; where the whole coverage of activities is shown in blue cells and the selected moves are highlighted in yellow-red colors for time window Sunday (top) and Monday (bottom). Counts are low as there are only two users in Glasgow.

Releasing trip counts with a small number of users is potentially dangerous using the fine-grained time windows where re-identification can happen by linking distinct travel segments. Another approach is to share 'moves' of groups of users with k-anonymity which requires individual movements to be indistinguishable from $k-1$ users. It reduces the chance of re-identification even if the attacker has partial information about an individual. Under the proposed framework, overlapping 'moves' are identified by the same 'cellid' in the 'agg_move' table on the structured layer. Noting that k-anonymity can be defined both spatially and temporally, Figure 7 demonstrates the 'move' cells shared by more than 2 users using our three cases in Glasgow within the 17:00-18:00 time window. The orange 'move' cells are identified as being shared by at least two users against the overall raster cells in blue. The associated underlying GPS records on the unstructured layer are further selected as shown in purple points. k-anonymity significantly reduces the volume of data available for sharing/publication especially with a small number of users.

An alternative generalization-based approach can be applied to the structured layer. With the move cells being annotated with roads through map matching with the road network, information such as travel time or speed can be aggregated to the road links/intersections. This allows the publication of average travel information instead of releasing raw GPS records. The released dataset is useful for planners focusing on traffic management. However, this strategy is still dangerous if the travel mode can be inferred from the travel speed, especially when the user group is small and speed on roads are dramatically different from user to user.

4.2 Masking GPS Traces

Anonymizing the 'stops' is also important as it is known that home/work locations (Golle & Partridge, 2009) and other sensitive locations (Sila-Nowicka & Thakuriah, 2016) can be used to infer an individual's identity. The framework generalizes the stops from locations to raster

cells which is a grid-masking technique. The framework achieves stop detection, trajectory segmentation and grid-masking in one step with no need for an extra anonymization process.

With enlarged cell sizes, stops and movements can be blurred into larger areas, shown in Figure 8. A trip, from 'home' to 'others', is detected using single (0.00091 decimal degree, about 60m), one and a half (about 85m) and double (about 115m) sized raster template respectively. To better preserve information about the trajectories while protecting privacy, Seidl et al. (2016) suggest a 30-50 meter cell size. Sila-Nowicka and Thakuriah (2016) advise a 500-meter cell size if home/work locations are involved. Adjusting cell sizes during the stop detection phase may affect the accuracy of the stop locations. Another shortcoming is that enlarging cell sizes does not ensure sensitive locations being sufficiently 'mixed' into the surrounding area, for example, a building in the countryside with no nearby neighbours.

4.3 Cloaking Sensitive 'Stops/Moves'

GPS data provides detailed trajectories meaning that an attacker can identify the individual through not only their frequently visited places but also their frequently used routes. For example, assume an attacker knows that User A follows a routine between home and work during weekdays. Given the travel count map at the bottom of Figure 6, the attacker can infer User A is highly likely to be traveling from the south. Additionally, some non-routine trips are also potentially risky to share. For User A again with previous guess plus knowing A visited a park at a weekend (upper Figure 6), the attacker can locate A's traces. We, therefore, test a strategy of cloaking both the top visited locations and most/least frequently used 'moves'. This strategy may reduce the utility of the data but makes it possible to share with a wider audience. The released data would still give insight into travel patterns.

4.3.1. Cloaking Sensitive 'Stops'

We cloak the top 10 sensitive locations (a different number of locations can be used) by firstly classifying stops based on their trip purposes. Then we calculate the minimum distance radius that would blur the stops in the 'home/work' category into the 10 nearest buildings and the stops in the 'other' category into the 10 nearest points of interest. Figure 9 shows an example for User A where four major stops (in red color cells) are blurred into the different blurring radii (in dark blue cells) where the final cells within the blurring radii (in light blue color cells) are blocked from the final data release.

This strategy intends to blur the stops into the environment with k-features. It protects the individual, e.g. A, by hiding his sensitive locations although A is potentially re-identifiable based on the combined cloaked locations given the temporal sequence of visits. Coarsening the spatial scales can help by aggregating travel origin/destination to census zones but may lose trip information. Other methods such as applying location swapping may disturb the semantic meanings of the trips.

4.3.2. Cloaking sensitive 'moves'

To cloak 'moves, we identify sensitive 'moves' as those 'move' cells that are traversed most frequently or very seldom. Two approaches are presented here to identify such moves. The first method ('TopLocation') selects the 'top' and 'low' use 'move' cells in relation to the frequency of the usage of their associated trip 'stops'. We locate the raster cells of the top visited stops in the aggregated stop table when the frequency is above a threshold t (15, 25, 50, 95 percentile) in the overall visit distribution. We then randomly select the trips associated with these stops for cloaking until the travel counts of their stops reach the overall average trip visits. In the second method ('TopMove'), we aggregate the number of unique trips travelled in a given 'move' cell then select those with trip counts above the 15, 25, 50, 95 percentile t of the overall 'move' cells trip counts. Trips that have over 50% of 'moves'

overlapping with the 'top' move cells enter a random selection ensures the top 'move' cells are not above the average trip counts. For both methods, we further cloak those 'move' cells that are associated with the stops visited by the users only once or the 'moves' travelled below twice as they indicates non-routine activities.

Figure 10 includes the cloaking results for User B zoomed into the user's main activity space symbolized with trip counts. We can see that a higher value of t releases data which more closely resembles the full data. A good choice of t would balance privacy protection and data utility such as $t = 75$ and $t = 50$ for 'TopLocation' and 'TopMove' methods respectively (how the framework accounts for information loss is included in the next section).

Cloaking moves has the potential to modify the overall travel pattern since users' most common movements are cloaked to prevent a frequent activity attack while cloaking unusual travel helps prevent re-identification by analyzing outliers. These processes make the anonymization result less suitable for travel pattern analysis.

