

Title	Geospatial data obfuscation methods applied to agricultural data
Authors	Nowbakht, Parvaneh
Publication date	2023
Original Citation	Nowbakht, P. 2023. Geospatial data obfuscation methods applied to agricultural data. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2023, Parvaneh Nowbakht. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-10-10 07:07:24
Item downloaded from	https://hdl.handle.net/10468/15470

Ollscoil na hÉireann, Corcaigh
National University of Ireland, Cork



Geospatial Data Obfuscation Methods Applied to
Agricultural Data

Thesis presented by
Parvaneh Nowbakht M.Sc.
for the degree of Doctor of Philosophy

University College Cork
DEPARTMENT OF GEOGRAPHY
Head of School/Department: Dr. Denis Linehan
Supervisors: Dr. Paul Holloway, Dr. Fiona Cawkwell
Dr. Lilian O'Sullivan and Dr. David Wall

2023

Research funded by Teagasc

Table of Contents

Chapter 1 General Introduction	1
1.1. Background	1
1.1.1. Data Privacy Protection.....	1
1.1.2. Agricultural Data	4
1.1.3. Agricultural Data and Geoprivacy Protection	5
1.1.4. Spatial Data	6
1.2. Challenges and Solutions	8
1.2.1. Existing methods and related work	8
1.2.2. Conceptualization of geoprivacy protection methods.....	8
1.3. Research Gap and aims.....	9
1.3.1. Research Aims	9
1.3.2. Research structure	12
Chapter 2 Point-Based Obfuscation Methods Review	13
2.1. Introduction	13
2.2. Aggregation	13
2.2.1. Spatial Aggregation	13
2.2.2. Areal Aggregation.....	14
2.2.3. Point Aggregation.....	15
2.3. Affine Transformation.....	17
2.4. Anonymization	19
2.5. Randomization	21
2.5.1. N*Rand and N*Dispersion methods	21
2.5.2. Donut method.....	23
2.5.3. θ -RAND and pinwheel methods (Trajectory)	23
2.5.4. Location Swapping	24
2.5.5. Nearest Neighbour Data Substitution (NeNDS)	25
2.6. Conclusion.....	26
Chapter 3 A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research.	28
3.1. Introduction	29
Existing methods and related work	32
3.1.1. Anonymization (Spatial k-anonymity Method).....	32
3.1.2. Randomization Methods.....	32
3.1.2.1. Rand Method	33

3.1.2.2. N*Rand Method	33
3.1.2.3. Donut and Density Method.....	35
3.1.2.4. Pinwheel Method	36
3.1.2.5. Hilbert Method.....	36
3.2. Data and Methods	38
3.2.1. Data Preparation	38
3.2.2. Obfuscation Methods	39
3.2.2.1. Donut-k and Density-k Methods.....	40
3.2.2.2. Modified AHilb and Donut-AHilb Methods.....	41
3.3. Evaluation Method.....	42
3.4. Results.....	45
3.4.1. Parameters determination.....	45
3.4.2. Distribution preservation Analysis	46
3.4.3. Correlation Alteration Analysis	47
3.4.4. Cluster Preservation Analysis.....	49
3.4.5. Displacement and K-anonymity Satisfaction Analysis	51
3.4.6. False identification and non-unique obfuscation analysis.....	53
3.4.7. Spatial information divergence analysis	53
3.5. Discussion.....	56
3.6. Conclusion.....	61
Chapter 4 Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data.....	63
4.1. Introduction	64
4.2. Data and Methods	68
4.2.1. Data preparation	68
4.2.2. Polygon-based Obfuscation Methods.....	68
4.2.2.1. Polygon-based obfuscation PDonut-k and PDensity-k methods	72
4.2.2.2. Polygon-based obfuscation PAHilb method	74
4.2.3. Evaluation methods	75
4.3. Results.....	77
4.3.1. Parameters determination.....	77
4.3.2. Distribution preservation Analysis	78
4.3.3. Correlation Alteration Analysis	79
4.3.4. Cluster Preservation Analysis.....	79
4.3.5. Displacement and K-anonymity Satisfaction Analysis	81
4.3.6. False identification and Non-unique Obfuscation Analysis	82

4.4. Discussion.....	83
4.4.1. False identification and Non-unique Obfuscation Analysis	84
4.4.2. Spatial Pattern Preservation	84
4.4.3. Geoprivacy Protection.....	85
4.4.4. Limitation and Suggestion.....	85
4.5. Conclusion.....	89
Chapter 5 Maintaining Environmental Context and Geoprivacy Protection in Agriculture	90
5.1. Introduction	91
5.2. Polygon-based Environmental Similarity Obfuscation Method (PESOM)	94
5.2.1. Number of features.....	94
5.2.1.1. Principal Component Analysis (PCA).....	95
5.2.2. Number of Clusters	96
5.3. Data and Results	97
5.3.1. Climate Data.....	97
5.3.2. Irish nutrient management planning online (NMP Online) Data.....	97
5.3.3. Parameters determination.....	98
5.3.4. Evaluation Metrics	101
5.3.5. Results	102
5.4. Discussion.....	104
5.5. Conclusion.....	108
Chapter 6 Incorporating contextual factors in agri-environmental research- options for obfuscation to support wider data application	109
6.1. Introduction	110
6.2. Materials and Methods.....	113
6.2.1. Study area and data	113
6.2.2. Obfuscation Methods (Rand and PESOM)	113
6.2.3. Machine Learning Methods (MLR and RF).....	114
6.3. Results.....	116
6.3.1. Original data.....	116
6.3.2. Obfuscated data	116
6.3.3. Scenario 1: using internal features.	118
6.3.4. Scenario 2: using internal and external features.	119
6.3.5. Scenario 3: using external features.....	122
6.4. Discussion.....	125
6.5. Conclusion.....	126
Chapter 7 General Discussion	127

7.1. Summary	127
7.2. Contribution	130
7.3. Research Constraints	136
7.4. Recommendation for further work.....	137
Appendices.....	140
Appendix 1	140
Appendix 2	143
Appendix 3	145
References:	146

List Of Tables

Table 1.1: Type of agricultural data and examples.....	5
Table 1.2: Three geometry features to represent real world spatial data.	7
Table 2.1: Healthcare data a) 4-anonymity b) 4-anonymity with 3-diversity (Machanavajjhala et al., 2006).....	20
Table 3.1: Outline of the seven obfuscation methods to generate the obfuscation area	39
Table 3.2: Description of the evaluation technique that have been used to evaluate and compare obfuscation method in terms of accuracy and privacy protection.....	43
Table 3.3: r_{min} calculated for different values of k_1 and r_{max} calculated for different values of k_2 when k_1 considered as six.	45
Table 3.4: Comparison of 2D KS-test and combination of 1D KS-test and correlation alteration of different obfuscation method	48
Table 3.5: Correlations and correlation alteration of spatial attributes with soil and herd related attributes using different methods.....	50
Table 3.6: List of 27 obfuscation methods that have been applied with the following parameters: $k = 6$, $N = 5$, $(\phi = 45^\circ / \phi = 225^\circ)$, $r_{min} = 20203$ m, $r_{max} = 36451$ m, $res = 521$, illustrating distribution preservation rank obtain from SK- test and correlation alteration.	51
Table 3.7: Statistical summary of point displacement for different obfuscation method	52
Table 3.8: The results of Spatial information divergence, Gdi and Ldi for different obfuscation method based on two spatial scales: Ldi2000 (fishnet cell size of 2000m and distance band of 23500m) and Ldi500(fishnet cell size of 500m and distance band of 900m).	55
Table 4.1: Outline of the six polygon-based obfuscation methods.....	72
Table 4.2: Description of the evaluation technique that have been used to evaluate and compare obfuscation method in terms of accuracy and privacy protection.....	76
Table 4.3: Comparison of 2D KS-test and combination of 1D KS-test and correlation alteration of different obfuscation method	79
Table 4.4: Correlations and correlation alteration of spatial attributes with soil and herd related attributes using different methods.....	80
Table 5.1: Principal Component Analyses of precipitation, minimum and maximum temperature features for Ireland 2016.	99
Table 5.2: Correlation matrix between 12 seasonal climate data including precipitation,	99
Table 5.4: Correlations and correlation alteration of spatial features with soil and herd related features using PESOM method.	103
Table 6.1: Result of different random forest model for original NSDB data set.	119
Table 6.2: Descriptive statistic for original test observed and RF3 model prediction of SOC.	119
Table 6.3: Comparison of RF model using internal (N, pH, P, K) and external features performance to predict SOC for original test dataset, obfuscated test dataset generated by PSEOM and Rand obfuscation methods.	121
Table 6.4: Comparison of descriptive statistic of original test observed and RF3 model prediction of SOC with model prediction SOC of obfuscated test data set from PSEOM and Rand obfuscation methods.	121
Table 6.5: Independent sample t-test results for predicted SOC of original dataset with predicted SOC of PESOM obfuscated test dataset and Rand obfuscated test dataset.	121

Table 6.6: Comparison of descriptive statistic of original test observed and MLR model prediction of SOC with model prediction SOC of obfuscated test data set from PSEOM and Rand obfuscation methods based on external features.....	123
Table 6.7: Comparison of MLR model using external features to predict SOC for original data, obfuscated data generated by PSEOM and Rand obfuscation methods.....	123

List Of Figures

Figure 1.1: Overview of research aim, questions.	11
Figure 2.1: Centroid (mean centre) and weighted mean centre of set of points.....	15
Figure 2.2: Three form of affine transformation (Armstrong et al., 1999).....	18
Figure 2.3: Obfuscation by enlarging the radius (a), shifting the centre (b), and reducing the radius (Ardagna et al., 2002).....	23
Figure 2.4: Substitution tree for Age attribute. (Parameswaran & Blough, 2008).....	26
Figure 3.1: Examples of obfuscated point generated by different methods: a) Rand, b) NRand and NMix c) Donut-k d) Density-k, e) Pinwheel225, f) Pinwheel45, g) AHilb and h) Donut-AHilb. Obfuscated area (grey) original point (black circle), 5 random point (black triangle), obfuscated point in NRand (white triangle) and in NMix (grey triangle), grey stars and black stars present existing point (parcels) in an inner and outer circle with $r_{min} = 291m$ and $r_{max} = 1088m$ radiuses around the original point based on 6-kanonmity in Density-k method. The numbers in g) and h) are Hilbert curve values, the obfuscation area for original point was generated considering first the button, top, right and left cells and then Hilbert curve value order.	34
Figure 3.2: Diagram to represent the relationship between field parcels and its unique centroid (point), ID and obfuscated point to retrieve field parcels after obfuscation.	38
Figure 3.3: a) r_{min} calculated for different values of k1, b) r_{max} calculated for different values of k2 when k1 considered as six	46
Figure 3.4: Rank of performance of different obfuscation method based on a) distribution preservation using 2D KS-test and b) combination of 1D KS-test with correlation.....	47
Figure 3.5: Original points (grey dot) and obfuscated points (black triangle) using a) NMix method and b) Density method in an area with radius 10000m	47
Figure 3.6: Total correlations alteration of spatial attributes with soil and herd related attributes for different methods.	49
Figure 3.7: Elbow method to find the optimal value of K	49
Figure 3.8: Comparison of statistical summary point displacement for different obfuscation method	51
Figure 3.9: 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.	54
Figure 3.10: The results of Spatial information divergence for different obfuscation method, a) Gdi, Ldi based on two spatial scales: b) Ldi2000 (fishnet cell size of 2000m and distance band of 23500m) and c) Ldi500(fishnet cell size of 500m and distance band of 900m).....	54
Figure 3.11: Estimated region of radius 500m after reverse geocoding for point 24 contains real points (triangles) and obfuscated points (small dots) using a) Density-Pinwheel45 method contains 8 real points and 5 obfuscated points. b) Density method contains 22 real points and 12 obfuscated points.	57
Figure 3.12: a) Obfuscated area contains k-1 other points (black triangle) to satisfy k-anonymity. b) Original field parcels with their centroids (grey area with black dots) and their obfuscated centres (black triangles) using density method in an area of 1000m radius, showing that obfuscated point might be in a another field parcels (risk of false identification), might be inside the other obfuscated area (black circle) or two obfuscated points might be inside the same original field parcels or obfuscated area (risk of non-unique obfuscation)	59

Figure 4.1: Conceptual diagram of obfuscation methods. a) and c) an example of a point-based Rand and Donut methods that transfers the original point (black dot in the centre) to a random location (red dot) inside a defined obfuscation area (yellow area), other points (black triangle). b) and d) an example of polygon based PRand and PDonut methods that transfer original location (represented here as the polygon with black dot in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area). The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon) are also represented in this diagram.....	69
Figure 4.2: Example of polygon based PAHilb (a) and PDonut_AHilb (b) obfuscation methods that transfer original location (represented here as the polygon with red star in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area). The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon).	71
Figure 4.3: a) r_{min} calculated for different values of k_1 and b) r_{max} calculated for different values of k_2 when k_1 considered as six.	77
Figure 4.4: Rank of performance of different obfuscation method based on distribution preservation using 2D KS-test and combination of 1D KS-test with correlation.	78
Figure 4.5: Elbow method to find the optimal value of number of cluster(k).	80
Figure 4.6: Comparison of polygon-based obfuscation methods based on a) TCA b) MCE, c) percentage of obfuscated polygons that satisfies k -anonymity for different values of k and d) percentage of obfuscated polygons with displacement less than different values of d	81
Figure 4.7: a) obfuscated polygon depicts as a circle with radius of an original polygon centroid to its farthest edge (or random edge). b) obfuscated polygon generated by randomly selecting polygon from a set of irregular polygons.....	86
Figure 5.1: Example of PESOM that transfer original location (represented here as the polygon with red star in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area), The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon).	95
Figure 5.2: PESOM method procedure	96
Figure 6.1: Correlation matrix between SOC from original NSDB dataset and a) soil related features including N, K, P and pH from original NSDB dataset. b) seven PCs climate data	116
Figure 6.2: Correlation matrix between 7 PCs climate data and SOC from obfuscated NSDB data generated by a) PESOM method and b) Rand method.	117
Figure 6.3: Comparison of performance of PESOM and Rand methods a) percentage of obfuscated location with displacement less than different values. b) percentage of obfuscated locations that satisfies k -anonymity for different values of k	118
Figure 6.4: RF3 model performance for test dataset. Scatterplot of Observed value of SOC vs predicted value.	120
Figure 6.5: RF model performance. Scatterplot of predicted value of SOC using original test dataset vs predicted SOC value using a) PESOM obfuscated test dataset and b) Rand obfuscated test dataset.	122
Figure 6.6: MLR model performance based on external features (climate features). Scatterplot of predicted value of SOC using original test dataset vs predicted SOC value using a) PESOM obfuscated test dataset and b) Rand obfuscated test dataset.....	124
Figure 7.1: Polygon-based obfuscation methods descriptions	135

ACRONYMS

Artificial Intelligent (AI)

Data Protection Commission (DPC)

Distribution Preservation Rank (DPR)

European Union (EU)

Food and Agriculture Organization (FAO)

General Data Protection Regulation (GDPR)

Geographical Information Science (GIScience)

Global Divergence Index (Gdi)

Global Positioning System (GPS)

Greenhouse Gas (GHG)

Irish nutrient management planning online (NMP Online)

Internet of Things (IOT)

Location-Based Services (LBS)

Local Divergence Index (Ldi)

Machine learning (ML)

Maximum Obfuscated Polygon (MOP)

Met Éireann Reanalysis (MÉRA)

Misclassification Error (MCE)

Multi Linear Regression (MLR)

Nearest Neighbour Data Substitution (NeNDS)

National Soil Database (NSDB)

polygon-based environmental similarity obfuscation method (PESOM)

Precision agriculture (PA)

Principal Component Analysis (PCA)

Random Forest (RF)

Soil Organic Carbon (SOC)

Social Research Ethics Committee (SREC)

Sustainable Development Goals (SDGs)

Sum of Square Error (SSE)

Total Correlation Alteration (TCA)

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Parvaneh Nowbakht

Acknowledgments

This thesis is a product of the help and support of many people. Firstly, I would like to thank my great supervisory team Dr. Paul Holloway, Dr. Fiona Cawkwell, Dr. Lilian O'Sullivan and Dr. David Wall for their encouragement, support, and guidance through all these years. I am grateful for their patience and positivity, criticism, and advice which made possible the completion of this project. I would like to express my special thanks to Paul and Lilian for their supervision of the nature of academic research and scientific writing.

I gratefully acknowledge the financial support of Teagasc for funding my research and three months of Higher Education Authority (HEA) extension financial support due to interruptions caused by the Covid-19 pandemic.

My dissertation would not have been possible without access to the (NMP online) datasets I have used. I am therefore indebted to all people from UCC and Teagasc for their help and assistance in the long process of providing a Data Sharing Agreement between UCC and Teagasc, General Data Protection Regulation (GDPR), Data Protection Impact Assessment (DPIA), Data Management Plan (DMP), and Ethical approval including Pat Murphy, Head of Environment Knowledge Transfer, Dr. Lilian O'Sullivan and Dr. David Wall from Teagasc, Dr. Paul Holloway, Dr. Aoife Coffey Research Data Coordinator, Catrina Solan Deputy Data Protection Officer, Liz Hales Ethics Support coordinator, Caitriona O'Sullivan Information Compliance Manager from UCC. Furthermore, I thank Padraig Foley Technologist - Nutrient Management Planning from Teagasc, Compass Informatics staff especially Ken Dowling IT Manager for providing the encrypted data through a secure data transfer process. Special thanks to Dr. Gourav Misra for providing the climate datasets (MÉRA) that were really important to the research.

I am also especially grateful for the support and help of the geography department staff including Agnes O'Leary, Suzanne Bermingham, and Mary Collins for administrative and technical support. In addition, I would like to thank Dr. Kieran Hickey and Dr. Denis Linehan, former and current head of the department. I would like to thank Teagasc staff, Hilary King and Katie Hetherington for administrative support and providing valuable training opportunities, workshops, and webinars. I am deeply grateful to the UCC Language Centre for offering the English for Postgraduate Studies courses. Thanks to Teagasc, the entire department, and UCC for providing me with a prestigious platform to learn and flourish as a

researcher. I also appreciate the unknown referee's valuable and profound comments and suggestions for my two published articles.

Finally, I thank my family. I wish to thank my Mum; Mahlegha, and my dad; Arghavan for being such an amazing source of inspiration and encouragement throughout my entire life. Huge thanks to my sisters and brothers especially Mehdi who took responsibility for taking care of our dad during his illness and gave me peace of mind to concentrate on my work. I also wish to thank my husband Farid who always be supportive and encouraging. Without his support and help during this journey, I would not be able to finish this dissertation. And huge thanks and appreciation to the most important people of my life my children Parmis and Parsa for being such supportive, patient, and understanding children. I dedicate this thesis to them.

Abstract

Geoprivacy protection is a controversial subject within the field of agri-environmental research. Agricultural transformation and digital farming are widely used in the agriculture industry. This enhances sustainable agriculture and subsequently sustainable development, and food security. At the same time, such technological advances can mitigate climate change by increasing agricultural productivity while protecting natural resources and reducing the environmental impact of agricultural activities. Digital farming generates a high volume of data including spatial data. This data can be used in data-driven modelling to improve agriculture systems and design and develop more sustainable agricultural policies and services to enhance sustainable agriculture. Sharing and making data accessible to a wide range of researchers and stakeholders is essential to gain maximum use of data. Data privacy and particularly geoprivacy protection is an essential factor when sharing agricultural data and needs urgent investigation.

To date, point-based obfuscation methods are the predominant approaches used to protect object confidentiality with spatial pattern and statistical accuracy preservation. In this research, functionality of point-based obfuscation methods for polygon nature objects was investigated. The terms “non-unique obfuscation” was introduced to the scientific literature for the first time. A high percentage of false-identification and non-unique obfuscation was recognized as the main drawback of point-based obfuscation when applied on polygon centroids as point spatial data.

Agricultural spatial data is often best represented as static polygon data with an association of attributes/properties of a field parcel or farm that includes coordinates, shape, size, topology, the relationship with the surrounding environment, and the impact of external factors on region characteristics that can be used to breach privacy. Agricultural spatial data with unique characteristics that distinguish them from other spatial data, therefore, require different geoprivacy protection techniques. Therefore, for the first time, to achieve a high level of geoprivacy protection, several polygon-based obfuscation methods including PN*Rand, PDonut-k, PDensity-k, PAHilb, PDonut_AHilb and PESOM methods were developed with consideration of these properties and avoidance of the occurrence of false-identification and non-unique obfuscation. The comparison of the performance of obfuscation methods, both point-based and polygon-based, in different aspects using various evaluation metrics indicated that the density-based obfuscation methods provided a better trade-off between level of confidentiality and accuracy. The results demonstrated that PESOM method maintained a high

level of geoprivacy protection and absolute environmental and climatic clustering preservation with no false-identification and non-unique obfuscation risk.

Following on from this, a case study was conducted to showcase the useability of obfuscated data to solve real world problems and highlight the importance of the choice of obfuscation method in outcome and suitability of obfuscated data for a certain purpose. Several evaluation metrics were developed to examine and assess the performance of obfuscation methods in terms of determining the level of privacy protection and spatial pattern preservation, and statistical accuracy. The results confirm the importance of choosing the right obfuscation method based on the influence of internal and external features on the results of the data-driven model. Therefore, the results of this research should be of wide interest to those working in GIScience, agri-environmental research, and computer science, and be of relevance to researchers and data managers.

Thesis Outputs

List of publication and output of this thesis are as follows:

- Chapter 3 is published as a peer-reviewed scientific article: Nowbakht, P., O'Sullivan, L., Cawkwell, F., Wall, D. P., & Holloway, P. (2022). A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research¹. *Transactions in GIS*, 26, 949– 979. <https://doi.org/10.1111/tgis.12892>
- Chapter 4 is published as a peer-reviewed scientific article: Nowbakht, P., O'Sullivan, L., Wall, D. P., & Holloway, P. (2023). Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data². *Transactions in GIS*, 27, 84– 104. <https://doi.org/10.1111/tgis.13009>
- A peer-reviewed conference paper and presentation: Nowbakht, Parvaneh, O'Sullivan, Lilian, Cawkwell, Fiona, Wall, David, & Holloway, Paul. (2021, April 6). A comparison of obfuscation methods for privacy protection: A case study using agri-environmental data in the Republic of Ireland². 29th Annual GIS Research UK Conference (GISRUK), Cardiff, Wales, UK (Online). <https://doi.org/10.5281/zenodo.4665691>
- A peer-reviewed abstract published as: Nowbakht, P. (2022). Geoprivacy protection of agricultural data², *The Boolean: Snapshots of Doctoral Research at University College Cork*, VI(1), pp. 179-183. doi: 10.33178/boolean.2022.1.29
- Chapter 5 is under review as a peer-reviewed scientific article: Nowbakht, P., O'Sullivan, L., Wall, D. P., & Holloway, P. (under review). Maintaining Environmental Context and Geoprivacy Protection in Agriculture. *Information Processing in Agriculture* journal.
- Chapter 6 is under review as a peer-reviewed scientific article: Nowbakht, P., O'Sullivan, L., Wall, D. P., & Holloway, P. (under review). Incorporating contextual factors in agri-environmental research- options for obfuscation to support wider data application. *Irish Journal of Agricultural and Food Research* journal.

¹ The Version of Record (VoR) of these manuscripts has been published. PN was lead author on these articles, developing the research question, methodology, analysis and results. As such there is no conflict for the content of this manuscripts to constitute part of this dissertation.

Role of Collaborators

The Ph.D. student Parvaneh Nowbakht was working under the supervision of Dr Paul Holloway and Dr Fiona Cawkwell from UCC and Dr Lilian O’Sullivan and Dr David Wall from Teagasc. This research has been funded by the Teagasc Walsh Scholarship programme [Grant 2018034].

All analysis presented within has been analysed by PN, with the exception of the climate data for 2015 -2016 and 2003-2004 that was obtained from de la Torre Cerro et al. (Under Review).

Chapter 1 General Introduction

1.1. Background

Agricultural spatial data are an important source of information in GIScience and agri-environmental research. To achieve sustainable agriculture and food security there is an urgent need to share agricultural data with a range of relevant stakeholders. Geoprivacy concerns while sharing agricultural data containing spatial data indicates the importance of data protection and need of a protection mechanism for such data. Geoprivacy protection techniques have been implemented in various disciplines especially in location-based services (Duckham & Kulik, 2005), healthcare (Hampton et al., 2010; Swanlund et al., 2020) and criminology (Kounadi & Leitner, 2015). But to date, geoprivacy protection approaches related to agricultural spatial data remain unexplored despite their unique properties and vital privacy concerns associated with this data.

1.1.1. Data Privacy Protection

The modern world is harnessing technology more than ever and advanced technology generates vast amounts of data. The advanced technology to collect, store and analyse data, (powerful, low-cost computers to process big data, cloud technology to store, machine learning and artificial intelligence for analysis) provide platforms to use and share such data via the internet which represents valuable information that can improve products and services. Therefore, data is a critical key asset that must be analysed to harness its potential to deliver on such product and service improvements. Availability of such data and open access data lead to a focus on methods that effectively use data and data-driven modelling. Montáns et al., (2019) have reviewed data-driven modelling in different fields of science and engineering such as physical and biological science, gas and oil industry, drug discovery and so on, which indicate the importance of using data to gain insight and generate knowledge. Lokers et al., (2016) examined three cases using big data in agri-environmental projects as interdisciplinary science to improve the handling of Big Data, including spatial-temporal data. Furthermore, combinations of data sources from different domains and backgrounds, or data sharing with

other organisation and researchers, offers further potential to increase and maximise the value of data.

Data therefore has intrinsic value and importantly, needs to be protected. Concerns about digital data transformation and personal data led to the adoption of the first national data protection law in 1973 in Sweden which only covered processing of personal data in traditional, computerised registers. There were not many material provisions regarding data handling or general principles of protection (Öman, 2004). With the advent of the Internet and its increasing use, the need for a mechanism for data protection became more important, which finally led to the implementation of European Union (EU) General Data Protection Regulation (GDPR) in 2018 (Lovell & Foy, 2018). The GDPR emphasises restrictions on personal data, in terms of collection, storage, usage and access. If data is accessible without permission, it might harm individuals or organizations in different ways such as by putting someone in an unsafe situation, impacting people's health, and risking financial loss or reputation. On the other hand, some data are considered private and sensitive data which by law must be protected. In addition to the right of taking legal action against a controller or a processor in case of infringing provisions of the GDPR, to ensure the implementation of the GDPR, regulatory authorities such as the Data Protection Commission (DPC) in Ireland have a wide range of remedial powers and sanctions such as:

- Issuing warnings and reprimands.
- Imposing a temporary or permanent ban on data processing.
- Ordering the rectification, restriction, or erasure of data; and
- Suspending data transfers to third countries.

Significant fines are also provided for companies that violate the provisions of the GDPR, for example, up to 4% of a company's global annual revenue or €20 million, whichever is greater².

According to article 4(1) GDPR personal data means “*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of*

² <https://www.itgovernance.eu/en-ie/dpa-and-gdpr-penalties-ie>

that natural person” and article 4(5) GDPR pseudonymisation means “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.”

Personal data can be identified by attributes such as name, identification number, location and quasi-attributes, a set of non-sensitive and non-identifier attributes that can be used to pinpoint individuals in the population (Machanavajjhala et al., 2006). Location information as an identifiable attribute can be in the form of text addresses or geographical data that can be converted to each other. Address matching refers to the process of transforming a text address into geographical coordinates and reverse the geocoding process, by converting the geographical coordinates into an approximate address on a map (Yildirim & Yomralioglu, 2004).

However, spatial data could be used to violate an object’s confidentiality. Therefore, in general, to comply with the GDPR, it is vital to design and develop techniques to protect the spatial identity while allowing data containing location information to be accessed and shared (Sui, 2014).

1.1.2. Agricultural Data

Agricultural sustainability is essential to the success of the UN Sustainable Development Goals (SDGs) (Bouma et al., 2019). The challenge of agricultural sustainability is to simultaneously achieve agricultural profitability; environmental quality; food sufficiency, quality and affordability (Laborde & Piñeiro, 2018). Sufficient food production for an increasing global population and at the same time, protecting and optimising use of the global natural resources such as soil and water resources are required to reach sustainable development (Godfray et al., 2010; Hrustek, 2020).

Agricultural activities are highly dependent on climate and environmental conditions. Climate change influences agriculture in different ways such as crop yield production, production quality, and agricultural pests. The impact is due to extreme weather events such as floods, droughts, and storms, seasonal and annual variability, increases in temperature and carbon dioxide (CO₂) levels, precipitation pattern changes, and reduction of water availability (Peng et al., 2020; Ray et al., 2019; Thornton & Lipper, 2014; Zhao et al., 2017). Therefore, climate change is an important factor for sustainable agriculture that increasingly requires adaptation strategies. Reciprocally, agricultural activities generate greenhouse gas (GHG) emissions including methane (CH₄), nitrous oxide (N₂O) and CO₂ which affect climate change. Land use management is a key factor to reduce atmospheric CO₂ concentration by increasing soil organic carbon (SOC) stocks which also play an important role in soil health, fertility, and crop production (FAO et al., 2022; Y. Wang et al., 2022).

An innovative solution is required in order to accomplish agricultural sustainability, and to protect soil from erosion and nutrient losses, to protect water and other natural resources while enhancing productivity. Precision agriculture (PA) is introduced as a management strategy to collect, process and analyse temporal, spatial and individual data and combine with other sources of data to improve agriculture sustainability (Colizzi et al., 2020). For example, using map or sensor-based variable-rate application for spraying fertilizers in fields can optimize the use of fertilizer, reducing cost and environmental impact (Ferris, 2017). In PA, information technology such as drones, robotics, machinery, automated hardware and software, sensors, Global Positioning System (GPS) guidance and digital equipment are widely used and generate vast amount of different types of data. Agricultural data can include agronomic data, farm business operations and management data, and data collected to provide general services

to agriculture as outlined in Table 1.1 which could be considered as personal and private data (Ferris, 2017; Jouanjean et al., 2020).

Table 1.1: Type of agricultural data and examples

Data type	Examples
Agronomic data Farm production process Tracking data (applied processes data)	Crop seed Dates of operations Water management Disease and pest management (type of herbicides, insecticide, fungicide used, and dates and location applied) Yield data Land data (Soil and fertility data, watershed, drainage, tillage practice) GIS GPS and field boundary data Livestock data (breed, genetics, feed, production)
Farm business operations and management data	Financial Tax Human resource Contracts Supply chain (partnerships, customer, and supplier information) Rolling and fixed asset data Machine operations data (fuel consumption, equipment function, reference) Reporting and compliance data (government policies, certification schemes)
Data collected to provide general services to agriculture	Climate and weather data Environmental and ecological data Commodity prices and market information

Agricultural data often contains spatial information associated to the field parcels with other types of information. Spatial data can be used to visualize this information, combine it with other sources of spatial data to gain insights into relationships and patterns to make better decisions, which might not be possible without it.

1.1.3. Agricultural Data and Geoprivacy Protection

Digital farming, including the use of robots, drones, internet of things (IOT) and sensors, artificial intelligent (AI) and data analysis can help farmers to improve their end products, lead to efficiencies that translate into cost reductions to maximize profits while

reducing the environmental impact. A further ancillary benefit is the simultaneously generated amount of agricultural data that can be implemented and analysed to further boost on aforementioned items. To increase farmers' willingness to adopt digital farming and share the generated data in their farms, the EU Code of Conduct on Agricultural Data Sharing (Manuel & Masseno, 2021) was launched which provided general principles for sharing agricultural data including data ownership, data access, control and portability, data protection and transparency, privacy and security, liability and intellectual property rights with focus on collection, storage, usage and sharing of data collected from farm. According to data protection and transparency principles, personal, or sensitive information should be pseudonymised and encrypted or protected with similar methods to reduce the risk of identification (Manuel & Masseno, 2021; Van Der Burg et al., 2021).

However, there is potential of breaching identification and violating confidentiality via agriculture spatial data. In addition to comply with the GDPR, in the case of agricultural data compliance with the EU Code of Conduct on the Sharing of Agricultural Data, is essential to protect agricultural data associated to location and to design and develop geoprivacy protection techniques suitable for such data.

1.1.4. Spatial Data

Spatial data describes and represents objects or events based on location information. Raster and vector data models are the two main models used to represent spatial data (Pingel, 2018).

A raster is a matrix or grid of cells (pixels) organized into rows and columns to represent the real-world features such as elevation, precipitation, etc. Each cell contains information related to the area it covers. Raster data models are mostly used to represent continuous attributes (e.g., elevation) but can also be used to represent discrete or categorical attributes such as land use. The cell value represents the average value of the area that cell covers for the continuous attributes and represents the majority value for discrete attributes (Longley et al., 2015; Pingel, 2018).

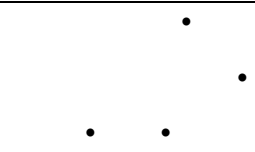
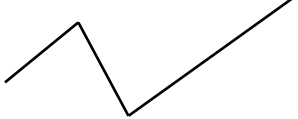
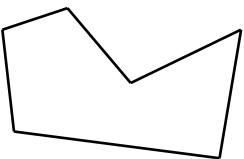
Vector models contain three geometry features (points, lines, and polygons) and are suitable to conceptualize discrete real-world objects such as buildings, roads, and lakes. A tabular form is used to store, analyse, and retrieve vector data which contains a pair of coordinates and other associated attributes. Points are the simplest feature with single

coordinate pairs (x, y) that represents either rows and columns in a projected coordinate system or latitude and longitude in a geographic coordinate system. Lines are defined by an ordered set of points (coordinate pairs). The starting or end point of a line is called node or vertex, and a segment line is selection of line between two ordered vertexes. A polygon is a closed curve defined by a set of connected lines whose start and end points are the same. The interior region of a polygon can be used to represent real-world regions such as lakes, or cities (Longley et al., 2015) (Table 1.2).

Advanced geospatial technology and a wide variety of spatial data have been used in a variety of applications to improve the quality of related services. For example, point data have been used in health monitoring and spatial crime analysis, while trajectory data contain rich information about people's travel patterns have been used to improve traffic forecasting and road planning. Geospatial analysis is used in the agriculture industry using spatial data collected on farms and generated by drones and satellites to crop yield monitoring, determine crop suitability, and boost crop production.

Examining, processing, and combining of spatial and spatiotemporal data with other sources of data enable us to discover underling spatial pattern, topology and relationships among objects that can answer complex questions that were otherwise not possible to answer a few years ago.

Table 1.2: Three geometry features to represent real world spatial data.

Spatial data type	Example	Vector model
Point	Building, people's location, farm	
Line	Road, trajectory of mobile user	
Polygon	Lake, farm, hospital	

1.2. Challenges and Solutions

1.2.1. Existing methods and related work

Geoprivacy refers to the “*individual rights to prevent disclosure of the location of one’s home, workplace, daily activities, or trips*” (Kwan et al., 2004 p. 15). However, there is potential of breaching identification and violating confidentiality via spatial data. Sensitive and private information can be retrieved via geographical coordinates by linking and matching this data with free online available data and services, social media, and websites. Geo-masking techniques, including spatial aggregation, affine transformation and randomization that aggregates, transforms or displaces the geographic location of an object (Armstrong et al., 1999b; Curtis et al., 2011; Kwan et al., 2004; Seidl et al., 2018), Spatial anonymization and obfuscation methods (Kern & Holz, 2013; Samarati & Sweeney, 1998; Zhong & Hengartner, 2008) have been developed and implemented within GIScience to reduce the risk of disclosure and protect sensitive data associated to location information by a process of degrading the quality of locational information while maintaining the spatial properties of geographical data (Armstrong et al., 1999b). Geoprivacy protection methods were widely used in various disciplines and applications, especially LBS, healthcare, and criminology where the location is represented as point spatial data. In Chapter 2 an extensive overview of the main methods that have been used in GIScience to-date is provided in more detail.

