THE TRADE-OFF BETWEEN PRIVACY AND GEOGRAPHIC DATA RESOLUTION. A CASE OF GPS TRAJECTORIES COMBINED WITH THE SOCIAL SURVEY RESULTS

K. Sila-Nowicka[a]*, P. Thakuriah[a]

[a] Urban Big Data Centre, University of Glasgow, Glasgow, G12 8RZ, United Kingdom
katarzyna.sila-nowicka, piyushimita.thakuriah@glasgow.ac.uk

Commission II, WG II/8

KEY WORDS: GPS trajectories, privacy, locational privacy, geoprivacy, anonymization, survey data

ABSTRACT:

Trajectory datasets are being generated in great volumes due to high levels of Global Positioning System (GPS) and Location-Based Services (LBS) use. Such data are increasingly being collected for a variety of academic, industrial and recreational reasons, sometimes together with other strands of personal data such as socio-demographic, social survey and other sensor data carried/worn by the person. In such cases, not only are movement data of a person available but also data on potentially a wide variety of other personal and household attributes. Making such person-level data available for analytics opens up the possibility of new directions in analysing, studying and understanding human behaviour, which is typically not possible with GPS trajectory datasets alone. At the same time, the GPS data should be released in a privacy-preserving way that takes into account the possibility of re-identification of individuals from quasi-identifiers available from other data strands. De-identification in these strands may be risked due to uniquely identifiable information on significant locations and other spatial behaviours and choices detected from GPS trajectories. Using a multimodal dataset that includes a GPS archive from 358 individuals, and by considering a number of alternative privacy-enhancing approaches, we look at the potential for privacy preservation when personally-identifiable data are available from multiple data strands, for the specific purpose of data to be released for transport research.

1. INTRODUCTION

GPS movement data have stimulated a great deal of research and development interest due to significant proliferation of mobile devices worldwide leading to voluminous amounts of mobility data that support smart location-based apps and Web 2.0 applications. Individual trajectory data have been increasingly collected for recreational and academic purposes and the possibility of linking an individual?s mobility data with additional sources of information potentially leads to being able to identify the locations visited by the individual (Seidl et al., 2015). Privacy of location also called geoprivacy or locational privacy is a person's right to protect his or her location information from disclosure, or to determine the extent to which the data can be shared (Duckham and Kulik, 2006; Cottrill and Thakuriah, 2015). Publishing and exploring such data is essential to improve transportation services and to better understand the dynamics of urban areas and regional economies.

This paper is motivated by two main considerations regarding the anonymisation of GPS trajectory data for publication. First, in the case of mobility services, the private information possessed by a traveler potentially comprises their identity, current location, origin and destination of travel, journey time, locational preferences and so on, requiring locational privacy preserving solutions from technological, legal, consumer-awareness building perspectives (Thakuriah and Geers, 2013). Among technological solutions, a topical approach utilised is to have privacy concepts built into physical and software systems as well as business processes from the ground up using Privacy Enhancing Technologies; however, commonly used approaches utilise pseudonyms, location accuracy degradation using geographical and temporal masking or cloaking, and various forms of encryption. However, the quality of specific privacy preserving approaches are use-dependent; as GPS data are obtained from on-line and off-line sources, different protection methods have to be designed for different purposes or end uses.

Our first objective is to examine these "fit-for-purpose" issues where different end uses require different levels of anonymisation and aggregation of the GPS trajectories. For example, for transport planning purposes we would need information about the movement on particular road network links; for the choice modelling we would need detailed and accurate information on the movements of individuals. We consider these two specific end uses of GPS trajectory data in transport research, i.e., road network-based analysis such as network models, and choice modelling, and examine the performance of alternative anonymisation methods for preventing privacy loss of GPS locations. We use a complex data source, the Integrated Multimedia City Data (iMCD) which includes a GPS dataset, for this purpose.

We examine approaches such as obfuscation of point data (geographic masking techniques), or generalisation (KDE-based methods, road-link assignment). Our goal is to examine correlation coefficients of spatial patterns between original and anonymised data and to identify approaches with which spatial patterns can be recreated as accurately as possible. To this end we use Pearsons correlation analysis (with adjustments made to the road network proximity and similarity between trips derived from GPS trajectories) to investigate the correlation between spatial patterns discovered from original and anonymised data.

The second consideration is that new challenges to privacy preservation are introduced in data-rich environments when there is linkage of information on the same individual from different data sources. This would require a comprehensive approach to locational privacy preservation in GPS data. Such an approach would require consideration of potentially multiple data strands or datasets in anonymising GPS data, rather than determining pri-

*Corresponding author

vacy preservation solutions based exclusively on the GPS data. This problem arises due to the presence of "quasi-identifiers" (e.g., age, gender, professional status, occupational category, size of family, ethnicity) that may work together to uniquely re-identify anonymised locations in GPS data such as home, work, places with significant stays and other points of interest. We make a preliminary effort in this paper to note when such issues may arise. Specifically, we address privacy issues regarding the identification of individuals in GPS trajectory datasets when these are linked to the external attributes obtained from social surveys as simple removing uniquely identifying information (ID, name) from the released data is not sufficient to prevent identification.