4.4. Calculating information loss under the framework

To assess threshold settings or comparing different anonymization methods, calculation of information loss is crucial. The framework provides convenient calculations of trip-based, spatial/temporal information loss. Trip-based information loss captures the percentage of trips that are eliminated from data publication (Formula 6). The anonymization also significantly affects the spatiotemporal coverage compared with the original. We take spatial aspect as a percentage change of a unique number of grid cells (Formula 7) since each cell represents the basic equal spatial unit covered by the GPS records. With the aggregated duration of each trip spent in every raster cell, the temporal aspect is calculated as a total trip duration loss (Formula 8).

$$InfoLoss_{trip} = 1 - \frac{\sum relased\ number\ trips}{\sum original\ number\ of\ trips} \quad (6)$$

$$InfoLoss_{spatial} = 1 - \frac{\sum relased\ unique\ cells}{\sum unique\ cell} \quad (7)$$

$$InfoLoss_{temporal} = 1 - \frac{\sum relased\ cells\ duration}{\sum original\ cells\ duration} \quad (8)$$

We can use this measurement of information loss to evaluate different settings of thresholds. For instance, in section 4.3.2, the parameter t in both 'TopLocation' and 'TopMove' methods. Figure 11 illustrates information loss from the three aspects for the two 'cloaking' methods performed on home/work trips (Mondays-Fridays). A higher threshold t reduces the set of raster cells for cloaking, hence the overall information loss declines. 'TopMove' method helps release more trips that covers larger spatial areas because 'TopLocation' method has the higher possibility of cloaking trips that access to the main stops such as home/work and grocery shops. Both methods give significant temporal information loss especially in Case B and C.

Figure 12, on the other hand, compares the information loss taking all the demonstrated methods in the previous sections. The best anonymization method in this scenario is to cloak the sensitive stops which preserve the majority of movement information. Other methods, such as k-anonymity, are less applicable to users with low overlapping spatiotemporal activity, cloaking movements with 'TopLocations' involves higher information loss if users display highly regular travel patterns.

5. Discussion and Conclusion

With the increasing availability of GPS records, questions of how to process them in order to understand meanings of the trips and how to share this information without risking privacy may not be treated as separate but intertwined topics. This paper describes a raster-based

semantic trajectory development and management framework with facilities for data anonymization and data sharing.

A raster-based stop detection algorithm, which samples higher dwelling time within raster cells with additional post-processing functions, is illustrated to have good performance for accurately detecting stops. This enables the construction of stop/move tables and supports complex semantic queries. Unlike other approaches where trajectory anonymization is detached from data processing, aggregating GPS records into raster templates does not introduce extra effort but is integrated into the framework. With GPS data organized into unstructured and structured layers, the framework supports data anonymization following, for example, k-anonymity, grid masking and spatial 'stop/move' cloaking methods. We also demonstrate its convenience in measuring trip, spatial and temporal information loss. Table 2 summarizes means of protection, and information loss. Other methods, such as generalization using KDE maps, speed/wait time on roads and aggregating O/D to census areas are supported by the framework but not reported with particular emphasis.

The proposed approach offers fast execution and minimal parameter and threshold setting. Regarding performance, the stop detection process does not require extra processing time as it is a common process in a GPS management framework. A time-consuming process is aggregating the GPS information to raster cells. If the cell size is small, the processing time will increase significantly when generating the structured layer. For choosing the cell size, we recommend taking the data anonymization into account where no less than 50-100 meters sized cells shall be considered. Assume peoples' major activity happens 5km around the home, around 100 raster cells per user shall be processed. Although taking time to generate, the process is partially combined with the stop detection and the structured layer will further facilitate low-cost multi-level data sharing which proves to be worthwhile.

There are limitations of the framework. Firstly, the cell size affects the stop detection result. The result shows that the accuracy of detection declines when enlarging the cell size while larger cell sizes give higher privacy protection. The stability of the method may benefit from testing under a larger ground truth. A hierarchical sampling procedure would help to relieve signal jump errors and unstable sampling frequencies. Secondly, although employing an overall raster template ensures consistent spatial granularity, there is still an extra step to perform a vector-based spatial function using the pre-generated raster index file. Such a file has to be updated if stops are detected through several sampling processes. From the aspect of anonymization methods, larger user groups would be valuable to test threshold settings such *k*-anonymity. As re-identification can be achieved on people's combined routines, methods targeted for combined activity attack is interesting to explore in the future. A further discussion on fitting anonymized GPS data to some specific data analysis is also highly valuable.

Acknowledgements

The authors would like to acknowledge the support of the Catch! project partners and to Innovate UK for funding the project (102426/53001-404133).

	Duration	**Sample Frequency On Moving**	**Number of Valid Days**	**Total Raw GPS Records**	**Extracted GPS Records**
User A	2016/07/13-2016/08/10	1-2 mins	27	29439	15656
User B	2016/11/12-2017/02/14	1-2 mins	42	13607	7780
User C	2016/04/12-2017/12/01	1-2 mins	64	24215	16720

Table 1. Summary of data cleaning of three users

Method	**Description**	**Protection**	**Information loss**
Travel Count	Aggregate on both temporal and special scales	Hides individual travel details from aggregated numbers	Loss of granularity of GPS information

***k*-anonymity**	Mix individual users in k-1 groups of other users	Hides individual in the crowd	High, if users' activities do not overlap in either/both spatial or/and temporal scale(s)
Cloaking Sensitive Stops	Apply different cloak radii based on semantic meanings of stops	Hide sensitive location from other urban features	Low, partial information loss on trips
Cloaking Sensitive Moves	Cloak frequent and non-routine movements	Hide high/ low-frequency activities	High, travel pattern influenced

Table 2. Example anonymization methods with descriptions

Reference

Abul, O., Bonchi, F., & Nanni, M. (2008, April). Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on* (pp. 376-385). Ieee.

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007, November). A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* (p. 22). ACM.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999, June). OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod record* (Vol. 28, No. 2, pp. 49-60). ACM.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine*, *18*(5), 497-525.

Ashbrook, D., & Starner, T. (2002). Learning significant locations and predicting user movement with GPS. In *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on* (pp. 101-108). IEEE.

Bogorny, V., Avancini, H., de Paula, B. C., Kuplich, C. R., & Alvares, L. O. (2011). Weka-STPM: a Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. *Transactions in GIS*, *15*(2), 227-248.

Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 160-172). Springer, Berlin, Heidelberg.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

Golle, P., & Partridge, K. (2009). On the anonymity of home/work location pairs. *Pervasive computing*, 390-397.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., ... & Miller, W. C. (2010). Mapping health data: improved privacy protection with donut method geomasking. *American journal of epidemiology*, *172*(9), 1062-1069.

Hinneburg, A., & Keim, D. A. (1998, August). An efficient approach to clustering in large multimedia databases with noise. In *KDD* (Vol. 98, pp. 58-65).

Huo, Z., Meng, X., Hu, H., & Huang, Y. (2012, April). You can walk alone: trajectory privacy-preserving through significant stays protection. In *International conference on database systems for advanced applications* (pp. 351-366). Springer, Berlin, Heidelberg.

Hwang, S., Hanke, T., & Evans, C. (2013, June). Automated extraction of community mobility measures from GPS stream data using temporal DBSCAN. In *International Conference on Computational Science and Its Applications* (pp. 86-98). Springer, Berlin, Heidelberg.

Kang, J. H., Welbourne, W., Stewart, B., & Borriello, G. (2004, October). Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*(pp. 110-118). ACM.