1.2.2. Conceptualization of geoprivacy protection methods

Obfuscation is a type of geoprivacy protection which was defined as a process of deliberately degrading the quality of location information to achieve a high level of geoprivacy protection (Duckham & Kulik, 2005). The obfuscation methods can be classified as vector-based or raster-based depending on the type of spatial data model and can be conceptualized into three types: point-based, trajectory-based, or polygon-based corresponding to the type of spatial data that were applied. To date, point-based methods are the predominant obfuscation methods that have been developed and implemented within GIScience. Trajectory data is spatiotemporal data that can be considered as a sequence of points in different time steps. Point-based obfuscation methods have been widely applied on spatiotemporal and trajectory data to obfuscate each point of trajectory data (Wightman et al., 2011; Zurbarán, Ávila, et al., 2014). Several point-based obfuscation methods were developed specifically for trajectory data with consideration of the properties of the spatiotemporal nature of trajectory data which can be considered as trajectory-based methods (Hwang et al., 2014; Kachore et al., 2015; Zurbarán,

González, et al., 2014). Despite the polygon nature of the real-world objects, to date, a polygon-based obfuscation method has not been found in the literature.

1.3. Research Gap and aims

Agricultural spatial data have unique characteristics that distinguishes them from other sources of spatial data, making it an important source of information in Geographical Information Science (GIScience) and agri-environmental research. Agricultural spatial data associated with a region of field parcels can be represented as a static polygon object. The properties of a field parcel represented as a spatial polygon contain spatial coordinates, shape, size, topology, and relationship with the surrounding environment and the impact of external factors on region characteristics. Therefore, in addition to the potential use of geographical location for field identification, other field properties can also be used towards a privacy violation. For example, an irregular field's shape can lead to easy identification of the field. Adjacent field parcels or topological relationships with other features such as roads, rivers, and mountains also can be used for field identification. To protect data confidentiality and reduce the risk of identification via location information, it is essential to obfuscate agricultural spatial data prior to sharing. Research on geoprivacy protection of agricultural spatial data or spatial data with similar properties has been very limited to date. This research seeks to bridge this gap and to contribute to the scientific field, through the development of obfuscation methods with consideration of polygon nature of agricultural spatial data.

1.3.1. Research Aims

The overarching aim of this research is to explore the characteristics of spatial agricultural data and examine, design, and develop obfuscation methods with consideration of these characteristics and properties. This can help to obtain a trade-off between geoprivacy protection and spatial pattern preservation and statistical accuracy. Agricultural spatial data is often static polygon data with coordinates, shape and size of the polygon, topology, and spatial relationship between adjacent polygons that can be used for identification and thereby breach privacy by sensitive information retrieval.

In addressing the primary research aim, this dissertation consists of several research questions that will be answered throughout the subsequent chapters and summarised in Figure 1.1:

- 1) How do existing point-based obfuscation methods perform on agricultural polygon data?
Through this research, I aim to identify possible drawbacks of using point-based obfuscation methods on polygon data. An extensive comparison of several point-based obfuscation methods on agricultural data (where a polygon centroid is considered as a representative point) is necessary to characterise the best obfuscation methods to provide the best trade-off between confidentiality and spatial pattern preservation along with other drawbacks potentially linked with point-based approaches.
- 2) Which properties of polygon data must be considered in development of polygon-based obfuscation methods? Considering that polygon-based obfuscation methods have not been found in the literature so far, the main aim of this study requires exploration and examination of polygon-based obfuscation requirements with consideration of polygon properties including the spatial coordinates, shape and size of the polygon, topology, and spatial relationship between adjacent polygons that can be used to identify real-world objects. Therefore, the aim of the chapter 4 and 5 is to explore, design and develop for the first time, polygon-based obfuscation methods that can take into account polygon properties and nature that can overcome any limitations found associated with the application of point-based obfuscation methods identified in the first part of this work.
- 3) Are proposed geometric-based obfuscation methods efficient in preserving contextual factors such as environmental and climate conditions in obfuscated data? Since agricultural activities both rely on environmental and climate conditions and affects the environment, it is important to understand to what extent obfuscated data generated by suitable obfuscation methods preserve the correlation between agricultural data and external environmental data. Importantly, it is necessary to determine whether new obfuscation methods that consider external properties of spatial data and the importance of contextual factors in agri-environmental research need to be developed.
- 4) How do choose an appropriate obfuscation method so that the research outcomes using obfuscated data, are the same or close to the research outcomes when the original data is used? Given the existence of considerable obfuscation methods, the challenge for data administrators is how to choose the appropriate obfuscation method regarding the purpose of the study and needs, investigate how the outcome of using obfuscated data is reliable and close to the outcome of the original data. To achieve highly accurate results in data-driven approaches it is important to consider the quality of data, type of modelling based upon available data and features selection. When considering if obfuscated data should be used instead of original data in addition to these factors, it is important to consider the level

of influence of a) internal and external features on research outcomes, b) priority of statistical accuracy and/or c) privacy protection in choosing the most appropriate obfuscation methods. In a final stage of this thesis a case study approach was presented to investigate the impact of the choice of the obfuscation method on the research outcome and usability of obfuscated data in relation to a real-world problem.

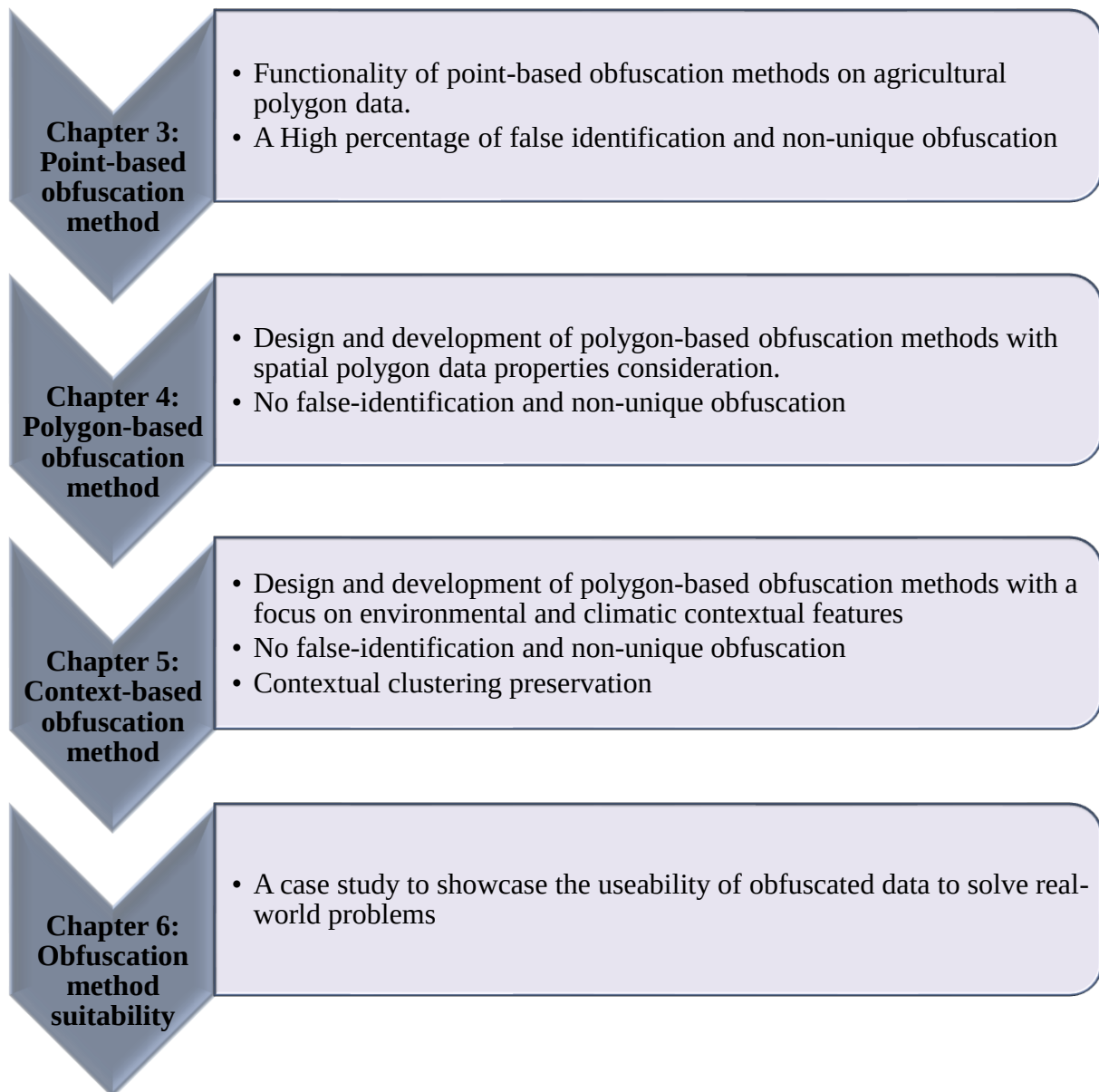


Figure 1.1: Overview of research aim, questions.

1.3.2. Research structure

The thesis is organised as follows:

- Chapter 2: “Point-Based Obfuscation Methods Review”, an overview of point-based obfuscation methods that are widely used in GIScience is provided.
- Chapter 3:” A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research”, explores how point-based obfuscation methods perform in terms of geoprivacy protection and statistical accuracy when applied on spatial agricultural data with polygon nature using centroid of polygon as a point representative. An extensive comparison of different point-based obfuscation methods is examined and evaluated to identify the opportunities and drawbacks of these methods when applying on agricultural data. In addition, the optimised size of the obfuscated area that can be attributed to the optimal level of privacy must be established.
- Chapter 4: “Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data”, the aim of this chapter is to develop polygon-based obfuscation methods that protect geoprivacy while maintaining statistical accuracy with consideration of polygon nature of agricultural data. These include spatial coordinates, shape, and size of the polygon and topological relationship as real-world objects that can be identified via these properties.
- Chapter 5: “Maintaining Environmental Context and Geoprivacy Protection in Agriculture”, the main aim of this chapter is to develop the obfuscation method in question 4.
- Chapter 6: “Incorporating contextual factors in agri-environmental research- options for obfuscation to support wider data application”, this chapter aims to compare the estimation of soil organic carbon (SOC) based on internal and external features using original data and obfuscated data generated by different obfuscation methods. For this purpose, a case study to show the usability of obfuscated data applied to a real-world problem is developed.
- Chapter 7: “General Discussions”, represents a summary of the work covered in previous chapters, highlighting important results and limitation. Suggestion and recommendations for future research based on the finding of this research are proposed.

Chapter 2 Point-Based Obfuscation Methods Review

2.1. Introduction

A point is the simplest geometry feature or vector data type used to represent a location. Each point can be considered as a location of buildings, farms, crime locations, healthcare services, etc, that each consists of a coordinate pair and the individual information associated to that location (e.g., attribute). The locations essentially reveal the identity of individuals and their information which might contain personal and sensitive data. To reduce the risk of identification and protect the data privacy for point-based data several obfuscation methods have been developed (e.g. Kwan, Casas and Schmitz, 2004; Marc, Gerard and Dale, 1999; Montazeri, Houmansadr and Pishro-Nik, 2016; Seidl, Jankowski and Clarke, 2017) which can be classified using the following conceptualisations:

- 1) Aggregation
- 2) Affine transformation
- 3) Anonymization
- 4) Randomization

2.2. Aggregation

2.2.1. Spatial Aggregation

In general aggregation is a common approach to preserve an individual's information by cumulating and aggregating the individual level data in a representative new object. In tabular form all records that match the specific condition or criteria can be aggregated and represented as a summary of those records (Murad et al., 2014; Zandbergen, 2014). For example, the statistical summary of farms that produce certain types of crops can be aggregated together to obfuscate private information. This can be controlled further, for example crop production below specific thresholds (e.g., less than 10 tonne) further obfuscates this data. In the context of geoprivacy, spatial aggregation is used to aggregate the individual's information in a specific region. There are two types of spatial aggregation, areal aggregation, and point aggregation.

2.2.2. Areal Aggregation

In this approach data are aggregated in a predefined areal unit such as block group, census tracts or zip/postal code. Census data widely apply this method to preserve the privacy and also release statistical reports for the required region (Kwan, Casas and Schmitz, 2004). Areal aggregation transforms a set of point identifiers (i.e., two coordinates) to a polygon's identifier. Then an appropriate point inside the polygon such as centroid (mean centre) of a polygon or weighted mean centre can be used to represent the polygon region (Figure 2.1). The spatial data distribution is a key factor to select the representative point inside the region. When spatial data is uniformly distributed in space, the mean centre would be an appropriate representative point, but for non-uniform distributions, a weighted mean centre would represent the region more accurately as it would be closer to the part with higher density. Areal aggregation also reduces the data dimensionality as it aggregates similar data with two dimensions (i.e., coordinates as an identifier) inside the specific region and represents it as one polygon identifier. It creates a statistical summary of data in that specific region (e.g., a representative point of areal aggregation contains a statistical summary of 200 individual points inside the specific region). The confidentiality of this approach depends on the population density and the required spatial resolution.

Curtis et al., (2011) aggregated a dataset containing breast cancer cases in women (all ages, race/ethnicities) occurring in the Los Angeles County Cancer Surveillance Program, between 1998 and 2007 in five ways including two with traditional boundaries (census tract and census block group), and three regular grids (of sizes 1, 0.75 and 0.5 km). Results suggest that across the scales of aggregation tested in this paper, even at the small grid, there was relatively little risk of reidentification. Areal aggregation might not provide a high level of spatial analysis and cluster detection due to the loss of individual information and spatial details (Hampton et al., 2010). The results highly depend on units of aggregation which indicates the importance of the choice of a unit of aggregation for different purposes.

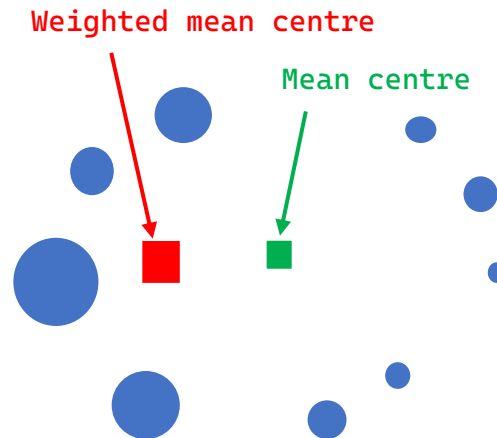


Figure 2.1: Centroid (mean centre) and weighted mean centre of set of points.

2.2.3. Point Aggregation

Geospatial data contains a location identifier as a primary key which can easily be used to detect and access the individual's private information. Point aggregation is a process that aggregates the subset of original data and represents it as a new location that satisfies both the required level of confidentiality and keeps the underlying patterns in the data. This subset can be defined as a fixed number of individuals or can be defined based on the data similarity and classification. Data similarity and pattern can be found in spatial data (e.g., upstream and downstream of a river, census tract, etc) or can be found in other attributes associated to spatial data (e.g., locations with similar weather conditions, land cover pattern and similarity, etc.). However, classified data might not be necessarily in the same geographical area.

Machine learning (ML) algorithms can be employed in geospatial data in order to detect data similarity and divide original datasets to subsets. When the subsets of original data are defined, then a set of new locations should be selected as a representative of each subset. The simplest method to compute the representative location of a subset is the average coordinate method. In this method the representative location coordinates of each subset is computed by taking average coordinates of all points inside the subset (Armstrong et al., 1999b). Location-allocation is another method to compute the representative location of each subset by minimizing the sum of distances from this location to original locations within each subset (Armstrong et al., 1999b).

Aggregation might satisfy the privacy requirements, particularly when it is applied on large number of individual level data, but it reduces the accuracy of the spatial analysis

especially for non-uniform spatial distribution. For example, in a region with a population of 20,000 and 20 patients the probability of randomly selecting a patient is $1/1000$. However, if the population is 2,000 then there is $1/100$ chance to randomly select a patient. This example indicates that when aggregation is applied on large number of individuals the probability of identification of individual level data decreases.

2.3. Affine Transformation

An Affine transformation is an alternative way to preserve location privacy while releasing individual level data for research and analyses. Transformation is a process of converting a set of locations to a new set of locations which keeps the original locations pattern. Affine transformation can be applied in three forms (translate, scale and rotation Figure 2.2) or combination of these three forms (Armstrong et al., 1999b; Kwan et al., 2004).

Translation alters original locations by a fixed increment (Figure 2.2a). Each location with coordinates of (x_i, y_i) moves to a new location $(x_i + \Delta x, y_i + \Delta y)$ where Δx and Δy are displacement constants. This approach preserves the location distribution, density, and locations pattern. It is important to choose the fixed increment in both coordinates, Δx and Δy to maintain the original location pattern.

Scale is another affine transformation that moves a location with coordinates of (x_i, y_i) to a new location $(\alpha x_i, \beta y_i)$ where α and β are scaling constants (Figure 2.2b). A trade-off between protection of geoprivacy and the accuracy of analytical results is a key issue to select the α and β . Large α and β transforms the original points to farther locations which might be out of study area boundaries, and significantly alters the location distribution and density, as well as consequently reducing the accuracy of analytical results. While for small α and β the location distribution and density are slightly changed with higher level of the accuracy but transformed locations are not far from the original locations with higher disclosure risk. This approach maintains the locations pattern but may change the location distribution and density.

In this approach each location is rotated by fixed angle about a pivot point (the centre point of any rotational system). A pivot point can be any points particularly origin of coordinate system. It moves a location with coordinates of (x_i, y_i) to a new location $(x_i \cos \theta - y_i \sin \theta, x_i \sin \theta + y_i \cos \theta)$ where $0 \leq \theta \leq 2\pi$ and the pivot point is the origin of coordinate system (Figure 2.2c). Rotation maintains the locations pattern, location distribution and density.

Armstrong et al., (1999) conducted a study of infant mortality rate in Des Moines, Iowa and applied affine transformation on infant deaths and births to preserve geoprivacy and reduce the risk of identification. The results of this study indicated that each of the affine transformations preserved (exactly) the existence of clusters, trends, and anisotropies

(direction-dependent spatial correlation). but did not preserve, even approximately, the location of clusters and event-geography relations.

However, Affine transformation with high risk of decoding through location re-engineering which cause the high risk of identification and low privacy protection is not widely used (J. Wang et al., 2022).

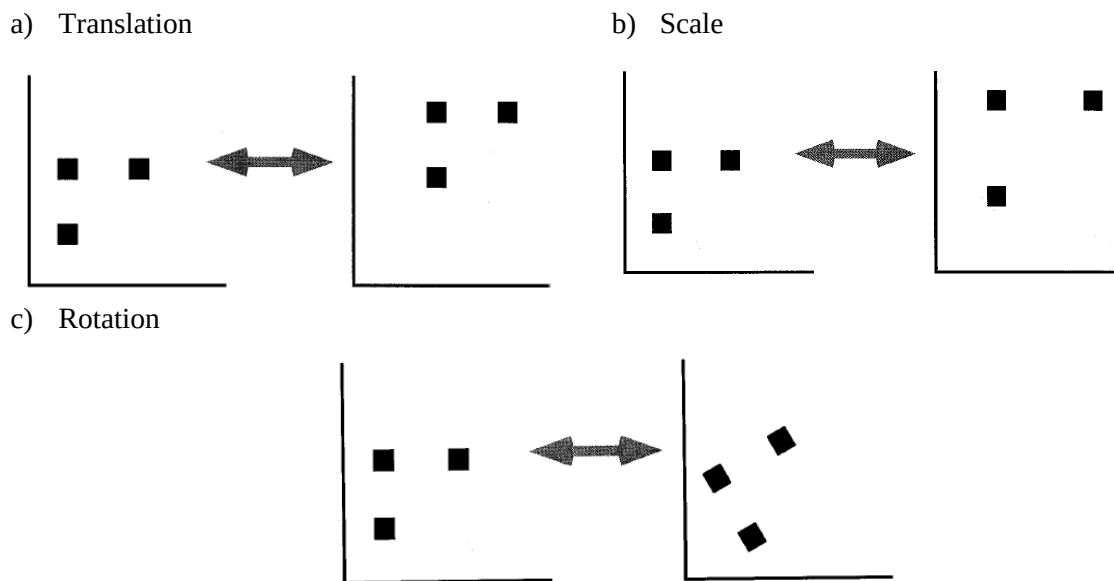


Figure 2.2: Three form of affine transformation (Armstrong et al., 1999)

2.4. Anonymization

Anonymization protects an individual's private or sensitive data by suppression, which is removing or encrypting the identifiable attributes such as name and ID in shared datasets or alternatively, using generalisation which involves replacing values of attributes with more general values or value ranges. For example, using the three first digits of zip code instead of all digits in Table 2.1a (Zang & Bolot, 2011). However the individual's sensitive attributes can be detected by quasi-identifiers, (Machanavajjhala et al., 2006). k -anonymity was defined to preserve the identification of any individual object using any combination of quasi-identifiers attributions by generating a group of at least $k - 1$ other individuals (Samarati & Sweeney, 1998). The sensitive information can still be leaked, though having external or background knowledge, which is known as a background knowledge attack (Rajendran et al., 2017; Samarati & Sweeney, 1998). For example, if a person knows his/her neighbour is a 45 old woman who has some health problem, by looking at the healthcare report available online that contains zip code, gender, age interval and disease (as a sensitive attribute) with 4-anonymity, he/she would be able to guess his/her neighbour's disease (Table 2.1a). Subsequently, a larger value of k might be needed to reduce the risk of background knowledge attacks, particularly when there is a wide diversity in a sensitive attribute e.g., type of disease.

Homogeneity attack is another drawback of the k -anonymity approach which occurs when there is no diversity in a sensitive attribute among k members of a group. For example, in Table 2.1a, there is no diversity in group 3 with sensitive condition 'Cancer', meaning that identifying and access to the sensitive information would be easily possible. To overcome this limitation, ℓ -diversity was introduced. This method groups points together in order to satisfy k -anonymity, whereby each group has at least $\ell - 1$ different sensitive values (Kern & Holz, 2013). In Table 2.1b, there are at least 3 different types of sensitive condition in each group to satisfy anonymity constrain.

Spatial k -anonymity exploits the concept of k -anonymity in order to protect the identity of user/location and reduce the risk of re-identification among set of locations. Spatial anonymization expands the area which contains other locations to make the individual's location unidentifiable (Zhong & Hengartner, 2008).

Spatial k -anonymity and location ℓ -diversity regarding the location privacy can be described as follows:

- Spatial k-anonymity: is a simple method which generates an area (neighbourhood) as an obfuscation area that contains the original point and $k - 1$ other points to protect the privacy of the original point. k is the minimum number of other locations that the original location is unidentifiable between them. The relationship between k and size of neighbourhood depends on population density (Montazeri et al., 2016).
- Location ℓ -diversity: is a method that generates at least $\ell - 1$ semantically different areas that satisfy spatial k-anonymity (Kounadi & Resch, 2018; Machanavajjhala et al., 2006; Rajendran et al., 2017). Xue, et al., (2009) defined a semantic location as a region on the map that typically gathers a user population that has similar contextual information, such as age, gender, activities, etc. Examples of semantic locations are schools, hospitals, golf courses, or the office of a specific company. This definition can be expanded to a region with similar characteristic properties such as land cover, geographical area, elevation, weather condition, etc. Examples of such semantic locations are residential area, county, forests, farms, etc.

Table 2.1: Healthcare data a) 4-anonymity b) 4-anonymity with 3-diversity (Machanavajjhala et al., 2006)

a)

		Non-Sensitive			Sensitive
		Zip Code	Age	Gender	Condition
Group 1	1	130**	<30	*	Heart Disease
	2	130**	<30	*	Heart Disease
	3	130**	<30	*	Viral Infection
	4	130**	<30	*	Viral Infection
Group 2	5	148**	>40	*	Cancer
	6	148**	>40	*	Heart Disease
	7	148**	>40	*	Viral Infection
	8	148**	>40	*	Viral Infection
Group 3	9	130**	3*	*	Cancer
	10	130**	3*	*	Cancer
	11	130**	3*	*	Cancer
	12	130**	3*	*	Cancer

b)

		Non-Sensitive			Sensitive
		Zip Code	Age	Gender	Condition
Group 1	1	1305	≤40	*	Heart Disease
	4	1305*	≤40	*	Heart Disease
	9	1305*	≤40	*	Viral Infection
	10	1305*	≤40	*	Cancer
Group 2	5	1485*	>40	*	Cancer
	6	1485*	>40	*	Heart Disease
	7	1485*	>40	*	Viral Infection
	8	1485*	>40	*	Viral Infection
Group 3	2	1306*	≤40	*	Heart Disease
	3	1306*	≤40	*	Viral Infection
	11	1306*	≤40	*	Cancer
	12	1306*	≤40	*	Cancer

Xue et al., (2009) used Location ℓ -diversity to improve user location privacy in Location-based Services (LBS) for a California dataset and synthetic data (randomly generated semantic locations in the map). Location Diversity improves Spatial k-anonymity by ensuring that each query can be associated with at least different semantic locations (e.g., school, shop, hospital, etc.). The algorithm was developed that generate groups of semantic locations, which preserve privacy and minimize the expected query processing and communication cost.

2.5. Randomization

Randomization adds noise to a region of interest or moves the original point to a new location in a random direction and random distance to protect the individual's geoprivacy and maintain the spatial patterns of points (Seidl et al., 2018). Randomization is a two stages procedure. Firstly, the obfuscation area is generated to satisfy k -anonymity and secondly the random point is selected inside this area.

2.5.1. N*Rand and N*Dispersion methods

N*Rand and N*Dispersion methods were introduced and applied in a study to generate an obfuscated path of a mobile user while using location based services (LBS) (Wightman et al., 2011). The performance of these methods was evaluated and compared in terms of average distance, maximum and minimum distance between the original paths and the obfuscated paths. N*Rand and N*Dispersion methods are point-based methods which when applied in a specific locations of user path to generate obfuscated path.

Rand is a simple randomize method which transforms the original point p to a random point inside a circular obfuscation area with the centre p and radius r . The choice of the random point as an obfuscation point depends on the size of the obfuscation area, which is dependent on spatial distribution, population density, attribute sensitivity and level of obfuscation (Murad et al., 2014; Wightman et al., 2011). The uniform distribution can be used to give the same probability to each point being selected (Wightman et al., 2011). Alternative non-uniform distributions such as Gaussian and Bimodal Gaussian distributions can be used to consider other criteria for selecting an obfuscated point (Murad et al., 2014; Zandbergen, 2014).

N*Rand methods including N*Rand and N*Mix are variations of Rand method. The obfuscated point is either the farthest or average point within N random points (or N random points from specific set of points) inside the obfuscated area to ensure that the obfuscated point is not too near or too far from the original point area.

In order to reduce the risk of identification, a new parameter N (number of random points) was introduced to the alteration Rand method called N*Rand methods including N*Rand and N*Mix (Wightman et al., 2011). The obfuscated point instead of a random point inside obfuscation area which might be too close or too far from original point would be the farthest or average point among N random points inside the obfuscated area. The obfuscated point in N*Mix method is the average of N random points (average of each coordinate) inside the circle

of (p, r) or obfuscation area. A disadvantage of this method is raised when N random points are uniformly distributed inside the obfuscation area and the obfuscation point, the average of these uniformly distributed points tends to be very close to an original point with high risk of identification. To avoid selecting N uniform random points inside the obfuscation area, this area can be divided to m parts (e.g. four quadrants) and applying NMix method in a randomly selected part or a part with higher population density (Wightman et al., 2011).

N*Dispersion method is another modified N*Rand method which Recursively applies one of N*Rand methods to shift the original point p to a new point p' inside an area of circle (p, r) , called shift the centre and again applies one of N*Rand method inside an area of circle (p', r') . The size of obfuscation areas (r and r') depends on the level of obfuscation and population density in these areas. This can be applied in S-E (Shift the centre & Enlarging the radius) when $r' > r$ or S_R (Shift the centre & reduce the radius) when $r' \leq r$ (Ardagna et al., 2002; Wightman et al., 2011). The terms of “Shift the centre”, “Enlarge the radius” and “Reduce the radius” are adapted from (Ardagna et al., 2002) to describe the double obfuscation. Ardagna *et al.*, (2002) has defined the location measurement of a user (cellular phone) based on sensing technologies as a circular area, Area (r, xc, yc) , centred on the geographical coordinates (xc, yc) and with radius r , which includes the real user's position (xu, yu) with probability $P((xu, yu) \in \text{Area}(r, xc, yc)) = 1$. Based on this definition three obfuscation methods illustrated in Figure 2.3 were introduced. Double obfuscation which is a two-step combination of these methods can be E-S, S-E, S-R and R-S.

Similar to the N*Rand method, Zurbarán, Ávila, et al., (2014) developed the Near-Rand method to generate an obfuscation point inside a large, squared shape as an obfuscation area (region of 100 km^2 in Takoradi, Ghana) instead of circular obfuscation area in N*Rand method. the obfuscated point would be an average of the N nearest points to the original points among p random points.

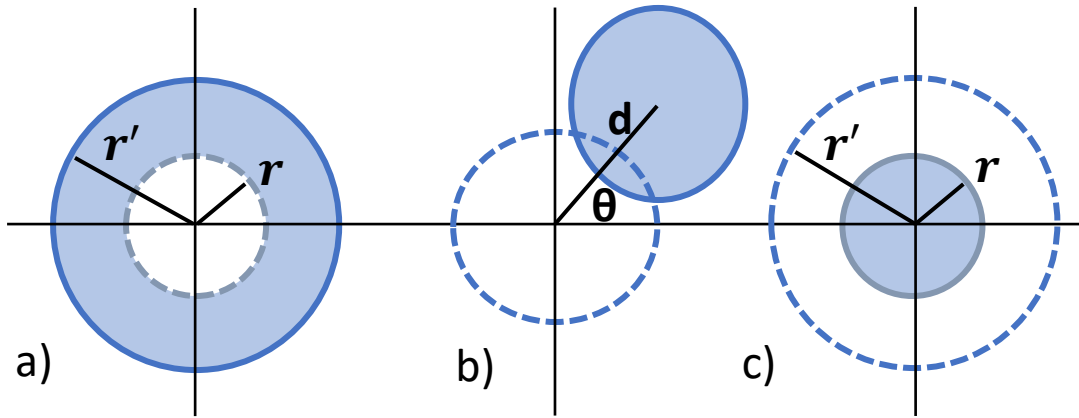


Figure 2.3: Obfuscation by enlarging the radius (a), shifting the centre (b), and reducing the radius (Ardagna et al., 2002).

2.5.2. Donut method

The Donut method is a modified randomize method similar to the Rand method. To ensure the obfuscation point would not be an original point (p) or too near p , two parameters minimum and maximum distance from original point ($rmin$, $rmax$) are defined to generate a donut shape with centre p and radiuses ($rmin$) and ($rmax$) as an obfuscation area. the obfuscated point can be generated using one of N*Rand methods (Hampton et al., 2010; Zandbergen, 2014).

The Donut method widely used in various studies such as healthcare and criminology to protect geoprivacy(Hampton et al., 2010; Lu et al., 2012; Manley, 2010). For example, Lu, Yorke and Zhan, (2012) focused on improving the Donut method by using the actual distribution of the residential addresses (or “risk location”) rather than the people (or “risk population”). In that study a total of 752 burglaries in the study area consisting of five block groups in San Antonio, Texas in 2006 were extracted and four different conditions based on risk location and risk population was considered to define the minimum and maximum distance or radius of donut shape. However, the results showed this method that considered risk location protected privacy more effectively than similar methods based on population size, with the obfuscated data tending to be clustered than the real case distribution.

2.5.3. θ -RAND and pinwheel methods (Trajectory)

The θ -RAND and pinwheel methods are modified randomize methods similar to the NRand method which is designed for applications that constantly keep track and update the

users' location to a service provider (Wightman et al., 2013, 2014; Zurbarán, González, et al., 2014). The θ -RAND generates n random points inside a random circle sector (instead of full circular area) with centre of original point p and radius r and then selects the farthest point or average points as an obfuscated point. The difference between the random angles θ_1 and $(\theta_1 + \theta_2)$ where θ_1 and θ_2 are between $[0^\circ, 180^\circ]$ is calculated to generate the random circle sector (Wightman et al., 2013).

In the pinwheel method the obfuscated point is a farthest point among n random points in a pinwheel shape with centre of original point p and radius r . Equation 2.1 generates the pinwheel shape for any α between $[0^\circ, 360^\circ]$

$$r_\alpha = \frac{(\alpha \bmod \phi)}{\phi} r \quad \text{Eq 2.1}$$

where ϕ indicates the degree between two sequential wings and r is a maximum radius of pinwheel. Wightman, Zurbaran and Santander, (2014) found that the average behaviour of the pinwheel method is heavily dependent on the value of ϕ and number of wings. The average performance of pinwheel for 191° , is very similar to the results of the RAND method, and for 45° the pinwheel method and θ -RAND shows similar performance.

2.5.4. Location Swapping

In the geo-masking and randomization methods described in previous sections, the original point is moved to a random point inside an obfuscated area. This obfuscated random point may be located outside of study area boundaries or fall inside an irrelevant area such as water bodies. Zurbarán et al.,(2018) applied the N-Rand-K method, a combination of N-Rand and k-anonymity on geo-referenced data collected from the Twitter® streaming API during the earthquake in Central Italy on August 24th of 2016. In this study the obfuscated points, fell outside the boundaries were not considered for aggregation, which affected the spatial analysis. Some of obfuscation points also fell in water bodies and inhabitable places. To overcome this drawback of geo-masking methods Zhang et al.,(2017b) introduce the location swapping method which is an improved geo-masking methods. Instead of selecting any random points inside the neighbourhood area of original point, this method selects a random point from a specific set of points (the specific sets might contain similar characteristic locations to the original point). In the Zhang et al (2017b) study, three US counties were selected for the analysis including Jackson County in Oregon, Travis County in Texas, and Wake County in North Carolina. These three counties were selected for geographic representation as well as for

the varying range of urban/rural population densities. The results of this study indicated that the location swapping method are more effective than Rand methods in prohibiting re-identification. The Nearest neighbour and road proximity analyses results indicated that cluster and distribution patterns are more similar to the pattern of unmasked data in the location swapping method. The land cover association analysis results revealed that the location swapping techniques can provide distribution of masked locations that are more similar in geography (land cover types) than Rand methods.

2.5.5. Nearest Neighbour Data Substitution (NeNDS)

NeNDS method was developed to protect individual private and sensitive clustered data in tabular form. There is a potential to modify and expand this method to apply in the context of geoprivacy (Parameswaran & D. Blough, 2005). NeNDS method is a two steps process. First, the dataset is partitioned into k similar groups or clusters called neighbourhood. Machine learning methods such as K -means clustering can be used to find the similarity among datasets. The data preservation is highly dependent on k (the number of neighbourhoods). A Larger k means a small number of elements in each neighbourhood, which can lead to having less privacy protection. While a smaller k and larger neighbourhood might change the correlation among data attributes.

In next step, to ensure each individual's attribute in the substituted dataset is different from individual's attribute in the original dataset a substitution process is applied for each cluster subject to following criteria:

1. swapping between two individuals attributes are not allowed.
2. substitution should be applied for any individual's attributes.
3. substitution is not allowed between identical individual attributes.

The tree traversal approach was suggested to obtain an optimum substitution pattern. The recursive process to create a tree for each attribute in each neighbourhood (cluster) and find the substitution sequence can be summarised as follows:

The root of tree set to the first element of a cluster. The children nodes set to the remaining elements in a cluster ordered from left to right based on their distance from their parent (left to right ordering). The distance between a parent and child called cost edge is assigned to the edge connection of the parent and child. Recursively each child becomes a parent of a subtree with the remaining elements which is not in its path. A depth first approach

from left to right is applied to expand each node, which means each time the leftmost child is expanded until to reach the leaf. The path with the least edge cost will be used in the substitution sequence. This approach significantly reduces the complexity of the search process due to the removal of those paths with a greater edge cost than the minimum edge cost of leftmost path.

Figure 2.4 shows this process for a cluster containing an Age attribute.