The remainder of this paper is structured as follows. Section 2 reviews existing anonymisation strategies, especially for the static GPS movement datasets. Section 3 provides details on a case study and the data used for this research. Section 4 shows the proposed methodology used for GPS trajectories anonymisation when the GPS data ought to be released to researchers together with an associated social survey data. Section 5 discusses the results and concludes the paper.

2. RELATAD WORK

In the recent years anonymisation of GPS traces has attracted significant attention due to the increasing use of location-aware devices such as smartphones or GPS trackers. This led to enormous increase in the volume of collected spatio-temporal data about individuals (Goel et al., 2012). Several approaches have been proposed to tackle the problem in a data publication perspective, in off-line manner, while the majority have been designed in the context of location-based services (Zheng and Zhou, 2011). According to Zheng and Zhou (2011), most of the privacy issues related to the GPS trajectories are connected with two types of LBS. These are either for collecting location data about the individuals (snapshots where the location information is accessed only when a user accesses the service) or continuous location data stream when a mobile user provides the location data. Privacy-preserving techniques for LBS (on-line) can be categorised into three groups:

- false locations
- space transformation
- spatial cloaking

The problem of location privacy has been well studied in the context of location-based services (Rossi et al., 2015; Olumofin and Tysowski, 2010; Gruteser and Grunwald, 2003; Ghinita and Damiani, 2009; Damiani et al., 2010; Hwang et al., 2014; Ghinita and Damiani, 2009; Ghinita, 2013; Jin et al., 2010), but mainly with a focus on on-line, service-focused anonymity. In this paper, we consider off-line and data-focused anonymity, as in the context of data publishing. Privacy-preserving techniques for trajectory publishing (off-line) vary from the ones for querying in LBS and can be categorised either into two or three groups. Two classifications are presented below. First way of classifying anonymisation methods suggested by Bonchi et al. (2011), divides these methods into: motion-patterns based methods and location-based methods. Where the first one investigates how potential attacker can predict locations of individuals, based on their mobility patterns, whereas the latter one aims to prevent from identifying users' sensitive locations (Abul et al., 2008; Monreale and Andrienko, 2010; Andrienko and Andrienko, 2009). The second classification looks at the methods from more technical point of view where these methods are divided, based on the techniques used to protect geoprivacy of the individuals:

- clustering-based (also called microaggregation)
- generalization- and grid-based (obfuscation methods: grid masking, random perturbation (in so called geomasking) and density based generalisation approaches)
- others such as suppression and condensation or space translation

Following the structure of the three groups of methods, related to the anonymisation, we first briefly present some examples of clustering-based methods. Then examples using generalisation and grid-based approaches. The last two parts of the review, concern examples of the suppression methods, and problems of adding quasi-identifiers or additional datasets to the GPS traces respectively.

The clustering-based approach utilizes the uncertainty of trajectory data to group k located in the similar space trajectories within the same time period to form so-called a k-anonymized aggregate trajectory Abul et al. (2008, 2010). Their approach *Never Walk Alone - NWA* defines an uncertain trajectory as a cylindrical volume with a defined radius, where two trajectories ale co-localised (Euclidean distance between them is less then designed radius). NWA anonymizes a given set of trajectories in three steps: preprocessing, clustering, and space translation. The disadvantage of this method is that it assumes that everything starts in the same time so it is more problematic to implement it for the data, as the trajectories certainly are from different time slots (Abul et al., 2008). Huo et al. (2012) propose to depersonalize only locations on the trajectories of individuals instead of standard way of anonymising a whole trajectory (*You Can Walk Alone (YCWA)*). The idea of YCWA, is to split trajectories into move, stay sequences and generalise each stop into a territory based on a generated split map. We would argue that the method presented by Huo et al. (2012), seems to be the most advanced from the clustering methods for anonymization purposes. They treat all the stop locations with an identical importance rather than weight them based on the significance or sensitivity of a place for individuals. We suggest to improve this part by using significant locations only, rather than all the available stops. As stated in (Golle and Partridge, 2009), the uniqueness of home and work pairs of locations can reveal the privacy of an individual - whereas unmasking 100 stops on the traffic lights during the journeys of an individual would not affect the privacy-related issues and could improve the dataset potentially released for transport studies.

There are various combination works that combine clustering methods with generalisation methods such as in the example presented by Monreale and Andrienko (2010). In their method they provide a very high level of privacy protection generalising the data to the point that it can be suitable at only to some kinds of analysis. In particular to analyse the flows between areas and to calculate basic statistics of the visits to these areas.

The next group of methods defined as a generalisation-based covers obfuscation methods such as grid masking, random perturbation (so-called geomasking) and also density based generalisation approaches. Random perturbation approaches for GPS anonymisation, are the ones which involve relocating the collected GPS points to preserve privacy of data subjects while maintaining the quality of data. There are various perturbation methods, listed in Figure 1 and they are mainly used in health research (Shi, X.

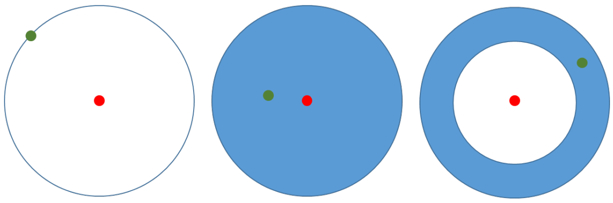

Figure 1: An example of geomasking methods, where the first from the left is a random directional and fixed radius masking way, the middle one is a random perturbation within a circle, and the last one represents donut masking.

