A comparison of obfuscation methods for privacy protection: A case study using agri-environmental data in the Republic of Ireland

Nowbakht, P.[*1,2]; O'Sullivan, L.[2]; Cawkwell, F.[1,3]; Wall, D.P.[2]; Holloway, P.[†1,3]

[1]Department of Geography, University College Cork, Ireland
[2]Teagasc - The Agriculture and Food Development Authority, ESLU, Johnstown Castle, Co Wexford Ireland
[3]Environmental Research Institute, University College Cork, Ireland

April 14-16, 2021

Summary

A major challenge of sharing spatially explicit agriculture data is to identify the trade-off between field parcel confidentiality and spatial pattern preservation. In this work, twenty-seven point-based obfuscation and evaluation methods were applied on agriculture data with a high point density. This study identified the optimal value of radiuses for Donut and Density methods to improve geoprivacy and statistical analysis accuracy. Modified AHilb and Donut-AHilb methods were developed to generate smaller and arbitrary obfuscation areas to improve location security. The Donut-AHilb method was found to be the best in spatial pattern preservation and satisfy larger k-anonymity but the risk of false identification and non-unique obfuscation were high. The term of non-unique obfuscated points was introduced which is important for static objects as two or more points might have the same obfuscated point.

KEYWORDS: Agriculture, Geoprivacy, Obfuscation, Spatial Analysis

1. Introduction

Advanced geospatial technology in data collection and storage has generated a significant amount of data in many fields and industries. While data sharing is essential to improve data usage efficiency and generate knowledge, data are prone to privacy violations, including data breaches. Since 1999, spatial anonymization, obfuscation and geo-masking methods have been developed and implemented to preserve the individual's privacy by a process of degrading the quality of locational information while maintaining the spatial properties of geographic data (Marc, Gerard and Dale, 1999). Obfuscation methods are widely used in some fields and sectors especially in location-based services (Duckham and Kulik, 2005), healthcare (Hampton et al., 2010) and criminology (Kounadi and Leitner, 2015) . Few studies have been developed and conducted that specifically examine obfuscation methods on agricultural data, despite the unique spatial patterns and privacy concerns associated with such features. A rare example of using obfuscation methods in agricultural research is a study which was carried out to investigate the impact of spatial association between diagnosis of Bovine tuberculosis (BTB) and exposure to the parasitic fluke *Fasciola* using 3,026 dairy herds in England and Wales (Claridge et al., 2012). In that example, random displacement within a buffer of 5 km was implemented to preserve farms confidentiality while visualizing the results as a map, without further investigation about the quality of the obfuscated data. However, there are several obfuscation methods that could be used within this context, but to-date they remain unexplored despite their ability to reduce risk of identification and spatial patterns preservation.

[*]111222849@umail.ucc.ie
[†]paul.holloway@ucc.ie

The polygon and static nature of field parcels, the highly confidential data, their related location and environmental data distinguish agricultural data from other data that obfuscation methods have been applied on. The *Irish online nutrient management plan (NMP Online)* dataset contains geographical, soil related and bovine animal herd related attributes of over 55,000 farms and is used to support farmers in the development of nutrient management plans for field parcels. It could be used by organisations for research and the development of more precision nutrient management planning support, however currently this information does not adhere to privacy regulation that sufficiently remove or obfuscate the personal information (i.e., location) associated with the data. Therefore, the aim of this study is to compare several point base obfuscation methods on *NMP online* data, as to how they performed in terms of security and accuracy, that are subject to stringent data protection regulations (Murphy *et al.*, 2015).

2. Data and Methods

After pre-processing of a subset of data from *NMP Online*, a total of 22425 field parcels information for 2016 with non-uniform distribution in Ireland were extracted and used in this study.
Various randomization methods with the k-anonymity concept were developed to reduce the risk of disclosure of agriculture data via geospatial attributes. Randomization methods transform the original point P to a random point or the farthest/middle point among *N* random points inside the obfuscation area and k-anonymity refers to ensuring that each point unidentifiable from k-1 other points in the obfuscation area. Table 1 outlines the seven overarching methods that were implemented in this study to generate the obfuscation area.