Krumm, J. (2007). Inference attacks on location tracks. *Pervasive computing*, 127-143.

Kwan, M. P., Casas, I., & Schmitz, B. (2004). Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?. *Cartographica: The International Journal of Geographic Information and Geovisualization*, *39*(2), 15-28.

Lei, P. R., Shen, T. J., Peng, W. C., & Su, J. (2011, June). Exploring spatial-temporal trajectory model for location prediction. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on* (Vol. 1, pp. 58-67). IEEE.

Leitner, M., & Curtis, A. (2006). A first step towards a framework for presenting the location of confidential point data on maps—results of an empirical perceptual study. *International Journal of Geographical Information Science*, *20*(7), 813-822.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W. Y. (2008, November). Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems* (p. 34). ACM.

Liao, L., Fox, D., & Kautz, H. (2007). Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, *26*(1), 119-134.

Liao, L., Patterson, D. J., Fox, D., & Kautz, H. (2007). Learning and inferring transportation routines. *Artificial Intelligence*, *171*(5-6), 311-331.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems*, *14*(4), 463-483.

Thierry, B., Chaix, B., & Kestens, Y. (2013). Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics*, *12*(1), 14.

Schoier, G., & Borruso, G. (2011). Individual movements and geographical data mining. Clustering algorithms for highlighting hotspots in personal navigation routes. *Computational Science and Its Applications-ICCSA 2011*, 454-465.

Schuessler, N., & Axhausen, K. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, (2105), 28-36.

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2017). Privacy and False Identification Risk in Geomasking Techniques. *Geographical Analysis*.

Seidl, D. E., Jankowski, P., & Tsou, M. H. (2016). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science*, *30*(4), 785-800.

Shi, X., Alford-Teaster, J., & Onega, T. (2009, August). Kernel density estimation with geographically masked points. In *Geoinformatics, 2009 17th International Conference on* (pp. 1-4). IEEE.

Sila-Nowicka, K., & Thakuriah, P. (2016). The trade-off between privacy and geographic data resolution. a case of GPS trajectories combined with the social survey results. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 535-542.

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, *30*(5), 881-906.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data & knowledge engineering*, *65*(1), 126-146.

Srinivasan, S., Bricka, S., & Bhat, C. (2009). Methodology for converting GPS navigational streams to the travel-diary data format. *Department of Civil and Coastal Engineering, University of Florida*.

Wang, Y., & McArthur, D. P. (2017). Linking Smartphone GPS Data with Transport Planning: A Framework of Data Aggregation and Anonymization for a Journey Planning App. GISRUK 2017.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K. (2013). Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(3), 49.

Yin, L., Wang, Q., Shaw, S. L., Fang, Z., Hu, J., Tao, Y., & Wang, W. (2015). Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PloS one*, *10*(10), e0140589.

Yuan, J., Zheng, Y., Zhang, L., Xie, X., & Sun, G. (2011, September). Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 109-118). ACM.

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, *2014*.

Zang, H., & Bolot, J. (2011, September). Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking* (pp. 145-156). ACM.

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22-34.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 29.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008, April). Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web* (pp. 247-256). ACM.

Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W. Y. (2011). Recommending friends and locations based on individual location history. *ACM Transactions on the Web (TWEB)*, *5*(1), 5.

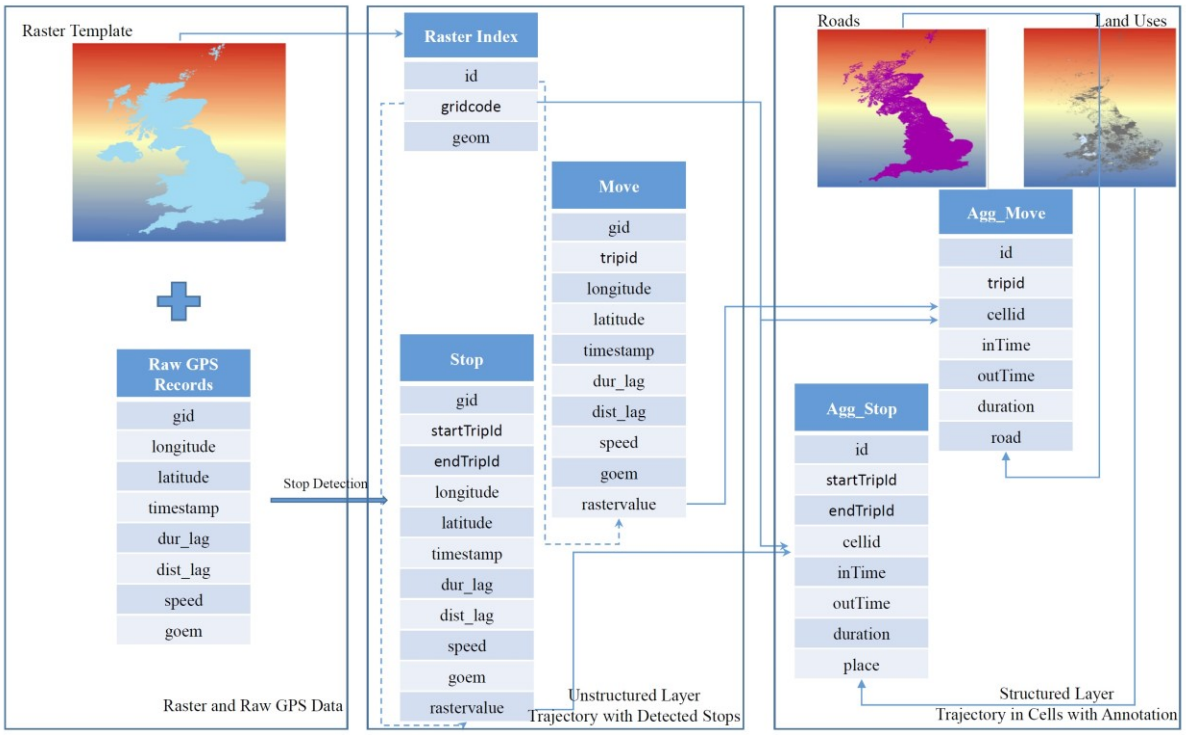

Figure 1. Overall stop/move trajectory data structure based on raster based stop detection and segmentation (Northern Ireland not included in the annotation process).