Alford-Teaster and Onega, 2009; Hoh et al., 2008; Hampton et al., 2010; Seidl et al., 2015).

Armstrong et al. (1999) introduce geomasking called then geographic masking techniques as the techniques with a high potential to improve the resolution of data published while still protecting privacy of individuals. Grid-masking is based on overlaying a grid of cells over the GPS trajectory, so each point is either snapped or transformed into a corresponding grid cell (Leitner and Curtis, 2006; Krumm, 2007; Seidl et al., 2015).

Another generalisation-based method was presented by (Poulis and Skiadopoulos, 2013) where they defined a sequence of locations as OID (quasi-identifier - as there is a double meaning of the quasi-identifiers in this paper we will use for the sequence of locations just the abbreviation to avoid misunderstanding), where they follow select-organize-anonymise paradigm to anonymise the data. The methods accounting for the sequences of points in many GPS trajectorie or individual ones are mainly based on clustering algorithms (Terrovitis and Mamoulis, 2008; Yarovoy et al., 2009). In generalisation-based methods there are mainly approaches that are blind for the sequences in the movement data (Nergiz et al., 2008; Monreale and Andrienko, 2010).

Suppression methods and space translation are usually used for economic censuses, where data reported for an areal unit with multiple establishments must be suppressed (confidential locations) to at least five locations when placed on maps (Leitner and Curtis, 2006).

Often using additional external information may help to identify individuals much easier than without this information. Having said so, we do not think here about only the private information about individuals but also just a standard GIS layer such as land use or Point of Interest database. Thanks to which significant places for individuals may reveal more individual's characteristics. Using data captured from the vehicles may show that fine-grained location traces reveal speed distribution and acceleration patterns that can be used to distinguish traces from different vehicle types (e.g., trucks and cars). Analysis of Zan et al. (2013) show that a type of a vehicle (car/van/track) can be identified with the accuracy of 96%. The authors show that it is possible to identify outlier driving patterns such as higher speed, which could be used to link anonymous segments of location traces and eventually recover complete trips. These examples shows the increased risks for an identification when information not related to individuals but available publicly is taken into consideration.

Having summarized these methods, that have been used in past to anonymise the static off-line GPS data, we have realised that there is no real review showing the use of the anonymous data and the extent of the lost information according to the purpose it may be used for. Further worth of mentioning is the fact spotted by (Nergiz et al., 2008) related to k-anonymity when sensitive information is present, stating that k-anonymity does not necessarily prevent the disclosure of the sensitive information. We provided an overview of research effort concerning how to anonymise a moving objects database. While only few papers have been published so far on this problem, much large body of work has been developed for location privacy in the on-line, dynamic context of location based services.

3. DATA AND CASE STUDY

Our location data come from the integrated Multimedia City Data (iMCD) project conducted in Glasgow by the Urban Big Data Cetnre, University of Glasgow, UK. It is a project where the 2095 participants of 1505 households in the Greater Glasgow Area took part in the social survey and 400 of them took part in the sensor project where they were carrying a GPS tracker recording locations every 5 seconds (TranSystem 747PorS) for a week. Furthermore the participant who took part in the sensor project were also asked to deliver an activity diary from the first full day of the study, to collect the data that could be treated as a ground truth in algorithms development.

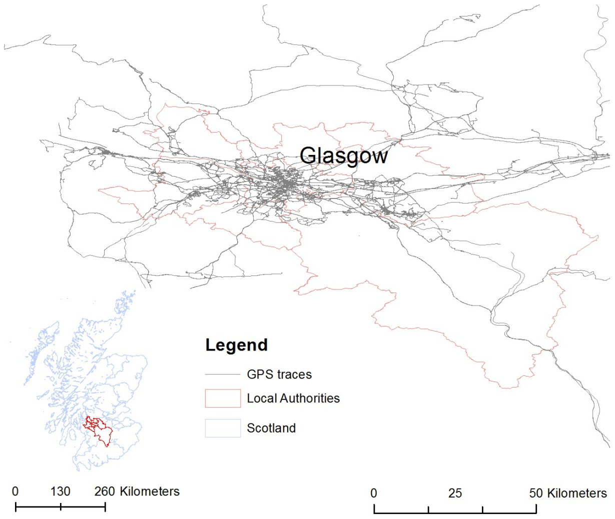

Figure 2: Study area of the iMCD project covered Glasgow City and six surrounding local authorities. The map presents the coverage of the study with GPS footprints of 358 users from whom we had obtained the valid data.

Our additional data come from the social behavioural survey for the iMCD project. With this dataset we are able to add to the GPS trajectories various attributes describing individuals based on their responses to the survey's questions. The specific aims of the survey were to collect data on:

- family type, household income etc. to allow understanding of the socio-economic background of the household;
- individual's patterns of activity and mobility;
- values, attitudes and perceptions on a range of topics relating to behaviours and daily activity;
- education and learning
- informal competencies, like financial understanding, IT skills etc.