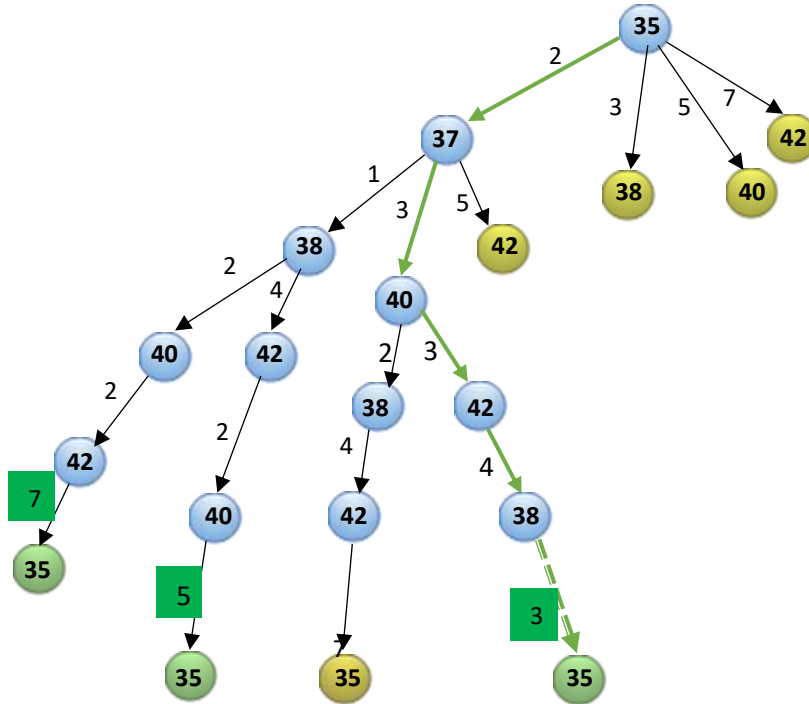


Figure 2.4: Substitution tree for Age attribute. (Parameswaran & Blough, 2008)

2.6. Conclusion

A major challenge of sharing spatially explicit agricultural and agri-environmental data is to identify the trade-off between field parcel confidentiality and spatial pattern preservation.

Although, spatial data can be represented in vector format as points or polygons, to-date, the predominant obfuscation methods have been developed and implemented on point features. In this chapter an extensive review of point-based obfuscation methods is provided to identify the most suitable methods for implement on centroid of polygon agricultural spatial data as point data. Due to the loss of individual information and location details in aggregation methods and maintaining the relationship and topology between a set of locations in affine transformation, which increases the risk of identification, these two types of methods were

not recognized as suitable methods in this research. Therefore, randomization and anonymization methods and combination of them which generate different shapes of obfuscation area, such as buffers, donuts, pinwheels, and squares was considered for further examination. In Chapter 3, the performance of 27 point-based obfuscation methods were evaluated on agricultural spatial data to examine the suitability of point-based method on polygon data. Furthermore, several evaluation methods were tested to measure the ability of each method to satisfy both privacy concerns and spatial pattern preservation.

Chapter 3 A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research.

This chapter also published as a peer-reviewed scientific article:

Nowbakht, P., O'Sullivan, L., Cawkwell, F., Wall, D. P., & Holloway, P. (2022). A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research. Transactions in GIS, 26, 949– 979. <https://doi.org/10.1111/tgis.12892>

3.1. Introduction

Advances in geospatial technology have generated a significant amount of data in many disciplines and industries. While data sharing is essential for many companies and organisations to improve efficiency and generate knowledge, data are prone to privacy violations, including data breaches. Geoprivacy concerns have been a challenge of GIScience research, particularly in relation to legal, ethical and technical aspects as spatial data can be used to identify individual objects and their private information (Sui, 2014). Such concerns have been propelled to the forefront of GIScience and the wider scientific fields in recent months, particularly in relation to contact tracing methods during the Covid19 pandemic health crisis (Kim and Kwan 2021).

Spatial anonymization, obfuscation and geo-masking methods have been implemented within GIScience for the last 20-30 years to preserve the individual's privacy by a process of degrading the quality of locational information while maintaining the spatial properties of geographical data (Armstrong et al., 1999b). Obfuscation methods are widely used in some fields and sectors especially in location-based services (Duckham & Kulik, 2005), healthcare (Hampton et al., 2010; Swanlund et al., 2020) and criminology (Kounadi & Leitner, 2015) whereby the performance of point-based methods to preserve object privacy and spatial pattern preservation is explored.

Sustainable land management is considered central to achieving the Sustainable Development Goals (SDGs) by 2030 (Bouma et al., 2019). Sustainable land use and management to meet the SDGs requires an innovative system that can protect the natural resource base while reaching productivity demands for food security. More targeted policies and practices rely in part, on leveraging data for development of innovative solutions for optimal farming. The use of GIScience and geospatial technologies has increased in agriculture and digital farming in recent years, generating, accessing and sharing large amounts of spatial and spatial-related data to enable farmers to make effective use of fertilizers and, pesticides, as well as effectively protecting the environment and resources, while increasing production, reducing costs, and supporting sustainable agriculture (Ferris, 2017). From a government perspective, the ability to access and use agriculture data collected on farms particularly spatial agri-environmental data, enable the design and implementation of better, more sustainable agricultural policies for farmers and society (Jouanjan et al., 2020). However, this relies on the access, availability and utilisation of data often collected for another purpose. There is a

direct relationship between farmer's willingness to adopt digital farming technology and the availability and accessibility of agricultural data. Fears and a lack of trust about data collection and sharing exists amongst farmers for number of reasons. One survey of 1000 farmers in Australia indicated that farmers lacked knowledge and understanding of the terms and conditions of their agreements with technology services providers (Wiseman et al., 2019). Furthermore, farmers were concerned about data ownership, preserving their data privacy and the use of their data for profit making and/or sharing with third parties (Wiseman et al., 2019). Increasing trust between farmers and data users is required, which has an implicit need for data protection techniques.

The privacy preservation and confidentiality of agricultural data is challenging. Private and personal data can be identified through breaches via farm's geographical coordinates even where other identifiers such as farm name or owner's name and address were anonymized or removed from dataset. Agricultural data can be identified via geographical coordinates by linking and matching this data with free online available data and services, social media and websites. For example, retrieving farm owner information using a farm address which can be extracted using freely available online reverse geocoding map services from geographic coordinates available in agriculture data. Linking personal information with agriculture data, could allow personal or commercially sensitive and private information to be retrieved (Kounadi et al., 2013). Farm information, such as type of crop, size of farm, type and number of animals, type of fertilizer, etc. may be used for interests other than the farmer's interests such as for informing insurance or, farm price or for marketing purposes. Equally, these data can be extracted by government agencies and used for regulatory purposes, for example, to verify compliance with environmental and animal welfare standards, or fiscal obligations. However, to-date, few studies have been developed and conducted that specifically examine geoprivacy on agricultural data, despite the unique spatial patterns and privacy concerns associated with such features which can be identified by their location information among other attributes.

A rare example of using obfuscation methods in agricultural research is a study by Claridge *et al.* (2012), 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. In this study, 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. Such a method represents

the simplistic end of the obfuscation continuum, meaning there are several obfuscation methods that could be used within an agricultural context, but to-date they remain unexplored despite their ability to reduce risk of identification and increase spatial pattern preservation. 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 upon that have a point base and dynamic characteristic such as location of crime, patient, or mobile users. Therefore, the aim of this study is to compare several point base obfuscation methods on agricultural data, focusing specifically on how the polygon nature of this spatial data deals with point-based obfuscation techniques and investigate how they perform in terms of privacy protection and accuracy.

Existing methods and related work

A point is the simplest geometry feature or vector data type used to represent a location. Each point represents a location, that consists of a coordinate pair and the individual information associated to that location. The locations can essentially reveal the identity of individuals' information which might contain personal and sensitive data. To reduce the risk of identification and to protect data privacy for point-based data several obfuscation methods have been developed (Armstrong et al., 1999a; Kwan et al., 2004; Montazeri et al., 2016; Seidl et al., 2018; J. Wang & Kwan, 2020; Zurbaran et al., 2020). Anonymization and Randomization are two types of point-based techniques that are used widely to protect geoprivacy while preserving statistical analysis.

3.1.1. Anonymization (Spatial k-anonymity Method)

Spatial k-anonymity generates an area (neighbourhood) as an obfuscation area that contains the original point and $k - 1$ other points to protect the privacy of the original point. k is the minimum number of other locations required so that the original location is unidentifiable between them (Figure 3.12a). The risk of identification is $1/k$, therefore larger values of k decrease the risk of identification and generally increase the size of neighbourhood, as the relationship between k and size of neighbourhood depends on point density (Cassa et al., 2006; Gruteser & Grunwald, 2003; Rajendran et al., 2017; Samarati & Sweeney, 1998; Zhong & Hengartner, 2008). Zurbarán et al (2018) applied the NRand-K method, a combination of NRand (explained further in the next section) and k-anonymity on geo-referenced data collected from Twitter during the earthquake in Central Italy on August 24th of 2016. The region was represented as a 2D grid of cells. The value of the radius of obfuscation area being dependent upon local point density. If the number of points in a gridded cell was less than value of k the maximum radius was determined otherwise the minimum radius was considered, the value of k could be arbitrarily set or set to the 75th percentile of the distribution of the resulting point count within each cell.

3.1.2. Randomization Methods

Geo-masking refers to a range of techniques that aggregates, transforms or displaces the geographic location of an individual in an unpredictable way to protect individual confidentiality, including spatial aggregation, affine transformation and randomization (Stinchcomb, 2004). Randomization adds noise to a region of interest or moves the original

point to a new location in a random direction and a random distance to protect the individual's geoprivacy while maintaining the spatial patterns of points (Murad *et al* 2010, Wightman *et al* 2011, Zandbergen 2014, Seidl *et al* 2017). Randomization is a two stages procedure. Firstly, the obfuscation area is generated and secondly the random point is selected inside this area.

Several methods have been developed in the literature to generate different shapes of obfuscation area, such as buffers, donuts, pinwheels, and squares (Kwan & Schwanen, 2018; Seidl *et al.*, 2015, 2018). Below we describe the main methods that have been used in GIScience to-date.

3.1.2.1. Rand Method

Rand is the simplest randomization method that transforms the original point P to a random point inside a circular obfuscation area with the centre P and radius r (Wightman *et al.*, 2011). One limitation of this method however is that the obfuscated point might be too near to or too far from an original point which increases the risk of identification or reduces the chance of preserving spatial patterns resulting in an increase in the amount of error introduced into the data set respectively (Figure 3.1a). To overcome this, methods that incorporate a minimum and maximum distance have been developed.

3.1.2.2. N*Rand Method

N*Rand methods including NRand and NMix are variations of Rand method. The obfuscated point is either the farthest or average point within N random points (or N random points from specific set of points) inside the obfuscated area (Figure 3.1b) to ensure that the obfuscated point is not too near or too far from the original point area. This method was used to send the obfuscated location of a mobile user to location based services (LBS) instead of real location to protect user privacy while receiving acceptable service (Wightman *et al.*, 2011). The probability of the obfuscated point being in a specific distance from the original point in N*Rand is p^N when this probability for Rand method is $0 < p < 1$. This can be explained using a binomial distribution model (Palmisano, 2018) where the probability function for accruing x out of N trials (x point among N Random point being in specific distance) is shown in Equation 3.1:

$$f(x) = \binom{N}{x} p^x (1 - p)^{N-x} \quad \text{for } x = 0, 1, 2, \dots \quad \text{Eq 3.1}$$

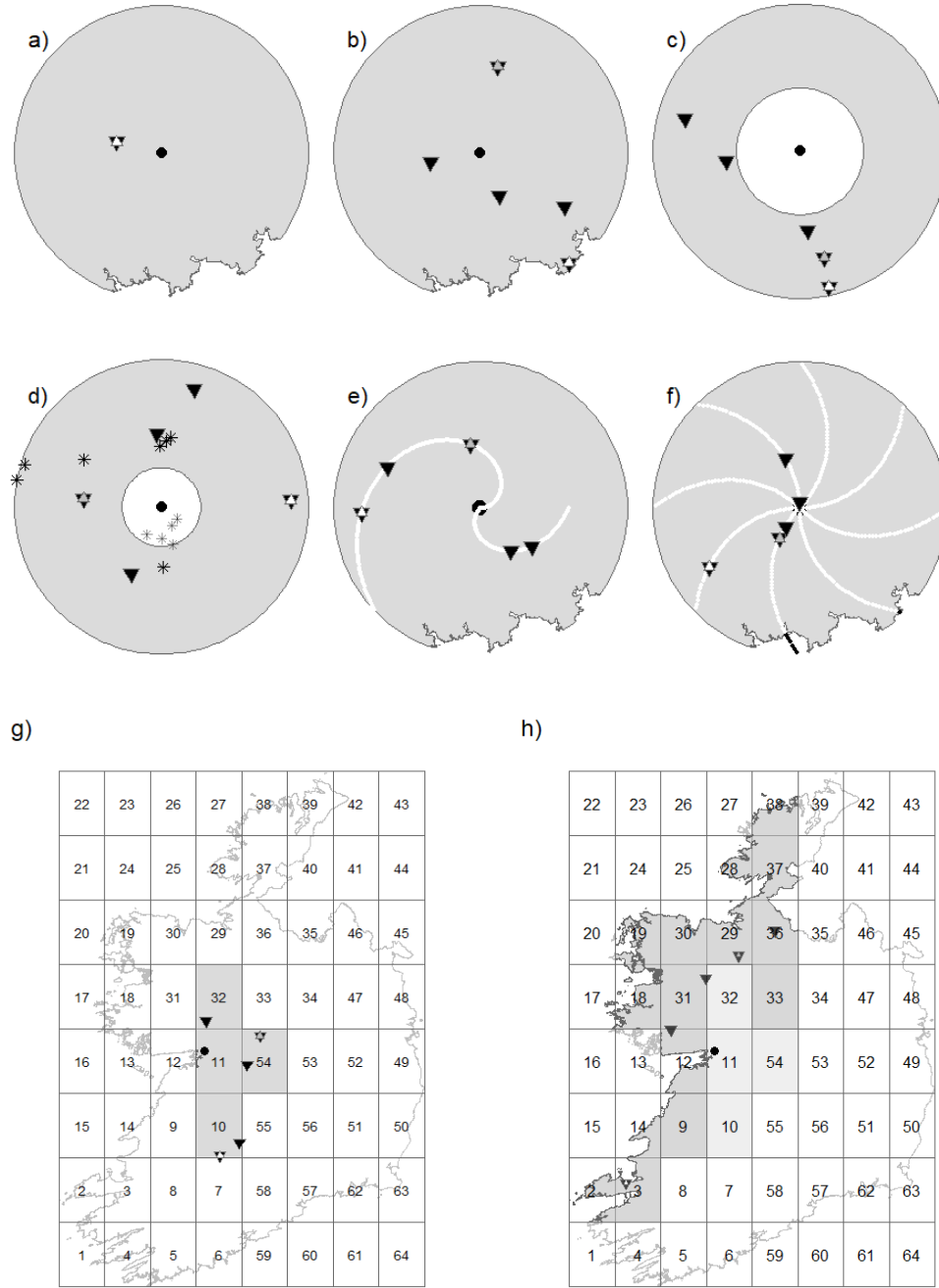


Figure 3.1: Examples of obfuscated point generated by different methods: a) Rand, b) NRand and NMix c) Donut-k d) Density-k, e) Pinwheel225, f) Pinwheel45, g) AHilb and h) Donut-AHilb. Obfuscated area (grey) original point (black circle), 5 random point (black triangle), obfuscated point in NRand (white triangle) and in NMix (grey triangle), grey stars and black stars present existing point (parcels) in an inner and outer circle with $r_{min} = 291m$ and $r_{max} = 1088m$ radiuses around the original point based on 6-kanonmity in Density-k method. The numbers in g) and h) are Hilbert curve values, the obfuscation area for original point was generated considering first the button, top, right and left cells and then Hilbert curve value order.

that means the chance of all N random points being too near the original point is p^N which is less than the chance of just one random point being too near the original point particularly for large value of N .

However, the second limitation of the Rand method may still exist whereby the obfuscation point might be too far from the original point. A modification of Rand methods called NMix method was defined as an average of N random points (average of each coordinates) inside the obfuscation area (Wightman et al., 2011). A potential disadvantage of this definition occurs particularly for large values of N (N random points might be in opposite direction) in which the obfuscated point tends to be very close to the original point or in the Donut-NMix method where this obfuscated point might be outside the donut shape. To overcome this limitation Wightman *et al* (2011) suggested to select the N random points inside a randomly selected quadrant of the obfuscation area. In this study the obfuscated point in NMix method is simply defined as a middle point from original point among N random points inside the obfuscation area.

3.1.2.3. Donut and Density Method

Donut Method is similar to Rand method, where the obfuscation area is a donut-shaped area, with centre P , constant radius $rmin$ and $rmax$ for all points (Stinchcomb, 2004). To determine the radius value in Rand method and inner and outer radiuses in Donut method, weighting by the average distance between subjects (points) which can be estimated by $\sqrt{1/point_density}$ (weighting by point density) were widely used to provide privacy protection while minimizing the introduced spatial error (Kwan et al., 2004; Lu et al., 2012; Stinchcomb, 2004). The point density can be the density of whole area or local area. Stinchcomb (2004) has suggested a constant weighting factor of 1-2 for $rmin$ and a constant weighting factor of 3-5 for $rmax$. Lu *et al* (2012) has compared, the constant weighting factor of a whole area point density with a varying weighting factor of local (block or street density) area point density on crime data in South-Central Texas that consisted of five block groups with different point densities. The weighting factors were considered between 0.6 and 1.4, with low weighting for higher-density areas and high weighting for lower-density areas. A weighting factor for $rmax$ three times that of the weighting factor of $rmin$ was also considered. However, the point-pattern analysis in this study showed no significant difference in geo-masking results as a result of applying a different weighting factor. Hampton *et al* (2010) adopted different approaches to determine $rmin$ and $rmax$ using underlying point density to

achieve predefined minimum and maximum k-anonymity. To examine the effect of size of obfuscation area on results, 10 different values of maximum k were examined while minimum k-anonymity was defined as a 10% of maximum k-anonymity. The results indicated that Donut method provided increasingly higher levels of privacy protection and lower cluster preservation as maximum k increased. Additional research was recommended to obtain the optimal parameters of the donut method to satisfy both higher level of privacy protection and cluster preservation.

3.1.2.4. Pinwheel Method

The Pinwheel method was designed for applications that constantly keep track and update the users' location to a service provider (Wightman *et al* 2013, 2014, Zurbarán *et al* 2014a, 2014b) and is a modified randomization method. In the pinwheel method the obfuscated point is a random point or either the farthest or the middle point among N random points in a pinwheel shape with centre of original point p and radius r (Figure 3.1e). Equation 3.2 generates the pinwheel shape for any α between $[0^\circ, 360^\circ]$

$$r_\alpha = \frac{(\alpha \bmod \phi)}{\phi} r_{max} \quad \text{Eq 3.2}$$

where ϕ indicates the degree between two sequential wings and r_{max} is a maximum radius of pinwheel. In the most recent adaptation of the Pinwheel method Zurbaran *et al* (2020) suggested multiplying r_α by a random number between $[0-1]$ in order to decrease the maximum radius of pinwheel and subsequently decrease the probability of obfuscated point be too far from original point.

3.1.2.5. Hilbert Method

The Hilbert curve is a fractal space-filling curve, which is used to reduce the multidimensional presentation to a one-dimensional presentation. Hilbert curve generation is a recursive algorithm, which divides the plane (surface region) into four parts. Each part is further sub-divided into four parts until the required curve resolution (resolution of original area) is obtained. The Hilbert curve order follows two rules to generate a specific unit shape (U shape) and rotation starting from the lower left corner (Breinholt & Schierz, 1998). Several obfuscation methods based on Hilbert curve were developed to generate obfuscation area by the sequential order of Hilbert curve value that satisfy k-anonymity (Ghinita *et al.*, 2007b, 2007a, 2008, 2010; Um *et al.*, 2010; Vu *et al.*, 2012). Um, Kim and Chang (2010) suggested

the AHilb method to generate a rectangular obfuscation area which included cells that follows the Hilbert curve order value and unconnected adjacent cells, to avoid the unnecessary extension of an obfuscation area. AHilb method was developed for mobile users using LBS. It generates a rectangular shaped obfuscation area because it incurs a small network overhead (when transmitted to the LBS) that might include cells without any locations/points.

3.2. Data and Methods

3.2.1. Data Preparation

The *Irish nutrient management planning online (NMP Online)* platform is designed to support farmers make decisions on nutrient management based upon soil and agronomic targets. It can deliver a fertiliser plan that will help target nutrients based upon soil fertility derived from soil test results. Nutrient management planning, whereby fertiliser is applied at the right rate, right time and in the right place will improve the production capacity on the farm, supporting profitability while enhancing environmental sustainability on the farm (Wall, D., & Plunkett, 2020). The *NMP Online* dataset could be used for research including in 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 that are subject to stringent data protection regulations (Murphy *et al* 2013, 2015) This study therefore set out to examine different obfuscation methods on the *NMP Online* dataset, evaluate and compare the results in terms of level of privacy protection and spatial patterns preservation.

The *NMP Online* dataset contains geographical, soil related and bovine animal herd related attributes of over 55,000 farms and field parcels. The centroid of each field parcel (polygon) is used as a point data to represent the field parcel and a unique key (plotUID) is generated based on its centroid to retrieve the information of each field parcel after obfuscation and avoid having different obfuscated points for a same point (Figure 3.2). After pre-processing of *NMP Online* data set a subset of 22425 field parcels information for 2016 with non-uniform distribution in Ireland were extracted and used in this study.

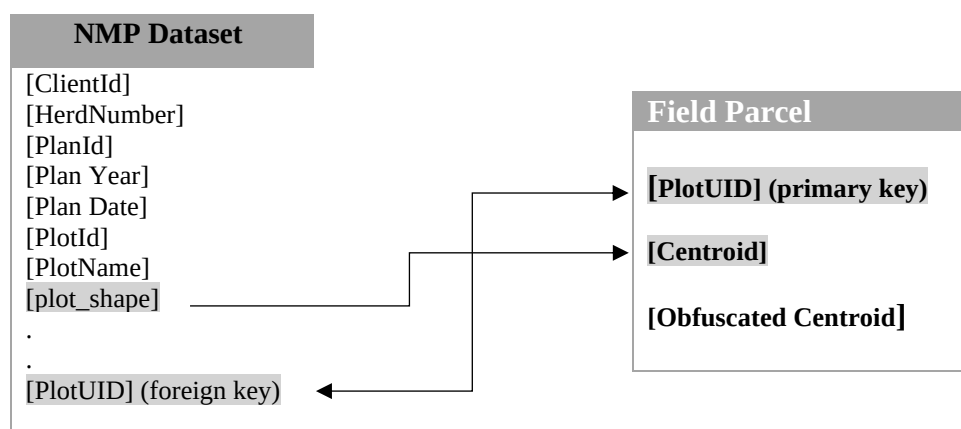


Figure 3.2: Diagram to represent the relationship between field parcels and its unique centroid (point), ID and obfuscated point to retrieve field parcels after obfuscation.

Table 3.1: 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 3.1a and 3.1b.	Point based
Donut	Dt	Obfuscation area is a donut-shaped area, with centre P, constant radius r_{min} and r_{max} for all points (maximum r_{min} and maximum r_{max} of all points to ensure that all obfuscated points satisfy k -anonymity) (Hampton et al., 2010; Zandbergen, 2014) Figure 3.1c.	Point based
Density	D	Using Donut method when radius r_{min} and r_{max} are vary depending on the point's density(Hampton et al., 2010) Figure 3.1d.	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 et al., 2010).	Area based
Donut-AHilb	Dt_AH	Obfuscation area is a difference of area that satisfy $2k$ -anonymity 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 & Grunwald, 2003).	Both

3.2.2. Obfuscation Methods

Various randomization methods with the k -anonymity concept were developed to reduce the risk of disclosure of agricultural data via geospatial attributes. Table 3.1 outlines the seven overarching methods including N*Rand (Rand, NRand and NMix), Donut, Density, Pinwheel, AHilb and Donut_AHilb methods that were implemented in this study to generate the obfuscation area (SI3.1-SI3.7, Appendix 1). Methods such as N*Rand were examined to investigate how selecting a random point inside the obfuscation area affects the methods performance. Different methods that generate different shapes of obfuscation area, such as buffer, donut, pinwheel, and square shape were examined to investigate the impact they might have on agriculture data. To investigate the impact of point density a modified version of NRank-K called Density- k and Donut- k were introduced in order to determine the size of donut shape (in Density and Donut method) and square size (in AHilb and Donut-AHilb method).

3.2.2.1. Donut-k and Density-k Methods

Donut-k and Density-k methods are qualitative techniques to determine the value of inner and outer radiuses in Donut and Density methods. As outlined above, quantitative approaches have been adopted to obtain $rmin$ and $rmax$ and in some cases maximum k-anonymity in previous research (Hampton et al., 2010; Kwan et al., 2004; Lu et al., 2012; Stinchcomb, 2004; Zurbarán et al., 2018). A challenge of this study is to develop and implement a qualitative approach 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 the smallest obfuscation area, that contains at least k-1 other locations. First, an iterative procedure was developed as follows: $rmin$ was set to minimum threshold with increments added at an increasing rate in each step until k-anonymity satisfaction. The procedure was repeated for $rmax$ which was initialized by obtained $rmin$. The choice of minimum threshold of $rmin$ and increasing rate are crucial. While a large value of both minimum threshold and increasing rate can reduce the computation time, it could result in missing the optimal $rmin$ that satisfies the k-anonymity. 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, a modified method based on varied increasing rate was developed and applied. The procedure can be described in algorithm 3.1.

Algorithm 3.1: Finding inner radius of donut shape (obfuscation area) for each point for a given k1 in Density-k and Donut-k methods

```
1. find_radius (pntt,k1)           // for each point (pnt) in dataset , k1: level of k-anonymity
1. output(radius)                 // optimal radius that satisfy the k-anonymity
2. rmin <- rmin_threshold         // initialize the rmin with zero or in polygon data with the distance of centroid of
3.                               // parcel with its nearest edge
4. min_rate <- 0.5                 // initialize the minimum threshold of increasing rate
5. rate <- 1000                   // initialize rate with a large value
6. while (rate > min_rate) do
7.   while (num_of_point < k1) do   // number of points in a buffer with radius rmin
8.     rmin <- rmin + rate
9.   end while
10.  rate <- rate/2                 // reduce the increasing rate to half
11. end while
12. return(rmin)
```

For different points radius r_{min} and r_{max} can vary depending on the point density (Density method) or can be constant for all points (maximum r_{min} and maximum r_{max} of all points to ensure that all obfuscated points satisfy k-anonymity in Donut method) (Figure 3.1c and 3.1d).

This procedure can be implemented for different values of k_1 first to identify the maximum k-anonymity with minimum r_{min} and then based on this r_{min} . This procedure is repeated for different values of k_2 to identify the maximum k-anonymity with minimum r_{max} , and consequently, identify the most effective Donut and Density methods for privacy and spatial pattern preservation. Larger values of k_1 and k_2 provide a higher level of geoprivacy protection but increase the radiuses and thereby degrade data quality. The optimal value of k_1 and k_2 can be considered when the growth of radiuses is significantly large.

The initial value for r_{min} can be set to zero for point base data while two scenarios can be adopted if the polygon nature of field parcel (plots) is considered, this initial value can be set to a distance of field parcel's centroid (p) to its farthest edge in scenario 1 or its nearest edge in scenario 2.

3.2.2.2. Modified AHilb and Donut-AHilb Methods

In this research to increase geoprivacy protection a modified AHilb method was applied by excluding the cells that contain no point so obfuscation area might be smaller and have arbitrary shape (Figure 3.1g). The modified method called the Donut-AHilb method is suggested in which the obfuscation area is a difference of area that satisfy maximum k-anonymity and area that satisfy minimum k-anonymity generated by AHilb method (Figure 3.1h).

3.3. Evaluation Method

Table 3.2 outlines a brief description of the evaluation methods implemented in this study. The distribution preservation can be evaluated using 2D KS-test by `peacock.test` package (Yuanhui, 2016) in R 4.0.2 (R Development Core Team, 2020). A limitation of this method is that the long processing time, which is very slow. To overcome this limitation the sum of 1D KS-test statistic of both spatial attributes and difference of correlation of spatial attributes of obfuscated and original points can be used as a quantitative metric to measure distribution preservation. The 1D KS-test statistic (D) is the maximum difference between the cumulative distribution function of two datasets to be compared. While a definition of the cumulative distribution function and therefore the multi-dimensional KS test of multi-dimensional space can be challenging, Peacock (1983) proposes the following. First the division of 2D space into four quadrants around any point and then the definition of a cumulative distribution function for each quadrant as well as generalizing the KS test for two-dimensional space. The main idea is to find the maximum differences between cumulative distribution function of two datasets in each quadrant around any point (x_i, y_i) and then find the maximum of these values.

K-means clustering is an unsupervised machine learning method which partitions the unlabelled data to a fixed predefined number of clusters (k) based on data similarity (Nainggolan et al., 2019). Three different distance metrics, Manhattan distance metrics, Minkowski distance metric and Euclidian distance metric can be used to measure the similarity within data. The k-means based on Euclidean distance metric has been shown to perform best (Singh et al., 2013), and as such we used it to measure similarity. The best number of clusters (k) occurs when the sum of the squared distance between the data point within a cluster and the cluster's centroid is at a minimum. In order to obtain the optimal value of k, the elbow method was used which applies the K-means clustering for different values of K and selects the one that the sum of square error (SSE) decline is significantly small. The resulting SSE plot for different K values makes it easy to choose the best K value based on these quantitative values.

Statistical summary of point displacement including min, max, mean and standard deviation which quantify the centre and spread (variability) of data were used when data is normally distributed (Wightman et al., 2011), for non-normal distribution min, max, median, 1st Q and 3rd Q (interquartile range, IQR) better explain the centre and spread of data. Larger values of min, max and median displacement provide better performance in terms of

geoprivacy protection but maintain less statistical accuracy. Similarly, larger IQR provides higher variability for selecting obfuscated points thus provides higher level of privacy protection. Accounting for these two effects, we can say that the best method performance to satisfy both privacy protection and accuracy could be seen when the min, max and median displacement are smaller with higher IQR.

Table 3.2: Description of the evaluation technique that have been used to evaluate and compare obfuscation method in terms of accuracy and privacy protection.

Evaluation technique	Description	Evaluation type
Location distribution preservation	Rank the methods from the best to worst based on 2D KS-test (statistical test to compare the original points and obfuscation points distribution) or 1D KS test 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	Number/percentage of misclassified obfuscation points called Misclassification Error (MCE) (Parameswaran and D. Blough, 2005)	Accuracy
Displacement statistical summary	Statistical summary of displacement including min, max, median and IQR	Accuracy
k-anonymity satisfaction	The number/percentage of obfuscated points which satisfied, k-anonymity for different values of k. (Kalnis <i>et al</i> 2007, Khoshgozaran <i>et al</i> 2011, Wang and Kwan 2020).	privacy protection
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 <i>et al.</i> , 2018) or are inside the area of other original point's region (when the polygon nature of field parcel is considered)	privacy protection
Non-unique obfuscation	Number of points with non-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)	privacy protection and Accuracy
Spatial information divergence (Gdi and Ldi)	The global divergence index (Gdi) is an average of three divergence indexes: the divergence of spatial mean, the orientation and major axis of the first standard deviational ellipse. The local divergence index (Ldi) represents the divergence of hotspot area (Kounadi & Leitner, 2015).	Accuracy

Spatial information divergence represents the dissimilarity of the obfuscated points' pattern with the original points' pattern using both global and local statistics through the global divergence index (Gdi) and local divergence index (Ldi) (Kounadi & Leitner, 2015). Gdi describes the divergence of obfuscated points' pattern to original points' pattern based on three characteristics; points pattern components including the spatial mean as a metric to measure

the divergence of centroid, orientation to measure the directional divergence, and major axis of the first standard deviational ellipse to measure of spatial dispersion. Ldi calculates the divergence of the obfuscated points' hotspots area from the original points' hotspots area using the Getis-Ord G_i^* statistic (Getis & Ord, 1992; Ord & Getis, 1995).

3.4. Results

3.4.1. Parameters determination

Table 3.3 shows the radius r_{min} and r_{max} for different values of k_1 and k_2 in Density-k method using scenario 2 where the initial value of r_{min} is set to a distance of a field parcel's centroid to its nearest edge. Table 3.3 and Figure 3.3a indicate r_{min} is increased slightly when k_1 is between four to six. A significant growth is observed when k_1 is increased from six to seven, which means the largest k_1 with minimum r_{min} is six, so six can be considered as an optimal k_1 with acceptable r_{min} . The next significant growth is observed when k_1 is increased from nine to ten, so nine also can be considered as another candidate for optimal k_1 . The choice between different candidates of optimal k_1 , in this case six or nine, is dependent on the priority of the study, either privacy protection or accuracy. After identifying the optimal value of k_1 , this procedure was repeated for different values of k_2 to identify r_{max} and optimal k_2 . Table 3.3 and Figure 3.3b show candidates of optimal k_2 for both candidate of optimal k_1 , six and eleven are two candidates for k_2 when optimal k_1 is six. Here, optimal k_1 and k_2 are both considered as six and based on these values the inner and outer radiuses in Density-k are between 50m to 20269m and 138m to 36804m respectively, therefore for Donut-k these values are maximum inner and outer radiuses of all points which are 20269m and 36804m to ensure all obfuscated points satisfy 6-anonymity.

Table 3.3: r_{min} calculated for different values of k_1 and r_{max} calculated for different values of k_2 when k_1 considered as six.

K1	r_{min}	K2	r_{max} (K1= 6)	r_{max} (k1 =9)
4	17970	4	36432	38306
5	19746	5	36801	38834
6	20269	6	36804	39202
7	23306	7	38306	39292
8	23587	8	38834	39357
9	23739	9	39202	41476
10	36258	10	39292	41618
11	36801	11	39357	41718
12	36805	12	41476	41843
13	37184	13	41618	41937

To determine the acceptable value of N in NRand and NMix based methods Wightman et al. (2011) examined the effect of different values of N and recommended a value of 4. The results showed improvement in the evaluated metrics (higher average displacement and acceptable standard deviation) by increasing N from one to four and remained steady for greater value of N , so the performance for value of 5 is similar to 4 with a bit more of computational

cost due to the extra trial. Here, a value of 5 was considered where just one middle point exists among 5 random points.