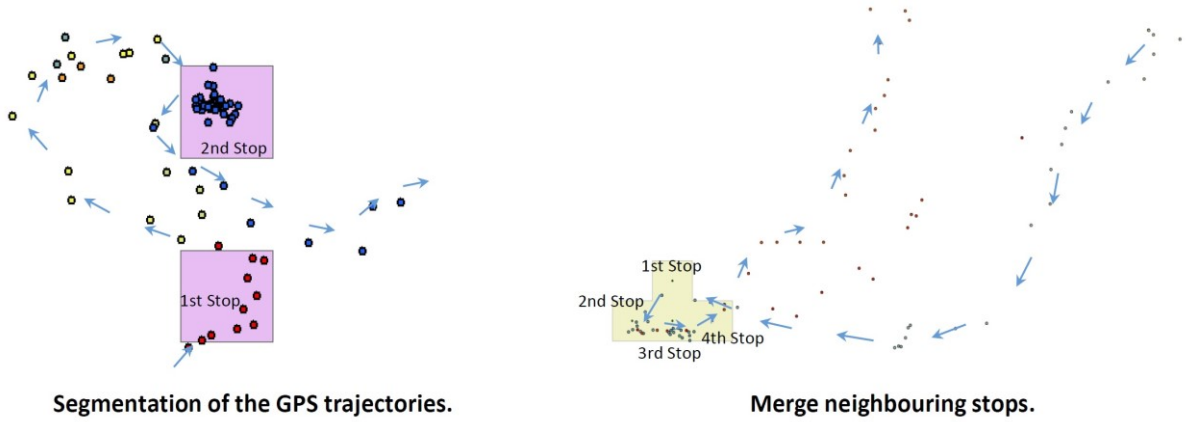

Figure 2. Illustrations of major post-processing functions.

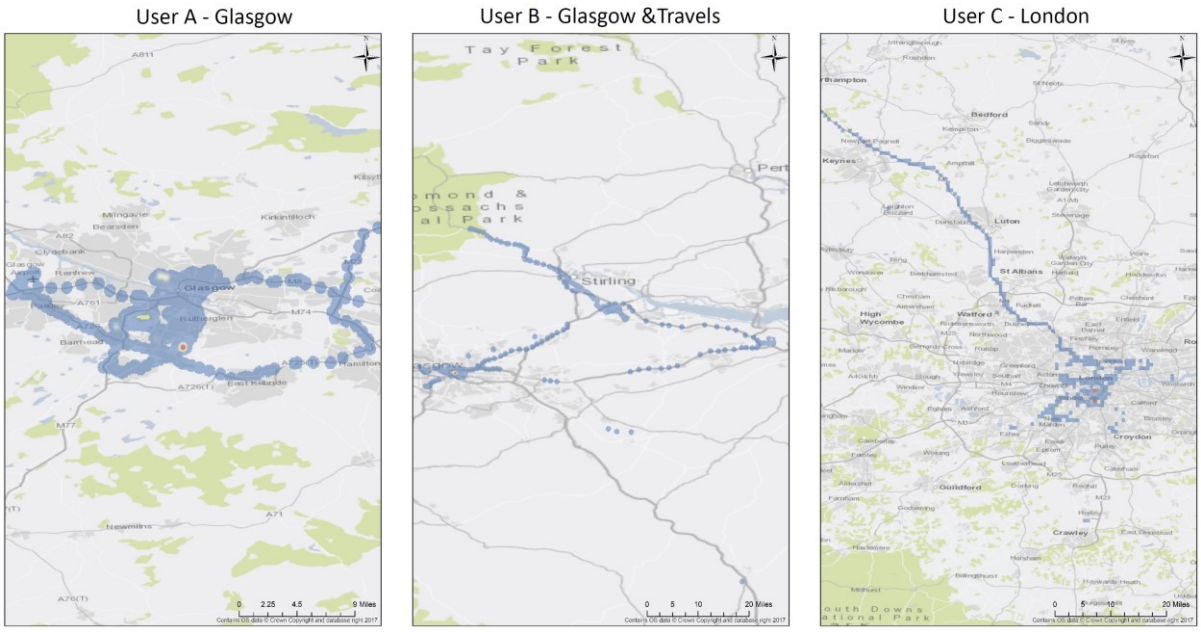

Figure 3. Kernel Density Estimation (KDE) surface of the three use cases.

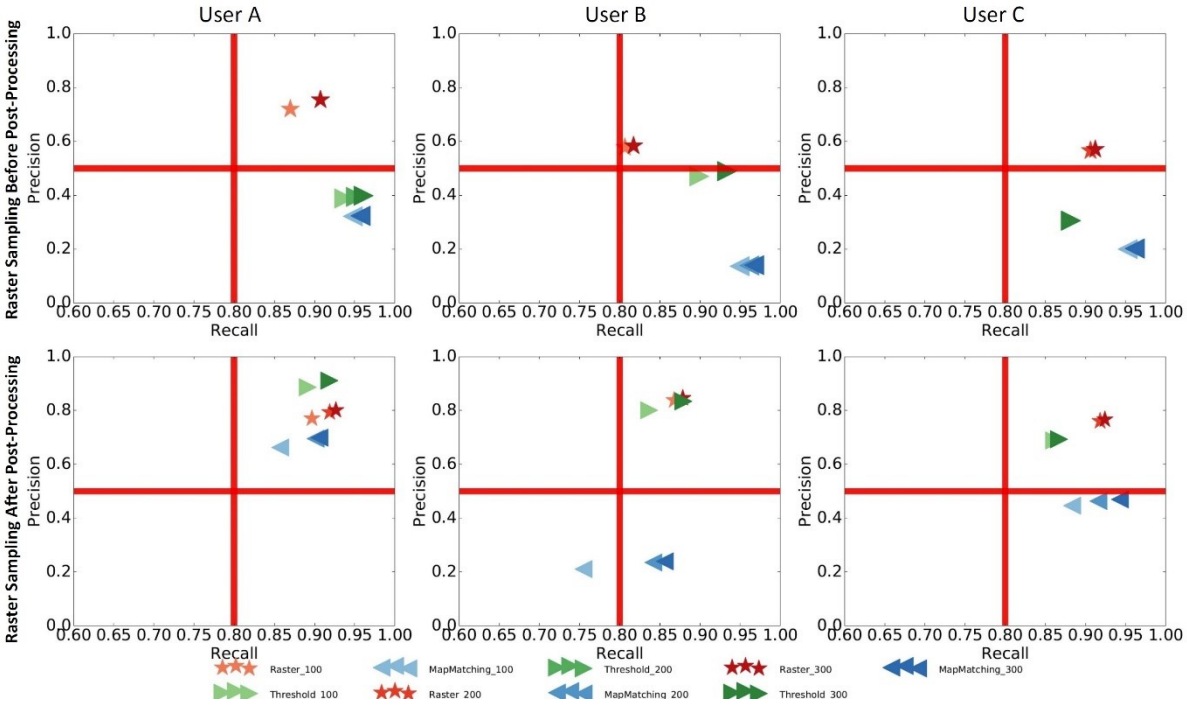

Figure 4. Precision/Recall plot for raster, threshold and map matching stop detection methods.

