When paths cross: maintaining utility of trajectory data in geomasking

Dara E. Seidl*

Recibido el 07 de enero de 2017; aceptado el 30 de mayo de 2017

Resumen

Con el aumento de la recolección de datos móviles a través del GPS y de otros servicios basados en la localización, se han realizado varios intentos de aplicar técnicas de geomascarización a los datos de rutas publicados para proteger la privacidad de las trayectorias. Sin embargo, la utilidad de los datos de la trayectoria mascarada y su valor para la investigación del transporte sigue en cuestión. Este estudio examina cómo la ruta inserida cambia cuando los datos de origen y destino están enmascarados para proteger la privacidad, así como calcula el anonimato de cada ruta recorrida usando una muestra de taxis de la ciudad de Nueva York. Se concluye que las rutas entre lugares enmascarados por perturbaciones aleatorias son significativamente diferentes de las rutas originales, así como que un producto de datos basados en análisis de red que suprime las rutas únicas es una solución viable para liberar estadísticas precisas de ruta y al mismo tiempo proteger la confidencialidad.

Palabras clave: *geoprivacy, GPS, trayectoria, enmascaramiento, ofuscación.*

Resumo

Com o aumento da coleta de dados móveis através do GPS e outros serviços baseados em localização, se tem realizado vários tentativas de aplicar técnicas de "geomasking" aos dados de rotas publicados para proteger a privacidade das trajetórias. Entretanto, a utilidade dos dados da trajetória mascarada e seu valor para pesquisas de transporte segue em questão. Este estudo examina como a rota inferida muda quando os dados de origem e destino estam mascarados para proteger a privacidade, assim como calcula o anonimato de cada rota percorrida por uma amostra de taxis da cidade de Nova York. Se determina que as rotas entre lugares mascarados por

* Department of Geography, San Diego State University, 5500 Campanile Drive, 92182 San Diego, CA, USA, e-mail:dseidl@mail.sdsu.edu

perturbações aleatórias são significativamente diferentes das rotas originais e que um produto de dados baseado na rede que suprime rotas únicas é uma solução viável para liberar estatísticas precisas de rota e proteger a confidencialidade.

Palavras chave: *geoprivacy, GPS, mascaramento, trajetória, ofuscação.*

Abstract

With an increase in mobile data collection through GPS and other location-based services, there have been a number of attempts to apply geomasking techniques to published route data in order to protect trajectory privacy. Yet, the utility of masked trajectory data and its value to transportation research remain in question. This study examines how the inferred route changes when origin and destination data are masked to protect privacy, as well as calculates the anonymity of each route traveled by a sample of New York City taxi cabs. It is determined that the routes between locations masked by random perturbation are significantly different from the original routes and that a network-based data product suppressing unique routes is a viable solution to release both accurate route statistics and protect confidentiality.

Key words: *geoprivacy, GPS, trajectory, geomasking, obfuscation.*

Introduction

Open data and replicable research are of great value to scientific advancement. The dual goals of geomasking techniques are to protect privacy and to preserve the utility of geographic data when sharing publicly or with other researchers. In their slight displacement of geographic point data, geomasking techniques often either protect privacy so well that data utility is hindered, or effectively maintain spatial distributions at the expense of anonymity, inadvertently leaving some personal data identifiable. Such is the case with traditional approaches to trajectory masking, which treat every waypoint as a location to be masked, thereby rendering the masked routes impractical for applications such as transportation planning, where accurate network traffic counts are important. This study therefore reevaluates previous approaches to masking high-frequency trajectory data and introduces a method to publish anonymized routes that are correctly matched to a street network.