Table 1 Outline of the seven obfuscation methods to generate the obfuscation area

Method	Method Code	Description	Type
N*Rand: Rand NRand NMix	N*R: R NR NM	Obfuscation area is a circle with the centre P and radius r. obfuscated point is a random point or the farthest/middle point among *N* random points inside the obfuscation area.(Wightman et al., 2011) Figure 1a and 1b.	Point based
Donut	Dt	Obfuscation area is a donut-shaped area, with centre P, constant radius rmin and rmax for all points (maximum rmin and maximum rmax of all points to ensure that all obfuscated points satisfy k-anonymity) (Hampton et al., 2010; Zandbergen, 2014) Figure 1c.	Point based
Density	D	Using Donut method when radius rmin and rmax are vary depending on the point's density(Hampton et al., 2010) Figure 1d.	Point based
Pinwheel	P	Obfuscation area is a pinwheel shape instead of circular area (Wightman et al., 2013).	Point based
AHilb	AH	Obfuscation area is a rectangular, including cells that follows the Hilbert curve order value and unconnected adjacent cells, to avoid the unnecessary extension of an obfuscation area(Um, Kim and Chang, 2010).	Area based
Donut-AHilb	Dt_AH	Obfuscation area is a difference of area that satisfy 2k-anonmity and area that satisfy k-anonymity generated by AHilb method.	Area based
k-anonymity	Ka	Obfuscation area contains the original point and k -1 other points to protect the privacy of the original point. k is the minimum number of other points that the original point is unidentifiable between them(Gruteser and Grunwald, 2003).	Both

For each distance-based method, we needed to compute optimal radiuses (rmin and rmax) based on a combination of Donut method and k-anonymity satisfaction to minimise the risk of identification and maximise spatial pattern preservation by generating a smallest obfuscation area, that contains at least k-1 other locations. An iterative procedure was developed as follows: rmin was set to minimum threshold and increased by increasing rate in each step until k-anonymity satisfaction. The procedure was repeated for rmax which was initialized by $r$min. There is a trade-off between computation time and accurate optimal radiuses using fixed increasing rate. To increase the processing speed and obtain more accurate optimal radiuses that satisfy the k-anonymity this method was modified based on varied increasing rate as follows:

1) Initialize the rmin and rmax with a minimum/maximum threshold.
2) Initialize minimum increasing rate with minimum threshold greater than zero and increasing rate with large value.
3) Increase the rmin by increasing rate in each step until k-anonymity satisfaction.
4) Reduce the increasing rate to half.
5) Repeat step 3 and 4 until increasing rate is less than minimum increasing rate.

This procedure can be implemented for different values of k to identify the maximum k-anonymity with minimum radiuses, and consequently, identify the most effective Donut and Density methods for privacy and spatial pattern preservation. In this study the maximin k-anonymity was obtained by this procedure is 6.
In this research to increase the geoprivacy protection a modified AHilb method is applied by excluding the cells that contain no point so obfuscation area might be smaller and have arbitrary shape.
These iterations led to 27 different variation of the seven overarching methods presented in Table 1, we evaluated these techniques using several evaluation metrics (Table 2).

Table 2 Description of the Evaluation technique that have been used to evaluate and compare obfuscation method in terms of accuracy and security.

Evaluation technique	Description	Evaluation type
Location distribution preservation	Rank the methods from the best to worst based on KS-test (statistical test to compare the original points and obfuscation points distribution) and correlation alteration of spatial attributes (DPR)	Accuracy
Correlation preservation	Total correlation alteration of location-related attributes between original and obfuscated points (TCA)	Accuracy
spatial cluster preservation	Total number of misclassified obfuscation points called Misclassification Error (MCE) (Parameswaran and D. Blough, 2005	Accuracy
k-anonymity satisfaction	The number/percentage of obfuscated points which satisfied k-anonymity for different values of k	Security
Displacement distribution	The number/percentage of points with different displacement	Accuracy
False identification	Number of points that their obfuscated points are another original point(Seidl, Jankowski and Clarke, 2017) or are inside the area of other original point's region (when the polygon nature of field parcel is considered)	Security
Non-unique obfuscation	Number of points with not unique obfuscated point (two or more points have a same obfuscated point or an obfuscated point is inside the area of others obfuscated point when polygon nature of field parcel is considered)	Security and Accuracy