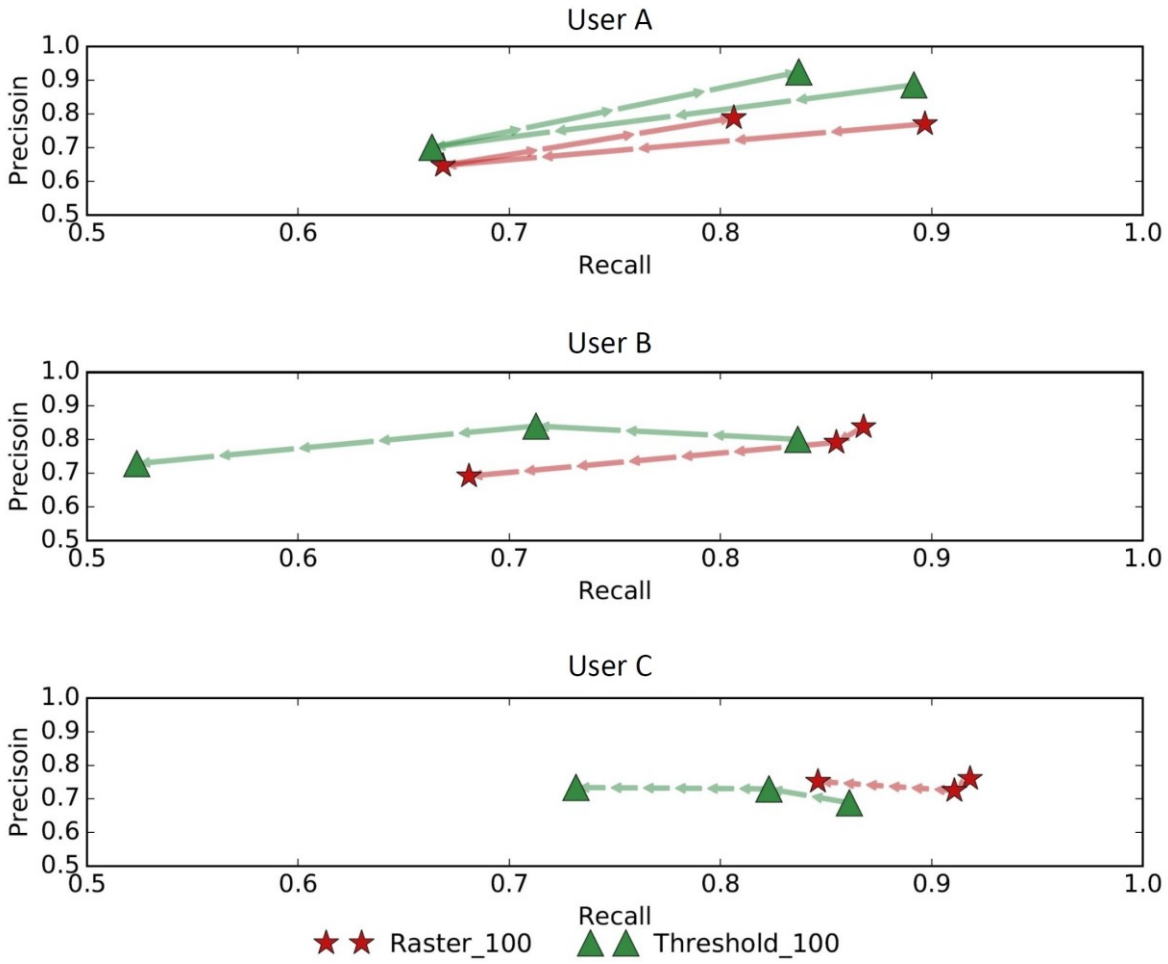

Figure 5. A comparison of precision/recall rates for the raster sampling method and threshold method with different cell sizes (0.00091 decimal degrees under WGS_1984) increased by a factor of 1.5 and 2 (roughly 60, 85 and 110 meters).