Geomasking techniques were introduced with discrete point data in mind, such as personal home locations associated with sensitive characteristics, disease cases, or crime incident locations. However, given the widespread adoption of location-based services and GPS, there has been growing recognition that personal trajectories also constitute a privacy risk. For instance, a prominent study finds that 95% of individuals can be uniquely identified from just four points over time (de Montjoye *et al.* 2013), and in 2009, more than 34,000 Americans were victims of stalking assisted by GPS (Baum *et al.* 2009). The risk to privacy for trajectory data stems from both the identifying characteristics of the stopover points, such as to a particular office building or clinic, and the uniqueness of the route.

Trajectory geomasking which displaces each waypoint presents several issues, which are addressed in this study. Figure 1 displays an example of GPS data at a one-second frequency along major roads. Even though the waypoints are masked, the density of the points still demonstrates that the points converge around the highways. Furthermore, the high frequency of the masked trajectory data provides hints about the masking technique used, as well as the displacement distance threshold, potentially leading to a reversal. Therefore, displacing each point may not be effective in obfuscating the exact route, especially if there is a high density of other travelers. Network analysis tools are also commonly available to help an adversary determine the fastest or most probable route between two general locations. For instance, the Google Directions API can be used to collect batch routing information for free.

At the same time, when trajectory waypoints or trip origins and destinations are masked, it can be challenging to collect accurate traffic counts and usage statistics of transportation facilities. If a researcher wishes to leverage anonymized origin-destination data for transportation planning, the results may not accurately portray routes traveled. There are unanswered questions about whether masking each waypoint unnecessarily protects privacy at the expense of better data utility. This study examines first how inferred route changes between origins and destinations when those locations are masked, and second, how anonymous route data may be published and shared with the suppression of unique route segments.

Related Work

Geomasking techniques were first introduced to assist researchers in publishing precise spatial data while respecting confidentiality through small displacements of point coordinates (Armstrong *et al.*, 1999). Existing techniques include random perturbation (Kwan *et al.*, 2004), donut masking (Hampton *et al.*, 2010), grid masking (Krumm, 2007), affine transformations (Armstrong *et al.*, 1999), Gaussian perturbation (Zandbergen, 2014), location swapping (Zhang *et al.*, 2017), Voronoi masking (Seidl *et al.*, 2015), and MGRS masking (Clarke, 2016). Variants of random perturbation see the most widespread usage, particularly outside academic environments in applications like the citizen science project iNaturalist and the bicycle rental site Spinlister, as well as within health-related research (Allshouse *et al.*, 2010; Shi *et al.*, 2009). This is likely due to the simplicity of the method; each coordinate set is displaced a random distance in a random direction within a selected distance threshold.

Over the past decade, greater attention has been devoted to obfuscating trajectory data in addition to discrete point data. Most of these trajectory masking studies treat each waypoint as a location to be masked. For example, Krumm (2007) applies Gaussian perturbation to GPS coordinates recorded at a frequency of six seconds apart to determine their resistance to home location-finding algorithms. The GEPETO spatial data management system of Gambs *et al.* (2010) also focuses anonymization efforts on waypoints of trajectory data, including random perturbation and downsampling of point data. Yang *et al.* (2015) perform semantic obfuscation to convert coordinates to anonymized semantic features. Seidl *et al.* (2016) apply random perturbation and grid masking to GPS data collected in household travel surveys, and Chen *et al.* (2013) introduce a local suppression method for publishing trajectory data. This too is based on suppressing sets of route coordinates that meet a pre-determined sensitivity threshold. Brito *et al.* (2015) recognize the challenge of anonymizing large mobility datasets and apply the scalable MapReduce paradigm in selecting quasi-identifiers for suppression.

All of these methods assume that the main data to be published are the waypoints themselves and not the route data already matched to a street network. To derive network statistics, a map matching algorithm would be necessary, rendering the masked trajectory data potentially unsuitable for this activity. Furthermore, there has been great progress in reconstructing personal route data even with sparse trajectory points (Krumm and Horvitz, 2006), and in modeling trip purposes from mobile data characteristics alone (Gong *et al.*, 2015). Masked trajectory data may not be immune to these more advanced trip reconstruction methods.