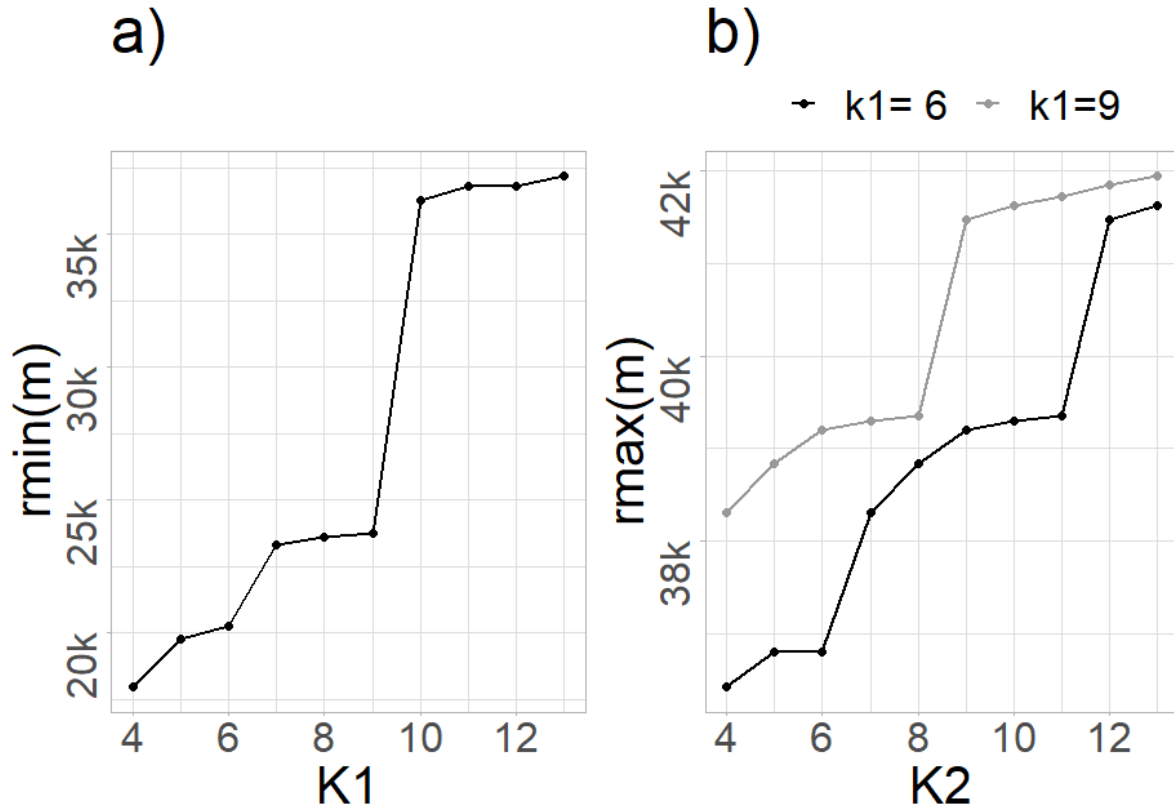


Figure 3.3: a) $rmin$ calculated for different values of k_1 , b) $rmax$ calculated for different values of k_2 when k_1 considered as six

3.4.2. Distribution preservation Analysis

Table 3.4 and Figure 3.4 confirm that the rank of obfuscation method based on 2D KS-test using peacock2 and Sum of 1D KS-test of both coordinates and absolute correlation alteration of spatial attributes (DPR) are the same. Density-based methods (Density, Density-Pinwheel and AHilb, Donut-AHilb) were the best performing with distribution preservation of one, while Pinwheel225-NRand (P225-NR) and NRand had the first and second highest distribution preservation rank of seven and six respectively (Figure 3.4). Figure 3.5 illustrate the impact of different methods (NMix and Density method) on obfuscated points distribution in an area of circle with 10000 m radius.

zero performing the best. N Rand and combined N Rand with Pinwheel (Pinw45-N Rand) method had the largest correlation alteration of 0.14 and 0.15 respectively, meaning density-based methods performed the best and the N Rand based method the worst in preserving correlation between attributes and spatial locations (Figure 3.6).

Table 3.4: Comparison of 2D KS-test and combination of 1D KS-test and correlation alteration of different obfuscation method

Method	Peacock2	KS-test (X)	KS-test (Y)	Correlation of obfuscated (X and Y)	Correlation alteration of X and Y	Sum of Ks-test X and Y and correlation alteration
R	0.1093423	0.042809	0.070546	0.2750719	0.005072	0.1184269
NR	0.1661093	0.06961	0.10511	0.287038	0.017038	0.191758
NM	0.1155407	0.042542	0.075184	0.2746979	0.004698	0.1224239
Dt_R	0.09841695	0.03835	0.063144	0.2661455	0.003855	0.1053485
Dt_NR	0.1179933	0.042809	0.075541	0.2708374	0.000837	0.1191874
Dt_NM	0.09948718	0.038261	0.062965	0.2710821	0.001082	0.1023081
D_R	0.003076923	0.0015608	0.0018729	0.2668849	0.003115	0.0065488
D_NR	0.003790412	0.002052	0.0020067	0.2668264	0.003174	0.0072323
D_NM	0.003522854	0.0021405	0.0020959	0.2664385	0.003562	0.0077979
P45_R	0.07567447	0.025552	0.049008	0.2718303	0.00183	0.0763903
P45_NR	0.1453735	0.056232	0.092352	0.2773338	0.007334	0.1559178
P45_NM	0.07487179	0.026176	0.048829	0.2682476	0.001752	0.0767574
P225_R	0.08187291	0.053512	0.040535	0.2678441	0.002156	0.0962029
P225_NR	0.194961	0.18979	0.047001	0.2925867	0.022587	0.2593777
P225_NM	0.07727982	0.014939	0.064794	0.2654318	0.004568	0.0843012
D_P45_R	0.003076923	0.0016054	0.0019621	0.2665158	0.003484	0.0070517
D_P45_NR	0.003790412	0.0020513	0.0020513	0.266033	0.003967	0.0080696
D_P45_NM	0.003344482	0.0021405	0.0017837	0.266244	0.003756	0.0076802
D_P225_R	0.007982163	0.0073133	0.0019621	0.2638011	0.006199	0.0154743
D_P225_NR	0.01092531	0.0096767	0.0049052	0.2623269	0.007673	0.022255
D_P225_NM	0.007892977	0.0069119	0.0016499	0.2642642	0.005736	0.0142976
AH_R	0.003701226	0.0016945	0.0021851	0.266585	0.003415	0.0072946
AH_NR	0.004503902	0.0020513	0.0029431	0.2669295	0.003071	0.0080649
AH_NM	0.003121516	0.0014716	0.0019175	0.2665913	0.003409	0.0067978
Dt_AH_R	0.005261984	0.0029877	0.0024526	0.2666228	0.003377	0.0088175
Dt_AH_NR	0.006688963	0.0029431	0.0042809	0.2663868	0.003613	0.0108372
Dt_AH_NM	0.004949833	0.0027648	0.0027648	0.2659197	0.00408	0.0096099

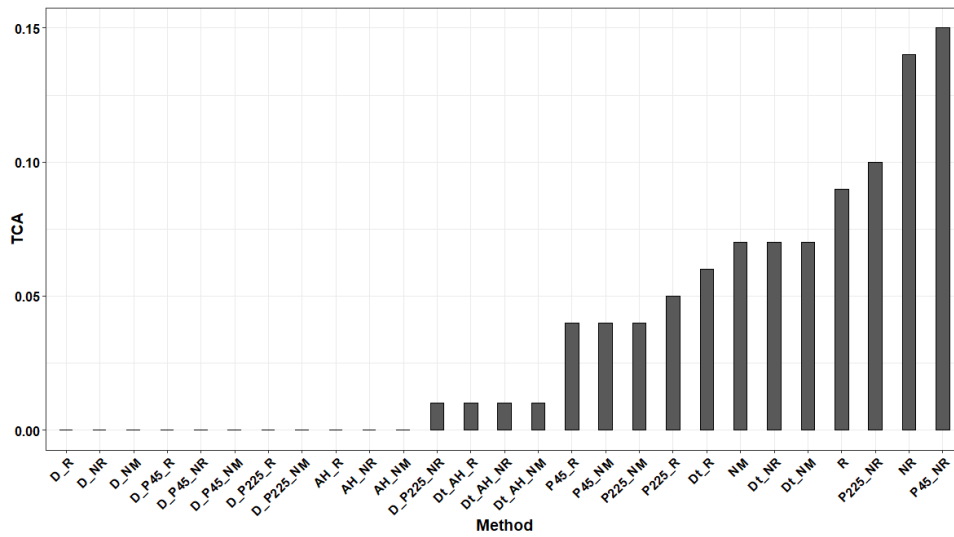


Figure 3.6: Total correlations alteration of spatial attributes with soil and herd related attributes for different methods.

3.4.4. Cluster Preservation Analysis

In cluster preservation analysis the optimal number of clusters was considered as five clusters which were obtained by the Elbow method. The SSE plot for different K values from 1 to 10 (Figure 3.7) shows as the value of k increases, the SSE decreases but adding more clusters after $k = 5$ does not provide a significant decrease in SSE. AHilb-Rand with 0.26 % MCE has the best clustering preservation and N Rand with 18.82% MCE has the worst (Figure 3.9c). Pinwheel225 shows better performance in cluster preservation than Pinwheel45, due to providing smaller obfuscated area and less variability to select obfuscated points.

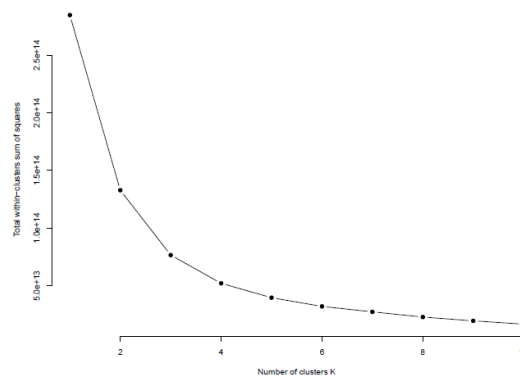


Figure 3.7: Elbow method to find the optimal value of K .

Table 3.5: Correlations and correlation alteration of spatial attributes with soil and herd related attributes using different methods.

Method	Spatial Attribute	Soil Related Attributes				Herd Related Attributes				TCA
		pH	P	K	Lime Required	Cow	Heifer	Other Animal	Concentrate Feed	
Original	X	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.09
	Y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
R	Obf x	0.19	-0.08	0.09	-0.04	-0.16	-0.05	0.14	-0.11	0.14
	Obf y	-0.04	-0.12	-0.09	0.15	-0.32	-0.22	0.11	-0.17	
NR	Obf x	0.18	-0.08	0.09	-0.04	-0.15	-0.05	0.13	-0.11	0.07
	Obf y	-0.04	-0.12	-0.09	0.15	-0.31	-0.22	0.12	-0.17	
NM	Obf x	0.20	-0.08	0.09	-0.05	-0.15	-0.05	0.14	-0.11	0.06
	Obf y	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
Dt_R	Obf x	0.20	-0.08	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0.07
	Obf y	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
Dt_NR	Obf x	0.19	-0.07	0.09	-0.05	-0.15	-0.05	0.14	-0.11	0.07
	Obf y	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
Dt_NM	Obf x	0.20	-0.08	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0
	Obfy	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.12	-0.17	
D_R	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obfy	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_NR	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obfy	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_NM	Obfx	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.04
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
P45_R	Obfx	0.20	-0.07	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0.15
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
P45_NR	Obf x	0.19	-0.08	0.08	-0.04	-0.15	-0.04	0.13	-0.11	0.04
	Obf y	-0.04	-0.11	-0.09	0.15	-0.31	-0.22	0.11	-0.17	
P45_NM	Obf x	0.21	-0.07	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0.05
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	
P225_R	Obf x	0.21	-0.07	0.09	-0.04	-0.16	-0.05	0.14	-0.11	0.1
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	
P225_NR	Obf x	0.20	-0.08	0.09	-0.04	-0.17	-0.06	0.14	-0.12	0.04
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.23	0.11	-0.17	
P225_NM	Obf x	0.21	-0.07	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.23	0.11	-0.17	
D_P45_R	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_P45_NR	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_P45_NM	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_P225_R	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.01
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
D_P225_NR	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	
D_P225_NM	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
AH_R	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
AH_NR	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
AH_NM	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.01
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
Dt_AH_R	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.01
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	
Dt-AH_NR	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.01
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	
Dt-AH_NM	Obf x	0.22	-0.07	0.10	-0.05	-0.16	-0.05	0.15	-0.11	0.01
	Obf y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.16	

Table 3.6: List of 27 obfuscation methods that have been applied with the following parameters: $k = 6$, $N = 5$, ($\phi = 45^\circ / \phi = 225^\circ$), $r_{min} = 20203$ m, $r_{max} = 36451$ 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

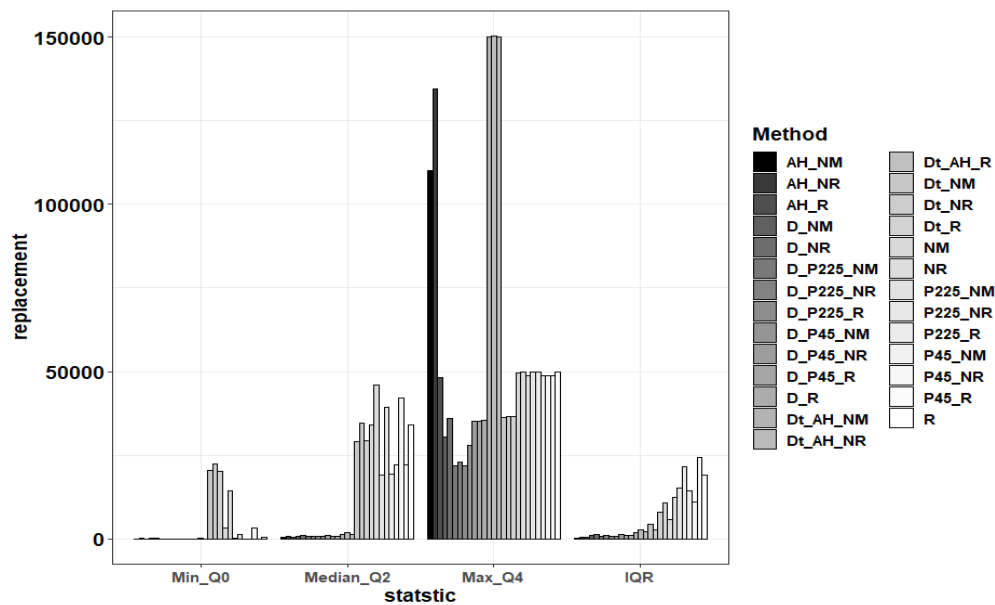


Figure 3.8: Comparison of statistical summary point displacement for different obfuscation method

3.4.5. Displacement and K-anonymity Satisfaction Analysis

According to the displacement statistical summary (Figure 3.8 and Table 3.7), density-based methods provide smaller min, median displacement. Therefore, while they provide better performance in terms of accuracy, with low IQR which indicates low range of the middle 50% of displacement, they provide less variability to select the obfuscated points which increase the risk of identification for 50% of points. In AHilb methods particularly

AHilb-NMix and AHilb-NRand with high max displacement and low IQR, while they provide less variability for 50% of points, they provide high variability for another 50% of points. Donut-AHilb method also provides high max and moderate min, median and IQR that means they provide moderate variability for 50% of points and high variability for the rest. Smaller min, max and median displacement and higher IQR results are indicative of better method performance for both geoprivacy protection and statistical accuracy.

Table 3.7: Statistical summary of point displacement for different obfuscation method

Method	Min (Q_0)	Q_1	Median (Q_2)	Q_3	Max(Q_4)	IQR($Q_3 - Q_1$)
R	471	23357	34035	42464	49996	19107
NR	14522	42567	46089	48365	49996	5799
NM	3305	28287	33956	38961	49516	10674
Dt_R	20197	25046	29262	33007	36448	7961
Dt_NR	22530	33082	34678	35686	36449	2604
Dt_NM	20489	26903	29221	31422	36398	4519
D_R	91	479	795	1517	35379	1038
D_NR	127	582	985	1865	35935	1283
D_NM	108	494	825	1553	30444	1059
P45_R	0	10000	22222	34444	48889	24444
P45_NR	3333	35556	42222	46667	48889	11111
P45_NM	0	14444	22222	28889	48889	14444
P225_R	0	8889	19333	30444	49778	21556
P225_NR	1333	30000	39333	45111	49778	15111
P225_NM	222	13111	19111	25556	48667	12444
D_P45_R	0	453	734	1391	35238	939
D_P45_NR	0	568	952	1792	35181	1225
D_P45_NM	0	468	756	1389	27985	921
D_P225_R	0	435	704	1318	21815	883
D_P225_NR	0	558	928	1732	23026	1174
D_P225_NM	0	447	708	1302	21885	855
AH_R	1	328	525	792	48197	464
AH_NR	202	642	852	1272	134461	630
AH_NM	55	393	531	733	109966	341
Dt_AH_R	15	749	1307	2810	150015	2060
Dt_AH_NR	104	1078	1781	3714	150304	2637
Dt_AH_NM	75	771	1308	2799	150068	2028

The results, as shown in (Figure 3.9d) indicate that Rand and Donut methods fully satisfy 6-anonymity and more than 99% satisfy 25-anonymity, which is the highest geoprivacy protection. Density, Density-Pinwheel45 and Density-Pinwheel225 almost fully (greater than 99.98%) satisfy 6-anonymity, 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-anonymity. AHilb method provides moderate geoprivacy protection between 60% of 6-anonymity to 5.88% of 25-anonymity. Donut-AHilb provides the best geoprivacy protection with 93.5% 6-anonymity to 30% 25-anonymity and almost 80% 10-anonymity. 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 grid-based obfuscation areas. Therefore, it is expected to provide higher percentage of 6-anonymity. Figure 3.9e 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.

3.4.6. False identification and non-unique obfuscation analysis

If field parcel polygon nature is not taken to account, non-unique obfuscation is zero for all obfuscation methods. The percentage of false identification for most of obfuscation methods is zero, with Density-Pinwheel and Donut-AHilb methods reporting very small false identification between 0.02% to 0.17%. 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 correlation preservation, no non-unique obfuscation and very low false identification can be considered as an appropriate method for data with high point density without considering the polygon nature of field parcels. On the other hand, if the polygon nature of field parcels is considered then a high percentage of false identification and non-unique obfuscation (20% and 50%) indicates the drawback of applying this method for data with high point density and polygon nature considered (Figure 3.9a, 3.9b).

3.4.7. Spatial information divergence analysis

In this study to calculate the Ldi, the point data was aggregated to a fishnet polygon with a cell size of 2000 m and hotspot analysis was applied with distance band of 23500 m. Results from global and local divergence indexes are shown in Table 3.8 and Figure 3.10 indicate the same trend across obfuscation methods. Low Gdi (between zero and 3.5) indicates that all methods preserve the point patterns globally while the variation of Ldi (between 0.8 and 63.5) shows different local point pattern preservation among methods. The density-based methods with less spatial divergence (Gdi less than 0.2, Ldi between 0.8 and 8.4) represent lowest dissimilarity of obfuscated points pattern to the original one followed by Donut method, N*Rand, Pinwheel45 and Pinwheel225. N*Rand and combined N*Rand with Pinwheel (Pinw45-N*Rand and Pinw225-N*Rand) methods have the largest spatial convergence (Gdi between 2.1 and 3.5, Ldi between 52 and 63.5) and represent the highest dissimilarity.

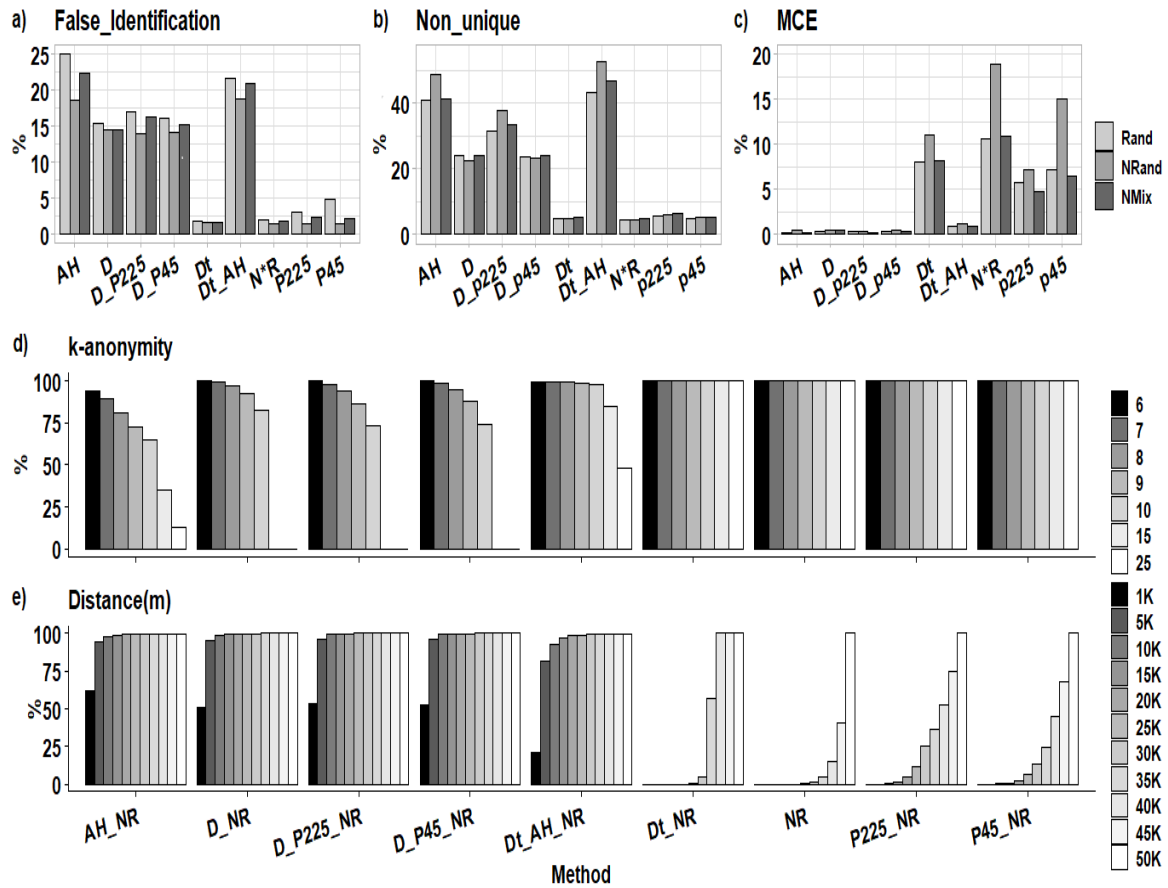


Figure 3.9: Comparison of 27 obfuscation methods based on a) False-identification b) non-unique obfuscation, c) MCE, d) percentage of obfuscated points that satisfies k-anonymity for different values of k and d) percentage of obfuscated points with displacement less than different values of d.

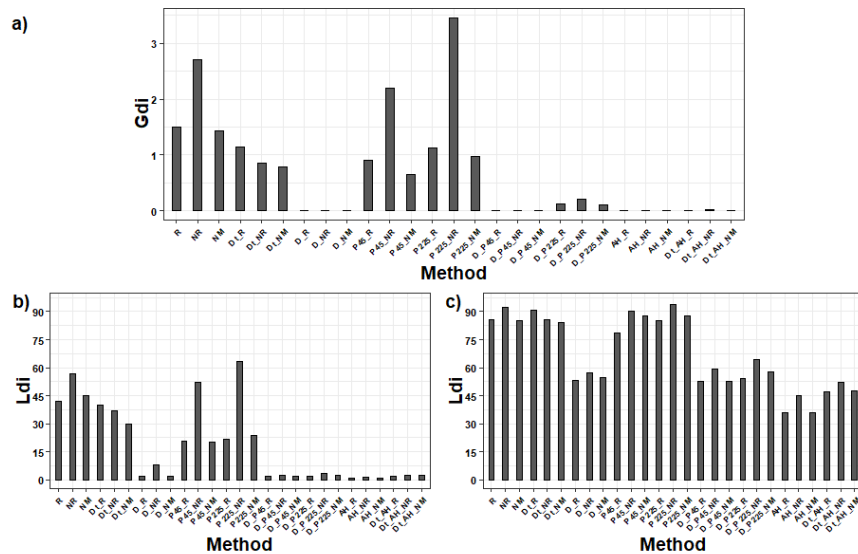


Figure 3.10: The results of Spatial information divergence for different obfuscation method, a) Gdi, Ldi based on two spatial scales: b) Ldi2000 (fishnet cell size of 2000m and distance band of 23500m) and c) Ldi500(fishnet cell size of 500m and distance band of 900m).

Table 3.8: The results of Spatial information divergence, Gdi and Ldi for different obfuscation method based on two spatial scales: Ldi2000 (fishnet cell size of 2000m and distance band of 23500m) and Ldi500(fishnet cell size of 500m and distance band of 900m).

Method	Mdi	Odi	Madi	Gdi	Ldi2000	Ldi500
R	1.71	0.83	1.96	1.5	41.93	85.63
NR	2.88	1.73	3.48	2.7	56.94	92.38
NM	1.6	0.9	1.82	1.44	45.2	85.22
Dt_R	1.38	0.41	1.62	1.14	39.95	90.98
Dt_NR	1.37	0.06	1.13	0.85	36.86	85.52
Dt_NM	1.19	0.19	0.95	0.78	29.79	84.36
D_R	0.01	0.02	0.01	0.01	1.9	53.18
D_NR	0	0.01	0.02	0.01	8.05	57.34
D_NM	0	0	0	0	1.89	54.91
P45_R	0.91	0.54	1.24	0.9	20.73	78.32
P45_NR	2.21	1.21	3.16	2.19	52.04	90.33
P45_NM	0.83	0.39	0.73	0.65	20.13	87.71
P225_R	2.45	0.46	0.44	1.12	21.81	84.97
P225_NR	9.51	0.69	0.19	3.46	63.49	93.95
P225_NM	1.77	0.2	0.93	0.97	23.89	87.63
D_P45_R	0.01	0	0.02	0.01	1.77	52.45
D_P45_NR	0	0.01	0.03	0.01	2.21	59.05
D_P45_NM	0.01	0	0	0	1.78	52.68
D_P225_R	0.26	0.04	0.06	0.12	2.03	54.19
D_P225_NR	0.41	0.09	0.09	0.2	3.25	64.22
D_P225_NM	0.24	0.03	0.03	0.1	2.47	57.94
AH_R	0.01	0.01	0.02	0.01	0.95	36.09
AH_NR	0.02	0.01	0.01	0.01	1.38	45.05
AH_NM	0.01	0	0	0	0.82	36
Dt_AH_R	0.02	0	0	0.01	2.09	46.96
Dt_AH_NR	0.04	0	0.01	0.02	2.54	51.91
Dt_AH_NM	0.02	0.01	0	0.01	2.18	47.59

3.5. Discussion

NMP Online has been developed to support farmers to achieve a balance between a high level of agricultural production and environmental protection. While the *NMP Online* map-based outputs enable advisers and farmers to understand and follow nutrient management plans efficiently (Murphy et al., 2015), geoprivacy concerns limit the potential for sharing this source of data among researchers, even towards the development of improved NMP applications. Moreover, in the GIScience literature, while several obfuscation methods have been developed, we do not have an extensive baseline comparison with which to effectively identify a suitable obfuscation method for a static polygon dataset with a high density. In an effort to overcome this limitation, here, the performance of 27 point-based obfuscation methods were evaluated on the subset of *NMP Online* dataset with high point density and non-uniform distribution. Furthermore, several evaluation methods were tested to measure the ability of each method to satisfy both privacy concerns and spatial pattern preservation.

We should point out that the obfuscation process in this case study is an offline process which should be done by data administrator or data provider prior to sharing data with users. The choice of obfuscation methods depends on the user specific role, need and priority of accuracy or privacy protection. The challenge for the data administrator is to investigate the need, interest and level of data vulnerability for different users to provide them the appropriate obfuscated data. Accuracy is the priority when sharing data to users from organizations, while the data privacy protection is the main issue when sharing data publicly. However, whilst a comparison of these obfuscation methods, as undertaken within this study, provides a baseline, specifying the parameters can be challenging. For example, determining the angles to use in pinwheel method, particularly when their impact on accuracy and privacy protection is explored (Figures 3.4, 3.8-9). Although methods such as Density-k and Donut-k methods calculate the best possible size of obfuscation methods for a certain level of k-anonymity and introduce a process to obtain the best candidate values for k, which subsequently reduces the amount of pre-processing decisions undertaken by the data administrator.

The ease of reverse geocoding as outlined in the introduction whereby geographical information in agricultural data could be readily linked with available online data, to easily enable adversaries to obtain the private and sensitive information related to a farm/farmer becomes very tedious when obfuscated data is shared and, in some cases, is impossible. After reverse geocoding adversaries could potentially only estimate an area that might contains the

real location of that farm. Figure 3.11 shows the estimated region of a radius (500m for obfuscated point 24) generated by Density-Pinwheel45 and Density methods which contains 8 and 22 real points respectively. A minimum of 8 and 22 times of linking and matching data is required to identify the real location of this point, although without having external or background knowledge such as the size of farm or crop type there is only 1/8 and 1/22 chance to correctly identify this point. If at this stage they have not been successful, then gradually expanding the estimated region is required which contains more real points, and in some cases this region might contain more than 100 actual farms. Besides this procedure being difficult and time-consuming, the possibility of correct identification is very low. However, questions pertaining to exactly how human attackers might attempt to re-identify the information persist and warrants future GIScience research (Martin & Schuurman, 2020; Swanlund & Schuurman, 2016).

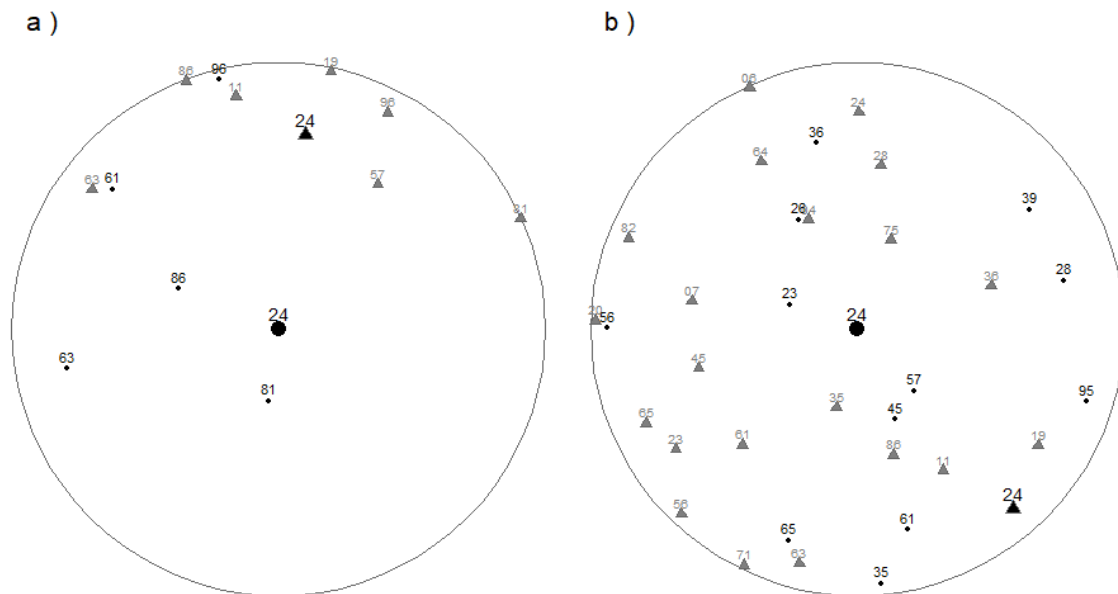


Figure 3.11: Estimated region of radius 500m after reverse geocoding for point 24 contains real points (triangles) and obfuscated points (small dots) using a) Density-Pinwheel45 method contains 8 real points and 5 obfuscated points. b) Density method contains 22 real points and 12 obfuscated points.

According to our results, the density-based methods reported the highest distribution preservation, cluster preservation, lowest Gdi, Ldi and correlation alteration while fully satisfying optimal k-anonymity, demonstrating that they appeared to be the best performance in terms of spatial pattern preservation and geoprivacy protection, followed by Pinwheel, Donut and finally N*Rand (Figure 3.4), 3.6, 3.9c, 3.9d, 3.10). These results are in agreement with the findings of Zurbaran *et al* (2020) who showed that the density-based, NRand-k

method outperformed a handful of other methods in terms of spatial pattern preservation followed by Pinwheel and Donut method applied on crime data. Similarly, Zurbarán *et al* (2018) showed that N Rand-k produced less spatial pattern alteration than N Rand methods with the same maximum radius and level of geoprivacy protection applied on spatial data collected from Twitter. Although, no significant differences are observed among density-based methods within all evaluation metrics, Donut-AHilb provides the best geoprivacy protection and AHilb method the worst (Figure 3.9d). Density-k method outperforms other methods and Donut-AHilb demonstrates the worst performance in terms of correlation alteration and cluster preservation, while Density-Pinwheel225 shows the worst distribution preservation, global, and local divergence due to generating small and asymmetrical obfuscated area (Figure 3.4, 3.6, 3.9c, 3.10).

The importance of false identification and non-unique obfuscation must be considered when exploring the polygon nature of field parcels. For example, when polygon data are considered as points, there remains the chance that the obfuscated point could still fall within an original parcel or area of obfuscated point (Figure 3.12b), negating the obfuscation process and exposing the data to a security breach. When we considered this in the accuracy metrics (Figure 3.9b) we found that density-based methods resulted in the highest percentage of false identification and non-unique obfuscation and could be easily adopted as a suitable method for such data without considering the polygon nature of the object. While the N*Rand method reported the worst performance in terms of spatial pattern preservation, it presents the lowest false identification and non-unique obfuscation due to generating a larger obfuscation area. Our suggestion to reduce both the false identification and non-unique obfuscation rate is to exclude the original points (area in polygon base) and obfuscated points (area of obfuscated points) from the obfuscation area generated by any obfuscation method. In polygon base consideration, the shape of the obfuscated polygon should also be acknowledged. One possible scenario is to keep the original shape by transforming the original polygon to an obfuscated centroid, but this might lead the attacker to easily identify the farm by its shape. Instead, we suggest researchers expand the obfuscated area to a circle with radius of a distance of field parcel's centroid to its farthest edge (Figure 3.12b). However, further research could be conducted to determine the impact of this suggestion on methods performance in terms of privacy protection and accuracy, across different GIScience thematic areas of research.

The results also indicate that among N*Rand methods (Rand, N Rand and N Mix), N Rand satisfies higher k-anonymity, all obfuscated points satisfy 25-anonymity with lower distribution

preservation, correlation alteration, and cluster preservation with higher global and local divergence while Rand and NMix performance show almost the same level of k-anonymity and spatial pattern preservation. This supports the findings of Wightman *et al* (2011), who showed similar results for trajectory data using different evaluation methods. Although N*Rand methods, particularly the NRand method, provide the highest percentage of high level k-anonymity (Figure 3.9d) and privacy protection, they also provide the highest percentage of large displacements (Figure 3.9e) and subsequently lowest distribution preservation and cluster preservation, highest correlation alteration and higher global and local divergence (Figure 3.10). The advantage of N*Rand is to provide low percentage of false identification and non-unique obfuscation when the polygon nature of an object is considered (Figure 3.9a, 3.9b). Subsequently, more research is needed to determine the impact of excluding original points (area in polygon base) and obfuscation points (area of obfuscated points) from obfuscation area generated by these methods.