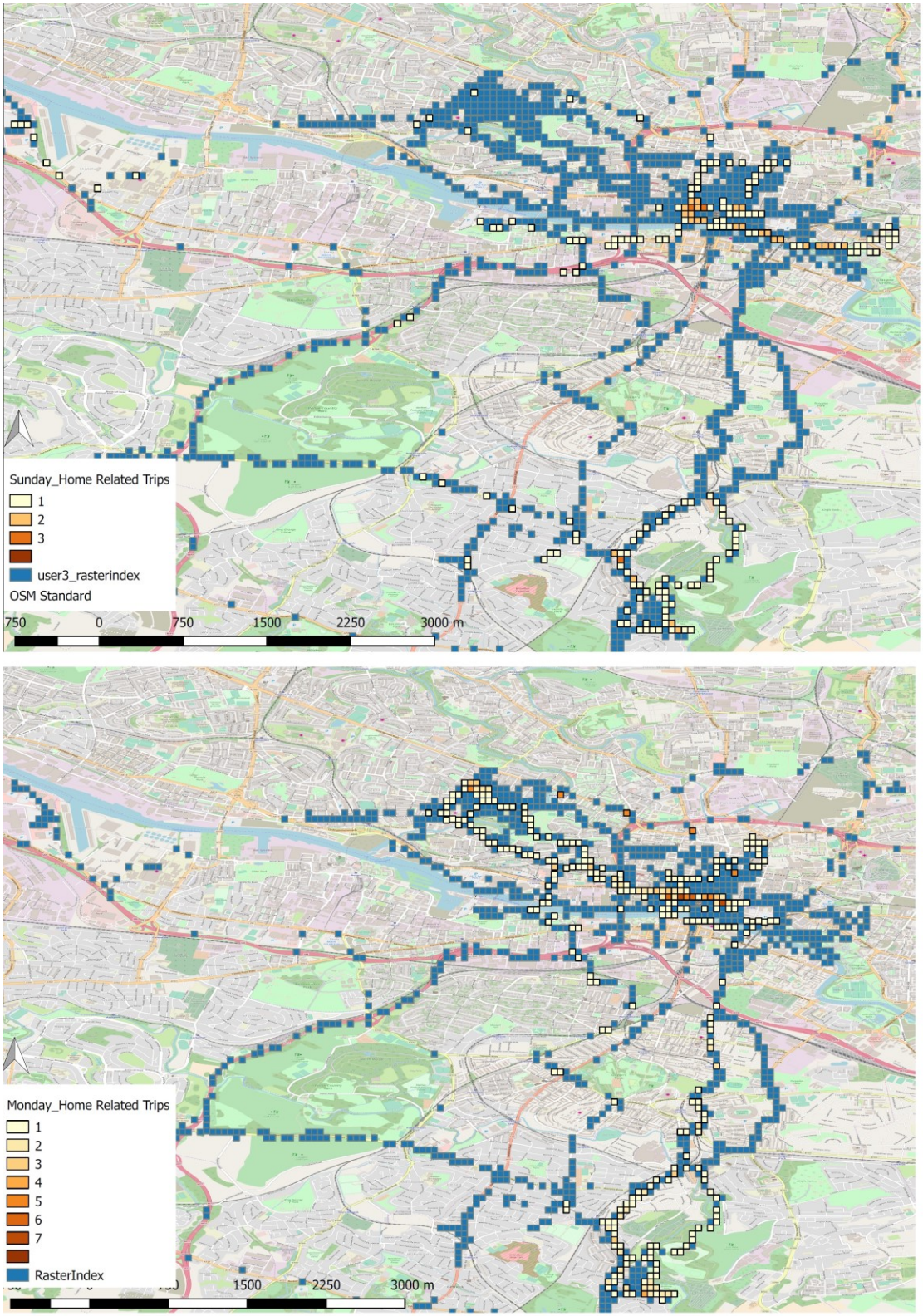

Figure 6. Example of extracting home related trip counts for Glasgow area on Sundays (Top) and Mondays (Bottom) symbolized in trip counts against the whole raster coverage (RasterIndex) in blue.

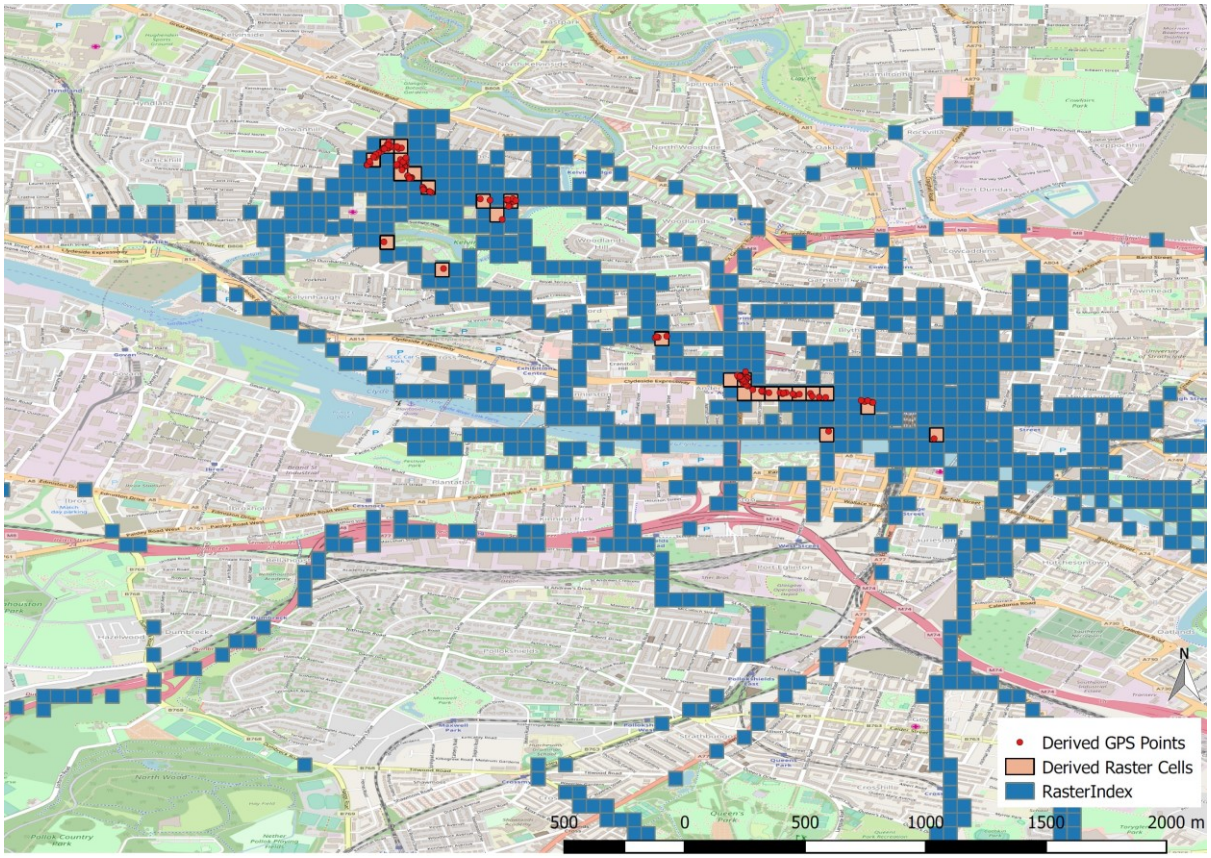

Figure 7. Shared ‘move’ cells confirming to 2-anonymity in Glasgow within 17:00-18:00 time window for use cases against the whole raster coverage (RasterIndex) in blue.