4. OUR APPROACH

Our line of research is about developing ad-hoc anonymisation techniques for the intended use of the data: for instance, with respect to a specific spatio-temporal analysis such as transport planning. We explore various methods for GPS data anonymisation and try to explain how the social survey responses may affect the privacy protection. Our main strength lies in anonymising, mainly significant locations to leave the rest of the data for the transportation planning purposes, traffic analysis etc. As stated in (Bonchi et al., 2011), most of the anonymisation methods are characterised with a huge of information loss once the data satisfy some concept of anonymity.

From the questions asked in the survey questionnaire, in the discussed iMCD project, we identified questions which could be used as the potential quasi-identifiers and increase the risk for re-identification of the individuals. Questions such as: about demographics, ethnicity or religious believes may not be disclosive from the aggregated social survey itself, but if connected to the GPS trajectories would reveal all the mobility behaviour of this particular person. In the table below we present a set of quasi-identifiers from the conducted social survey, which could affect the privacy levels. We divide these identifiers into two groups: spatial and non-spatial.

Table 1: Potential quasi-identifiers from the iMCD social survey.

Non-spatial quasi-identifiers	Spatial quasi-identifiers
Ethnicity	Kid's schools names
Income	Use of public transport
Language course	Way of travel to work
Employment and support allowance	Pet ownership
Sexual orientation	Walking with pet
Job description	
Disability	

Further in this section, we provide the information on how to process classify and anonymise GPS trajectories. First though we would want to draw the attention of the reader, to possible risks for divulging the privacy when releasing the data in conjunction with a social survey. With the potential identifiers listed in Table 1, few possible scenarios in which these variables could help to re-identify the survey's respondents, are presented. With the detailed information about ethnicity (e.g.Irish, Gypsy / Traveller, Polish, Bangladeshi or Scottish Caribbean) and the daily patterns of these individuals uncovered from GPS trajectories (visiting for example Polish grocery shops or the Polish Community house), we could with a high probability identify a certain individual. Furthermore the information about an evening language course, sexual orientation or employment and support allowance, could reveal certain individuals' identity, if they have visited a foreign language centre, a job centre or a gay club during the survey duration (language courses could be on weekly bases, and to get a job seeker allowance, a person needs to visit the centre at least every two weeks). A particular job description, such as a driver or a postal worker, when combined with the classified GPS trajectories, may easily reveal the identity of an individual, as there are ways of inferring from raw GPS traces some occupations, based on the possible movement patterns. Moreover, declaring the disability in the survey may be detected from the traces when a potential survey respondent uses, for example wheelchair ramps. These are non-spatial potential quasi-identifiers which could increase the risk of re-identification of the survey respondents. The second group of identifiers is more related to the space as contains information about particular locations. School locations may be identified from GPS trajectories, therefore if a parent gives a lift to a kid and leaves a GPS footprint which could help to re-identify this particular respondent. Another good example would be a question in which the respondents needed to declare whether they have a pet and whether they regularly walk with it. If they do it can be identified from the GPS trajectories and the person could be identified.

According to Huo et al. (2012) collection and publication of people's everyday trajectories pose serious threats on people's personal privacy. Collecting these GPS traces with a social survey increases the disclosure process significantly. There are many questions that reveal individuals privacy when combined with GPS trajectories therefore we suggest excluding from GPS trajectories stop locations that are not classified as significant but cover sensitive locations such as a nursery, school or church.

To proceed with the anonymisation process we first processed GPS movement data using the algorithms developed by Sila-Nowicka et al. (2015).

- GPS processing
- Identification of significant locations
- Spatial cloaking of significant locations - various ways to prevent identification based on the built environment

The data were classified into adequate travel modes and the trip purposes were detected and classified as individuals home and significant locations need to be anonymised to protect the privacy of an individual. The main advantage of our work that we account for the significance of the stay points where they are important and more sensitive than ordinary location samples. As mentioned above, we divide our methods according to the purpose they may be used for. If used for transport planning purposes, we suggest using either the first method (spatial cloaking) we describe or the third (GPS road link counts). For urban planning purposes we can use all the possible methods that preserve the information about densities and spatial patterns. For spatial choice modelling where the most detailed and restrictive data would be required we would suggest to provide instead of location, trip chaining dataset with attributes defining the surrounding in a geographical and socio-demographical way.

We first propose to use one of the geomasking methods. As we are restrained to a road network, we delete significant locations from the GPS trajectories, instead of changing and randomly assigning the location. Therefore, we obtain a new dataset in which all the movements are presented as the raw (or classified into travel modes and activity locations) data and just the network-based area around the significant locations is obfuscated. The classified data are presented on Figure 3 and the the geomasked dataset is presented in Figure 4.

In the next step we generalised GPS trajectories using Kernel Density Estimation method. This way we show the spatial pattern of individuals, rather than reveal their full set of GPS locations. The KDE map is constructed based on the anonymised trajectory from Figure 3. To investigate how accurate the spatial pattern of the data is preserved, we calculate Pearson's correlation coefficients between unmasked and masked data. For the classified GPS trajectories we calculate separate coefficients for all the derived from raw GPS data, segments with classified travel modes and the results are listed in Table 2.