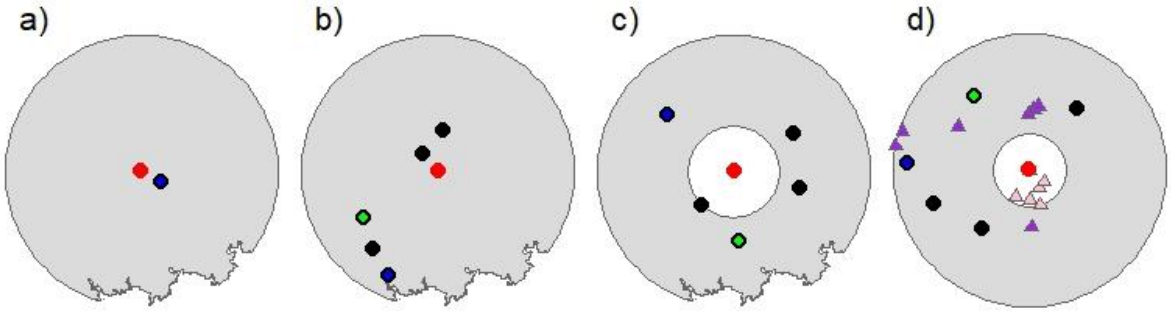

Figure 1 Examples of obfuscated point generated by different methods: a) Rand, b) NRand and NMix c) Donut d) Density. Obfuscated area (grey) original point (red), 5 random points (black), obfuscated point in NRand (blue) and in NMix (green), pink and purple triangles present existing points (parcels) in an inner and outer circle with $rmin$ = 291m and $rmax$ = 1088m radiuses around the original point based on 6-kanonmity in Density method.

3. Results

Density based methods (Density, density-pinwheel and AHilb, Donut-AHilb) were the best performing with distribution preservation of one, zero correlation alteration and very small MCE. Alternatively NRand had the second highest distribution preservation rank of six, correlation alteration of 0.14 and 18.82% MCE as the worst performing method in terms of accuracy (Table 3, Figure 2 and Figure 3c).

Table 3 list of 27 obfuscation methods that have been applied with the following parameters: k = 6, N = 5, (ϕ = 45°/ ϕ = 225°), $rmin$ =20203 m, $rmax$ = 36451 m, res = 521, illustrating distribution preservation rank obtain from SK- test and correlation alteration.

Performance order with respect to Distribution preservation (DPR) and Correlation alteration (TCA)						
Method	Rand (R)		NRand (NR)		NMix (NM)	
	DPR	TCA	DPR	TCA	DPR	TCA
N*Rand (N*R)	4	0.09	6	0.14	4	0.07
Donut (Dt)	3	0.06	4	0.07	3	0.07
Density (D)	1	0	1	0	1	0
Pinw45 (P45)	2	0.04	5	0.15	2	0.04
Pinw225 (P225)	3	0.05	7	0.1	2	0.04
Density_Pinw45 (D_P45)	1	0	1	0	1	0
Density_Pinw225 (D_P225)	1	0	1	0.01	1	0
AHilb (AH)	1	0	1	0	1	0
Donut_AHilb (Dt_AH)	1	0.01	1	0.01	1	0.01

The results, as shown in (Figure 3d) indicate that Rand and Donut methods 100% satisfy 6-anonmyity and more than 99% satisfy 25-anonmyity which is the highest security protection. Density, density-pinw45 and density-pinw225 almost fully (greater than 99.98%) satisfy 6-anonmity, however, almost

none of them satisfy values greater than 15- anonymity, although this is an artefact due to the fact that the obfuscation area was generated based on 6-anonmity. AHilb method provides moderate security protection between 60% of 6-anonmity to 5.88% of 25-anonmity. Donut-AHilb provides the best security protection with 93.5% 6-anonmity to 30% 25-anonmity and almost 80% 10-anoonmity. It is important to bear in mind that the percentage of points that satisfy k-anonymity was calculated based on circular obfuscation area while AHilb method generates gridded based obfuscation area. Therefore, it is expected to provide higher percentage of 6-anonmity. Figure 3e illustrates that 40% to 60% of obfuscated points generated by density-based methods have less than 1000m displacement and nearly 95% have less than 5000m displacement. Taken together, these results suggest that the Donut-AHilb method with a high percentage of high k-anonymity, less displacement, low MCE high level of distribution and corelation preservation can be considered as an appropriate method for data with high point density. On the other hand, a high percentage of false identification and non-unique obfuscation (20% and 50%) indicates the drawback of these methods for data with high location density and polygon nature consideration (Figure 3a, 3b). All statistical analysis, obfuscation and evaluation methods were coded using R programming language.