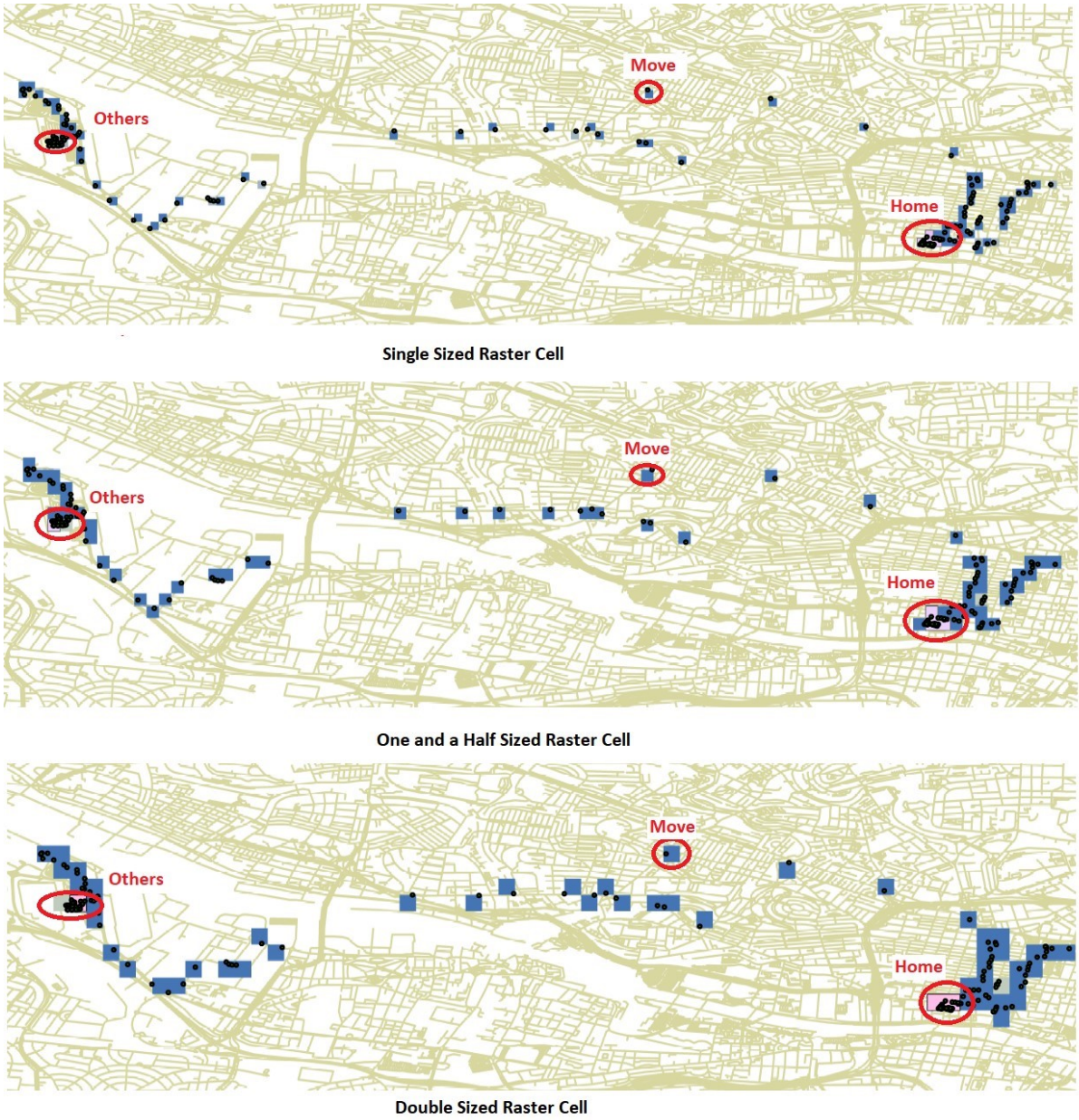

Figure 8. Grid-masking with different cell sizes.

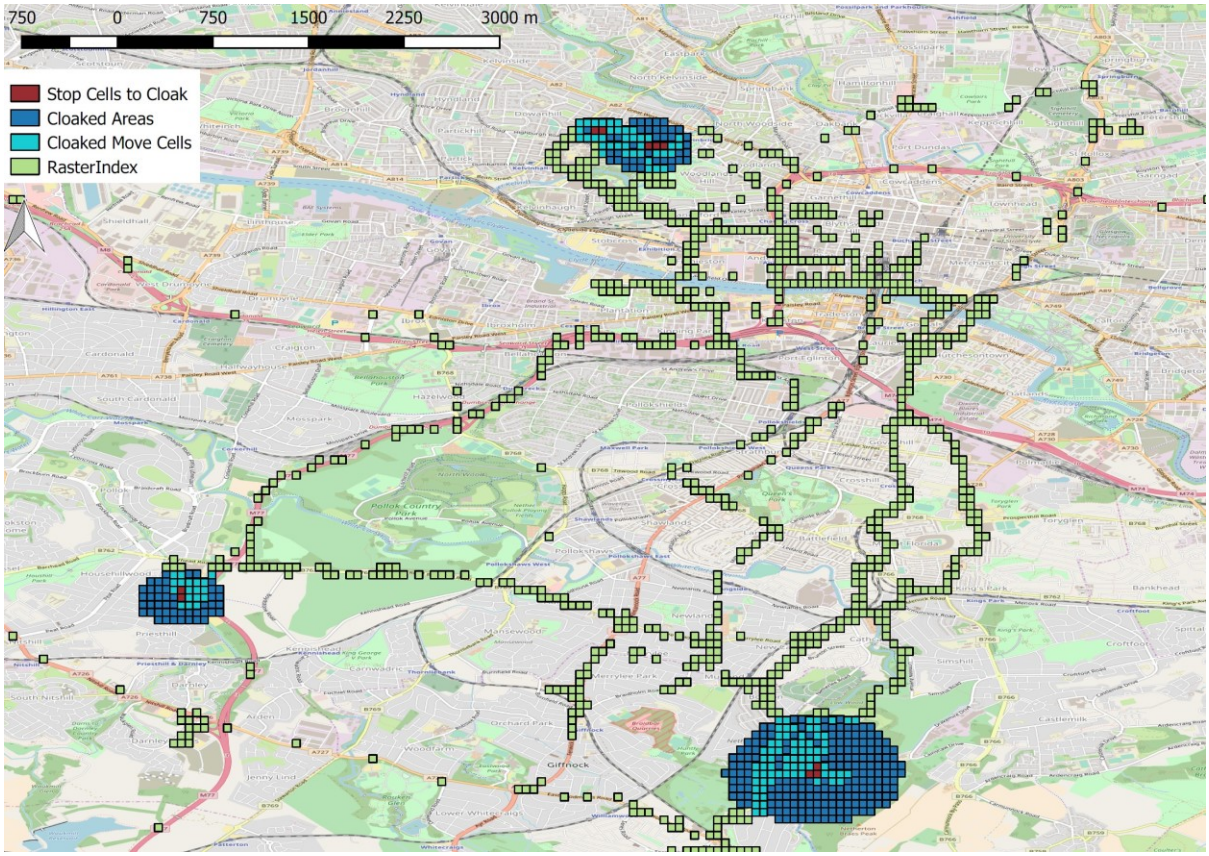

Figure 9. Spatial cloaking frequently visited stops illustrated in User A example in Glasgow area.