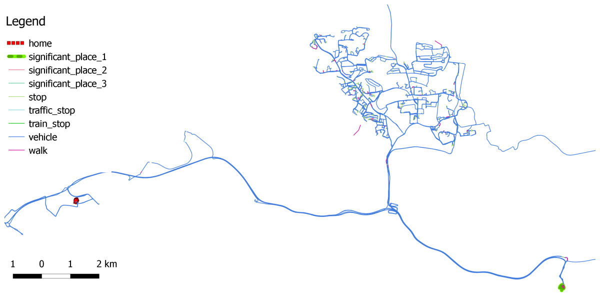

Figure 3: Classified GPS trajectory of an individual who carried a GPS tracker for a week.

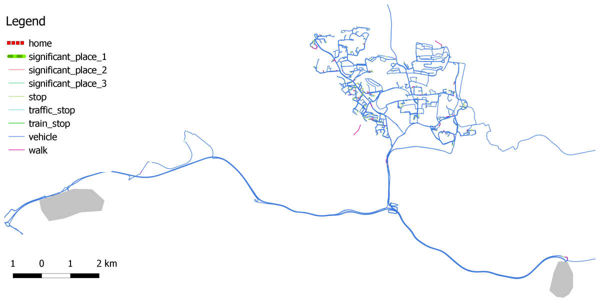

Figure 4: Spatial cloaking - GPS points inside the boundary created from 300-500 meters road network area around the significant locations (plus possible connection to the main road) to preserve privacy.

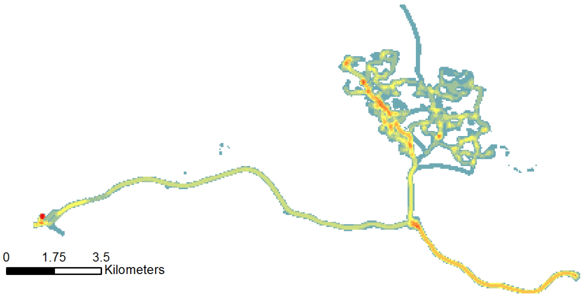

Figure 5: Trajectory generalisation - KDE for line segments derived from GPS trjectories. Kernel bandwidth 100m for the cell 50m. This example is derived from the anonymised GPS trace where all the significant locations were cloaked.

Table 2: Pearson's correlation coefficients for various travel modes. The results are based on the sample of 5 individuals.

Segments of GPS trajectories classified into travel modes	Pearson's correlation coefficient
Driving	0.942
Walking	0.852
Bus	0.938
Train	1.000
Stops	0.954
Traffic stops	0.982
Home	0.000
Significant places	0.000

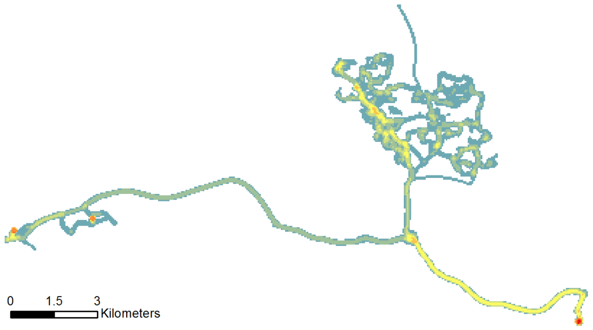

Figure 6: Trajectory generalisation - KDE for line segments derived from GPS trjectories. Kernel bandwidth 100m for the cell 50m. Trajectory without anonymised significant locations.

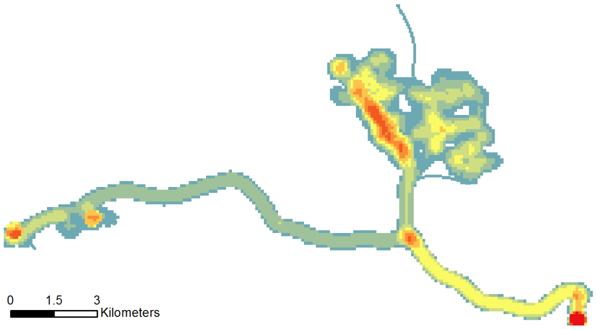

Figure 7: Trajectory generalisation - KDE for line segments derived from GPS trajectories. Kernel bandwidth 300m and the cell 100m. Similarly to the Figure above, this generalisation is prepared from the raw dataset for one individual user, therefore significant locations which were not eliminated can be identified. The resolution of the map may preserve from identification of these places, nevertheless in our opinion they should be removed before.

To increase the privacy protection we could use KDE to generalise the previously cloaked GPS trajectories. An example is presented in Figure 5.

Different Kernel bandwidths may result in the different anonymisation effects. The bigger the bandwith, the less unique locations can be identified. The example of these two approaches is presented in Figure 6 and Figure 7. In comparison to 6, Figure 5 shows, how cloaking the significant locations increases the visual importance of other locations in individuals' daily routines.

The next approach uses GPS trajectories to assign them to the nearest road links, after they have been map-matched. This method is similar to the one presented by Lin et al. (2009), where they replace specific locations with a road ID, moving object ID and the direction. Presentation like is sufficient to derive trajectories or traffic flow information. In our case we first cloaked all the sensitive locations. Then, to prevent from the re-identification of individuals we publish only values for the road-links where, more than two separate users conducted a trip. Figure 8 presents the classified GPS-based movement segments into travel modes and stops. These segments are being map matched to the road segments and then counted. Furthermore information such as speed, movement direction or the average number of stops or traffic stops can be counted and assigned to the road-link (see Figure 9).