Privacy in geomasking procedures is typically measured using the construct of spatial *k*-anonymity (Zandbergen, 2014). In computer science, *k*-anonymity requires that each data subject be part of a database containing at least *k* subjects with matching characteristics (Sweeney, 2002). As an extension of this concept, trajectory *k*-anonymity requires that each trajectory be attributable to at least *k*-1 others, a statistic that necessitates some calculation of route similarity or clustering (Nergiz *et al.*, 2009). Xiong *et al.* (2014) calculate privacy risk factors for GPS data by comparing the spatial entropy vectors of real-time mobility traces to historical distributions. In environments where GPS data may be inaccurate, a raster-based clustering method may be appropriate (Meratnia and de By, 2002). However, if the mobile data are already matched to a street network, the calculations may be performed using a vector-based intersection.

For testing the preservation of spatial distributions following geomasking, Armstrong *et al.* (1999) identify five principal dimensions to uphold: pairwise relations, event-geography relations, clusters, trends, and anisotropies. These categories are based on masking discrete point data, rather than trajectories. Trajectory data may instead be useful for travel times, travel distances, routing information, facilities encountered, or overall trip density. For example, Seidl *et al.* (2016) examine spatial pattern preservation of trajectories using Pearson's correlations between kernel density rasters of original and masked trajectory data. Whether masked trajectories will be adequate or if correct routing information is necessary depends on the end goal of the data user.

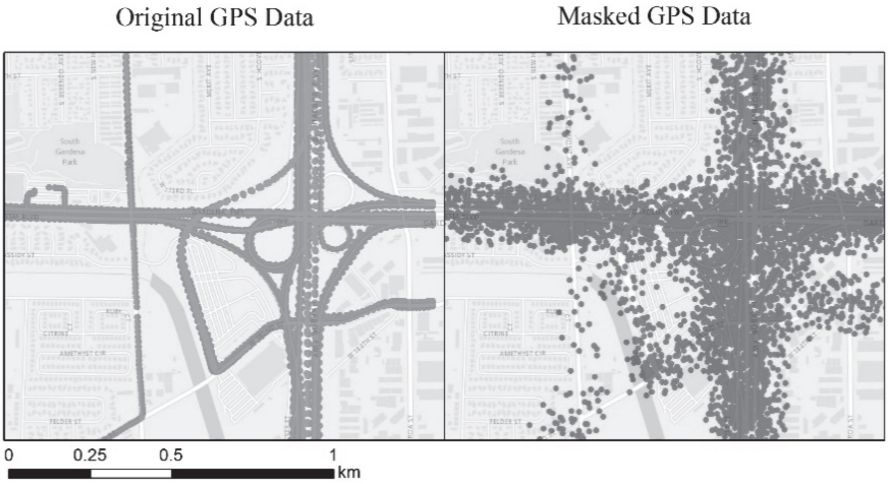

Figure 1. Original GPS data of 1-second frequency and masked GPS points with an outer threshold of 100 meters.