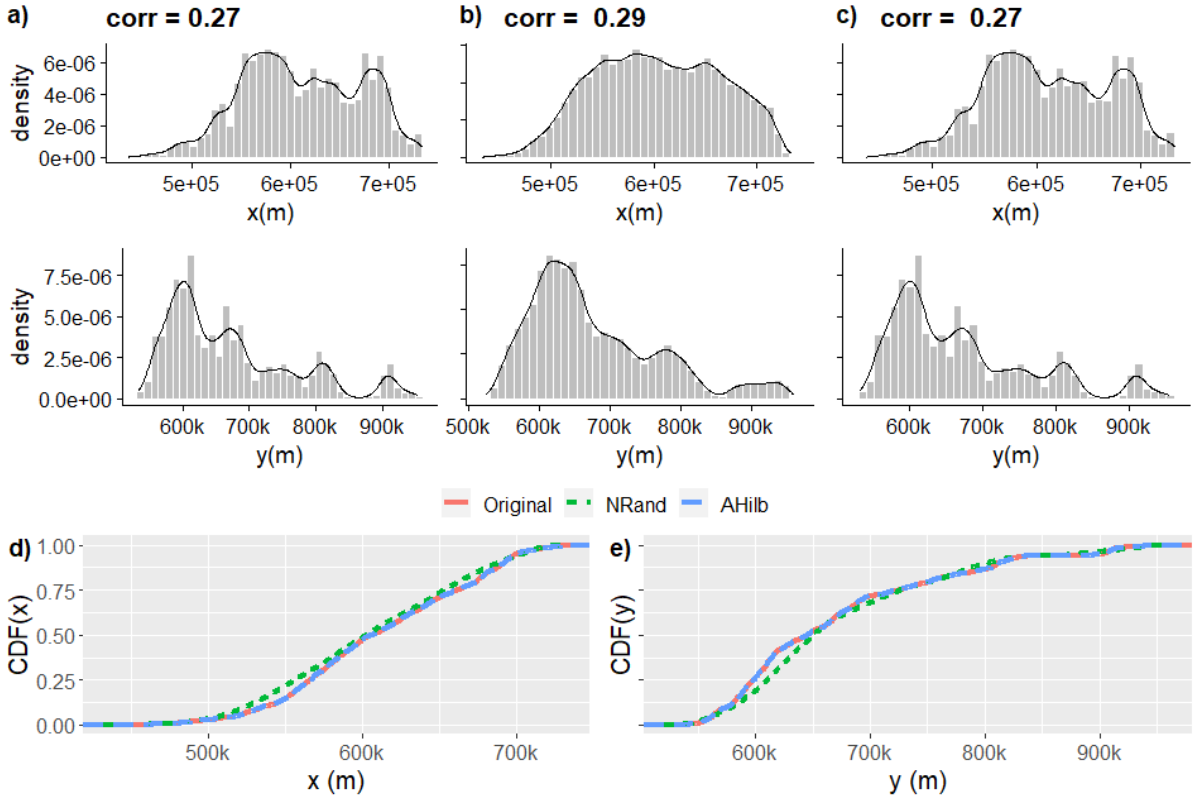

Figure 2 Geographical attributes distribution and their correlations of a) original points and obfuscated points generated by b) NRand and c) AHilb methods indicates that AHilb perfectly preserve distribution of both directions with no correlation alteration while distribution of both directions of NRand method are clearly different from the original points and correlation is changed from 0.27 to 0.29. Cumulative distribution function (CDF) plots of spatial attributes d) and e) which are used in 2D KS test to statistically test if the original and obfuscated points have the same distribution also confirm these results.