The newly created dataset, such as this one, can be used for transportation planning purposes without revealing any substantial private information. The dataset can be created at different levels of aggregation: an aggregated level where we do have only information on the travel counts per road link with a direction attribute; and a disaggregated structure. The latter option provides the information of individuals passing through particular road links at certain times of a day. This method requires though proper spatial cloaking and not revealing the road links where there was only one person on the road.

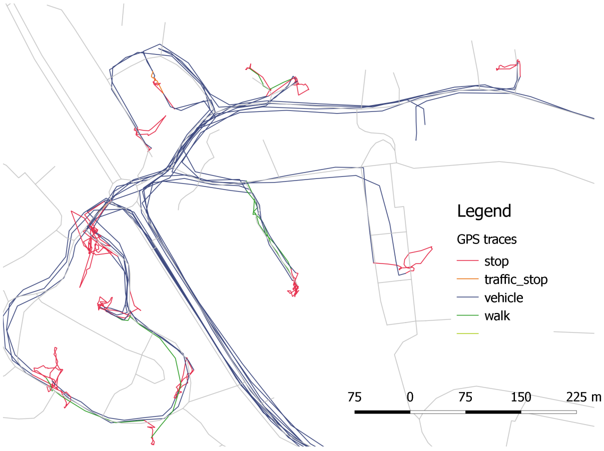

Figure 8: Classified GPS trajectories.

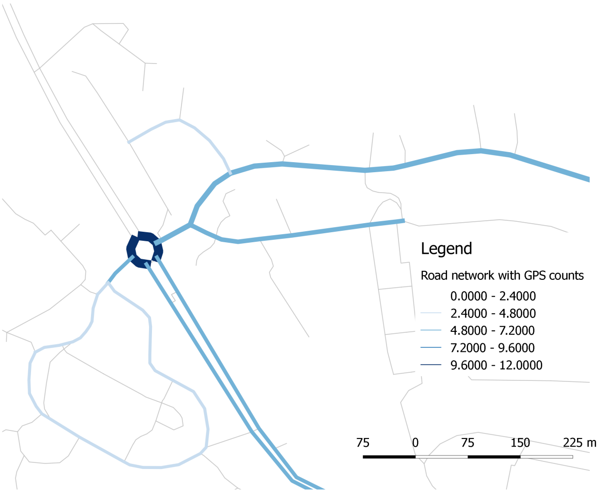

Figure 9: Road link generalisation.

Another discussed approach covers grid-masking algorithm where we show the potential of this method for anonymisation purposes. Authors such as Leitner and Curtis (2006) conclude that there is a threshold cell size, for both privacy and masking, above which larger cell sizes cause the unmasked pattern to be perceived differently. According to Seidl et al. (2015) the best solution would be using a cell size of 30 meters to keep the highest amount of information about human movements. Such a small grid cell could though affect the privacy reliability, therefore usually the cells should be bigger. Shi et al. (2015) and Seidl et al. (2015) analysed the Pearson's correlation coefficients, to detect the similarities in spatial patterns between unmasked and masked trajectory data. They suggested that using a size of a cell equal 100 meters, gives sufficiently good results for anonymisation of the GPS data and preservation of the spatial pattern. We tested the method for our dataset and got similar results. The results from testing the Pearon's correlation coefficients are presented in 3. To calculate the correlations we followed a method presented by Seidl et al. (2015), and calculated the difference between un-masked and masked trajectories. The high correlation coefficient shows the high level of preserved spatial pattern in comparison to the unmasked data. Additionally Leitner and Curtis (2006) noticed, there is a threshold cell size for both privacy and masking, above which larger cell sizes cause the unmasked pattern to be perceived differently.

Table 3: Pearson's correlation coefficients for various methods and on different levels of the accuracy.

Method		
Spatial Cloaking	0.954	0.521
Grid-based	0.812	0.417

The results from Table 3 are based on the Kernel density estimations for two different cell sizes and bandwidths. As expected the results with the smaller grid cells return higher coefficient values signifying better spatial pattern representation.

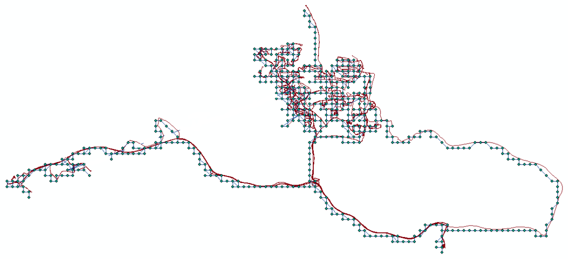

Figure 10: Grid-based masking - points inside 250*250 m grid cells are recalculated as centroids of the particular cells and the new trajectories are derived.

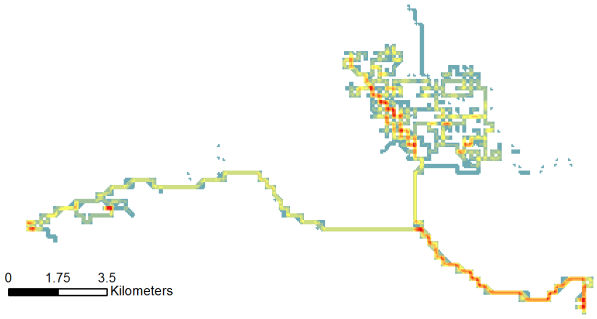

Figure 11: Generalisation of grid-masked trajectories - KDE for the trajectories derived using grid-based-masking.