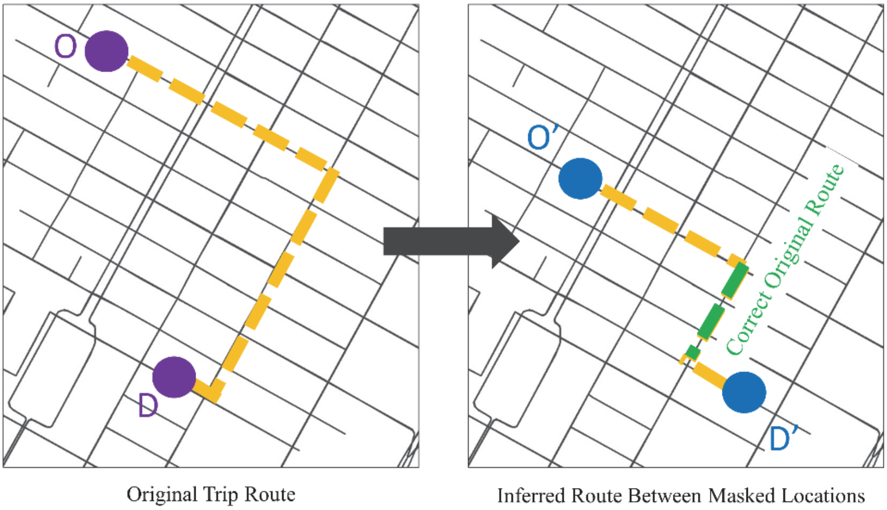

Figure 2. Route of lowest impedance change when origin and destination are masked with a distance threshold of 100 meters.

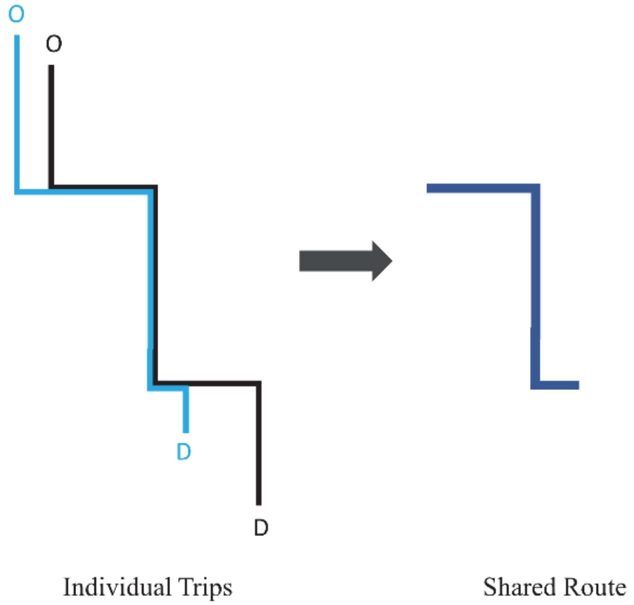

Figure 3. Shared route segments for two simultaneous trips.

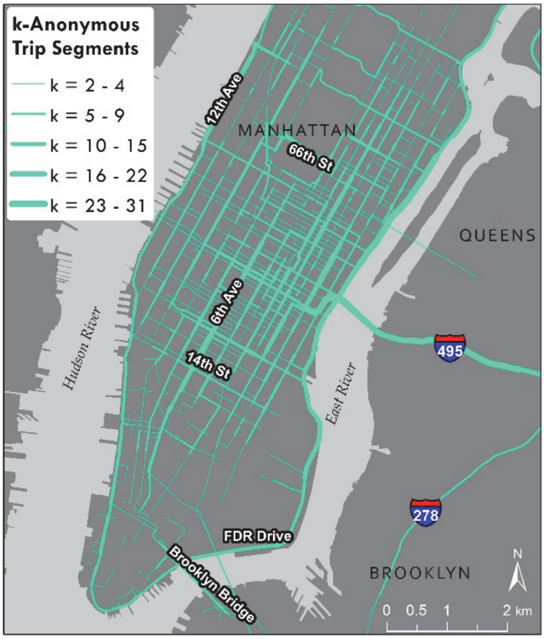

Figure 4. *k*-Anonymous trip segments from sample taxi data.

Conceptualization

The first component of this study examines how probable routes change when the origins and destinations of trip data are masked. Origin and destination data are often published online without corresponding route information. For example, the United States Census Bureau releases origin-destination employment statistics, and taxi commissions across major cities often release taxi trip data in the form of stop-over locations. These publically available datasets may complement other transportation planning data in modeling of routes of least impedance between the stop locations. However, if the origins and destinations are masked, the inferred route could follow a completely different network route. Figure 2 displays the effect of masking an origin and destination from O to O' and D to D' through random perturbation within a distance threshold of 100 meters. Even if the displacement distance of the locations is very slight, the inferred route may change substantially, particularly among dense street networks. This study thus examines how the fastest route between a sample of origin and destination data changes when those locations are masked.