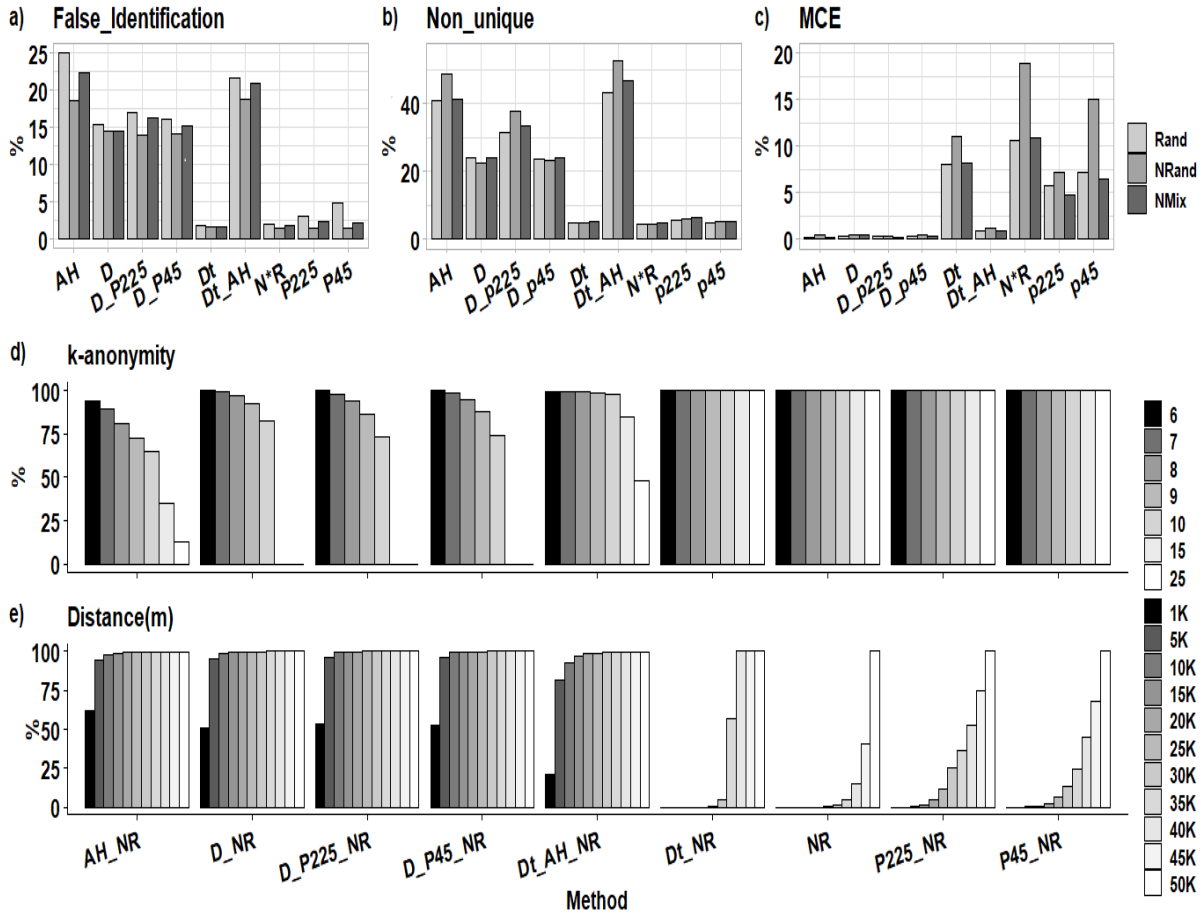

Figure 3 Comparison of 27 obfuscation methods based on a) False-identification b) non-unique obfuscation, c) MCE, d) percentage of obfuscated points that satisfies k-anonymity for different values of k and d) percentage of obfuscated points with displacement less than different values of d.

4. Conclusion

The main contribution of this paper is to review different obfuscation methods and apply modified and combined methods to identify their suitability for agriculture data to improve their performance. Different evaluation methods in terms of security and accuracy can lead to the identification of the most appropriate obfuscation method. According to this investigation obfuscation methods with smaller location distribution, correlation alteration and less MCE provide better performance in term of accuracy. Methods with a higher percentage of points satisfy higher k-anonymity and higher percentage of lower points displacement provides better trade-off between security and accuracy. False identification and non-unique obfuscation are also two important factors for evaluating obfuscation methods which have not received much attention in previous research. This study suggests that excluding the original points (area in polygon base), obfuscated points (area of obfuscated points) from obfuscation area generated by obfuscation method can reduce occurrence of false identification and non-unique obfuscation, although further research is required to examine the impact of this suggestion on methods performance in terms of security and accuracy.

5. Acknowledgements

The first author is funded by Teagasc -The Agriculture and Food Development Authority. A joint project between Teagasc and UCC (Walsh Scholarships Ref Number 2018034).