The Figure 13 shows the grid-based masking algorithm results. This method varies from the standard grid-based algorithms, as there is no information about the movement preserved, only the information about the individuals locations counted per particular grid cell is stored. In case of a small grid cell the spatial cloaking of significant locations is needed, in a bigger grid cell such as 300 - 500 meters there is rather no need for cloaking any information from the trajectories.

5. RESULTS AND CONCLUSIONS

A successful method of GPS data anonynmisation should be the one, which maximizes the the geoprivacy of individuals, allowing

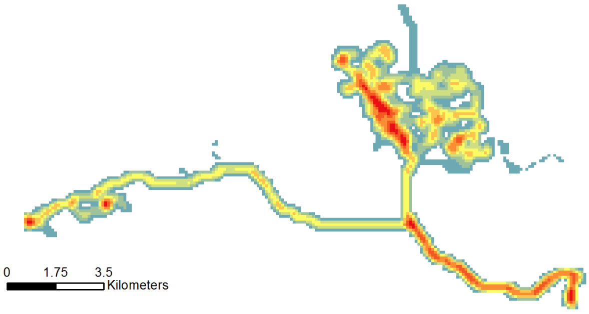

Figure 12: Generalisation of grid-masked trajectories - KDE for the trajectories derived using grid-based-masking.

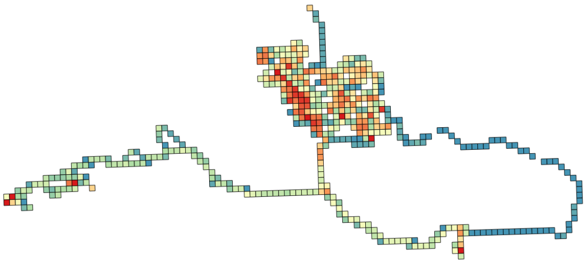

Figure 13: Generalisation of the GPS trajectories - points inside the particular cell are counted and treated as the density of particular grid showing density of movements of individuals.

though on the preservation of spatial patterns of human movement. The balance might be different though for different research questions. In this paper we presented various methods used in the literature along with the suggestions which methods could be used for different research questions. For transport modelling, and transportation studies in general we can use aggregated data which show the road attendance per hour, week or any other assigned period of time. Another option for transportation researchers is to analyse GPS traces as raw as possible, to be able to identify small variations in speed or acceleration to detect or example traffic incidents. Therefore the method of spatial cloaking where all the sensitive or significant locations from an individual's trajectory are eliminated and the rest of the data is classified into travel modes, provides unprecedented and rich dataset for transport studies.

We have claimed to be able to present a trade-off between the data privacy and the resolution of the GPS-based movement data. To this end, we provided a list of possible threats when using unusual quasi-identifiers from social surveys, in conjunction with the anonymised GPS data. All the questions which can be associated with spatial dimension can be treated as potentially dangerous for the privacy preservation. As mentioned before, information about the religion can have a spatial aspect. Therefore, when we aim to release the dataset which combines both: social survey and GPS movement data, we need to pre-analyse it, based on the questions we designed in the survey. In the UK social survey's results are usually published at the local authority level, or at the lower resolution with the restricted access to the data (i.e. research purposes only). This spatial resolution satisfy the anonymisation for survey data but might not prevent re-identification of particular individuals when we have access to the both datasets. Releasing the survey data with anonymised GPS data can be framed into a releasing framework:

- When realising the spatially cloaked trajectories, all the significant or sensitive locations have to be eliminated from the GPS trajectory;
- When realising the data, as the generalised rasters we ought to eliminate sensitive locations as well, to prevent from privacy revealing issues;
- Using road links, as a method to release GPS movements we need to cut road links when there is only one individual moving along this particular road segment, or to cloak the final bit of the trajectory when it reaches a significant location (mainly home and work pairs)
- Grid masking (Figures 10-13) shows to be a valid method for urban planning purposes where the attendance in places is noted and the movement directions can be detected per particular grid cell; depending on the grid cell size, cloaking might be needed. If the cell size is 500 meters there would be a need to probably eliminate only the most important locations for particular individuals instead of all the sensitive locations that could create potential risk of divulging one's geoprivacy.

These are preliminary analysis and apart from the Pearson's correlation coefficients we need to incorporate in the study k - anonymity concept to verify the presented results. This study is exploratory and investigates 'what if' scenarios, in situations when the GPS data are realised not only with simple quasi-identifiers such age or gender but also the ones which are either spatially related or more personal and may increase the identification of individuals.

ACKNOWLEDGEMENTS (OPTIONAL)

The data used for this research are part of the data collection funded under the ESRC grant agreement: ?Urban Big Data? number ES/L011921/1.

References

Abul, O., Bonchi, F. and Nanni, M., 2008. Never walk alone: Uncertainty for anonymity in moving objects databases. Data Engineering, 2008. ICDE . . . .

Abul, O., Bonchi, F. and Nanni, M., 2010. Anonymization of moving objects databases by clustering and perturbation. Information Systems 35(8), pp. 884–910.