This study also tests a method of publishing trajectory data that remains k-anonymous. Instead of releasing individual trip data, the entire network of shared and simultaneous routes is published. Any network segment with fewer than k travelers is not released. A visual example of a shared route between two trips is shown in Figure 3. The advantage of this process is that complete and accurate usage statistics are made available along network segments with at least k travelers, without revealing unique origins and destinations that may make individuals identifiable and thereby violating confidentiality. However, the utility of such data will depend on if the resulting transportation statistics differ from the original statistics including unique trips. This study therefore tests whether the spatial pattern of k-anonymous routes differs from a complete dataset that includes unique routes.

Methods

This study utilizes taxi cab origin and destination data to test the extent to which geomasking these locations impacts inferred trajectories and whether *k*-anonymous shared route segments differ from full trajectory data with unique routes. Though they may offer clues to the identity of the passengers, origin and destination data are commonly released by taxi commissions and provide a wealth of potentially useful mobility data. In this study, data were obtained from the New York City Taxi and Limousine Commission (TLC), which publishes origin and destination data from all taxi trips in New York. Trips were sampled from the two-minute period of 9:00 AM to 9:02 AM in the morning rush hour of Friday, June 10, 2016 to capture a total of 625 co-occurring taxi trips with recorded origin and destination coordinates within New York City. The Google Directions API was then used to estimate the fastest and, therefore, the probable routes between the 625 sets of taxi origins and destinations. The Google Directions API produces an overview polyline of the resulting trip for each routing request. The polylines follow the same street network and therefore correspond well with other co-located trips, sharing the same vertices. If the trajectory waypoints originate from GPS devices, a higher level of coordinate variation necessitates some map matching procedure to positively associate them with a particular street network segment.

For the first part of this study, the taxi origin-destination data are masked three times by random perturbation: at distance thresholds of 50, 100, and 150 meters. Random perturbation is selected because of its common application as a masking technique, and the distance thresholds are centered around 100 meters for a reasonable balance demonstrated in previous work between anonymity and accuracy (Zandbergen, 2014). Seidl *et al.* (2016) find that 250 meters is too large of a distance threshold to preserve spatial distributions in two urban areas. It is acknowledged that while the mean trip length of the taxi data is 5 kilometers, the dimensions of a Manhattan city block are approximately 80 by 274 meters, meaning that masking within these distance thresholds is likely to displace the origin-destination data to new blocks. Following these three iterations of geomasking, the inferred route is then calculated again using the Google Directions API for each of the three new sets of origins and destinations. To compare the new routes to the original routes, three sets of intersections are performed with a distance tolerance of 3 meters, and the length of the intersections is calculated. The Wilcoxon signed ranks test is then applied to test whether the network trip distance at the 50-, 100-, and 150-meter masking thresholds is significantly different from the original trip length in pairwise comparisons. This test is an appropriate choice for the continuous dependent variable of trip length, since the data do not follow a normal distribution and are from matched populations comprised of each origin and destination set.

For the second component of this study, a similar intersection procedure is performed on the original unmasked route data with itself. This results in a calculation of the length of each trip shared with another taxi cab during the sampled time period. The resulting data product of this intersection is a polyline dataset of a New York City street network of segments with at least k=2 taxis sharing a trip route. For comparing this output of anonymized routes to the original routes, line density rasters are generated with a cell size and search radius both of 100 meters. These parameters are selected given that a resolution of 1:100,000 is appropriate for mapping the overall New York City area. The density rasters then become inputs for a Pearson's correlation between the two density surfaces, as Seidl *et al.* (2016) use to compare spatial distributions between original and masked trajectory data.