6. Biography

Parvaneh Nowbakht is a PhD candidate in the Department of Geography at University College Cork, and Teagasc Walsh Scholar with Teagasc Agriculture and Food Development Authority of Ireland in the Environment, Soils and Land Use Department, Johnstown Castle, Co. Wexford. Parvaneh's research interests include data scientist and machine learning in general, spatial analysis of environmental and climate data, creating tools for improving geoprivacy.

Lilian O'Sullivan is a research officer with Teagasc Agriculture and Food Development Authority of Ireland in the Environment, Soils and Land Use Department, Johnstown Castle, Co. Wexford. Lilian's research is focussed on sustainable soils and land use where she uses a suite of tools including spatial analysis to investigate opportunities for sustainability, particularly in agricultural landscapes through integrated land use and management.

Fiona Cawkwell is a lecturer in Remote Sensing in the Department of Geography at University College Cork and a principal investigator in the Environmental Research Institute at University College Cork. Fiona's research interests include the use of time series of optical and microwave satellite imagery to evaluate changes in the natural landscape; the use of imagery as an input to machine learning approaches for optimising vegetation growth and biomass estimation; and the impacts of climate change on natural systems

David Wall is a senior research officer with Teagasc Agricultural and Food Development Authority of Ireland in the Environment, Soils and Land Use Department, Johnstown Castle, Co. Wexford. David's research interests include biogeochemical cycling of nutrients and tracking nutrient fate in agro-ecosystems; soil specific and spatially explicit nutrient management planning tools and advice for the farmers and farm advisory personnel; development soil quality indicators and monitoring schemas suitable for different scales.

Paul Holloway is a lecturer in Geographic Information Science and Systems in the Department of Geography at University College Cork, a principal investigator in the Environmental Research Institute at University College Cork and the Vice President of the Irish Organisation of Geographic Information. Paul's research interests include using GIScience and spatial analysis to address a suite of ecological, environmental, and geographical issues, including, incorporating movement within species distribution models; investigating habitat selection of mobile animals; and using spatial statistics and simulation to investigate the effects of climate change in natural and agriculture systems.

References

Duckham, M. and Kulik, L., 2005. *A Formal Model of Obfuscation and Negotiation for Location Privacy*. [online] Available at: <http://www.geosensor.net/papers/duckham05.PERVASIVE.pdf> [Accessed 31 Oct. 2018].

Gruteser, M. and Grunwald, D., 2003. Anonymous usage of location-based services through spatial and temporal cloaking. *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, MobiSys 2003*, 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.1062–1069.

Kounadi, O. and Leitner, M., 2015. Spatial Information Divergence: Using Global and Local Indices to Compare Geographical Masks Applied to Crime Data. *Transactions in GIS*.

Marc, P.A., Gerard, R. and Dale, L.Z., 1999. Geographically masking health data to preserve

confidentiality. *Statistics in Medicine*, [online] 18(5), pp.497–525. Available at: <http://dx.doi.org/10.1002/(SICI)1097-0258(19990315)18:5%3C497::AID-SIM45%3E3.0.CO;2-#>.

Murphy, P., Lalor, S.T.J., Mechan, S., Plunkett, M., and Wall, D.P. (2015) Nutrient Management planning (NMP) Online, - an integrated tool for adaptive nutrient management planning. Teagasc *Soil Fertility Conference 2015* proceedings, 16th October 2015, Clonmel, Co. Tipperary, Ireland, pp.8-9. https://www.teagasc.ie/media/website/publications/2015/Teagasc-Soil-Fertility.pdf

Seidl, D.E., Jankowski, P. and Clarke, K.C., 2017. Privacy and False Identification Risk in Geomasking Techniques. pp.1–18.

Um, J.H., Kim, H.D. and Chang, J.W., 2010. An advanced cloaking algorithm using Hilbert curves for anonymous location based service. *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, pp.1093–1098.

Wightman, P., Coronell, W., Jabba, D., Jimeno, M. and Labrador, M., 2011. Evaluation of location obfuscation techniques for privacy in location based information systems. In: *2011 IEEE Latin-American Conference on Communications, LATINCOM 2011 - Conference Proceedings*.

Wightman, P., Zurbarán, M., Zurek, E., Salazar, A., Jabba, D. and Jimeno, M., 2013. θ-Rand : Random Noise-based Location Obfuscation Based on Circle Sectors. pp.100–104.

Zandbergen, P.A., 2014. Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine*, 2014, pp.1–14.