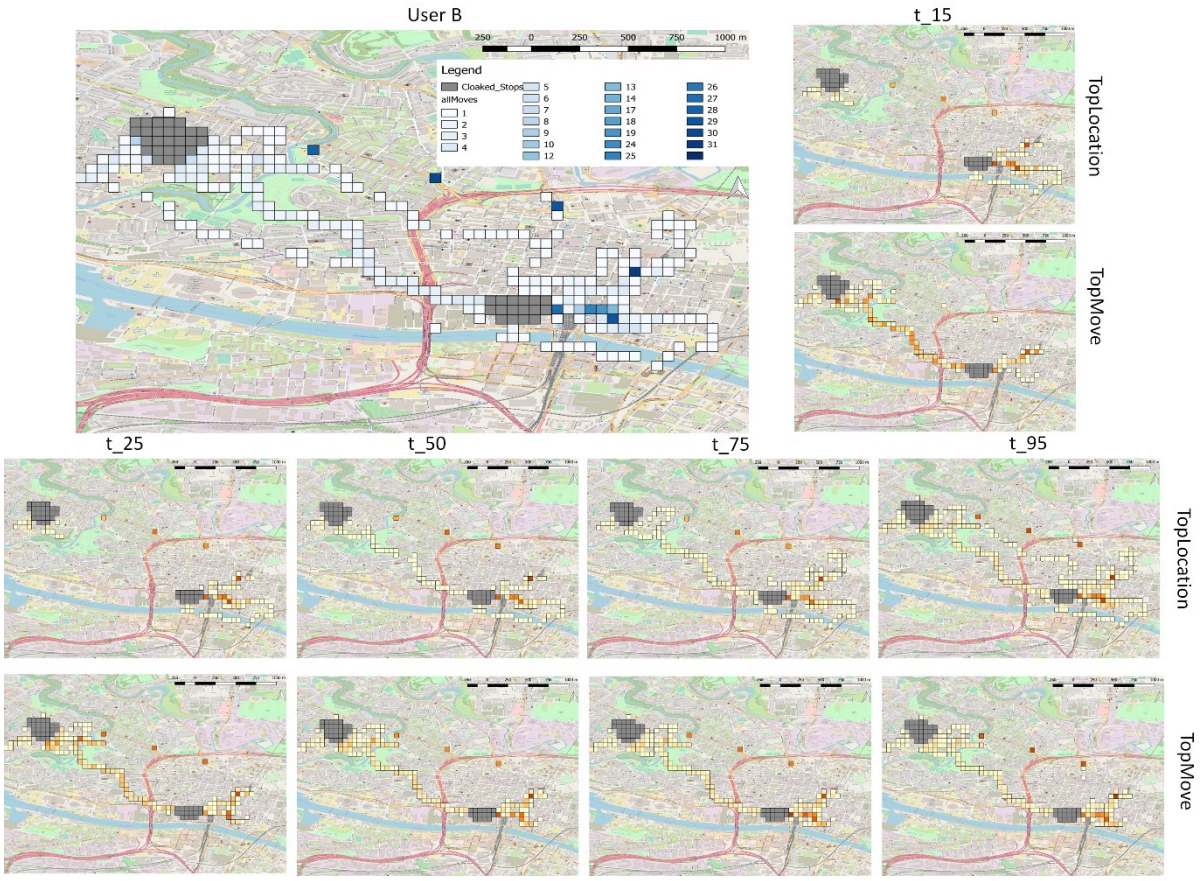

Figure 10. Cloaking results for the two ‘moves’ cloaking methods combined with ‘stop’ cloaking in use cases b for different values of t

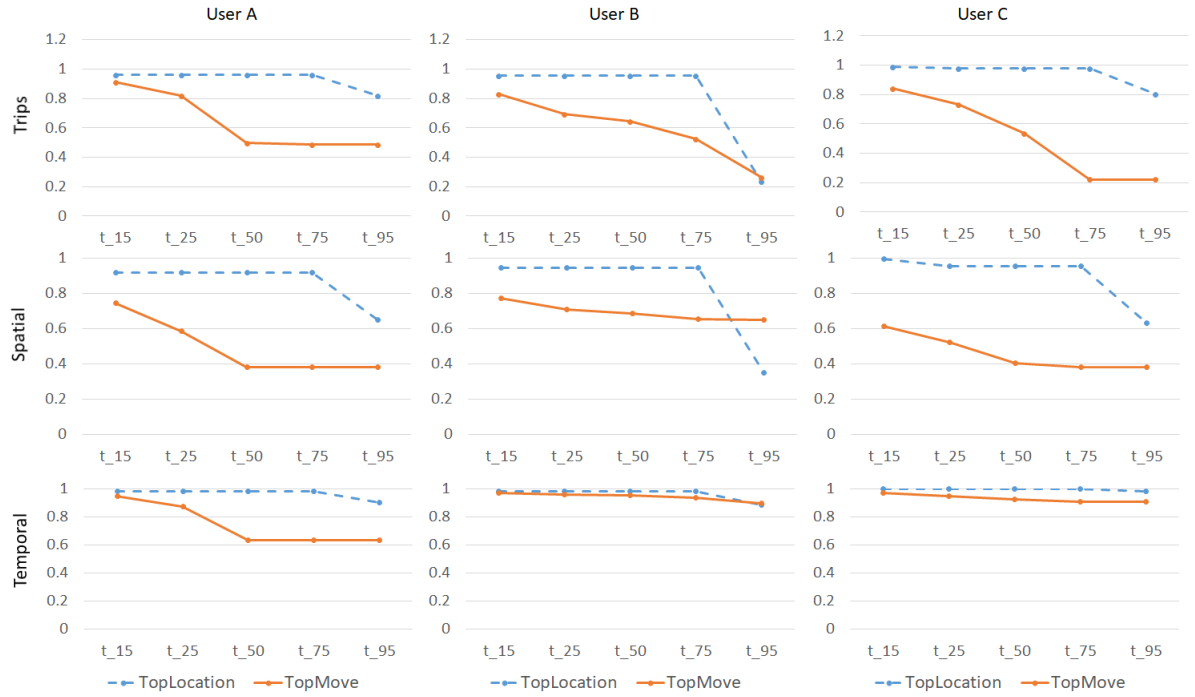

Figure 11. Information loss for TopLocation and TopMove anonymization with different *t* from trip, spatial and temporal aspects.

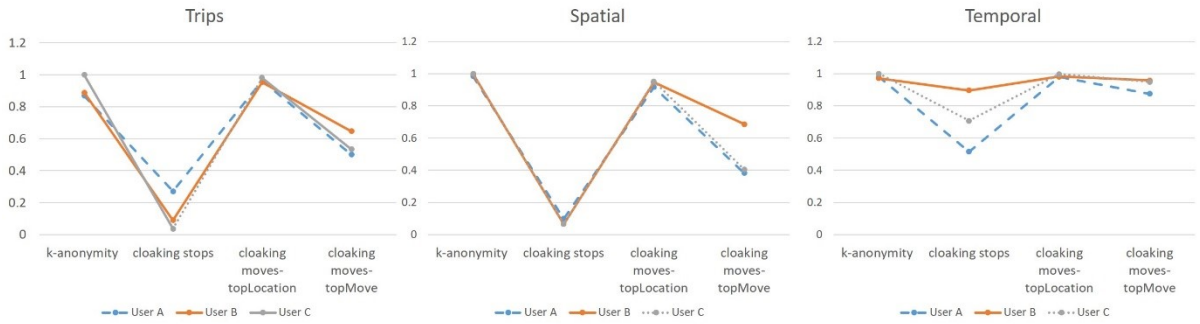

Figure 12. Overall information loss comparing demonstrated methods from trip, spatial and temporal aspects.