Results and discussion

This study finds that when taxi origins and destinations are masked through random perturbation, as the distance threshold increases, the mean inferred route distance increases, and the mean length of the route shared with the original data decreases. These results are summarized in Table 1. The Wilcoxon signed ranks test for trip length found that each of the masked data sets is significantly different (p =0.00) from the mean trip length results for the unmasked origin-destination data. Not only are inferred trips longer in masked data, but they share less of the route with the original trips. In the 150-meter masked set, the inferred taxi route shared only 67% on average the same route as the unmasked trip, and 9% of the trips shared no part of the route with their unmasked counterparts. The difference is even present at the 50-meter masking threshold, where 3% of the trips had a completely new route, and on average, 15% of the original trip routes were not maintained in the masked version. These results suggest that applying random perturbation to trip origins and destinations, even at these smaller distance thresholds, is not an appropriate substitute for research requiring accurate street network statistics.

Table 1
Length and correspondence of trip routes following random perturbation of origins and destinations

	Original	*50m*	*100m*	*150m*
Mean trip length (km)	*5.01*	5.06	5.11	5.14
Wilcoxon signed ranks p-value	—	0.00*	0.00*	0.00*
Mean shared route length (km)	—	4.52	4.14	3.89
Mean % of route shared with original	—	84.62	74.74	67.08
% of trips with a completely new route	—	2.72	6.40	9.12

The second objective of this study is to compare the calculated network of *k*-anonymous routes derived from the non-unique trip data to the original set of taxi trajectories. From the intersected polyline dataset, 618 out of 625 trips (98.9%) shared a portion of the route with at least one other taxi. There were thus 7 taxi rides that were completely unique, sharing no part of the trip with other taxis during the time period. On average, each taxi trip shared 88.2% of its total trip distance along anonymous routes, sharing its path with at least one other cab. Figure 4 illustrates the completeness of the *k*-anonymous taxi network segments during this two-minute stretch in the New York City area, particularly in Midtown. Furthermore, there was no great loss of spatial information between the original taxi trip routes and the anonymous routes shown in Figure 4. The line density rasters generated from the polyline route datasets demonstrate a Pearson's correlation of 0.89 with p= 0.00. This means that the travel patterns represented in the original data continue to be well-represented in the anonymous trip data along street networks.

Conclusion

Building on previous attempts by others to apply anonymization techniques to trajectory data, this study is a response to concerns as to whether geomasking each waypoint in trajectory data is necessary or even protective of privacy. Figure 1 above highlights how random perturbation of high-frequency GPS data may actually enhance personal route disclosure along popular highways and provide clues as to the distance threshold of masking. Advanced techniques of route reconstruction may further chip away at the anonymity masking seeks to provide, particularly in areas of low road density, where there may be one most probable route. A second concern is what kind of research may benefit from masked trajectory data. Transportation planners must in general search elsewhere for more accurate traffic counts and route statistics in traffic models.

Any publisher of spatial data concerning human subjects must achieve a delicate balance between data utility and personal anonymity. When striking this balance appears too challenging, alternative confidentiality solutions are often selected over geomasking techniques, such as secured data enclaves, software agents, or aggregation to large administrative units. This study re-evaluates previous work on geomasking trajectory data and proposes an alternative solution for releasing anonymous route data that also maintains accurate transportation network statistics. First, this study finds that geomasking taxi origin and destination data through random perturbation leads to significantly different trajectories compared to the inferred original routes between the locations. This can impede any analysis requiring accurate transportation statistics. Second, it is observed that in just two minutes of urban New York taxi data, 98.9% of taxi trajectories shared a route with at least one other cab, revealing high *k*-anonymity in the unmasked dataset. Furthermore, the suppression of unique routes in the resulting network demonstrates little difference in line density estimation compared to the full route dataset with a Pearson correlation of 0.89. The release of a fully *k*-anonymous network of shared personal routes helps to better balance between protecting privacy and preserving the spatial information of trajectory data.