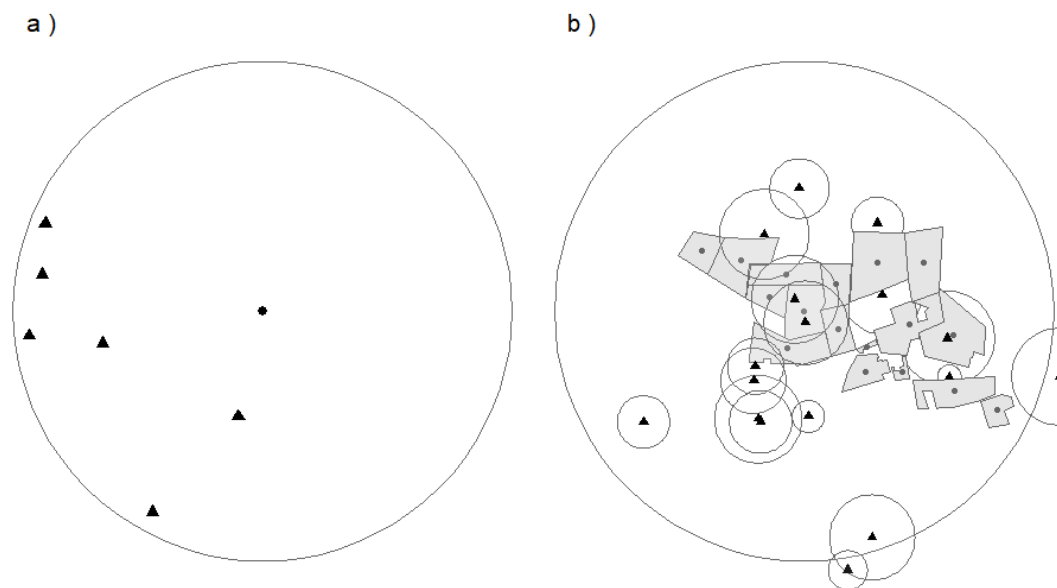


Figure 3.12: a) Obfuscated area contains $k-1$ other points (black triangle) to satisfy k -anonymity. b) Original field parcels with their centroids (grey area with black dots) and their obfuscated centres (black triangles) using density method in an area of 1000m radius, showing that obfuscated point might be in a another field parcels (risk of false identification), might be inside the other obfuscated area (black circle) or two obfuscated points might be inside the same original field parcels or obfuscated area (risk of non-unique obfuscation)

Many of these obfuscation methods require user-decisions to specify the parameters. To examine the impact of this, we explore the impact the number of wings included in the Pinwheel method had on performance by using two different angles between two sequential wings. The results of two different angles of 45° and 225° show no significant differences are observed within different evaluation metrics, just that the distribution preservation, global and local divergence of angle 225° is slightly worse than angle 45° . Pinwheel225 shows better performance in cluster preservation and correlation alteration than Pinwheel45 due to providing smaller and asymmetrical obfuscated area and less variability to select obfuscated points, these results are consistent with those obtained by Zurbaran *et al* (2020) who applied the Pinwheel method with 30° and 100° on crime data. However, just based on the comparison of two angles, a definite conclusion may not be accurate. The method should be replicated with other angles on other datasets for further research to reach a more robust conclusion. Similarly, it is well established in GIScience that the neighbourhood size influences the results of local spatial statistics (Ghimire, Rogan, & Miller, 2010; Miller & Hanham, 2011; Holloway & Miller, 2015). To explore the effect of spatial scale on local divergence, the Ldi, was also calculated using a much finer consideration of neighbourhood (fishnet cell size of 500m and distance band of 900m) (Figure 3.10c). While results indicate high Ldi (between 36 to 94) and dissimilarity across all methods, the same trend is observed. The density-based methods with less local spatial divergence (between 36 and 64) represent lowest dissimilarity, N Rand and combined N Rand with Pinwheel (Pinw225-N Rand) methods having the largest local spatial convergence (92 and 94) and the highest dissimilarity. While we might expect finer spatial neighbourhoods to produce more local divergence and exploration of more scales to provide a wider range of results, the fact that results are broadly consistent across both a coarse and fine consideration of scale provides us with confidence in our results.

In addition to the polygon nature of field parcels (objects) another important difference between agri-environmental data with other GIScience data that previous obfuscation methods were applied on, is the location related information they contain such as fertilizer compounds, herd, soil and climate data. It is important to explore how the obfuscation methods preserve the correlation between these location-related attributes with spatial coordinates in the process of obfuscation. In this study the total correlation alteration (TCA) between soil-related (pH, N, P, K) and herd-related attributes (which is not changed in process of obfuscation) with spatial coordinates have been considered as an important metric to measure the accuracy of obfuscation methods on *NMP Online* data. These variables are internal to *NMP Online* so in

addition to further research outlined above, a next logical step will be to test these results with other external relevant datasets, which will change in the process of obfuscation, such as soil type or climate data to determine the wider applicability of methods for use with agri-environmental data. The application of geographic context to spatial data or the development of context-aware methods are burgeoning research areas in other GIScience and thematic domains (Kwan & Schwanen, 2018; Brum-Bastos, Łoś, Long, Nelson, & Demšar, 2021) meaning such research would be timely.

Data sharing is necessary to generate knowledge, however geoprivacy concerns present a major challenge for GIScience. *NMP Online* is a spatially explicit tool used by farmers and advisors to develop nutrient management plans for Irish farms. Even with such tools, there remains an urgent need to increase nutrient efficiency in agriculture and to nutrient limit losses to the environment, particularly under increasingly stringent environmental legislation. To unlock the potential of existing agricultural data, such as *NMP Online*, the trade-off between field parcel confidentiality and spatial pattern preservation must be determined. Through this work, we have identified the obfuscation techniques to achieve the best results in relation to spatial pattern preservation and geoprivacy protection. This increases scope to use the data for research purposes with a view to increasing sustainability in the agricultural sector.

3.6. Conclusion

The main contribution of this paper was 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 privacy protection and accuracy can lead to the identification of the most appropriate obfuscation method. According to this investigation, obfuscation methods that report smaller location distribution, correlation alteration and less MCE, global and local divergence provide better performance in term of accuracy such as density-based methods: AHilb, Density-k and Density-Pinwheel methods (Figure 3.4, 3.6, 3.9c and Table 3.6).

Methods with a higher percentage of points satisfy higher k-anonymity and higher percentage of lower points displacement minimising trade-off between privacy protection and accuracy such as Donut-AHilb and Density-Pinwheel (Figure 3.9d, 3.9e). Finally, we highlight the importance of maintaining geographic context within the obfuscation methods, identifying those certain methods, such as density-based methods, Density-k and AHilb methods alter the correlation between geographical variables less than others such as NRand, combined method

Pinwheel-NRand and Donut-k method (Figure 3.6 and Table 3.6). Such consideration of the space through which data exists is an important consideration for all GIScience obfuscation methods. Altogether, we have identified the obfuscation techniques to achieve the best results in relation to spatial pattern preservation and geoprivacy protection using agricultural data. Moving forward, this increases scope to use the data for research purposes with a view to increasing sustainability in the agriculture sector.

False identification and non-unique obfuscation are two important factors for evaluating obfuscation methods which have not received much attention in previous research. High percentage of false identification and non-unique obfuscation is identified as a main drawback of point-based obfuscation method when applied on polygon data. Excluding the original points (area in polygon base) and obfuscated points (area of obfuscated points) from the obfuscation area generated by the obfuscation method can be considered to reduce the occurrence of false identification and non-unique obfuscation, In Chapter 4, this suggestion is considered in design and development of polygon-based methods and further research is carried out to examine the impact of this suggestion on methods performance in terms of privacy protection and accuracy.

Chapter 4 Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data

This chapter also published as a peer-reviewed scientific article:

Nowbakht, P., O'Sullivan, L., Wall, D. P., & Holloway, P. (2023). Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data. Transactions in GIS, 27, 84– 104. <https://doi.org/10.1111/tgis.13009>

4.1. Introduction

Achieving food security with growing global population alongside the concomitant impacts of climate change is one of the most pressing and challenging global concerns (Fujimori et al., 2019), one of the most significant discussions at the 2021 COP26 in Glasgow, UK (Moosmann et al., 2021) and an essential component of the 2030 Sustainable Development Goals (SDGs) (Bouma et al., 2019). The growing world population, demands increased agriculture production to reach food security (Godfray et al., 2010; Hrustek, 2020). At the same time, protecting the natural resource base, and mitigating climate change are key frontiers of the agriculture industry, requiring an innovative solution (Hrustek, 2020). Digital and smart farming is designed to achieve sustainable agriculture and obtain more production, while simultaneously preserving sustainability and reducing the environmental impact of agriculture activities (Ferris, 2017; Hrustek, 2020; Laborde & Piñeiro, 2018; Soomro, 2015). Digital farming generates a significant amount of valuable data, much of which is collected from farms that is linked to the Global Positioning System (GPS) location. For example, auto-steering tractors fitted with GPS units reduce overlapping of passes while planting seeds or applying chemicals and fertilizers, which reduces the cost, fuel consumption and subsequently protects the environment while increasing crop production (Bora et al., 2012; Fielke et al., 2019; Top & Kesk, 2019).

The available spatial agriculture data if combined with climate, environmental and other spatial data can support informed decision making at farm level, regional and national scales and enables the government to design and develop more sustainable agricultural policies and services in the agriculture sector and society (Jouanjean et al., 2020). However, farmers are concerned about data ownership, preserving their data privacy and the use of their data for profit making and/or sharing with third parties. Specifically, farmers' lack of trust and knowledge about confidentially risk of data collected on farms directly influence their motivation to adopt digital farming technology (Wiseman et al., 2019). Ultimately this has a negative impact on the availability and accessibility of agricultural data that are valuable for research, innovation and agricultural policymaking (Wiseman et al., 2019). Increased awareness and understanding of the terms and conditions of data usage and data sharing agreements along with General Data Protection Regulation (GDPR), privacy and data protection laws can increase the farmer's willingness to adopt digital farming (Wiseman et al., 2019). To support sustainable land management by leveraging existing agricultural data, in

addition to increased awareness, and understanding of protective legislative supports, reducing the risk of identification is another key consideration.

Within GIScience, the predominant method to negate privacy restrictions is to obfuscate the data, preventing object identification and retrieval of private information (Sui, 2014). Obfuscation methods degrade the quality of the spatial data to protect confidentiality while preserving spatial patterns and statistical analysis, with anonymization and randomization the two most popular approaches used in GIScience (Armstrong, Rushton, & Zimmerman, 1999; Kwan, Casas, & Schmitz, 2004; Duckham & Kulik, 2005; Hampton et al., 2010; Kounadi & Leitner, 2015; Montazeri, Houmansadr, & Pishro-Nik, 2016; Seidl, Jankowski, & Clarke, 2017; Wang & Kwan, 2020; Zandbergen, & Brussoni, 2020; Zurbaran, Salazar, Brovelli, & Wightman, 2020).

Agricultural data often contains spatial information which have the potential for reidentification of the field parcels and disclosure of private information such as size of farm, type of crop, type and number of animals, type of fertilizer, type and number of equipment, financial information, which may be used for interests other than the farmer's interests such as for informing insurance, farm price, or for marketing purposes. If compromised, these data could be used for targeted marketing purposes or exploited for business development which may not be of benefit for farmers, even if their data is underpinning such efforts. Equally, these data can be extracted by government agencies and used for regulatory purposes, for example, to verify compliance with environmental and animal welfare standards, or fiscal obligations. Also, data could be used to assess farmer compliance by non-state entities, again without agreement of the data owner. Moreover, these data can give insight into the economic potential of privately owned land and could be used for land valuation or other economic assessments that may not be desired by the landowner. As demands on land are increasing and interest in the potential of the land to deliver ecosystem services, such as carbon sequestration, is growing, interest in the land base from sectors outside of agri-food has increased. Farmers are aware of the role of their data in the development of market based initiatives or other business developments and are keen that their contribution is not overlooked or used without their prior knowledge. Farmer information can be leaked through spatial coordinates or through external or background knowledge, which is known as a background knowledge attack (Rajendran et al., 2017; Samarati & Sweeney, 1998). For example, in polygon objects such as field parcels, an attacker can find the shape, approximate polygon size of a field parcel, adjacent field parcels or topological relationships with other

features such as roads or, rivers, using freely available online data and services that can lead the polygon being identified. To reduce the risk of a background knowledge attack, the shape and size of obfuscated polygon should not be the same as original polygon. Therefore, to visualize the obfuscated data on a map, the obfuscated polygon can be depicted as a square or circle, with length or radius equal to the original polygon's centroid to its farthest edge (or random edge). The topological relation of the obfuscated data also should not be the same as original data to decrease the risk of identification, but the obfuscated polygons should not overlap each other or other features and should not have gaps with other features on the map to maintain rules of topology and generate realistic map visualization.

Most of the obfuscation methods implemented in GIScience represent points, such as the location of crime (Leitner & Curtis, 2004; Thompson et al., 2015; Zurbaran et al., 2020). However, in some cases the world phenomena are best represented using polygons, such as a field parcel. A polygon is a geometric feature or vector type that represents a closed curve which is defined by a set of connected lines whose start and end points are the same. A polygon divides the space into inside and outside regions, with the interior region representing real-world regions such as lakes, cities and field parcels (Longley et al., 2015). Although, data can be represented in vector format as points or polygons, to-date, the predominant obfuscation methods have been derived on point features.

The predominant methods used within GIScience to-date include anonymization and randomization. Randomization is a two steps producer that first generates an obfuscation area and then transform the original location to a random location or the farthest/middle location among N random location inside the obfuscation area (Murad *et al* 2010, Wightman *et al* 2011, Zandbergen 2014, Seidl *et al* 2017). The generated Obfuscation area can be in different shapes, such as buffers, donuts, pinwheels, or grids (Kwan & Schwanen, 2018; Seidl et al., 2015, 2018). The size of the obfuscation area can be constant for all locations in a dataset or can be different depending upon local location density in density-based methods. The Hilbert curve has also been used, which is a fractal space-filling curve and one dimensional representation of the multidimensional space (Breinholt & Schierz, 1998). Here, the obfuscation area is usually presented as an irregular grid generated by following the Hilbert value ordering with k-anonymity satisfaction consideration (Ghinita et al., 2007b, 2007a, 2008, 2010; Um et al., 2010; Vu et al., 2012). To support this, Um, Kim and Chang (2010) suggested the use of unconnected adjacent cells followed by cells that follows the Hilbert value order to avoid unnecessary extension of obfuscation area.

To explore the performance of point-based obfuscation methods on polygon nature data, the centroid of a polygon can be considered as point data. Recent research by Nowbakht et al., (2022) developed and tested several anonymization and randomization techniques on the centroid of field parcels. Results highlighted a high percentage of false identification and non-unique obfuscation, indicating the drawback of point-based obfuscation methods when applied to geographic features best represented as polygons. Seidl et al. (2017) addressed the risk of false identification for point features which occurs when the obfuscated point mistakenly link to another original point, while Nowbakht et al., (2022) generalized false identification for polygon features when the obfuscated point was inside the area of another original polygons, and more specifically when there was an intersection between the original and obfuscated polygons. While this study identified the importance of false identification and non-unique obfuscation, as well as the main drawbacks of using point-based obfuscation methods on polygon nature objects (i.e., occurrence of non-unique obfuscation and false-identification) using polygon-based obfuscation methods in their own right still remains unexplored.

The aim of this study is therefore to develop polygon-based obfuscation methods that preserve geoprivacy and spatial data properties when considering polygon nature of agricultural data. To develop polygon-based obfuscation methods, it is important to consider properties of spatial polygon objects including the spatial coordinates, shape, and size of the polygon and topological relationship as real-world objects can be identified via these properties.

4.2. Data and Methods

4.2.1. Data preparation

The *Irish nutrient management planning online (NMP Online)* is a spatially explicit tool that was designed to develop nutrient management plans for Irish farms based upon soil and agronomic targets. It can help farmers to improve the production capacity of the farm by a fertiliser plan that helps target nutrients based upon soil fertility derived from soil test results, while achieving high levels of environment protection (Wall, D., & Plunkett, 2020). NMP online contains farm information including spatial data on field boundaries, type of crop, size of farm, livestock information, type of fertilizer, and soil analysis for over 55,000 farms. This is used to support nutrient management planning, but cannot be shared in their current format as they may compromise farmers' privacy (Murphy *et al* 2013, 2015).

This study was conducted to examine different polygon-based obfuscation methods on the subset of *NMP Online* dataset, with high density and non-uniform distribution of 22,425 field parcels for 2016 in Ireland and investigate how the polygon-based obfuscation method might overcome the limitation of point-based obfuscation method when applied on polygon nature objects. The *NMP Online* dataset contains polygon geographical data which uniquely identifiable by its centroid, to uniquely identify each polygon after obfuscation and avoid generating multiple obfuscated polygons for a same polygon a unique key (plotUID) is assigned to each polygon. A mock dataset of 5000 field parcels with similar distribution in Co.Cork, Ireland also provided and used in this study to support our findings and reproducible research (Appendix 3).

4.2.2. Polygon-based Obfuscation Methods

Polygon-based obfuscation methods, which are an adaptive form of point-based obfuscation methods, were developed to overcome the limitation of point-based obfuscation methods applied on polygon objects. This is a two steps procedure which first consists of the obfuscation area generation and secondly the construction of a random polygon inside this obfuscation area (Figure 4.1b, d).

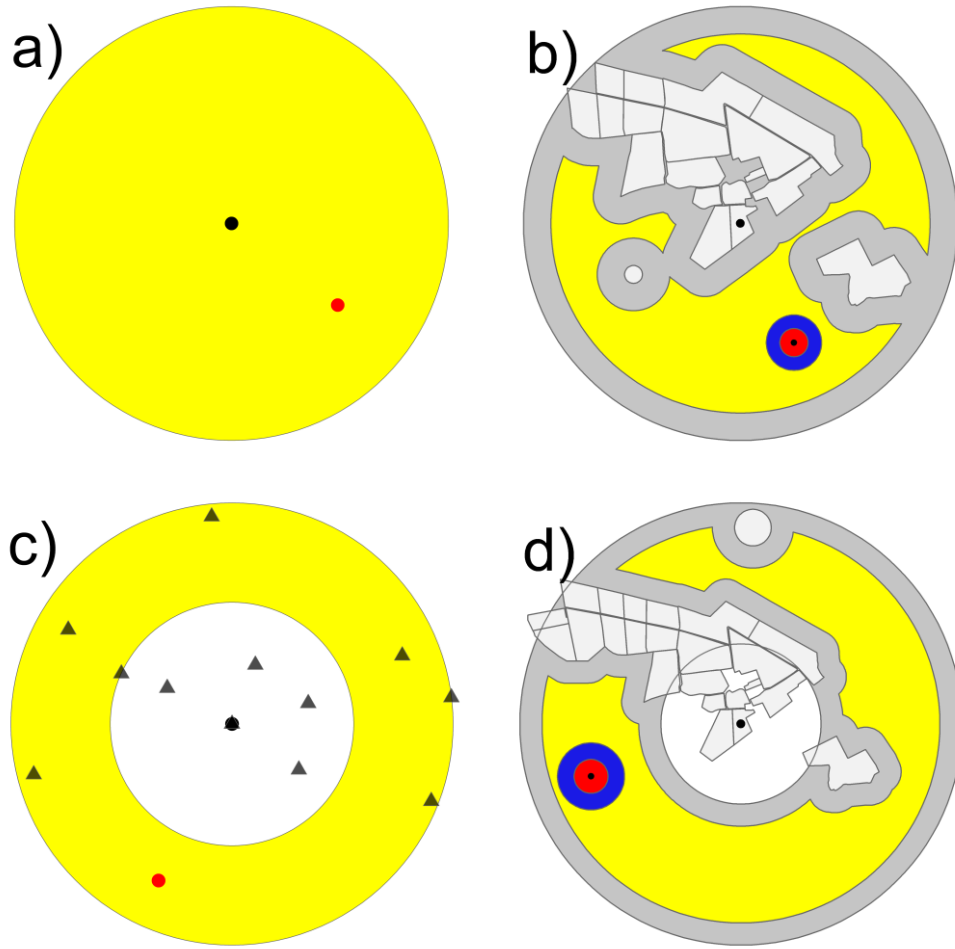


Figure 4.1: Conceptual diagram of obfuscation methods. a) and c) an example of a point-based Rand and Donut methods that transfers the original point (black dot in the centre) to a random location (red dot) inside a defined obfuscation area (yellow area), other points (black triangle). b) and d) an example of polygon based PRand and PDonut methods that transfer original location (represented here as the polygon with black dot in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area). The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon) are also represented in this diagram.

Firstly, to generate polygon-based obfuscation area, the obfuscation area was created using one of the point-based obfuscation methods on the polygon centroid. All existing polygons and obfuscated polygons inside the overall obfuscation area were then unioned and defined as a sensitive area, within which the new obfuscated point cannot be located. To provide a higher level of privacy protection, the obfuscation area should be large enough (at least k times of original or obfuscated polygon area) to provide higher variability to randomly select the obfuscated polygon.

Secondly, to construct the obfuscated polygon, we defined a buffer with a random point as a centre of obfuscated polygon, with a radius of the original polygon's centroid to its farthest edge as a maximum obfuscated polygon (MOP). This distance was chosen as it covers the entirety of the obfuscated polygon. The randomly generated obfuscated point should not be in a sensitive area, and importantly the distance between random point and border of sensitive area should be larger than radius of MOP. For this reason, the buffer zone with size of radius of MOP was defined as an extended sensitive area. To ensure that there was no overlap between obfuscated polygon and sensitive area, the extended sensitive area was removed from the obfuscation area. In other words, the polygon-based obfuscated area was generated by excluding the extended sensitive area from the obfuscation area generated by the point-based obfuscation methods for the polygon centroid (Figure 4.1b). This approach was used in PAHilb method and is outlined in algorithms (SI4.1 and SI4.2, Appendix 2). To reduce the risk of identification via the shape and size of the polygon, these features need to be modified. The MOP can be considered as an obfuscated polygon. Here, obfuscated polygon was constructed by a circle with radius equal to distance of polygon's centre to its random edge (Figure 4.1b, Figure 4.2). We opted to stop the obfuscation process here as to fully explore the impact of our novel methods, but we revisit and recommend the need to construct more realistic looking obfuscated polygon with considering shape, size, and topological relations of obfuscated polygons in the discussion. Figure 4.1 illustrates the conceptualization of obfuscation methods to compare the point-based obfuscation methods and polygon-based methods.

Another approach was examined to generate polygon-based obfuscation area, non-sensitive study area was defined by excluding all polygons from the study area in advance in an offline step. Then the obfuscation area was generated for each polygon from the non-sensitive study area and non-sensitive study area were updated by excluding each new extended obfuscated polygon. This approach is suitable when obfuscation area is large such as PNRand and PDonut methods (algorithm 4.1). Table 4.1 outlines the five polygon-based obfuscation methods that were used in this research to generate the obfuscation area, including PN*Rand (PNRand and PNMix), PDonut-k, PDensity-k, PAHilb and PDonut_AHilb.

Algorithm 4.1: PDonut Method: polygon-based obfuscated area generation

```
2. Input (sensitive,org_poly, rmin, rmax)    //org_poly: polygon in dataset, sensitive: sensitive area,
    //rmin and rmax: estimated radii of obfuscation area
3. Output (poly_obf_area, sensitive)        //polygon-based obfuscation area and sensitive area
4. pnt <- polygon centroid
13. rMOP <- farthest_edge                    //distance of centroid of polygon and its farthest edge
14. r2 <- rmax
15. buf1 <- buffer(pnt, rmin)
16. poly_obf_area <- empty                  // polygon-based obfuscation area
17. while (area(poly_obf_area) < k *area (org_poly)) do
18.     r2 <- r2 + 2 * rMOP                //gradually increase the size of obfuscation area
19.     buf2 <- buffer(pnt, r2)
20.     obf_area <- Difference(buf2, buf1)    //generate point-base obfuscation area
21.     extended_sensitive <- buffer(sensitive, rMOP) //buffering the sensitive area by rMOP
22.     poly_obf_area <- Difference (obf_area, extended_sensitive)
23. end while
24. obf_pnt <- random point inside poly_obf_area //random point inside polygon-based
    //obfuscation area as a centroid of obfuscated polygon
25. MOP <- buffer(obf_pnt, rMOP)            //buffering the obfuscated point by rMOP to generate MOP
26. sensitive <- union(sensitive, MOP)      //add MOP to sensitive area
27. return(poly_obf_area, sensitive)
```

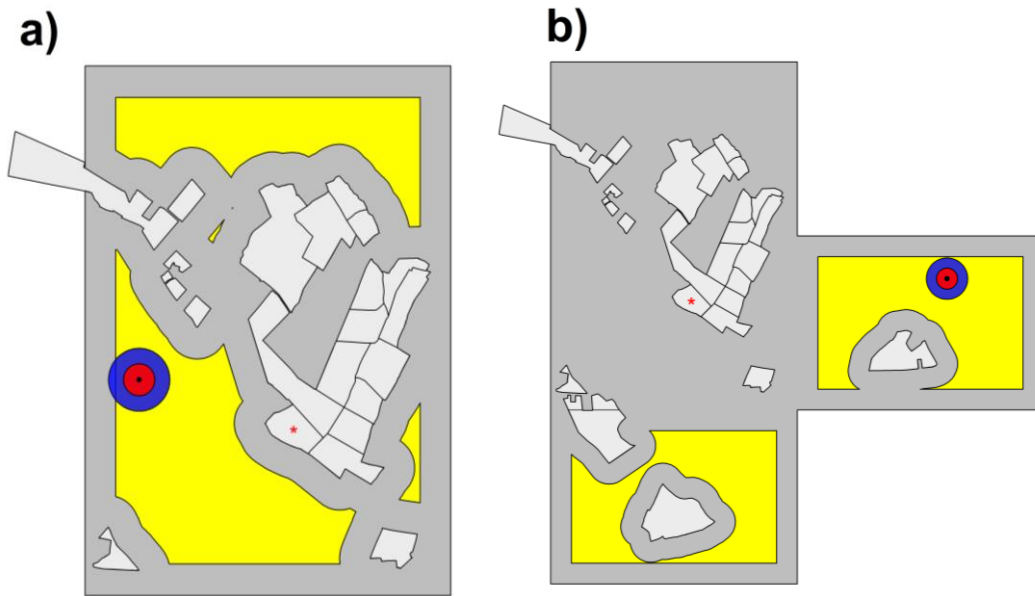


Figure 4.2: Example of polygon based PAHilb (a) and PDonut_AHilb (b) obfuscation methods that transfer original location (represented here as the polygon with red star in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area). The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon).

Table 4.1: Outline of the six polygon-based obfuscation methods

Method	Method Code	Description
PN*Rand: PRand PNRand PNMix	PN*R: PR PNR PNM	Obfuscation area is a circle around centroid of polygon and radius r excluding extended sensitive area. Obfuscated polygon is a circle around random point or the farthest/middle point among N random points inside the obfuscation area.
PDonut	PDt_R	Obfuscation area is a donut-shaped area, around centroid of polygon, excluding extended sensitive area. Constant size of obfuscation area for whole dataset.
PDensity	PD_R	Donut shape obfuscation area with varying size which the size depends on local location density.
PAHilb_NMix	PAH_NM	Obfuscation area is an area, including adjacent cells contain part of polygons inside the current cell, unconnected adjacent cells and cells follows the Hilbert curve order value, excluding sensitive area. obfuscated location is a middle point among N random location inside the obfuscation area.
PDonut-AHilb	PDt_AH	Obfuscation area is a difference of area that satisfy k_2 -anonymity and area that satisfy k_1 -anonymity generated by PAHilb method.
Pk-anonymity	PKa	Obfuscation area contains the original polygon and $k - 1$ other polygons to protect the privacy of the original polygon. k is the minimum number of other polygons that the original one is unidentifiable between them(Gruteser & Grunwald, 2003).

4.2.2.1. Polygon-based obfuscation PDonut-k and PDensity-k methods

Several point-based approaches have been adopted to obtain the size of obfuscation area considering k -anonymity concept and taking the underlying point density into account (Lu et al., 2012; Stinchcomb, 2004). For example, Kwan (2004) suggested dividing population (point) densities into ten equal interval classes and assigning the weighting factor by one from maximum density to 5.5 for lowest density.

The size of the obfuscation area can be constant for all locations in a dataset or can be different depending upon local location density in density-based methods. Differences in density of facilities between urban and rural areas, for example, create differential risks of disclosure between urban and rural areas; low densities in rural areas increase the risk of reidentification.

Donut-k and Density-k are point-based qualitative techniques were designed to maximize the geoprivacy protection and spatial pattern preservation by generating the smallest obfuscation area as far as possible (Nowbakht et al., 2022). It is important to bear in mind that when the spatial data are represented as polygons then the k-anonymity satisfaction means that the obfuscation area contains at least k-1 other polygons such that the original polygon is unidentifiable among them. Polygon-based quantitative approaches (PDonut-k and PDensity-k) were developed and summarized in algorithm 4.2 to determine the size of donut shape (value of inner and outer radii associated with k1 and k2 anonymity). The aim of this algorithm is to find the minimum value of radius of obfuscation area that satisfy k-anonymity which means for a specific value of k-anonymity if the radius decreases the k-anonymity is violated.

Algorithm 4.2: *Pfind_radius* Finding radius of buffer for each polygon-based on k-anonymity for use in PDensity-k and PDonut-k methods

```

1. input (org_poly, k1)           // org_poly: polygon in dataset, k1: level of k-anonymity
2. output(radius)                 // optimal radius that satisfy the k-anonymity
3. rmin <- rmin_threshold         // initialize the rmin with the distance of centroid of
4.                                // polygon and its farthest/nearest edge
5. min_rate <- 0.5                 // initialize the minimum threshold of increasing rate
6. rate <- 1000                   // initialize rate with a large value
7. pnt <- polygon centroid
8. while (rate > min_rate) do
9.   while (num_of_point < k1) do // number of points in a buffer with radius rmin
10.    rmin <- rmin + rate
11.   end while
12.   rate <- rate/2                // reduce the increasing rate to half
13. end while
14. //increase rmin that the buffer of size rmin contains k1 polygons
15. buf <- buffer(pnt, rmin)
16. lst_points <- list of points inside a buf
17. rmin <- max(distance between each point inside a buf and pnt + point's farthest edge)
18. return(rmin)

```

To obtain optimal values of $k1$ and $k2$ and subsequently the associated optimal values of r_{min} and r_{max} , first, this procedure can be implemented for different values of $k1$ (e.g. from 3 to 15) to obtain the optimal value of r_{min} for each value of $k1$. The optimal r_{min} is plotted against the $k1$ which shows r_{min} is increased when $k1$ is increased. The first significant growth of r_{min} with the corresponding value of $k1$ can be considered as an optimal value of $k1$ with associated optimal value of r_{min} . This procedure can be repeated for different values of $k2$ to

identify the maximum k-anonymity with optimal r_{max} . Larger values of k_1 and k_2 provide a higher level of geoprivacy but increase the radii and thereby degrade data quality. The optimal value of k_1 and k_2 can be considered when the growth of radii is significantly large. This process was suggested by Nowbakht et al., (2022) to provide higher levels of geoprivacy and spatial pattern preservation for point-based objects.

Similarly, this process can be implemented for polygon nature objects using the polygon-based procedure outlined in Algorithm 4.2. The initial value of r_{min} can be set to a distance of polygon centroid (p) to its farthest or nearest edge.

4.2.2.2. Polygon-based obfuscation PAHilb method

A polygon-based AHilb method (PAHilb) was developed to generate an obfuscation area using the Hilbert curve and k-anonymity concept. To generate an obfuscation area that satisfies k-anonymity and higher variability, the obfuscation area should be integrated and large enough for selecting the obfuscation location. This study suggests the generation of an obfuscation area which includes cells that follows three rules until k-anonymity satisfaction and the area of obfuscation area is larger than k times of area of original polygon. These three rules are: 1) eight adjacent cells that might contain part of the polygons which their centroids are inside the current cell, counting the number of complete polygons inside these cells, 2) four adjacent cells, counting the number of complete polygons inside these cells, and 3) cells which follow the Hilbert curve value order, counting the number of complete polygons inside these cells. The obfuscation area is then obtained by removing extended sensitive area inside this area. This is a time-consuming algorithm, so to reduce the processing time it is important to consider the following: ensure each cell is visited once and the area of obfuscation area is not exceeded by the threshold area. PAHilb method are summarized in algorithms (SI4.1 and SI4.2, Appendix 2). Figure 4.2 illustrates the obfuscation area and obfuscated polygon generated by obfuscation PAHilb and PDonut-AHilb methods.

4.2.3. Evaluation methods

Obfuscation methods change the geographical locations to maintain geoprivacy. As a result of this transformation, spatial clustering, location distribution and correlation between geographical attributes and other location-related attributes might be changed. Furthermore, several evaluation methods were employed to measure these alterations and compare the performance of obfuscation methods. Table 4.2 outlines a brief description of these evaluation methods. Some of these evaluation methods work the same for polygon objects and can be applied on the centroid of polygons. However, it is important to consider the polygon nature of object in some of evaluation methods such as k-anonymity satisfaction, false identification and non-unique obfuscation. To calculate the distance between two polygons several polygon distance methods can be used including centroid distance, minimum distance and Hausdorff distance (Long et al., 2018; Maciąg, 2017). In this study the centroid distance was used to calculate the distance between original polygon and its obfuscated polygon. Although the difference of distance calculated by centroid distance method and other methods is more likely less than 300 meters, as polygon-based obfuscation method transfers the centroid of original polygon to a random point inside obfuscated area as a centroid of obfuscated polygon and the obfuscated polygon was constructed inside the MOP with the radius of distance of polygon centroid from its farthest edge which in 98% of instances are less than 300 m.

The choice of obfuscation methods depends on the user specific role, need, the purpose of research and priority of accuracy or privacy. Therefore, different evaluation metrics were represented here to identify the appropriate obfuscation method for specific purpose while ensuring the higher level of privacy protection. For example, distribution and clustering preservation are essential when agricultural data combines with environmental and climate data to explore crop type suitability and crop production prediction. TCA is required when the internal attributes are the main focus of statistical analysis. The importance of false-identification and non-unique obfuscation was considered in design and development of polygon-based method that guarantee the false-identification and non-unique obfuscation are not occurred.

Cluster preservation analysis can be implemented using k-means clustering which is an iterative algorithm that partitions the unlabelled dataset into predefined k number of distinct groups. Each member of a group just belongs to one group with similar characteristic. Following the recommendation of Singh, Yadav, & Rana, (2013), Euclidean distance metric

can be used to measure data similarity. Different distance metrics such as Cosine similarity, Manhattan distance, Minkowski distance, and Euclidian distance metric can be used to measure the similarity within data. Cosine similarity measures the cosine value of the angle between two data points and magnitude is not taken to account which is perform well for clustering high dimensional data where magnitude is not important. Manhattan distance metric is sum of the absolute differences between the two vectors and Euclidean distance metric is the most common measure of distance which measures the straight-line distance between two points, works well for low dimensional data. The k-means based on Euclidean distance metric has been shown to perform best (Singh et al., 2013) for 2 or 3-dimensional data. The time complexity of the k-means clustering algorithm is approximately linear and it is suitable for clustering a large-scale dataset (Hamfelt et al., 2011). To determine the optimal number of clusters (k) the elbow method can be used. This minimizes the sum of the squared distance between the data point within a cluster and the cluster's centroid.

Table 4.2: Description of the evaluation technique that have been used to evaluate and compare obfuscation method in terms of accuracy and privacy protection.

Evaluation technique	Description	Evaluation type
Location distribution preservation	Rank the methods from the best to worst based on 2D KS-test (statistical test to compare the original points and obfuscation points distribution) or sum of 1D KS test 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	Number/percentage of misclassified obfuscation points called Misclassification Error (MCE) (Parameswaran and D. Blough, 2005)	Accuracy
k-anonymity satisfaction	The number/percentage of obfuscated points/polygons which satisfied k-anonymity for different values of k. (Kalnis <i>et al</i> 2007, Khoshgozaran <i>et al</i> 2011, Wang and Kwan 2020).	Privacy protection
Displacement distribution	The number/percentage of locations with different displacement	Accuracy
False identification	Number of points/polygons that their obfuscated points/polygons are another original point/polygons(Seidl et al., 2018) or there is an intersection between obfuscated polygon and other original polygons (when the polygon nature of object is considered)	Privacy protection
Non-unique obfuscation	Number of points/polygons with non-unique obfuscated point/polygon (two or more points/polygons have a same obfuscated point/polygons or there is intersection between an obfuscated polygon and others obfuscated polygons when polygon nature of object is considered)	Privacy protection and Accuracy

4.3. Results

4.3.1. Parameters determination

To determine the optimal value of k-anonymity and subsequently the inner and outer radii (r_{min} and r_{max}) for density-based methods, the PDensity-K was implemented for different values of k_1 where the r_{min} was initialized to a distance of a field parcel's centroid to its nearest edge. The results indicate that there has been a slight increase in r_{min} for k_1 between four to six (Figure 4.3a), while a significant growth is observed when k_1 reaches, seven. This means, $k_1=6$ is the optimal value of k_1 with associated optimal value of r_{min} . The next significant growth is visible when k_1 extends from nine to ten, so another candidate value of k_1 is nine. The choice between k_1 candidates, depends on the priority of the study. If the priority is to protect privacy, the larger value of k_1 is better. Alternatively, if the priority of the study is accuracy the best choice is to select the smallest one. In this case study, we considered the smallest value of k_1 , which is six. Based on the value of six for k_1 the PDensity-K was applied for different values of k_2 . Figure 4.3b suggests two values for k_2 , six and eleven. Again, to provide a larger obfuscation area and have more variability to select random obfuscated polygon, value of eleven is preferred. The corresponding values for inner and outer radii of donut shape obfuscated area in PDensity-K are between 85m to 20540m and 235m to 39417m respectively and for PDonut-K these values are 20540m and 39417m when $k_1 = 6$ and $k_2 = 11$.