Andrienko, G. and Andrienko, N., 2009. Movement data anonymity through generalization. Proceedings of the 2nd . . . .

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

Bonchi, F., Lakshmanan, L. V. and Wang, H. W., 2011. Trajectory anonymity in publishing personal mobility data. ACM SIGKDD Explorations Newsletter 13(1), pp. 30.

Cottrill, C. D. and Thakuriah, P. V., 2015. Location privacy preferences: A survey-based analysis of consumer awareness, trade-off and decision-making. Transportation Research Part C: Emerging Technologies 56, pp. 132 – 148.

Damiani, M., Bertino, E. and Silvestri, C., 2010. The PROBE Framework for the Personalized Cloaking of Private Locations. Transactions on Data Privacy 3(2), pp. 123–148.

Duckham, M. and Kulik, L., 2006. Dynamic and mobile gis: Investigating changes in space and time. CRC Press, chapter Location Privacy and Location-Aware Computing, pp. 35–51.

Ghinita, G., 2013. Privacy For Location Based Services. Synthesis Lectures on Information Security, Privacy, ...

Ghinita, G. and Damiani, M., 2009. Preventing velocity-based linkage attacks in location-aware applications. Proceedings of the 17th ....

Goel, P., Kulik, L. and Kotagiri, R., 2012. Privacy aware trajectory determination in road traffic networks. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12, ACM Press, New York, New York, USA, p. 406.

Golle, P. and Partridge, K., 2009. Pervasive Computing. Lecture Notes in Computer Science, Vol. 5538, Springer Berlin Heidelberg, Berlin, Heidelberg.

Gruteser, M. and Grunwald, D., 2003. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In: Proceedings of the 1st international conference on Mobile systems, applications and services - MobiSys '03, ACM Press, New York, New York, USA, pp. 31–42.

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

Hoh, B., Gruteser, M., Herring, R. and Ban, J., 2008. Virtual trip lines for distributed privacy-preserving traffic monitoring. Proceedings of the 6th ....

Huo, Z., Meng, X., Hu, H. and Huang, Y., 2012. You can walk alone: trajectory privacy-preserving through significant stays protection. Database Systems for Advanced ....

Hwang, R.-H., Hsueh, Y.-L. and Chung, H.-W., 2014. A Novel Time-Obfuscated Algorithm for Trajectory Privacy Protection. IEEE Transactions on Services Computing 7(2), pp. 126–139.

Jin, W., LeFevre, K. and Patel, J., 2010. An online framework for publishing privacy-sensitive location traces. Proceedings of the Ninth ACM International ....

Krumm, J., 2007. Inference attacks on location tracks. Pervasive Computing.

Leitner, M. and Curtis, A., 2006. A first step towards a framework for presenting the location of confidential point data on mapsresults of an empirical perceptual study. International Journal of Geographical Information Science 20(7), pp. 813–822.

Lin, D., Gurung, S., Jiang, W. and Hurson, A., 2009. Privacy-Preserving Location Publishing under Road-Network Constraints. researchgate.net.

Monreale, A. and Andrienko, G., 2010. Movement Data Anonymity through Generalization. Transactions on Data ....

Nergiz, M., Atzori, M. and Saygin, Y., 2008. Towards trajectory anonymization: a generalization-based approach. ... of the SIGSPATIAL ACM GIS 2008 ....

Olumofin, F. and Tysowski, P., 2010. Achieving efficient query privacy for location based services. Privacy Enhancing ....

Poulis, G. and Skiadopoulos, S., 2013. Select-organize-anonymize: A framework for trajectory data anonymization. ... (ICDMW), 2013 IEEE ....

Rossi, L., Walker, J. and Musolesi, M., 2015. Spatio-Temporal Techniques for User Identification by means of GPS Mobility Data. arXiv: 1501.06814v3 [cs.CR] 4(1), pp. 1–11.

Seidl, D., Jankowski, P. and Tsou, M., 2015. Privacy and spatial pattern preservation in masked GPS trajectory data. International Journal of ....

Shi, L., Chi, G., Liu, X. and Liu, Y., 2015. Human mobility patterns in different communities: a mobile phone data-based social network approach. Annals of GIS.

Shi, X. Alford-Teaster, J. and Onega, T., 2009. Kernel density estimation with geographically masked points. In: 2009 17th International Conference on Geoinformatics, IEEE, pp. 1–4.

Sila-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U. and Fotheringham, A. S., 2015. Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science pp. 1–26.

Terrovitis, M. and Mamoulis, N., 2008. Privacy preservation in the publication of trajectories. Mobile Data Management, 2008. ....

Thakuriah, P. V. and Geers, D. G., 2013. Transportation and Information: Trends in Technology and Policy. Springer.

Yarovoy, R., Bonchi, F., Lakshmanan, L. V. S. and Wang, W. H., 2009. Anonymizing moving objects. In: Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '09, ACM Press, New York, New York, USA, p. 72.

Zan, B., Sun, Z., Gruteser, M. and Ban, X., 2013. Linking anonymous location traces through driving characteristics. In: Proceedings of the third ACM conference on Data and application security and privacy - CODASPY '13, ACM Press, New York, New York, USA, p. 293.

Zheng, Y. and Zhou, X., 2011. Computing with Spatial Trajectories. In: Media, p. 327.