References

Allshouse, W.B.; Fitch, M.K.; Hampton, K.H.; Gesink, D.C.; Doherty, I.A.; Leone, P.A.; Serre, M.L.; Miller, W.C. (2010). "Geomasking sensitive health data and privacy protection: an evaluation using an E911 database", *Geocarto International*, 25(6):443-452.

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

Baum, K.; Catalano, S.; Rand, M. and Rose, K. (2009). *Stalking Victimization in the United States*, Washington, DC; U.S. Department of Justice, Bureau of Justice Statistics.

Brito, F.T.; Araújo Neto, A.C.; Costa, C.F.; Mendonça, A.L.C., and Machado, J.C. (2015). "A distributed approach for privacy preservation in the publication of trajectory data", *Proceedings of the ACM SIGSPATIAL 2nd Workshop on Privacy in Geographic Information Collection and Analysis (GeoPrivacy '15)* Bellevue, WA, November.

Chen, R.; Fung, B.C.; Mohammed, N.; Desai, B.C. and Wang, K. (2013). "Privacy-preserving trajectory data publishing by local suppression", *Information Sciences*, 231:83-97.

Clarke, K.C. (2016). "A multiscale masking method for point geographic data", *International Journal of Geographical Information Science*, 30(2):300-315.

de Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M. and Blondel, V.D. (2013). "Unique in the crowd: the privacy bounds of human mobility", *Scientific Reports* 3, available in <http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html>.

Gambs, S.; Killijian, M.O. and Cortez, M.N.d.P. (2010). "GEPETO: A GEoPrivacy-Enhancing Toolkit", *IEEE Computer Society AINA Workshops*, pp. 1071-1076.

Gong, L.; Liu, X.; Wu, L. and Liu, Y. (2016). Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, *43*(2), pp. 103-114.

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):1062-1069.

Krumm, J. (2007). "Inference attacks on location tracks", *5th International Conference, Proceedings, PERVASIVE 2007* Toronto, Canada, May 13-16, pp. 127-143.

Krumm, J. and Horvitz, E. (2006). "Predestination: Inferring destinations from partial trajectories", *International Conference on Ubiquitous Computing,* Springer Berlin Heidelberg, pp. 243-260.

Kwan, M-P.; Casas, I., and Schmitz, B.C. (2004). Protection of geoprivacy and accuracy of spatial -information: how effective are geographical masks?, *Cartographica* 39(2):15-28.

Meratnia, N. and de By, R.A. (2002). "Aggregation and comparison of trajectories", *Proceedings of the ACM Symposium on Advances in Geographic Information System*, pp. 49-54.

Nergiz, M.E.; Atzori, M.; Saygin, Y. and Güç, B. (2009). "Towards trajectory anonymization: a generalization-based approach", *Transactions on Data Privacy* 2, pp. 47-75.

Shi, X.; Alford-Teaster, J., and Onega, T. (2009). "Kernel density estimation with geographically masked points", *Proceedings of the 17th International Conference on Geoinformatics,* August.

Seidl, D.E.; Paulus, G.; Jankowski, P. and Regenfelder, M. (2015). "Spatial obfuscation methods for privacy protection of household-level data", *Applied Geography*, 63:253-263.

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

Sweeney, L. (2002). "*k*-Anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557-570.

Xiong, J.; Xiong, J. and Claramunt, C. (2014). "A spatial entropy-based approach to improve mobile risk-based authentication", *ACM SIGSPATIAL International Workshop on Privacy in Geographic Information Collection and Analysis (GeoPrivacy '14)*, Dallas/Ft. Worth, TX, November 2014.

Yang, J.; Zhu, Z.; Seiter, J., and Tröster, G. (2015). "Informative yet unrevealing: semantic obfuscation for location based services", *Proceedings of the ACM SIGSPATIAL 2nd Workshop on Privacy in Geographic Information Collection and Analysis (GeoPrivacy '15)* Bellevue, WA, November 2015.

Zandbergen, P.A. (2014). "Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data", *Advances in Medicine,* pp. 1-14.

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