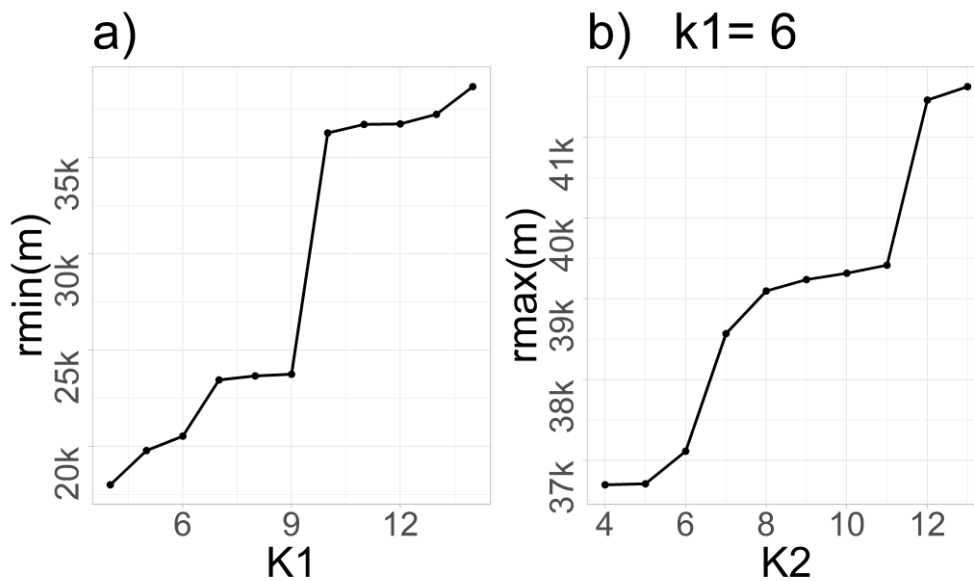


Figure 4.3: a) r_{min} calculated for different values of k_1 and b) r_{max} calculated for different values of k_2 when k_1 considered as six.

4.3.2. Distribution preservation Analysis

The centroid of the polygon objects distribution represents the polygon distribution. The results obtained from distribution preservation analysis of polygons centroid indicate that the correlation of spatial attributes for original data is 0.26609. PNRand provides maximum correlation alteration between spatial attributes of original and obfuscated data, maximum 1D KS-test statistic of both spatial attributes. This means PNRand offers the lowest distribution preservation for both spatial attributes and consequently largest value of sum of 1D KS-test statistic of both spatial attributes and correlation alteration between spatial attributes of obfuscated and original polygons' centroid (DPR) (Table 4.3). The rank of polygon-based obfuscation methods based on the 2D KS-test and DPR as a metric to measure distribution preservation are similar (Figure 4.4) which support the idea of (Nowbakht et al., 2022) that the 1D KS-test statistic and correlation alteration is an alternative approach of 2D KS-test to perform distribution preservation analysis. Table 4.3 and Figure 4.4 show that density-based methods (PAHilb, PDensity and PDonut-AHilb) demonstrate best distribution preservation followed by PNMix, PDonut and PNRand methods.

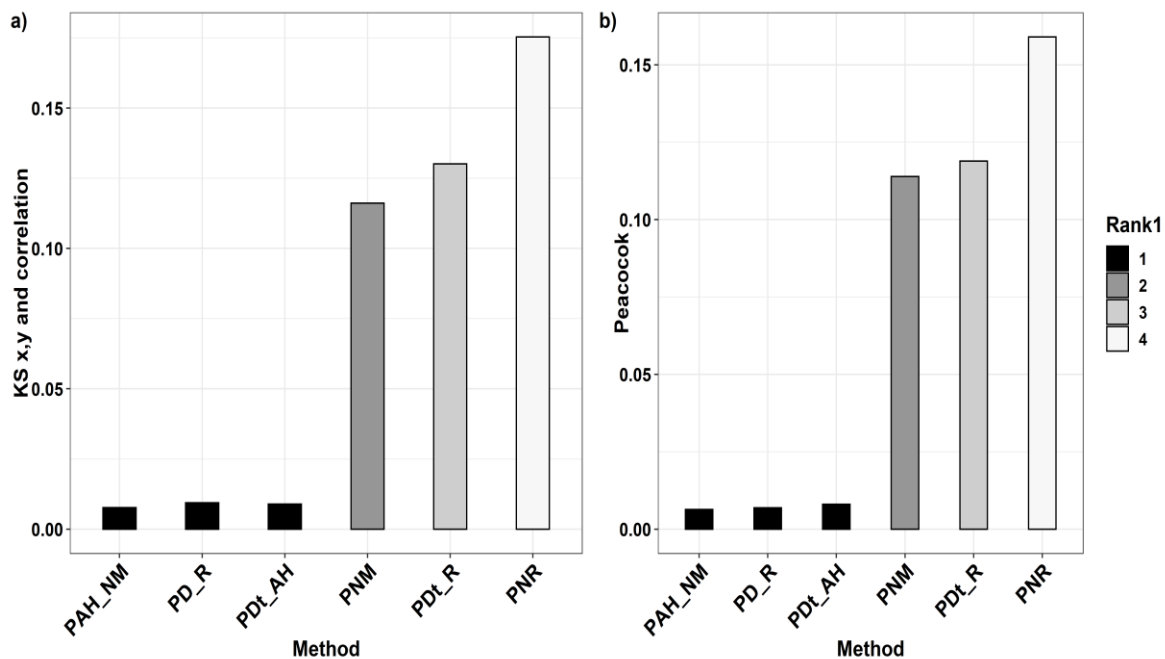


Figure 4.4: Rank of performance of different obfuscation method based on distribution preservation using 2D KS-test and combination of 1D KS-test with correlation.

Table 4.3: Comparison of 2D KS-test and combination of 1D KS-test and correlation alteration of different obfuscation method

Method	Peacock2	KS-test (X)	KS-test (Y)	Correlation of obfuscated (X and Y)	Correlation alteration of X and Y	Sum of Ks-test for X, Y and correlation alteration (DPR)
PNR	0.15897	0.06100	0.10185	0.27854	0.01245	0.17530
PNM	0.11389	0.04187	0.07309	0.26727	0.00118	0.11614
PDt_R	0.11889	0.04736	0.07474	0.27412	0.00803	0.13013
PD_R	0.00691	0.00357	0.00473	0.26718	0.00109	0.00939
PAH_NM	0.00638	0.00352	0.00410	0.26606	0.00003	0.00765
PDt_AH	0.00803	0.00410	0.00459	0.26585	0.00024	0.00893

4.3.3. Correlation Alteration Analysis

The centroid of the polygon object can be used to determine the correlation between the spatial attributes of polygon objects with location-related attributes associated with the object.

Table 4.4 shows the correlation between spatial attributes (centroid) and soil- related attributes, spatial attributes and herd-related attributes associated with the field parcel, and the sum of absolute alteration of these correlations between original and obfuscated data (TCA) using different methods. The results indicate that density-based methods (PAHilb, PDonut-AHilb and PDensity) with zero correlation alteration performed the best and PDonut and PNMix methods with 0.6 TCA showed moderate performance. PNRand with 0.12 TCA showed the worst performance in terms of correlation preservation (Figure 4.6a).

4.3.4. Cluster Preservation Analysis

Cluster preservation analysis of polygon objects can be performed based on the centroid of polygons. A polygon can be considered as a member of a cluster if its centroid is member of a cluster. The optimal number of clusters using the elbow method was identified as five (Figure 4.5). The results as shown in Figure 4.6b indicate that density-based methods (PAHilb, PDonut-AHilb and PDensity) with 0.21%, 0.3% and 0.61% MCE respectively have the best cluster preservation performance and PNRand method with 15.67% MCE has the worst while PNMix and PDonut method with similar outcomes have moderate cluster preservation.

Table 4.4: Correlations and correlation alteration of spatial attributes with soil and herd related attributes using different methods.

Method	Geographical Attributes	Soil Related Attributes				Herd Related Attributes				TCA
		pH	P	K	Lime Required	Cow	Heifer	Other Animals	Concentrate Feed	
Original	X	0.22	-0.07	0.1	-0.05	-0.16	-0.05	0.15	-0.11	
	Y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
PNR	Obfuscated x	0.19	-0.07	0.09	-0.05	-0.15	-0.04	0.13	-0.1	0.12
	Obfuscated y	-0.04	-0.12	-0.09	0.15	-0.32	-0.22	0.12	-0.17	
PNM	Obfuscated x	0.2	-0.07	0.1	-0.05	-0.15	-0.04	0.14	-0.11	0.06
	Obfuscated y	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
PDt-R	Obfuscated x	0.2	-0.07	0.09	-0.05	-0.16	-0.05	0.14	-0.11	0.06
	Obfuscated y	-0.04	-0.12	-0.09	0.16	-0.32	-0.22	0.12	-0.17	
PD-R	Obfuscated x	0.22	-0.07	0.1	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obfuscated y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
PAH-NM	Obfuscated x	0.22	-0.07	0.1	-0.05	-0.16	-0.05	0.15	-0.11	0
	Obfuscated y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
PDt-AH	Obfuscated x	0.22	-0.07	0.1	-0.05	-0.16	-0.05	-0.15	-0.11	0
	Obfuscated y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	

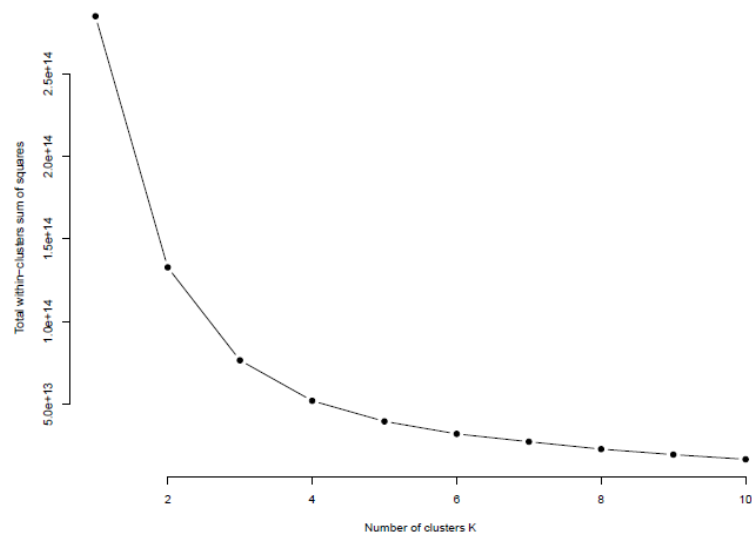


Figure 4.5: Elbow method to find the optimal value of number of cluster(k).

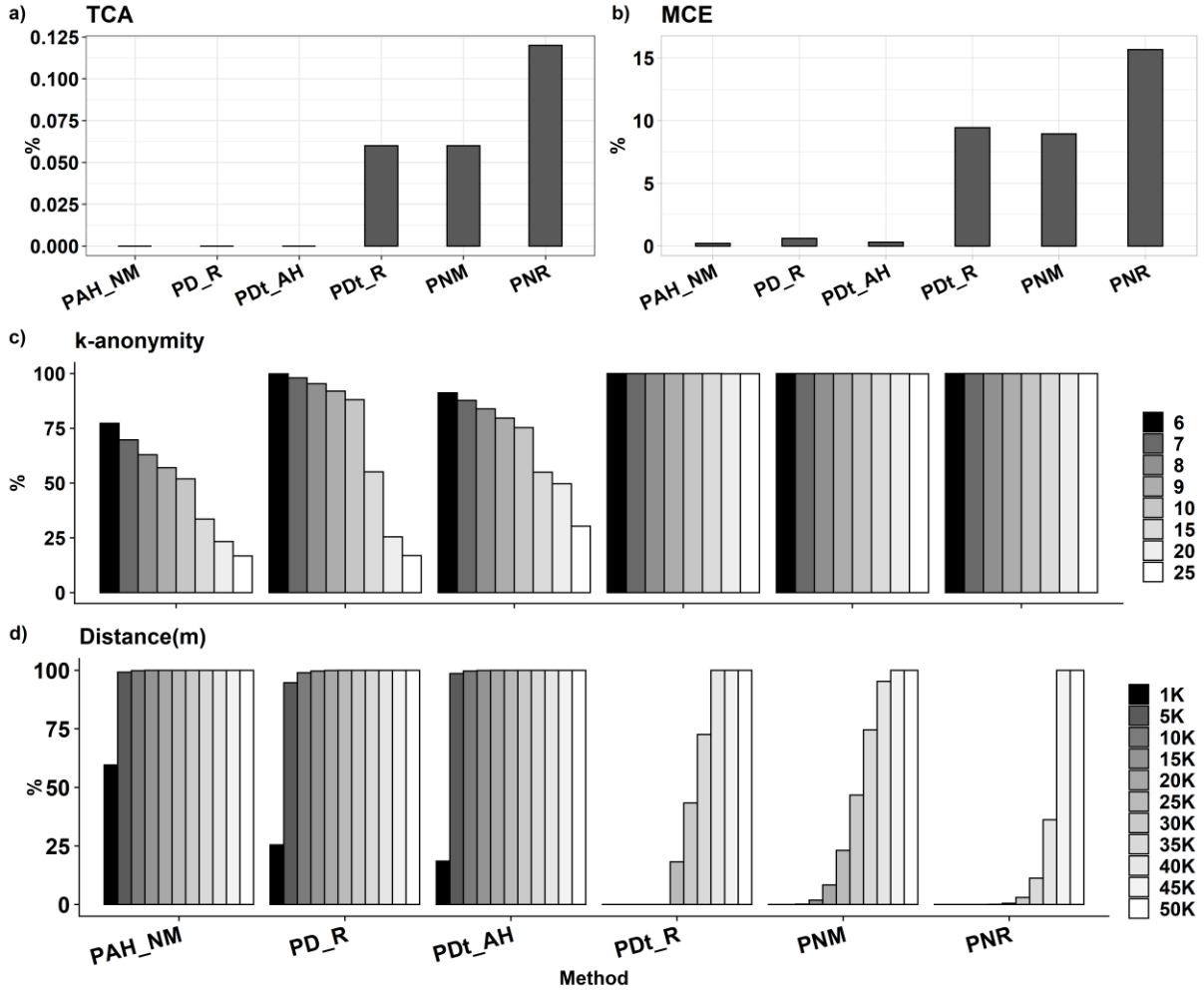


Figure 4.6: Comparison of polygon-based obfuscation methods based on a) TCA b) MCE, c) percentage of obfuscated polygons that satisfies k-anonymity for different values of k and d) percentage of obfuscated polygons with displacement less than different values of d.

4.3.5. Displacement and K-anonymity Satisfaction Analysis

Location displacement refers to distance between original location and obfuscated location, where location could be a point or polygon. polygon displacement can be defined as a displacement of a polygon centroid that is considered representative of a polygon.

The percentage of locations with certain displacement is an applicable metric to measure the level of privacy and accuracy. Since the maximum displacement is 50,000 m, to perform location displacement analysis the moderate size of intervals of 5000 m (from 1000-5000, 5001-10,000, ...) has been considered which represent variation of results. Figure 4.6d reveals that in density-based methods (PDonut-AHilb, PDensity and PAHilb) 19% to 60% of location displacement are less than 1000 m and almost all of location displacements (99% to 95%) are

less than 5000 m indicating the best accuracy with negative effect on privacy preservation. PNRand that provides very small percentage of locations displacement smaller than 30,000m to maintain higher geoprivacy protection but less accuracy.

The percentage of obfuscated locations generated by each obfuscation methods which satisfy k -anonymity for different value of k can be used as a metric to measure the level of privacy protection of obfuscation methods. The method that provides high percentage of obfuscated locations satisfy higher k -anonymity shows better performance in terms of privacy preservation.

According to Figure 4.6c PNRand, PNMix and PDonut methods provide the highest geoprivacy protection which almost fully satisfy 25-anonymity (100%, 99.84%, 99.92%). PDensity and PDonut-AHilb with (99.83% and 77.26%) satisfy 6-anonymity, 88% and 75% satisfying 10- anonymity, 17% and 30% satisfy 25-anonmity offers moderate privacy protection. PAHilb method provides 17% 25-anonymity, and more than 50% satisfy 10-anonymity and 77% satisfy 6-anonymity which demonstrate the worst in geoprivacy protection. Although, higher percentage of 6-anonymity is expected for PAHilb and PDonut-AHilb methods due to difference of shape of generated obfuscation area with circular area used to calculate the percentage of k -anonymity (Nowbakht et al., 2022).

4.3.6. False identification and Non-unique Obfuscation Analysis

High percentage of false identification and non-unique obfuscation has been identified as a main drawback of point-based obfuscation methods applied on the centroid of polygon as a point features (Nowbakht et al., 2022). Therefore, polygon-based obfuscation methods were designed and developed to overcome these limitations and certify that the risk of false-identification and non-unique obfuscation is zero by excluding the extended sensitive area from obfuscation area. However, the false identification and non-unique obfuscation are zero for all polygon-based methods.

4.4. Discussion

The main aim of this research was to design and develop an effective algorithm to obfuscate a spatial object that is represented as a polygon, in this case field parcels. In doing so, we aimed to overcome the main drawback of point-based obfuscation methods when applied on centroid of polygon objects, whereby we accounted for geographic context in the obfuscated locations and report no false-identification and non-unique obfuscation.

Several evaluation methods in terms of accuracy and privacy protection were adapted to evaluate and identify the most suitable obfuscation method. Evaluation methods were applied on the centroid of polygons as representative points such as distribution preservation, correlation alteration, cluster preservation and displacement analysis. Other evaluation methods were applied on the polygons, including K-anonymity satisfaction, false-identification, and non-unique obfuscation.

Similar research was conducted by Wang et al., (2022) to provide a guideline for choosing a proper obfuscation method among several random perturbation methods with different parameters, such as using different distribution function and circle or donut shape obfuscation area based on a suggested evaluation framework. The simulated data were generated in random, regular, and clustered spatial patterns to provide a guideline for different point spatial patterns. The first step to use this guideline is to identify the point dataset spatial patterns and then follow the guideline to identify the recommended obfuscation method for obtain more proper results regarding the certain evaluation metric. Although they examined different obfuscation and evaluation methods on different data type, the results are consistent with our finding and results on NMP data as a clustered data. The results confirm that spatial k-anonymity increases as the masking radius increases for all suggested methods and three spatial patterns simulated data. Larger radii indicate larger errors (or lower analytical accuracy), which shows the important need of mechanism to identify the trade-off between value of k-anonymity and radius. Nowbakht et al.,(2022) introduced the qualitative algorithm to determine the optimal value of k-anonymity with corresponding optimal value of radius for point data and here the adaptive algorithm was presented for polygon data. The importance of false identification also has been raised to evaluate and compare the obfuscation methods.

4.4.1. False identification and Non-unique Obfuscation Analysis

False identification and non-unique obfuscation are two important factors for evaluating obfuscation methods particularly for static polygon objects such as agricultural data. Occurrence of false-identification and non-unique obfuscation have a significant impact on data accuracy and privacy protection. This might lead to a reduction in the number of locations in obfuscated data or, to change the distribution of locations and local statistical results which affect data accuracy. Moreover, false-identification and non-unique obfuscation can negate geoprivacy protection by identifying an object (field parcel) even with incorrect information which might be misused against the owner. McLafferty (2004) introduced the term ‘false-identification’ and Seidl et al., (2017) outlined the importance of that regarding U.S. legislation and judicial processes. Subsequently, a topographical framework has been introduced to evaluate risk of correct and false-identification by considering parcel and polygon nature of existing maps which might be used to match and link the point-based obfuscated location. A high percentage of false identification and non-unique obfuscation has been identified as the main drawbacks of point-based obfuscation method when applying on centroid of polygon as a point spatial data (Nowbakht et al., 2022). Polygon-based obfuscation methods were designed and developed to solve this drawback. The main idea is to eliminate extended sensitive area from the obfuscated area generated by point-based methods. This guaranteed that there was no intersection between the obfuscated polygon and sensitive area. As a result of this, false-identification and non-unique obfuscation were zero for all polygon-based methods used in this study. As such, this approach should support any geoprivacy work whereby polygon features are the focus of the research across the GIScience discipline.

4.4.2. Spatial Pattern Preservation

Density-based methods (PAHilb, PDonut-AHilb and PDensity) performed well, reporting the best distribution preservation, cluster preservation, lowest correlation alteration, providing the best pattern preservation, followed by PDonut and PNMix with similar performance and finally PNRand method (Figure 4.4, 4.5a, b). While density-based methods demonstrate high and similar performance in terms of privacy protection, PAHilb method shows the worst geoprivacy protection and PDonut-AHilb with higher percentage of higher k-anonymity provides the best privacy protection (Figure 4.6c) for such data with high location density and non-uniform distribution. Zurbaran *et al* (2020) and Zurbarán *et al* (2018) also reported similar

findings for point-based methods that density-based method provided better performance in terms of spatial pattern preservation followed by Donut and NRand methods.

4.4.3. Geoprivacy Protection

PDonut, PNMix and PNRand with fully 25-anonymity satisfaction represent the best geoprivacy protection. While PDensity demonstrates high pattern preservation, it shows moderate geoprivacy protection with fully optimal k-anonymity and more than 50% 15-anonymity, PDonut with fully satisfying 25-anonymity which is the highest geoprivacy protection provides moderate level of distribution, clustering preservation, correlation alteration (Figure 4.4, 4.5a, b, c). Therefore, these findings suggest that PDonut-AHilb method provides the best trade- off between field parcel confidentiality and spatial pattern preservation for non-uniform polygon distribution. These findings concur with those obtained using our mock data provided in (Appendix 3) and also are consistent with previous research that has focused on points (Nowbakht et al., 2022; Zurbaran et al., 2020; 2018). While several studies confirm that density-based methods perform well in terms of privacy protection and accuracy (Cassa et al., 2006; Gruteser & Grunwald, 2003; Rajendran et al., 2017; Samarati & Sweeney, 1998; Zhong & Hengartner, 2008), since the performance of density-based methods highly depends on location density as the optimal value of k-anonymity and radii of donut shape were drawn out from data. This suggests that future studies looking to exploit obfuscation should explore density-based methods.

4.4.4. Limitation and Suggestion

Map visualization is another challenge of polygon-based obfuscation methods as the polygon can be identified via its shape, size and topological relation. The polygon-based obfuscation methods in this study generate the obfuscated area based on polygons centroid to its farthest edge to ensure no false-identification and non-unique obfuscation. This decision was made to ensure our results could focus on the novel undertakings of obfuscating polygons and ensuring non-unique obfuscation. However, we suggest that future research should explore how best, to visualize the obfuscated polygon on a map. In this study a simple depiction as a square or circle with length or radius of an original polygon centroid to its farthest edge (or random edge) as a size of obfuscated polygon was applied which may not resemble reality as effectively as possible (Figure 4.7a, b). Alternatively, randomly selecting a polygon shape from a set of irregular polygons can be used to generate an obfuscated shape. For example, this approach was used to create polygons for mock dataset (Figure 4.7b). Random generation

techniques have been developed to construct random polygons for a given set of points which have been used in a wide range of applications such as image processing, edge detection, animation and graphical application to generate natural objects such as clouds, land formations, biological or hydrological phenomena (Auer & Held, 1996; Dailey & Whitfield, 2008). To improve the quality of map visualization and generate more realistic output, random polygon generation techniques can be used to construct arbitrary shape as an obfuscated polygon, but further research should be conducted on how these arbitrary polygons fill the study area and how the obfuscated polygons relate with each other and satisfy the rules of topology.

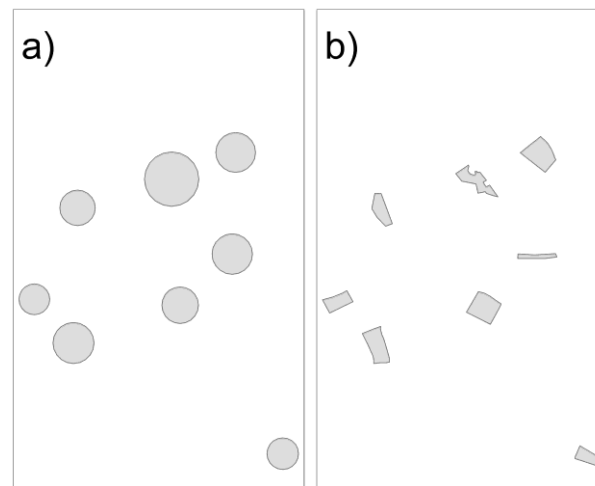


Figure 4.7: a) obfuscated polygon depicts as a circle with radius of an original polygon centroid to its farthest edge (or random edge). b) obfuscated polygon generated by randomly selecting polygon from a set of irregular polygons.

The polygon-based obfuscation methods were developed to obfuscate each polygon (field parcel) individually to ensure that the topological relation between obfuscated polygon is not preserved to reduce the risk of identification via some background knowledge about adjacent polygons. An alternative method can be developed to obfuscate the randomly selecting a set of adjacent polygons. The topology of this set of adjacent polygons can be represented as a graph structure (using Graph theory) which maintains the relationship between adjacent polygons and then randomly changes the adjacent relationship in graph structure to construct the topology of set of adjacent obfuscated polygons. Polygon-based methods guarantee that there is no non-unique obfuscation and subsequently no duplication and overlap between obfuscated polygons which satisfy the rules of no duplication and overlap of topology, but further research is recommended to explore how to fill the gap between sets of adjacent obfuscated polygons to satisfy the rule of no gap between adjacent polygons of topology.

In this study, TCA between soil related (pH, N, P, K) and herd-related attributes with spatial coordinates represent how obfuscation methods maintain the relationship of these internal environmental attributes with geographical coordinates (Figure 4.6a). This is important for studies using such data, as it means the obfuscation methods do not introduce large uncertainty into the spatial configuration of internal attributes. However, it is important to investigate the impact of obfuscation methods on relationship between external environmental and climate data with spatial coordinates. Moreover, despite the nature of polygon-based methods that guarantee no false-identification, there is a chance of occurrence of false-identification for field parcels are not in the dataset. One of the drawbacks of these obfuscation methods is that the obfuscated polygon might be inside an unrealistic region such as city centre, water body, school, etc. This has been noted in point-based obfuscation methods (Elkhodr et al., 2014; Zurbarán et al., 2018), meaning, further research is recommended on how obfuscation methods can preserve any possible clustering based on environmental data (soil type) and climate data (temperature and precipitation) which is important when exploring agri-environmental questions such as the suitability of crop type with environmental situation.

One solution to such obfuscation challenges is to generate the obfuscation area from semantic locations inside study area. Semantic locations can be defined as regions on the map with similar contextual information such as age, gender, activities, etc for example, school and hospital (Xue et al., 2009), but for agricultural data can be referred to regions with similar characteristic such as water body, desert, grassland, etc and unrealistic locations referred to regions where are not semantic. For example, grassland can be considered as a semantic to agricultural data and water body with different characteristic are unrealistic or non-semantic. Therefore, the obfuscation area can be generated by excluding the unrealistic regions from obfuscation area generated by any of polygon-based obfuscation methods introduced in this study. Alternatively, the obfuscation area can be considered the semantic regions inside the obfuscation area. It is important to note that the obfuscation area should be large enough (at least k times of original or obfuscated polygon area) to provide higher variability to select random obfuscated polygon to improve privacy protection. Such further research was beyond the scope of this research and is a direct result of the novel method applied within this research to account for non-unique obfuscation of polygons. Despite this, further studies, should aim to incorporate, semantic data into account.

Although polygon-based obfuscation methods designed to protect geoprivacy of polygon data, there is a possibility of applying them for point data. In real world, point data is inside its

surrounding area which can be considered as a polygon data such as location of patient in a hospital or location of moving vehicle in a military zone. Beside that there are various polygon features around point data that might be sensitive or non-realistic regions which should be excluded from obfuscation area for example, hospital as a sensitive region for patient and lake as a non-realistic region for moving vehicle. Indeed, further research are needed to explore how polygon-based obfuscation methods perform for data point in terms of improving privacy preservation, accuracy and computation time compared to existing point-based obfuscation methods (Damiani et al., 2009; Elkhodr et al., 2014; Scheider et al., 2020).

4.5. Conclusion

In this research we have identified and evaluated an algorithm that can overcome the challenges of false-identification and non-unique obfuscation for polygon nature objects. The results increase levels of geoprivacy that has general relevance for the wider GIScience community. In relation to the data used in this research, the results can help to allay concerns of data owners, such as farmers, which could in turn support the use of data for more informed decision making at all scales. This has relevance for sustainable land management, which is central to meeting the key societal challenges of our time.

The results of this study indicate that density-based methods with smaller location distribution alteration, TCA and MCE perform better in terms of accuracy. Density-based methods particularly PDonut-AHilb method have been found to perform the best to achieve the trade-off between field parcel confidentiality and spatial pattern preservation.

Shape, size, and topology of obfuscated polygon are important factors for map visualization, further research is recommended to improve map visualization quality while reducing the risk of identification. A further study with more focus on the relationship between external environmental and climate data with spatial coordinates and how obfuscation methods maintain this relationship or preserve any possible clustering based on environmental climate data is therefore suggested. In the next Chapter, a new polygon-based obfuscation method is designed and developed with consideration of importance of contextual factors such as environmental and climate conditions in agri-environmental research to preserve geoprivacy and contextual correlation between agricultural data and external environmental data.

Chapter 5 Maintaining Environmental Context and Geoprivacy Protection in Agriculture

*This chapter is under review as a peer-reviewed scientific article in
Information Processing in Agriculture journal*

5.1. Introduction

Climate change poses a threat to achieving global food security and sustainable development. Food security and agriculture are considered an essential component to attaining the 2030 agenda of the UN Sustainable Development Goals (SDGs) (Bouma et al., 2019). According to the Food and Agriculture Organization (FAO), the global population is expected to reach around ten billion by 2050, while world hunger increased due to the COVID-19 pandemic crisis from 8.9% to 9.3% from 2019 to 2020 (FAO et al., 2022). It is predicted that nearly 670 million people, or 8% of world population will still experience hunger by 2030, which was the SDG target for the end of hunger (FAO et al., 2022). The impact of climate change on agriculture and subsequently on food security was the most significant discussion at the 2021 COP26 (Moosmann et al., 2021), while at COP27 the official Food and Agriculture Pavilion was introduced to indicate the importance of agrifood system transformation as a part of the solution to the climate crisis³. Climate change affects agriculture in diverse ways. In terms of crop yield production, projections indicate global reductions in key commodities such as, maize by 7.4%, wheat by 6.0%, rice by 3.25% and soybean by 3.1% for a one degree-Celsius increase in global mean temperature (Zhao et al., 2017). In relation to agriculture production quality, temperature and precipitation changes reduce the consumable food calories (Ray et al., 2019), while global climate change is expected to significantly affect the population dynamics of insect pests and diseases and thus reduce agriculture production (Mafie, 2022). The impact of climate change on agriculture depends on geographical location and topography. For example, it is projected that climate change affects agriculture in southern Europe more than northern Europe. As a result, a reduction in the crop yield is observed in southern Europe due to increased temperature and reduced precipitation (McEldowney, 2020). Therefore, understanding the spatial components of agricultural systems is of key importance to achieving agricultural targets, especially given the fact that the Covid19 crisis and the war in Ukraine have put extended pressures on the need to find solutions for the climate crisis, food security, and agricultural sustainability.

Weather conditions and soil type play a vital role in agriculture. Agricultural activities such as predicting crop type suitability, growth and production, effective use of water, fertilizers and pesticides are highly dependent on environmental and climate conditions with different geospatial patterns and variability (Ganesan et al., 2021). Combinations of agricultural data

³ <https://www.fao.org/newsroom/detail/cop27-first-official-food-and-agriculture-pavilion/en>

collected on farms with environmental and climate data from different sources can be used to support farmers to make more informed decisions about farm production and use of resources, while at regional and national levels empowers governments to foster agricultural policies and services that support sustainable land management (Jouanjean et al., 2020).

Despite the advancement of technology within agriculture research to collect, store and analyse farm data, data sharing and access are a major challenge to fully exploit the digital potential of these data. The risk of confidentiality and privacy violation is a significant concern in agriculture. Privacy violations can have different and even destructive effects such as identification, data misuse, data control, and social engineering which can have a direct impact on farmer willingness to share agriculture data (Wiseman et al., 2019). Agricultural data often contain sensitive and private farm information such as size of farm, crop type, crop yield, type of fertilizer, type and number of animals, type and quantity of equipment and financial information which can be retrieved by its geospatial information and violate privacy legitimization. The conflict between sharing data and data privacy urgently requires mechanisms to protect privacy while sharing data to leverage its potential to contribute to the ambition of sustainable agriculture. Removing identification attributes such as ID number and name can reduce the risk of identification, but the spatial location of the farm is often sufficient for identification, meaning geospatial methods of ensuring privacy are needed.

Obfuscation techniques were developed by altering the location of an object while preserving spatial pattern and properties to satisfy this purpose (Armstrong et al., 1999b). To protect geoprivacy and simultaneously maintain spatial pattern and statistical accuracy of areal objects (herein defined as polygons) polygon-based obfuscation methods were introduced with consideration of the spatial properties including the spatial coordinates, shape, and size of the polygon, as real-world objects can be identified via these properties (Nowbakht et al., 2023). These methods were designed to avoid false-identification and non-unique obfuscation which are important for static polygon objects.

However, to date most obfuscation methods rely solely on distance from original feature referred to as the geometric-based techniques. These techniques do not consider the underlying environmental conditions which might generate an obfuscated location inside an irrelevant region such as, a water body, desert, etc. For example, Zurbarán et al., (2018) identified several obfuscated locations representing Twitter users in Italy, who were reported in the surrounding sea, which was unrealistic. Elkhodr et al., (2014) refer to this as a fake obfuscated location and

introduced a semantic obfuscation method (S-obfuscation) reliant on an ontological classification of geographic locations based on geographical knowledge to reduce generating fake obfuscated locations. In relation to agricultural data, environmental and climate conditions have a significant impact on agriculture and vice versa, meaning to maximise data privacy and sharing, effective methods of incorporating environmental context into the obfuscation methods are needed.

The main purpose of this chapter is to explore the extent to which agricultural data can still be utilised by preserving environmental factors via polygon-based obfuscation method development that preserves the environmental condition in the generated obfuscated data. Therefore, this research examines the emerging role of contextual factors in a new obfuscation method.

5.2. Polygon-based Environmental Similarity Obfuscation Method (PESOM)

A polygon-based environmental similarity obfuscation method (PESOM) was developed using environmental and climate data to generate obfuscation areas that are located in regions with similar environmental and climate conditions. The main idea is to cluster the study area based on environmental and climate data using a machine learning clustering algorithm. Unsupervised clustering methods such as k-means clustering, search for underlying patterns of a dataset and partitions the data based on their similarity. Temporal and spatial variability of climate data can be used to cluster the study area based on the similarity of their climate condition. A cluster that contains the original location can be considered as a primary obfuscation area to ensure that the obfuscated location is in the same environmental and climate condition to preserve environmental properties. However, according to definitions provided by Nowbakht et al., (2023) and Chapter 4, a new obfuscated area can then be constructed by excluding the extended sensitive area from the primary obfuscated area to avoid occurrence of false-identification and non-unique obfuscation and guarantee that there is no overlap between obfuscated polygon and sensitive area. Obfuscated polygon can be constructed using the same procedure used in Chapter 4 (Figure 5.1).

To apply unsupervised clustering there are two issues that should be considered: first determine the number of features and second the number of clusters. Therefore, PESOM is implemented in several stages this procedure illustrated in Figure 5.2.

5.2.1. Number of features

Spatial climate data including precipitation, maximum and minimum temperature with different temporal time resolutions (e.g. hourly, daily, monthly and annual) are features that affect agricultural properties such as crop type and crop production (Dong et al., 2021). In general, a small temporal resolution means a large number of features for clustering which might cause overfitting while a large temporal resolution means a small number of features which might reduce the accuracy of clustering. Using monthly climate data including precipitation, maximum temperature, and minimum temperature ($3 \times 12 = 36$) with 36 features might result in overfitting and annual climate data with just three features cannot explain the characteristics of different areas and the similarity between them. It is proposed to use the seasonal climate data during a specific year (Akrami et al., 2022; Bethere et al., 2017),

($4 \times 3 = 12$) with twelve features for applying K-means clustering. Twelve features are a suitable number of features and represents the variation of climate condition during the year.

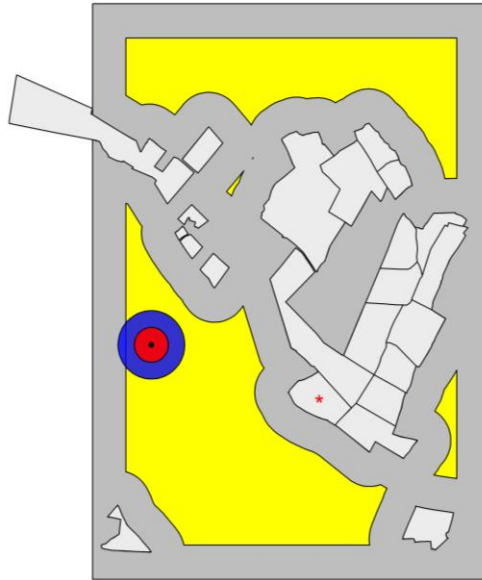


Figure 5.1: Example of PESOM that transfer original location (represented here as the polygon with red star in the centre) to a random location (red circle) inside a defined obfuscation area (yellow area). The sensitive area (grey polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon).

5.2.1.1. Principal Component Analysis (PCA)

In Machine Learning, attributes or variables are called features. According to Hughes phenomenon (Hughes, 1968) for a sample with a fixed number of observations, the performance of machine learning model increases if the number of features increases; however, after a certain point the model tends to over fit, and the performance decreases. In clustering, multicollinearity can be problematic as colinear features (highly correlated) may overweight certain features over other features to identify underlying patterns in the dataset. PCA is a dimensionality reduction technique which transforms a set of correlated features into a series of uncorrelated features with lower dimension, so called principal components. The principal component is a linear combination of original correlated features. The aim of using PCA is to obtain a minimum number of PCs that explains the maximum percentage of the variance present in the original set of features.

5.2.2. Number of Clusters

In general, the optimal number of clusters required in k-means clustering which can be obtained using the elbow method. But in this case the largest number of clusters is required where the area of each cluster and subsequently obfuscation area is minimized. To estimate the number of clusters there are several functions such as square root, cube root and log of number of observations. Carlis and Bruso (2012) suggest the RSQRT function (Eq 5.1) which refers to a recursive square root of number of observations or features. The choice of number of observations or features dependent on the scale of features, when all features are numeric, the number of observations is used otherwise the number of features is used.

$$\text{RSQRT}(x) = \{x^{2^{-r}}; r = 1, 2, \dots, r_{\max}; x^{2^{-r_{\max}}} < 2.25\} \quad \text{Eq 5.1}$$

The largest number of clusters obtained by RSQRT function can be considered as a number of clusters for clustering the study area.

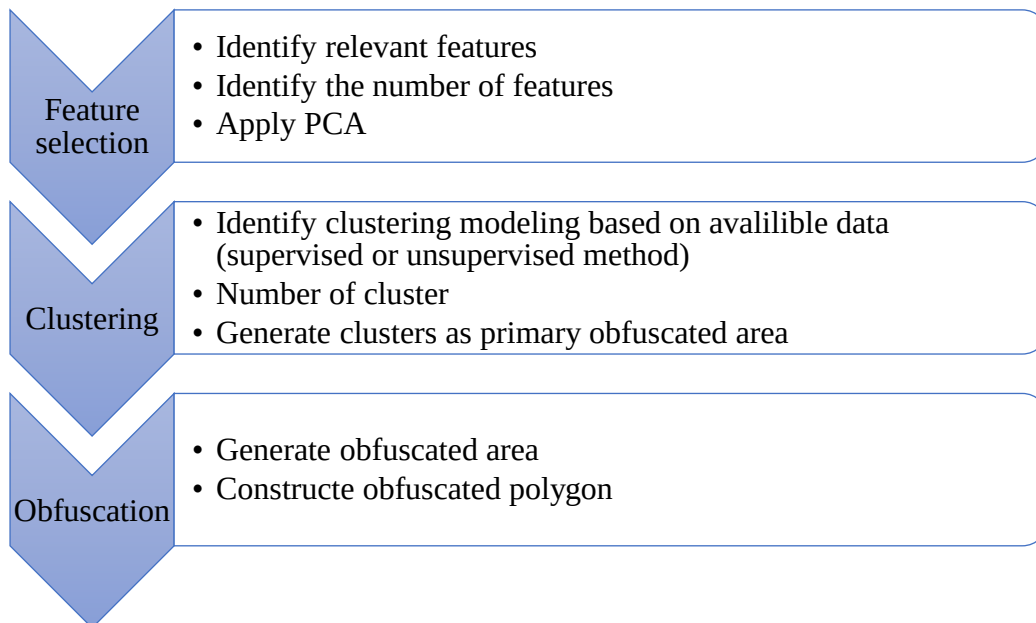


Figure 5.2: PESOM method procedure

5.3. Data and Results

5.3.1. Climate Data

Met Éireann Reanalysis (MÉRA) data which consists of a 38-year very high-resolution regional climate reanalysis over the period of 1981 to 2019 for Ireland (Gleeson et al., 2017; Gleeson & Whelan, 2020) was obtained from de la Torre Cerro et al. (in press) at a pre-processed 5km resolution. The hourly precipitation, maximum and minimum temperature from December 2015 to November 2016 (one year) was used to calculate seasonal variables of precipitation, maximum temperature, and minimum temperature for country. The seasons for Ireland was considered as follows: spring from 1st March 2016 to 31st May 2016, summer from 1st June 2016 to 31st August 2016, autumn from 1st September 2016 to 31st November 2016 and winter from 1st December 2015 to 28th February 2016. The seasonal climate data were standardized prior to further processing. Although normal distribution of the dataset is required to apply PCA, according to Altman & Bland, (1995) and Ghasemi & Zahediasl, (2012) the distribution of these datasets was ignored due to having large number of observations. Jolliffe & Cadima, (2016) also pointed out that there is no need for distributional assumption for PCA as a descriptive tool.

5.3.2. Irish nutrient management planning online (NMP Online) Data

The *Irish nutrient management planning online (NMP Online)* platform was designed and developed to support informed decision-making on farms in relation to fertiliser and nutrient management strategies to improve profitable farming. The nutrient management plan is important to boost productivity while protecting the environment specifically concerning water quality and gaseous emissions (Wall, D., & Plunkett, 2020). The *NMP online* platform provides a valuable dataset including spatial data on field boundaries, type of crop, size of farm, livestock information, type of fertilizer and soil analysis for over 55,000 farms, which can be used for research to achieve more precision nutrient management planning support and environmental protection. To use this data for research purposes, it is vital to obfuscate spatial data first in order to prevent data confidentiality and geoprivacy issues (Murphy *et al* 2013, 2015).

A unique key (plotUID) is generated and assign to each field parcel as unique identifier to retrieve the information of each field parcel after obfuscation and avoid multiple obfuscated

locations for a same location (Figure 3.2). After pre-processing of *NMP Online* data set a subset of a total of 22425 field parcels information for 2016 were extracted and used in this research. The PESOM obfuscation method was implemented on this subset of NMP online dataset with high density and non-uniform distribution to explore the capability of this method to preserve environmental and climate condition while protecting geoprivacy.

5.3.3. Parameters determination

Table 5.2 shows correlation matrix between 12 seasonal climate features. There is a strong positive correlation between some of features e.g., between spring minimum temperature and summer, autumn, and winter minimum temperature (more than 0.9) while between some features there is very weak correlation or even no correlation such as correlation between summer minimum temperature and autumn precipitation is 0.09 and there is no association between autumn minimum temperature and winter precipitation. To avoid multicollinearity effects and dimensional reduction, PCA was applied on seasonal climate features in study area. The top seven principal components (PCs) explain 48% to 99% of total variance present in the original features (Table 5.1 and Figure 5.3) was selected for the further process.

The study area (Ireland) with 5km grid-cell size contains 3440 cells meaning that there are 3440 observations of each feature. RSQRT function was applied on 3440 observations with seven numerical features (PC1, PC2, ..., PC7) to obtain potential number of clusters for k-means clustering. $RSQRT(3440) = (58.6, 7.6, 2.7, 1.6)$, the largest number of clusters, 59 was selected to cluster study area and generate optimal obfuscation area. Figure 5.4 illustrates the clustered study area of Ireland based on seven PCs with eight clusters and 59 clusters to show the contrast between size of cluster and number of clusters.

Table 5.1: Principal Component Analyses of precipitation, minimum and maximum temperature features for Ireland 2016.

Principal component	Eigenvalue	Proportion of variance (%)	Cumulative variance explained (%)
PCA1	5.80	48.35	48.35
PCA2	3.91	32.58	80.93
PCA3	0.99	8.27	89.20
PCA4	0.72	6.00	95.20
PCA5	0.21	1.73	96.93
PCA6	0.18	1.47	98.40
PCA7	0.08	0.66	99.07
PCA8	0.06	0.53	99.59
PCA9	0.02	0.18	99.77
PCA10	0.01	0.11	99.88
PCA11	0.01	0.08	99.96
PCA12	0.01	0.04	100.00

Table 5.2: Correlation matrix between 12 seasonal climate data including precipitation, minimum and maximum temperature for Ireland.

		Spring			Summer			Autumn			Winter		
		Rain	Tmax	Tmin	Rain	Tmax	Tmin	Rain	Tmax	Tmin	Rain	Tmax	Tmin
Spring	Rain	1	-0.35	-0.35	0.55	-0.36	-0.56	0.55	-0.76	-0.41	0.66	-0.60	-0.32
	Tmax		1	-0.24	0.21	0.84	0.05	-0.46	0.49	-0.28	-0.31	0.25	-0.24
	Tmin			1	-0.26	-0.49	0.93	0.34	0.54	0.97	0.06	0.81	0.97
Summer	Rain				1	-0.30	-0.39	0.40	-0.64	-0.37	0.18	-0.58	-0.34
	Tmax					1	-0.17	-0.69	0.40	-0.46	-0.48	-0.01	-0.50
	Tmin						1	0.09	0.77	0.92	-0.16	0.91	0.90
Autumn	Rain							1	-0.38	0.23	0.70	-0.04	0.35
	Tmax								1	0.58	-0.45	0.86	0.53
	Tmin									1	0.00	0.83	0.97
Winter	Rain										1	-0.10	0.15
	Tmax											1	0.84
	Tmin												1

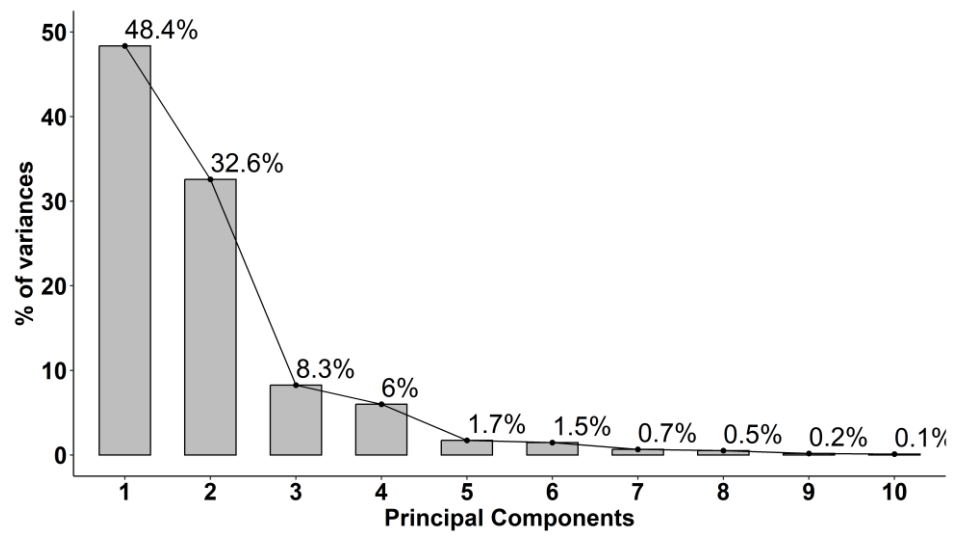
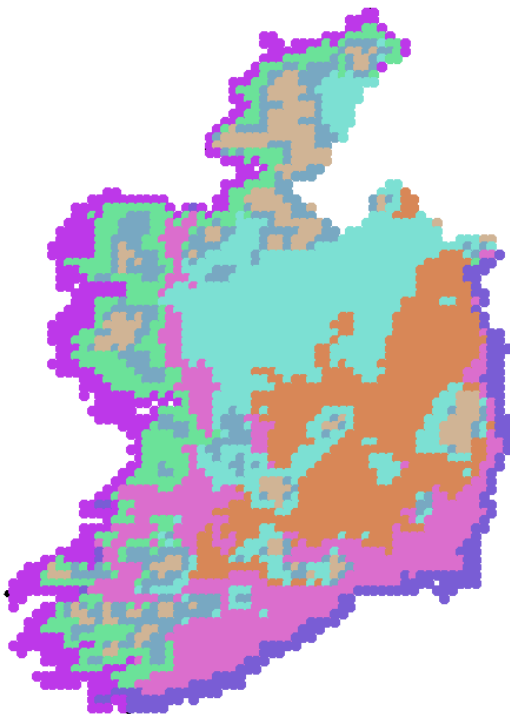


Figure 5.3: Scree Plot to determine the number of Principal Components based on % of variance explained by each PCs.

a)



b)

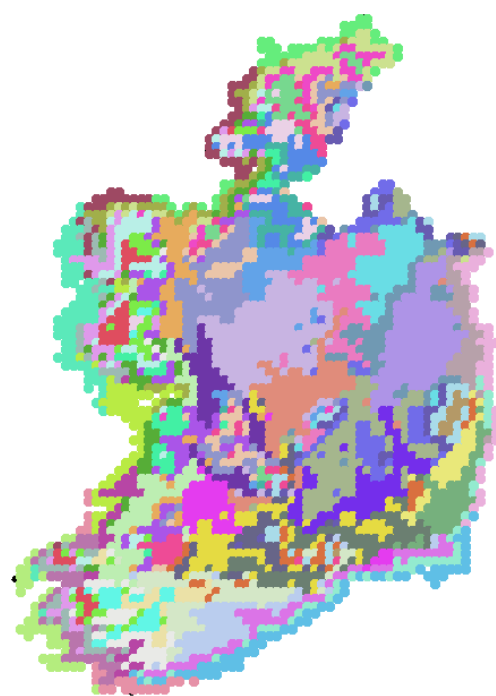


Figure 5.4: Clustered study area (Ireland) based on seven PCs a) eight clusters b) 59 clusters.

5.3.4. Evaluation Metrics

There are several determinant factors in choosing the obfuscation methods such as the user's specific role and need, the purpose of research, and the priority of accuracy or privacy protection. For example, in research related to environment and weather conditions such as exploring the impact of agriculture on the environment or environment on agriculture, crop suitability and crop production prediction, environmental and climatic condition preservation are essential.

The polygon-based obfuscation method alters original spatial data to new locations with new shapes, size, and new topography to reduce risk of identification. To evaluate the performance of obfuscation methods, the alteration of these factors should be considered. Nowbakht et al., (2023) highlighted several evaluation metrics with consideration of polygon nature of objects to identify the suitability of different obfuscation methods for specific cases. In certain evaluation techniques such as correlation alteration, cluster preservation, and displacement analysis the centroid of polygons as representative points was being used. The distance between the original point and obfuscated point was defined as a point displacement, for polygon nature objects the location displacement was referred to as the displacement of a polygon centroid. The brief description of evaluation metrics is outlined in Table 5.3.

Table 5.3: Description of the evaluation metrics to measure the obfuscation method capability in terms of accuracy and privacy protection.

Evaluation metric	Description	Evaluation type
Correlation preservation	Total correlation alteration of location-related features between original and obfuscated location (TCA)	Accuracy
Spatial cluster preservation	Percentage of misclassified obfuscated location called Misclassification Error (MCE) (Parameswaran and D. Blough, 2005)	Accuracy
k-anonymity satisfaction	The percentage of obfuscated location which satisfied k-anonymity for different values of k. (Kalnis <i>et al</i> 2007, Khoshgozaran <i>et al</i> 2011, Wang and Kwan 2020).	Privacy protection
Displacement distribution	The percentage of locations with different displacement	Accuracy
False-identification	Number of locations that there is an intersection between obfuscated location and other original location	Privacy protection
Non-unique obfuscation	Number of locations with non-unique obfuscated point/polygon (two or more points/polygons have a same obfuscated point/polygons or there is intersection between an obfuscated polygon and others obfuscated polygons when polygon nature of object is considered)	Privacy protection and Accuracy

5.3.5. Results

The results show that PESOM provides high geoprivacy protection that almost fully satisfy 100-anonymity (98.13%) and 75% satisfy 1000-anonymity with negative effect on accuracy. That means 98% of obfuscated field parcels were in a region with 100 other field parcels which reduce the risk of identification by 1/100 and 75% of them were in a region of 1000 field parcels with very low risk of identification of 1/1000. A consequence of the lower risk of identification is generating a larger region around the original location which was transferred to the obfuscated location which causes greater location displacement and reduces the statistical accuracy. Almost 50% of displacement is greater than 40,000 m and less than 10% of displacement is less than 20,000 m which show low statistical accuracy (Figure 5.5). Figure 5.5 indicates the negative relationship between percentage of k-anonymity satisfaction and percentage of displacement.

There is no false-identification and non-unique obfuscation were reported as expected due to polygon-based development of PESOM by excluding the extended sensitive area from the primary obfuscation area.

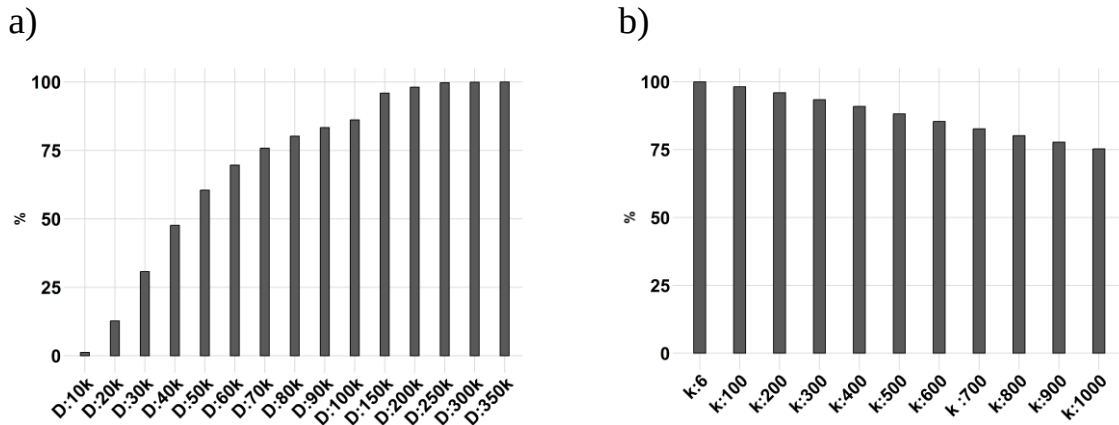


Figure 5.5: a) percentage of obfuscated polygons with displacement less than different values of d using PESOM method. b) percentage of obfuscated polygons that satisfies k-anonymity for different values of k.

The correlation between geographical features and internal soil related features (pH, P, K, Lime-required), herd related features and sum of absolute difference of location-related features between original and obfuscated location (TCA) is shown in Table 5.4. TCA is a metric to show how the relationship between location and its internal features is changed in process of obfuscation, the lower value shows less alteration and therefore, higher correlation

preservation. PESOM method provides relatively low correlation preservation with TCA of 1.5.

PESOM method provides perfect environmental and climate cluster preservation, due to the nature of method that guarantees that the obfuscated location will be in the same environmental and climate condition as the original location. That means there is no misclassified obfuscated location, and the MCE is zero. As a results of this, the distribution of features which was used to cluster the study area (here seven PCs) is not changed. Figure 5.6 shows the similarity of distribution of seven PCs which was used to cluster the study area for original and obfuscated data. Nowbakht et al., (2023) identified that density-based methods provide the best trade-off between field parcel confidentiality and statistical accuracy. The MCE rate of those method based on environmental clustering is 20% which means 20% of obfuscated location are in different environmental and climatic conditions than the original locations.

Table 5.3: Correlations and correlation alteration of spatial features with soil and herd related features using PESOM method.

Method	Geographical Features	Soil Related Features				Herd Related Features				TCA
		pH	P	K	Lime	Cow	Heifer	Other Animals	Concentrate Feed	
Original	X	0.22	-0.07	0.1	-0.05	-0.16	-0.05	0.15	-0.11	
	Y	-0.05	-0.12	-0.09	0.16	-0.32	-0.22	0.11	-0.17	
PESOM	Obfuscated x	0.15	-0.02	0.09	-0.07	-0.06	0.01	0.08	-0.05	1.5
	Obfuscated y	-0.10	-0.12	0.08	0.15	-0.29	-0.2	0.08	-0.17	

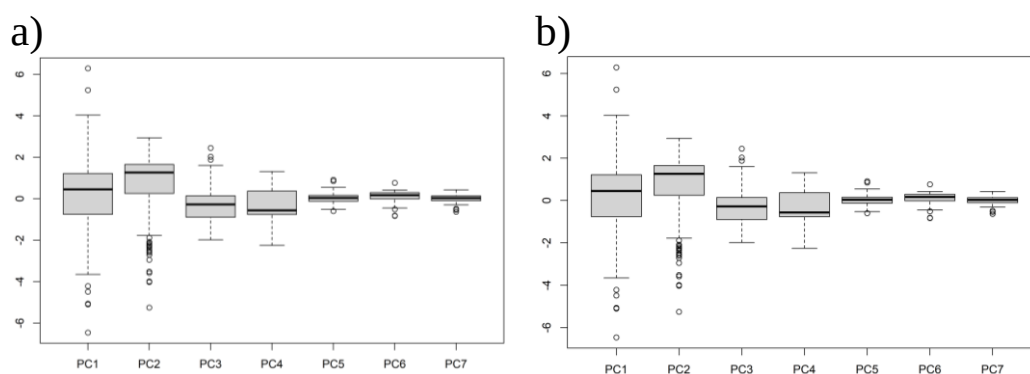


Figure 5.6: Boxplots of seven PCs used for clustering study area a) original data b) obfuscated data.

5.4. Discussion

Implicitly, the impact of agriculture on the environment and conversely the environment on agriculture is indicative of the importance of environmental and climatic conditions preservation of shared agricultural data which are prone to privacy violation via their spatial data. The polygon nature of agricultural data (fields parcel/farms) make them identifiable by their location coordinates, shape, size, topology, and spatial relationship between adjacent polygons. Several polygon-based obfuscation methods were designed and developed with consideration of these properties ensuring no occurrence of false-identification and non-unique obfuscation (Nowbakht et al., 2023). These obfuscation methods were geometric-based methods without consideration of underlying environmental condition. In this study PESOM, a polygon-based obfuscation method with consideration of underlying environmental and climatic conditions was developed. The generated obfuscated data with high geoprivacy protection and perfect environmental and climatic clustering preservation is suitable for agri-environmental and climate change research and can be shared with a wide range of researchers and even released for open access.

The absence of non-unique obfuscation, guarantees there is no duplication and overlap between obfuscated locations. As a result, the number of obfuscated data remains constant which is crucial for static objects such as agricultural data to maintain general statistical accuracy. No false-identification and non-unique obfuscation increase the data confidentiality due to a reduced possibility of linking incorrect information to an individual object (in this instance field parcels). PESOM provides a large level of displacements and consequently large TCA, which indicates the alteration of relationship between internal features (soil-related and herd-related features) and spatial coordinates. Geoprivacy protection is reported to be extremely high in obfuscated data. Although the disclosure risk is dependent on clustering parameters (type and number of features and number of clusters). A similar study was carried out using environmental features for the epidemiological analysis by Richter, (2018), who developed a verified neighbour approach by using environmental features to limit the potential randomly obfuscated point to those with similar environmental conditions. For example, the obfuscation area was clustered based on the water system supplier. The results showed a high level of clustering preservation based on external features and a high-level of geoprivacy protection; however, in contrast to our findings it provided low displacement due to generating integrated clusters, which indicates the effect of external features on obfuscation method performance and also using a nearest neighbour approach to randomly select the obfuscation

location. Likewise, Houfah-Khoufah & Touya, (2020) conducted a study to compare the performance of different obfuscation methods particularly in regard to cluster preservation on simulated COVID-19 data from Paris. A cluster-oriented method was introduced using the DBSCAN algorithm to cluster the data points based on internal features and generate the polygon extent of cluster as obfuscated area. Although clustering was not based on environmental (external) features and applied on different spatial data type, the results are consistent with our finding that cluster-oriented method provides perfect cluster preservation and high level of geoprivacy protection. It must be borne in mind in process of obfuscation, the external features might be changed which are highly important contextual factors in agri-environmental research. The PESOM is a context-aware method that considers the importance of the contribution of these contextual factors in agri-environment research and guarantees to not alter these factors in obfuscated data (Brum-Bastos et al., 2021).

The objective of this study was to generate obfuscated data that can be suitable for agri-environmental research. Although density-based methods were reported with sufficient privacy protection and statistical accuracy (Nowbakht et al., 2023), a 20% misclassified rate is observed regarding environmental and climatic clustering used in PESOM which reduces the suitability of generated obfuscated data in environmental and weather-related research. The PESOM method with a high level of privacy protection and no misclassified environmental cluster provides suitable obfuscated data that has scope to be used in future in research where the climate and environmental factors are important, such as agricultural research. For example, crop suitability and productivity, locust outbreaks and identifying suitable sites for example for aquaculture development which are highly dependent on environmental and climate conditions (Mechiche-Alami & Abdi, 2020; O'Mahony & Cerro, 2021; Peng et al., 2020). Therefore, while the PESOM method can obfuscate farms with a suitable level of privacy, questions remain as to whether these data can be used within statistical models that incorporate internal and external environmental factors, warranting future research.

PESOM satisfies high level of geoprivacy protection with no false-identification and non-unique obfuscation risk for existing data and absolute environmental and climatic clustering preservation. It provides a low level of correlation preservation; large location displacement and subsequently low local analytical accuracy can be expected. Therefore, it is suitable for large scale research such as national and global scale. The effectiveness of PESOM method at local scale research is reduced for example, in research to explore the effect of

fertilizer in a local region. Although, a large number of clusters was used to generate optimal primary obfuscation area and subsequently, low displacement, clusters might not be integrated and could contain several regions at long distances. Therefore, randomly selected obfuscated location might be in the furthest region which causes a large displacement. Scale can significantly affect results many phenomena, including agriculture (Mao et al., 2021; Merlos & Hijmans, 2020). The environmental and climate data used was relatively fine scale (~4 km), but even with a fine spatial resolution this data still covers a relatively large area. Therefore, further research is recommended to explore the impact of scale on efficiency of PESOM method.

To improve the statistical accuracy while maintaining absolute environmental and climatic clustering, PESOM and existing obfuscation methods particularly density-based methods can be integrated, or the threshold distance can be considered to reduce the possibility of large displacement. Figure 5.7a shows the negative effect of generating a large, obfuscated area and subsequently large displacement which is inevitable in PESOM, for location (field parcel) number 6507 belonging to cluster 24 with 1760 other field parcels. To reduce the probability of large displacement, the obfuscated area can be limited to a buffer zone with a radius of 4000m around the original location as a threshold with 115 other locations (Figure 5.7b) which still provides a high level of k-anonymity (115-anonymity) along with maintaining environmental conditions. However, further research should be extended to explore how to determine the density-based methods parameters or threshold distances with regard to the level of geoprivacy protection. An alternative suggestion is to generate smaller and integrated obfuscation areas by sorting the separated parts of primary obfuscation area (cluster) in terms of closeness to original location and include them to obfuscation area until satisfaction of certain level of k-anonymity.

In this study climate data and unsupervised machine learning were used for clustering the study area, the choice of features selected with the technique to cluster or classify the study area dependent upon the availability of data, labelled data, specific interest, and problem to be solved. For example, the study area could be clustered based on soil type available in, soil information dataset for research associated with the effect of soil type in agriculture production. Therefore, this method is flexible enough to be adopted across multiple case studies and integrated into various case studies. Moving forward, research is needed to investigate the importance of different data formats (i.e., continuous vs categorical), as well as the importance of the interdependence of environmental variables.

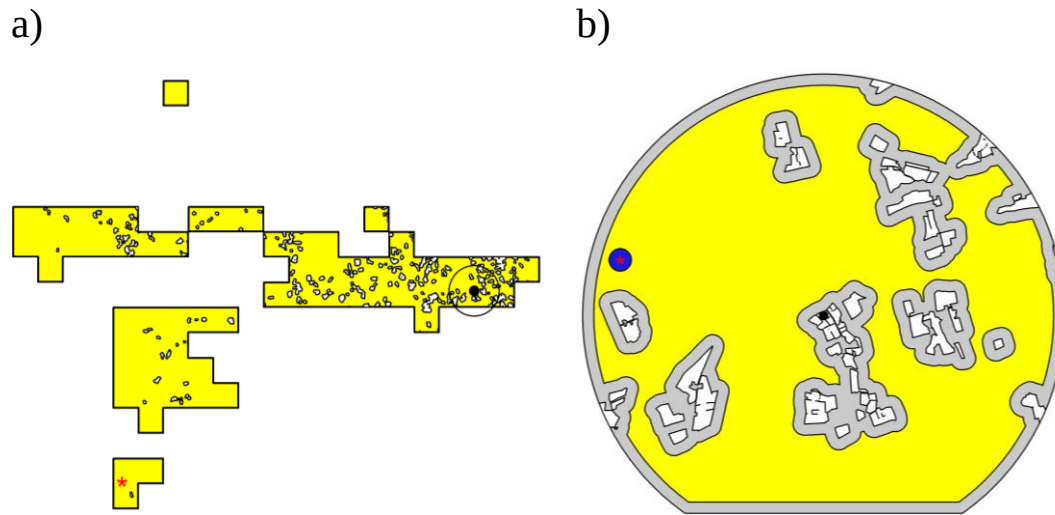


Figure 5.7: a) Obfuscated location (red star) and obfuscated area (yellow area) for field parcel number 6507 (black dot) belonging to cluster 24 with 1760 other field parcels (white shapes) using PESOM method. The region bounded by the threshold of 4000 m is illustrated by a black circle. b) Limited obfuscated area with a threshold of 4000 m (yellow area), the sensitive area (115 white polygons) extended sensitive polygon (dark grey area) and maximum obfuscated polygon (MOP) (blue circle around obfuscate polygon (red star)) are also represented in this diagram.

Although the PESOM is a polygon-based method, a modified point-based environmental similarity obfuscation methods (ESOM) can be used for phenomena best represented as a point such as spatial health data. Environmental and climate condition have effective influence on many phenomena and issues like Environmental determinants of disease. An extensive review of the interaction between COVID-19 incidence and both the natural and human environmental factors was conducted and highlighted the impact of meteorological factors such as humidity, wind speed, precipitation, and most important temperature on COVID-19 incidence. Air pollution was identified as a factor in increasing the risk of COVID-19 incidence, on the other hand, COVID-19 lockdown leads to the reduction of air pollution (Han, Yin, Wu, Wang, & Li, 2023). Therefore, it is important to consider underlying environmental and climate conditions of obfuscated and shared spatial data particularly agricultural spatial data to increase confidentiality and analytical accuracy.

5.5. Conclusion

The main contribution of this chapter is to develop the polygon-based obfuscation method with consideration of importance of contextual factors in agri-environmental research. Weather condition with vital impact on agricultural activities was considered as external features. PESOM was designed to generate obfuscated data with the same climate condition as original data. The PESOM method provides a high level of geoprivacy protection and no misclassified environmental cluster with relatively large displacement. However, further research is required to explore how combined PESOM with existing obfuscation methods or the threshold distance can improve method performance in terms of geoprivacy protection and statistical accuracy while maintaining absolute environmental and climatic clustering. A further study with more focus on the role of different data formats (i.e., continuous vs categorical) is recommended.

To showcase the reliability of obfuscated data to solve real-world problems and how the outcomes using obfuscated data are close to the outcome using original data a case study is presented in the next Chapter. The influence of internal and external features on application outcomes to identify the appropriate obfuscation methods for certain purposes is investigated.

Chapter 6 Incorporating contextual factors in agri-environmental research- options for obfuscation to support wider data application

*This chapter is under review as a peer-reviewed scientific article in
Irish Journal of Agricultural and Food Research journal*

6.1. Introduction

Sustainable agriculture is an essential component of the 2030 agenda of the UN Sustainable Development Goals (SDGs) (Bouma et al., 2019). The objective of sustainable agriculture is to optimize the use of natural and environmental resources to increase agriculture productivity to satisfy food security. Simultaneously, agriculture affects climate change in different ways including greenhouse gas (GHG) emissions, soil and water quality, and biodiversity with ~10 % of the European Union's (EU's) GHG derived from agricultural activities (EEA, 2023; McEldowney, 2020). GHG emissions have various sources such as methane (CH_4), nitrous oxide (N_2O) and carbon dioxide (CO_2). Methane is produced by livestock digestion processes and stored animal manure, N_2O is derived from organic and mineral nitrogen fertilisers while (CO_2) is associated with burning fossil fuel and land-use changes (Lynch et al., 2021). Although agricultural activities generate emissions, land use and management can help to reduce atmospheric CO_2 concentrations by increasing carbon storage in trees and carbon in soils (McEldowney, 2020; Xu et al., 2011). For example, grassland soils can hold higher soil organic carbon (SOC) than arable soils (Simo et al., 2019) the latter having higher carbon losses due to the mineralisation of carbon associated with the ploughing management regime. The carbon in soils is reported as more than three times that of atmospheric carbon and more than four times the aboveground biomass carbon, meaning soil carbon has significant potential to partially offset agricultural greenhouse gas emissions (Lal, 2004; Schulte et al., 2016). This reveals the importance of estimating the spatial and vertical SOC distribution to support sustainable land management and to negate the environmental impact of certain land use and management activities.

To mitigate agricultural GHG particularly CO_2 , producing more diverse crops, rotating crop types, and implementing a mixed land use such as agroforestry practice are suggested to increase carbon storage in trees and soils (McEldowney, 2020). Increasing carbon content of soil and carbon sequestration, improves soil fertility, crop productivity and mitigates soil erosion which can help to achieve sustainable agriculture by informed land management (McEldowney, 2020). Measuring SOC and monitoring the changes over time is a challenging issue. Direct measurement and sampling strategy to collect and assess the high spatial and vertical variability of SOC are time consuming and expensive (Smith et al., 2020).

Agri-environmental and climate data availability with machine learning approaches can be used to identify potential patterns and relationship between SOC and management and

environmental conditions such as soil type and climate conditions to estimate the SOC. Multi Linear Regression (MLR) and Random Forest (RF) models are widely used to predict the crop suitability, crop yield production and SOC based on environmental data including climate, soil, water, and fertilization data from various source such as soil sampling, weather station data, remote sensing and satellite data (Ganesan et al., 2021; Jeong et al., 2016; Sothe et al., 2022).

Achieving sustainable agriculture and sustainable land management requires an innovative system. Adopting digital farming can increase the agriculture productivity while preserving natural resources, optimize use of fertilizers and pesticides and reduce the impact of agricultural activities on environment (Ferris, 2017; Hrustek, 2020; Laborde & Piñeiro, 2018; Holloway et al., 2018). Advancement of digital farming generates vast amounts of agricultural data including geographical data. Performing agricultural spatial data analysis can enable researchers to gain insight into data, find spatial patterns, and interact between different types of phenomena. For example, the interaction of environmental condition on agriculture and vice versa, which consequently improve decision-making in agricultural systems.

However, despite its potential, geographical data can be used to identify the field parcels and disclosure of private information such as size of farm, type of crop, type and number of animals, type of fertilizer, type and number of equipment, financial information, which might be misused and used against interest of the data owner or used for interests of others than data owner. Therefore, geoprivacy protection is a significant concern when sharing agricultural data. Obfuscation is one of the predominant techniques designed and developed to alter the location data to reduce risk of disclosure while maintaining statistical accuracy and spatial patterns (Armstrong et al., 1999b; Swanlund et al., 2020; J. Wang & Kwan, 2020). The challenge is to identify the appropriate obfuscation method for a certain purpose and need while considering the trade-off between location confidentiality and spatial pattern preservation. Wang et al. (2022) provided a guideline to identify a proper obfuscation method from several popular obfuscation methods regarding a certain purpose of study, although further research is suggested to consider more evaluation metrics and context of the study area.

However, obfuscation alters the location which might change the locations distribution, topology, relationship between different features including internal and external features and clustering. Different obfuscation methods affect distinct aspects of spatial data. Although various evaluation metrics were applied to evaluate the obfuscated data in a certain aspect, the question is how obfuscated data is reliable to solve real-world problems and how the outcomes

of applications using obfuscated data is close to the outcomes using original data. The second issue is how the influence of internal and external features on application outcome can lead to identifying the appropriate obfuscation methods. Therefore, the objective of this case study is to show the usability of obfuscated data in addressing a real-world problem. For this purpose, estimation of SOC based on available internal and external features in Ireland was examined, first with the original data and then with the obfuscated data. The rationale behind this case study is such that if the spatial relationships can be replicated, then high risk data can be shared to support research purposes.

6.2. Materials and Methods

6.2.1. Study area and data

The National Soil Database (NSDB) is a national database of soil geochemistry for Ireland of major nutrients, major and trace elements (Fay et al., 2007). The project was conducted during 2003 to 2005, to complete the soil information geochemistry including data point maps and spatial distribution maps at the national scale (Fay et al., 2007). To measure SOC concentration for the near-surface soils (0-10cm) and other geochemistry features such as N, P and K the sampling strategy was adapted to collect 1015 samples across Ireland except south-eastern part which was already sampled in 1995. The samples were collected from predetermined sites from the national grid system with grid resolution of 10×10 km. Two samples were taken from each 100 km^2 grid, one from northeast and be another from centre of grid. SOC was predicted based on soil-related features measured from samples and climate features including precipitation, minimum and maximum temperature from December 2003 to November 2004. Climate data was obtained from de la Torre Cerro et al. (in press) at a pre-processed 5km resolution of Met Éireann Reanalysis (MÉRA) data. The NSDB point-based data can be used to demonstrate the effect of point-based obfuscation methods on SOC estimation. After pre-processing 931 points of NSDB data were extracted to use in this case study.

To estimate and predict SOC based on available data including soil-related features (internal) and climate features (external) three scenarios can be adopted. In scenario 1, just internal features (soil-related features) were considered. In scenario 2, the best model to predict the SOC, regard to the level of impact of internal or external features on model performance was identified. Scenario 3 can be considered to explore models just using external features to predict the SOC.

6.2.2. Obfuscation Methods (Rand and PESOM)

Anonymization and randomization are two types of techniques that are used widely to protect geoprivacy while preserving statistical analysis and spatial patterns. Anonymization exploits the k -anonymity concept that generates obfuscation area that contains at least $k - 1$ other locations to ensure the original location is not identifiable and reduce the risk of identification by $1/k$ (Gruteser & Grunwald, 2003). Randomization transfers the original location to a random or farthest/middle location among N random locations in a predefined

obfuscation area (Murad et al., 2014; Seidl et al., 2018; Wightman et al., 2011; Zandbergen, 2014). The predominant obfuscation methods designed and developed to date are geometric-based methods that rely on distance from the original location without considering the underlying environmental condition such as Rand method. Due to the influence of agriculture on the environment and conversely environment and climate conditions on agriculture, a polygon-based environmental similarity obfuscation method (PESOM) was developed to preserve geoprivacy while maintaining the environmental and climate condition of the original data (Nowbakht et al., under review). PESOM is a cluster-oriented method using machine learning clustering algorithms to cluster the study area based on external features as an obfuscation area where the obfuscated location would be a random location inside the obfuscation area. To cluster the study area based on external features it is important to consider the underlying patterns of external features, identifying more influence features, clustering algorithm and its parameters such as number of clusters. Prior to applying the clustering algorithm, principal component analysis (PCA) was used to avoid multicollinearity, overweighting certain features over other features and for dimensionality reduction. Nowbakht et al., (under review) has provided extensive discussion relating to feature selection and selecting number of clusters using k-means clustering.

To investigate the useability of obfuscated data generated by different obfuscation methods for a certain purpose, two obfuscation methods were selected in this case study, PESOM and Rand Method. Although, in general spatial agricultural data contain geographical locations and areal regions which require polygon-based obfuscation method, in this case study due to the fact that NSDB is a point spatial dataset, the PESOM method was modified to work with points.

Obfuscation methods change the geographic locations to preserve geoprivacy and reduce the risk of identification. As a result of this transformation, location distribution, spatial clustering and correlation between geographical attributes and other location-related features including external features such as environmental and climate features might be changed.

6.2.3. Machine Learning Methods (MLR and RF)

MLR is a statistical technique to predict the dependent features based on the linear relationship between the dependent feature and two or more independent or explanatory features. MLR provides the overall explanatory variation of the model and the level of

contribution of each independent feature on the total variance explained, it is easy to implement and interpret.

RF is a supervised machine learning technique used in both classification and regression problem. In RF, Bootstrapping technique is used to construct a large number of decision trees as a forest on randomly selected observations and features of training dataset to improve the prediction accuracy. The outcome is the majority vote of the forest for classification and average for regression.

6.3. Results

6.3.1. Original data

PCA was applied on seasonal climate features in study area for dimensional reduction and to avoid impact of multicollinearity. The top seven principal components (PCs) were selected for the further process.

SOC is reported to have high correlation with N and moderate to very weak correlation with pH, P and K respectively (Figure 6.1a). This result is to be expected as N additions to soil increase crop yield which can increase the plant residue returns to the soil and build SOC (Powlson et al., 2011). The results of the correlation analysis also show moderate correlation between SOC with two components of climate features PC2 and PC1 (-0.40 and 0.21) and very weak correlation with the rest (Figure 6.1b). These results suggest that multilinear regression model can be used to predict the SOC based on soil fertilisers and climate features.

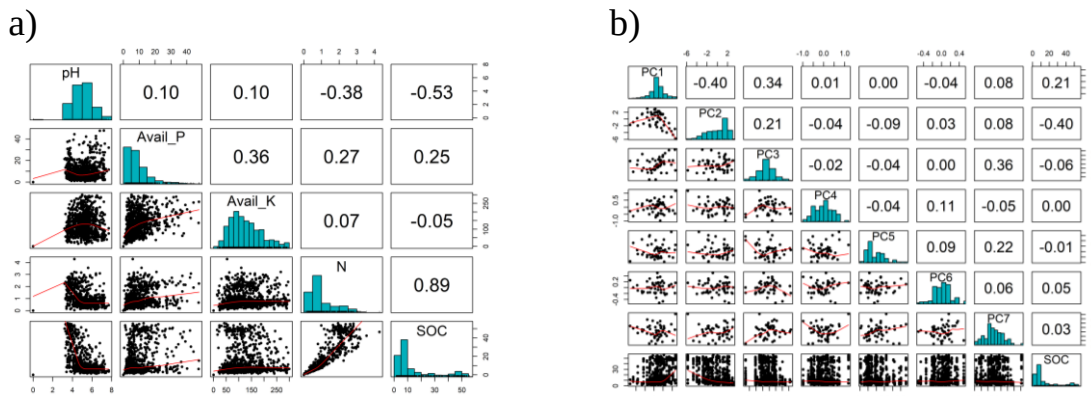


Figure 6.1: Correlation matrix between SOC from original NSDB dataset and a) soil related features including N, K, P and pH from original NSDB dataset. b) seven PCs climate data

6.3.2. Obfuscated data

Comparison of Figure 6.2a, b and Figure 6.1b show the correlation between SOC and climate features were not changed for PESOM obfuscated data. In contrast, the Rand Method changed the correlation between SOC and climate features. This indicates the prediction of SOC using Rand obfuscated data might not be as accurate as using PESOM obfuscated data.

Both methods demonstrate a high level of privacy protection, 90% of obfuscated locations satisfy 20-anonymity that means for 90% of obfuscated locations, the chance of identification is 1/20 indicating obfuscated location is in a region with 20 other locations. The

PESOM method offers a higher level of privacy protection for 75% of obfuscated locations which the risk of identification declines to 1/50 (the obfuscation location is in a region with 50 other locations). While the Rand method provides this level of protection just for 45% of obfuscated locations (Figure 6.3a).

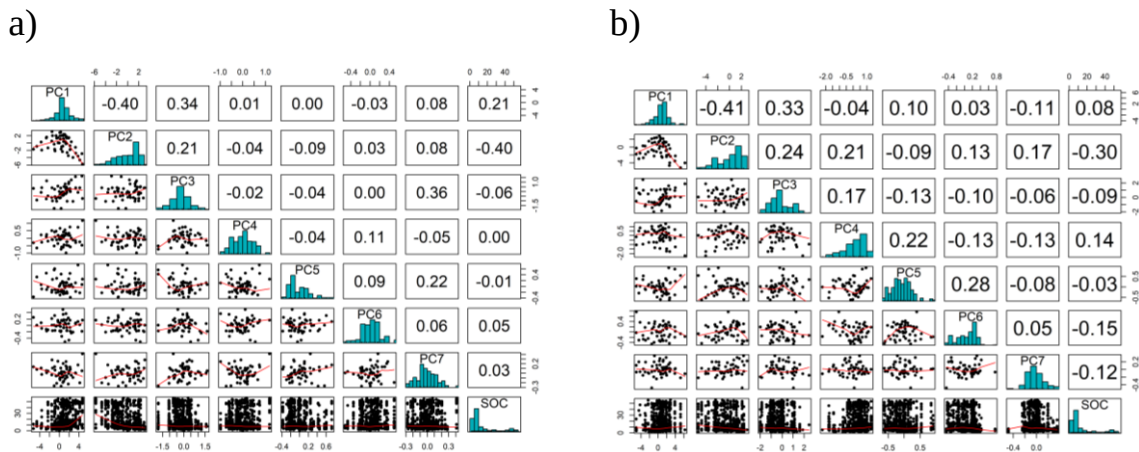


Figure 6.2: Correlation matrix between 7 PCs climate data and SOC from obfuscated NSDB data generated by a) PESOM method and b) Rand method.

Displacement refers to the obfuscated location distance transformation from the original location. The percentage of displacements for certain values is shown in Figure 6.3b indicates that all obfuscated locations generated by Rand methods have less than 50 km displacement while 60% of obfuscated points generated by PESOM displacement are more than 50 km and more than 70% of locations are transferred to an obfuscated location between 100km to 400km displacement. Larger location displacement reduces the statistical accuracy. While, in general the Rand method provides better performance in terms of statistical accuracy with a high level of privacy protection, further research is required to examine the suitability of obfuscated methods for certain purposes for example, predicting SOC.

PESOM method provides perfect environmental and climate cluster preservation with zero misclassified error (MCE), due to the nature of method that guarantees the obfuscated location is in the same environmental and climate condition as the original location. The MCE for Rand method based on the same climate clustering is reported to be 97% which means 97% of obfuscated locations are in different climatical condition from the original locations.

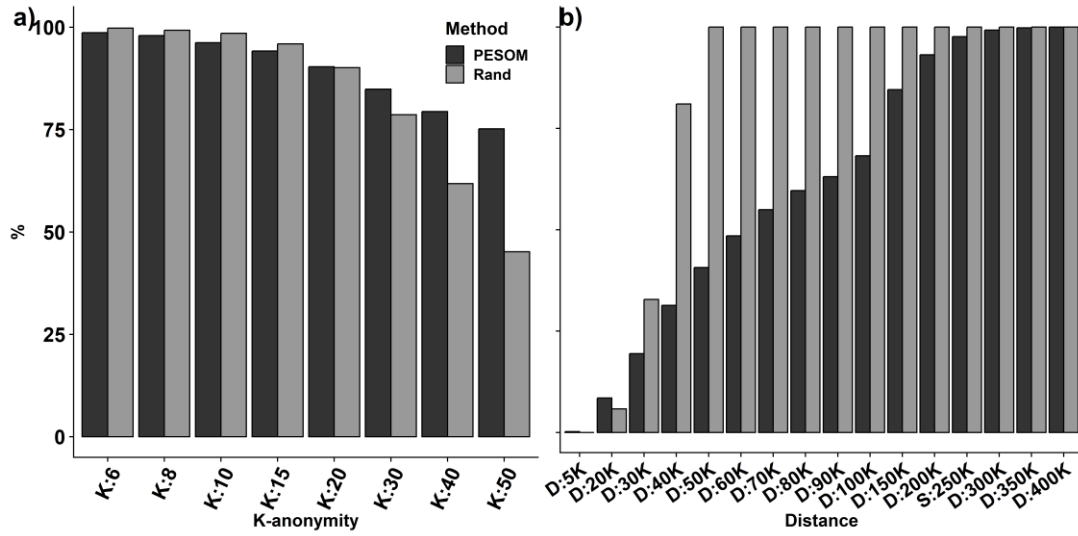


Figure 6.3: Comparison of performance of PESOM and Rand methods a) percentage of obfuscated location with displacement less than different values. b) percentage of obfuscated locations that satisfies k-anonymity for different values of k

6.3.3. Scenario 1: using internal features.

To estimate and predict SOC based on soil related features (internal features) and climate features (external) a stepwise selection method was used to construct the multiple linear regression model. The adjusted R-squared value of 0.85 shows a high level of explanatory of SOC. The PCs (linear combination of climate features) were not significantly associated with SOC ($p > 0.05$). The soil-related features were reported to be statistically significantly associated with SOC ($p \leq 0.05$). The Equation 6.1 can be used to estimate the SOC which for one percent increase in pH and K the SOC decrease by -4.4 and -0.03 respectively. There are 0.26 and 15.52 increase in SOC for one unit increase in P and N respectively.

$$\widehat{SOC} = 23.45 - 4.41 pH + 0.26 P - 0.03 k + 15.52 N \quad Eq\ 6.1$$

As this multiple linear regression model relies on soil-related features collected from sample data points that are internal features, the obfuscation process does not affect them. Therefore, estimation of SOC based on internal features is the same after obfuscation regardless of type of obfuscation method.

6.3.4. Scenario 2: using internal and external features.

Several Random Forest (RF) models were trained using a training dataset (a randomly selected subset of original NSDB dataset) to estimate SOC based on both internal and external features which explained more than 91% of variation of SOC. The RF3 model used six of the most important features including soil related and two PCs climate features reported the best results with 91.43% explained the variance of SOC (Table 6.1) with RMSE of 3.26 based on the test dataset that had not been included in model training. The descriptive statistics of the observed and predicted SOC of RF3 model for test dataset are highly similar. Although the minimum predicted value is higher and maximum of predicted value is lower than observed value, the mean is very close with lower standard deviation (Table 6.2). Comparison of the observed and predicted SOC of RF3 model and deviation from the 1:1 line (true model) indicates that for small value of SOC, less than 20 the predicted value match observed value well and close to the 1:1 line. For SOC, values between 20 to 40 are over-predicted and for larger values greater than 40 are under-predicted (Figure 6.4).

Table 6.1: Result of different random forest model for original NSDB data set.

Model	% Var explained	Root Mean squared residuals of train dataset	Root mean squared error (RMSE) for test dataset	features
RF1	91.13	4.42	3.34	PC1, PC2, PC3, PC4, PC5, PC6, PC7, N, pH, P, K
RF2	91.03	4.45	3.24	N, pH, P, K
RF3	91.43	4.34	3.26	N, pH, P, K, PC2, PC1

A t-test based upon a test dataset (subset of NSDB data) on NSDB measured SOC against RF predicted SOC shows no significant mean difference observed ($p = 0.44 > 0.05$) and mean difference is 0.18.

Table 6.2: Descriptive statistic for original test observed and RF3 model prediction of SOC.

	Test Data Set	
	Observe SOC	Predict SOC
Min	1.44	2.10
Max	55.8	49.62
Mean	13.7	13.88
SD	14.7	13.96
Sample size	186	186

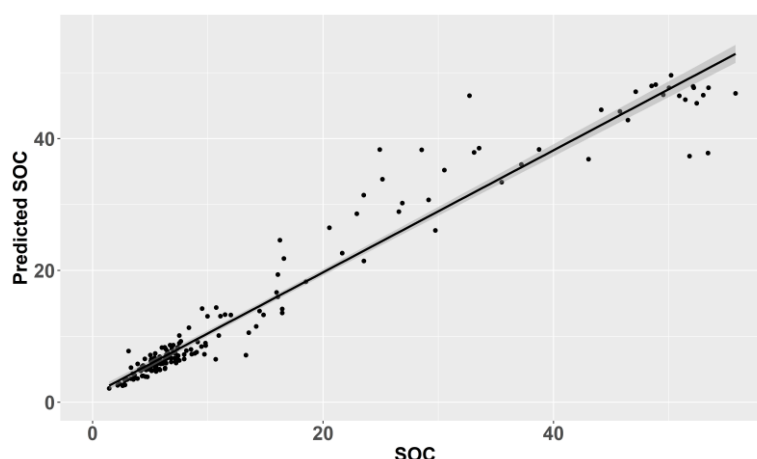


Figure 6.4: RF3 model performance for test dataset. Scatterplot of Observed value of SOC vs predicted value.

Table 6.3 shows the summary of the implementation of RF model on PESOM, and Rand obfuscated data for the same training dataset. Since the most important features to estimate the SOC are soil-related features which are internal features and climate features have little impact on the RF model performance, there is no substantial difference observed between RF model performance using PESOM obfuscated data and Rand obfuscated data.

The descriptive statistics of the predicted SOC from RF model using obfuscated data generated by PESOM is the same as the descriptive statistics of the predicted SOC using original dataset as expected. Some overestimation is observed when RF model implemented on obfuscation Rand method output (Table 6.4). Comparison of the predicted SOC of RF model using the original data and predicted SOC of RF model using PESOM and Rand obfuscated data and deviation from the 1:1 line indicate that although both obfuscated data generate similar predicted SOC to the original predicted SOC, the predicted SOC from PESOM obfuscated data perfectly fall in the 1:1 line compare to the Rand obfuscated data predicted SOC that highlights some minimal deviation (Figure 6.5).

An independent-sample t-test was conducted to compare the predicted SOC from original data with PESOM obfuscated data and Rand obfuscated data. There was no significant difference reported between predicted SOC from original data with either obfuscated predicted SOC (Table 6.5) which indicates that using obfuscated data instead of original data can predict SOC as accurately as predicted SOC from original data. The mean difference between predicted SOC from original data and PESOM obfuscated data (-0.02) is very small and smaller than the mean difference between predicted SOC from original data and Rand obfuscated data

(0.09). This shows that using PESOM obfuscated data produce better results than Rand obfuscated data to predict SOC which depends on climate condition.

Table 6.3: Comparison of RF model using internal (N, pH, P, K) and external features performance to predict SOC for original test dataset, obfuscated test dataset generated by PSEOM and Rand obfuscation methods.

Model/ data	% Var explained	Root Mean squared residuals of train dataset	Root mean squared error (RMSE) for test dataset	External Features
RF3(original)	91.43	4.34	3.26	PC2, PC1
RF(PEMOS)	91.42	4.34	3.28	PC2, PC1
RF(Rand)	91.29	4.38	3.44	PC2, PC4

Table 6.4: Comparison of descriptive statistic of original test observed and RF3 model prediction of SOC with model prediction SOC of obfuscated test data set from PSEOM and Rand obfuscation methods.

	Original test Data		Obfuscated test Data (PBOM)	Obfuscated test Data (Rand)
	Observe SOC	Predict SOC	Predict SOC	Predict SOC
Min	1.44	2.10	2.11	2.05
Max	55.8	49.62	49.92	48.29
Mean	13.7	13.88	13.90	13.8
SD	14.7	13.96	13.93	13.84
Sample size	186	186	186	186

Table 6.5: Independent sample t-test results for predicted SOC of original dataset with predicted SOC of PESOM obfuscated test dataset and Rand obfuscated test dataset.

Model/ data	t-value	Confidence interval	Mean difference	p-value
RF(PEMOS)	-0.99	(-0.06, 0.02)	-0.02	0.33
RF(Rand)	1.28	(-0.04 ,0.22)	0.086	0.20

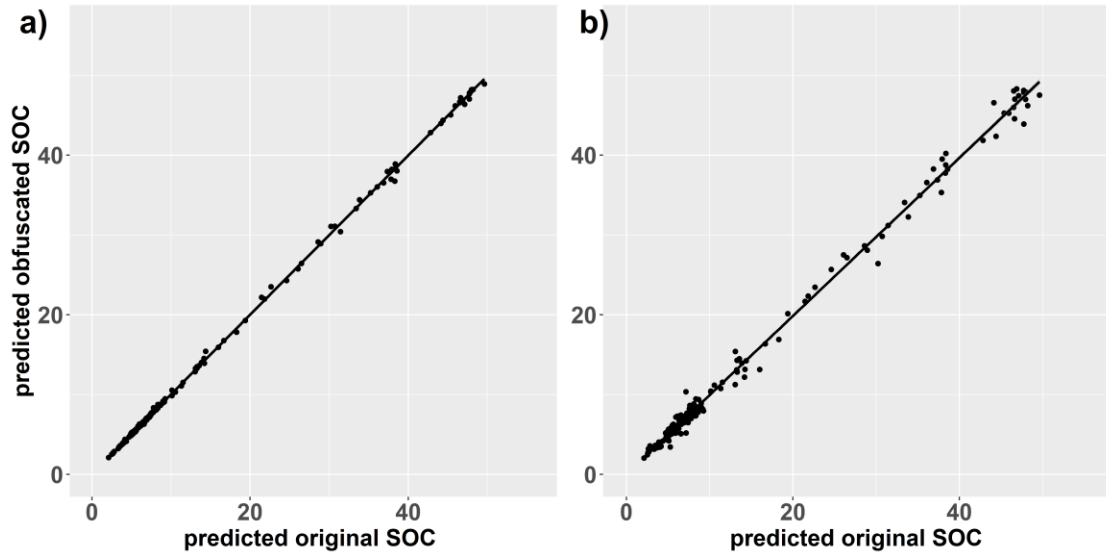


Figure 6.5: RF model performance. Scatterplot of predicted value of SOC using original test dataset vs predicted SOC value using a) PESOM obfuscated test dataset and b) Rand obfuscated test dataset.

6.3.5. Scenario 3: using external features.

The MLR model can be constructed just using climate data, Figure 6.1a shows the moderate to weak correlation between SOC and PC1 and PC2. The results of MLR model to estimate the SOC based on climate data indicates that adjusted R-squared value was 0.15 that shows a low level of explanatory power to estimate SOC. Just PC2 demonstrates statistically significant impact on SOC ($p < 0.05$) with f-statistic value of 131.8. the suggested LR model can be represented in Equation 6.2.

$$\widehat{SOC} = 14.8 - 2.9 PC2 \quad Eq\ 6.2$$

The PESOM obfuscation method preserves the correlation between SOC and climate features (external features) (Figure 6.2a and Figure 6.1b). Therefore, the MLR model using PESOM obfuscated data was the same as the MLR using original data and can be represented in Equation 6.2 too. The MLR model for Rand obfuscated data can be modelled in Equation 6.3.

$$\widehat{SOC} = 14 - 2.5 PC2 + 4.3 PC4 \quad Eq\ 6.3$$

The descriptive statistics of the predicted SOC from MLR model using original dataset and PESOM obfuscated data are exactly the same with over-prediction of minimum SOC to 5.86 and under prediction of maximum SOC to 31.43 (Table 6.6).

Table 6.6: Comparison of descriptive statistic of original test observed and MLR model prediction of SOC with model prediction SOC of obfuscated test data set from PSEOM and Rand obfuscation methods based on external features.

	Original test data		PBOM Obfuscated test data	Rand Obfuscated test data
	Observe SOC	Predicted SOC	Predict SOC	Predict SOC
Min	1.44	5.86	5.86	4.63
Max	55.8	31.43	31.43	27.16
Mean	13.7	14.85	14.85	14.46
SD	14.7	6.16	6.16	5.27
Sample size	186	186	186	186

Implementation of MLR model on Rand obfuscated data to predict SOC provide over-prediction and under-prediction the minimum and maximum SOC to 4.63 and 27.16 respectively (Table 6.6) and different to the descriptive statistic of the predicted SOC from original data. The summary of implementation of MLR model on PESOM, and Rand obfuscated data for the same training dataset is shown in Table 6.7. The results indicate that the MLR model on PESOM obfuscated data is exactly the same as the MLR model on original data and the MLR model on Rand obfuscated data reduce the level of explanatory of SOC to 12%. Although the MLR models capture a low level of variation of SOC (15% and 12%), since just external features were used to predict SOC, they demonstrate the impact of the obfuscation method on the outcome clearly.

Table 6.7: Comparison of MLR model using external features to predict SOC for original data, obfuscated data generated by PSEOM and Rand obfuscation methods.

Model/ data	Adjusted R-squared	F-statistic	Root mean squared error (RMSE) for test dataset	External Features
LR(original)	0.15	131.8	13.26	PC2
LR(PEMOS)	0.15	131.8	13.26	PC2
MLR(Rand)	0.12	51.47	13.12	PC2, PC4

The PESOM obfuscation methods provide perfectly obfuscated data to predict SOC based on external features as the predicted SOC is exactly the same as predicted SOC from

original data with almost no deviation from the 1:1 line (Figure 6.6a). The SOC prediction from original data and Rand obfuscated data are clearly different and a large deviation from 1:1 line confirms the difference (Figure 6.6b).

Results of independent-sample t-test confirms that the predicted SOC from original data and PESOM are significantly the same and no difference is observed (mean difference is zero). While the t-test results for predicted SOC from original data and Rand obfuscated data shows no significant difference ($p = 0.26 > 0.05$), the mean difference is reported to be 0.4 which indicates that the PESOM obfuscated data is a better option to be used for predicting SOC based on external features.

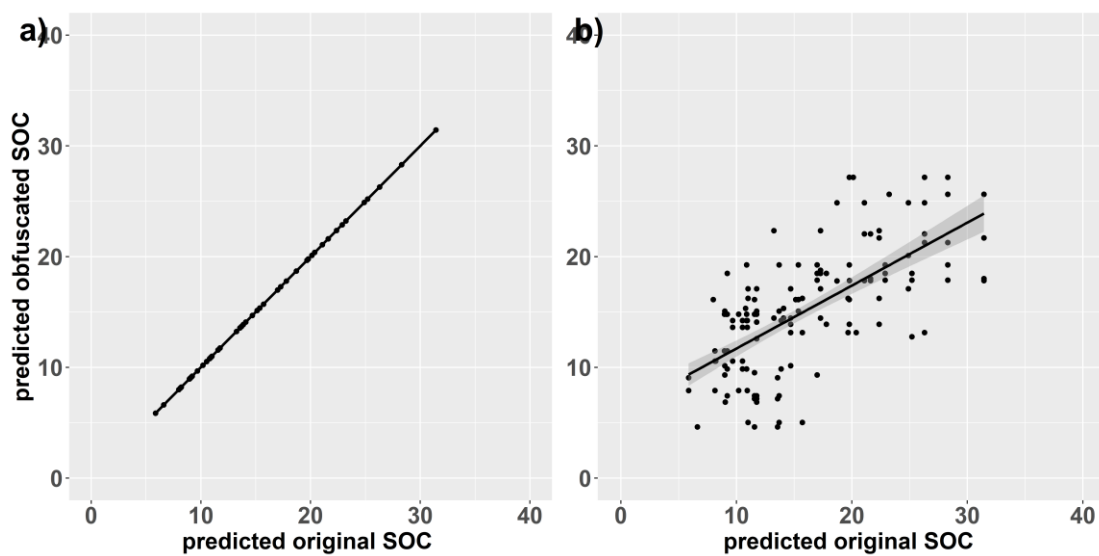


Figure 6.6: MLR model performance based on external features (climate features). Scatterplot of predicted value of SOC using original test dataset vs predicted SOC value using a) PESOM obfuscated test dataset and b) Rand obfuscated test dataset.

6.4. Discussion

The case study was designed to determine the effect of different generated obfuscated data to solve real world problem. In this case study two obfuscation methods were adopted to be implemented on NSDB data to predict SOC based on internal (soil-related) and external (climate) features and explore the importance of choice of obfuscation method.

Comparison of SOC estimation based on three scenarios indicate the importance of features selection on accuracy of statistical or machine learning models. According to this investigation, soil-related features which indicate the soil chemical properties with higher correlation with SOC are the most important features to predict SOC. As indicated above, this can be explained by biogeochemical processes that occur between elements in the soil, such as between carbon and nitrogen. While external features slightly impact the SOC estimation, the combination of internal and external features improve the results. The results are consistent with Li et al., (2020) findings which showed that soil chemical properties played a significant role in predicting SOC and climate features played relatively small roles.

The choice of predicting model is another factor to obtain more accurate results which highly depends on type of available data, needs, facilities and equipment, and further research is suggested to investigate the impact of different predicting model on accuracy of prediction. While the choice of features and predicting models are essential factors to achieve high accuracy, when the use of original data is not allowed another factor arises in how to choose the obfuscation methods to obtain an accurate estimation. The results of this case study indicate that the SOC estimation using PESOM obfuscated data for the same features and model provides more accurate estimation than Rand obfuscated data. It is important to note that estimation based on obfuscated data is more accurate when it is close to estimation based on original data, not the observed values. Therefore, the aim is to generate obfuscated data that can proceed the same as the original data that means if the accuracy of the model using the original data is low the suitable obfuscation method is one that provides approximately the same low accuracy, not that obfuscation method provides close estimation to real value and high accuracy.

Internal features were the most important factors to predict SOC which were not changed in the process of obfuscation. MLR model using internal features to predict SOC provides high level of accuracy with easy interpretation which can help different stakeholder to support farmers in relation to nutrient management to enhance profits while it can support

government to adopt informed agricultural policies to achieve sustainable land management. External features were reported to have slight impact on SOC. Even so, still the results indicate that the obfuscated data generated by a cluster-oriented method produced more significant SOC estimation than geometric-based obfuscation method to estimate SOC using original data.

However, the objective of study and problem to be solved, influence of internal and external features on research outcome, priority of statistical accuracy or privacy protection are components that should be considered when choosing the most appropriate obfuscation methods. For example, clustering preservation is the most important factor when environmental data have the major influence on statistical analysis which indicate that the cluster-oriented method would be the best choice or the density-based methods such as Density and AHilb methods (Nowbakht et al., 2022) are more appropriate methods when the distribution preservation is the main focus of study. It is essential to note that in this research climate data was used to cluster the study area and soil-related features was obtained from sampling and considered as internal features which do not change in the process of obfuscation. Thus, obfuscation data might not be consistent with the geochemical map. There is a possibility that obfuscated locations with soil-related values are located in regions with inappropriate soil types and geochemical conditions. Therefore, to generate obfuscated data that is more relevant to geochemical research, the clustering should be based on soil type, land cover or a combination of soil type and climate condition. This indicates the importance of the research objective in feature selection for the clustering of the study area.

6.5. Conclusion

The results of this case study show the importance of features selection, predicting models and choice of obfuscation methods when the use of original data is prohibited, to archive high accuracy and improve the model outcomes. While the combination of internal and external features improves the results, in the choice of obfuscation method, it is essential to consider the degree of these features' impact on the accuracy. In application or research with low or no external features impact, the obfuscation method with high geoprivacy preservation is the most appropriate obfuscation method. Hence, clarity on research priorities is essential to help determine the appropriate methodological approaches to obfuscate spatial data to achieve the optimal geoprivacy protection and spatial patterns and statistical accuracy.

Chapter 7 General Discussion

Geoprivacy protection is a significant concern when sharing agricultural data with geographical information. Sharing and access to agricultural data is essential to optimise use of these data and to integrate them with other data sources such as environmental and climate data in agri-environmental research to support sustainable agriculture and land management. To reduce the risk of identification and privacy violation, agricultural spatial data must be obfuscated prior to sharing. In Chapter 2, an overview of point-based obfuscation methods which are the most used techniques were presented. But to date, geoprivacy protection approaches related to agricultural spatial data remain unexplored despite their unique properties and vital privacy concerns associated with this data. Agriculture spatial data have unique characteristics that distinguish them from other spatial data. Agricultural spatial data is often static polygon data with an association of attributes/properties of a field parcel or farm that includes coordinates, shape, size, topology, the relationship with the surrounding environment and the impact of external factors on region characteristics. However, most obfuscation methods developed to date do not consider a) the areal nature of field-level data and b) the environment at the original and obfuscated sites. In this dissertation, several key research questions were identified related to agriculture data geoprivacy protection which are outlined here.

7.1. Summary

- 1) Are existing point-based obfuscation methods suitable for protecting the geoprivacy of agricultural data with polygon nature? In other words, is there any significant limitation to applying the point-based obfuscation method on polygon nature data?

In Chapter 3, a comparison of 27 modified and combined point-based obfuscation methods (where a polygon centroid is considered as a representative point) was carried out using different evaluation metrics to measure the effectiveness of the obfuscation methods to reduce the risk of disclosure while providing a high level of statistical accuracy and spatial pattern preservation. The term “non-unique obfuscation” was defined for point spatial data when an obfuscated point associated with two or more original points. Then non-unique obfuscation was redefined for polygon spatial data with consideration of the polygon nature and properties of such spatial data when there are intersections between two or more obfuscated locations. Findings reported (Chapter 3, section 3.6) a high chance of non-unique obfuscation occurrence when point-based obfuscation methods were applied on centroid of polygon spatial data

particularly for high density spatial data with large polygons. The definition of false identification which refers to when the obfuscated point mistakenly links to other original points (McLafferty, 2004; Seidl et al., 2018) was generalized for polygon spatial data when there are intersections between the original and obfuscated polygons. However, the results showed there is very little chance of false-identification and non-unique obfuscation for point spatial data occurring, particularly for low density location distribution data (Chapter 3, section 3.6). In contrast, there was a high percentage of false-identification and non-unique obfuscation reported when point-based obfuscation methods were used to obfuscate polygon nature spatial data which indicates a lack of the suitability of these methods to be used for obfuscating polygon spatial data (Chapter 3). Therefore, false-identification and non-unique obfuscation was identified as important factors to be considered in the obfuscation process, particularly for static polygon spatial data in terms of statistical accuracy and privacy protection.

A high percentage of false-identification and non-unique obfuscation could reduce the number of locations in obfuscated data or change the spatial distribution, and local and national statistical results which affect data accuracy. Furthermore, geoprivacy protection might be violated by identifying an object (field parcel) even with incorrect information which might be misused against the owner. The limitations of point-based obfuscation methods when applied on the centroid of polygons determined the need for specific obfuscation methods to deal with the polygon nature of agricultural data, which led into the next research question which asks:

- 2) In the design and development of polygon-based obfuscation methods to achieve a high level of geoprivacy protection and statistical accuracy, which properties of polygon data are crucial to be considered?

In Chapter 4, five novel adaptive polygon-based obfuscation methods were designed and developed to overcome the limitation of point-based obfuscation methods when applied on centroid of polygon as a point spatial data. For this purpose, it was suggested to eliminate the extended sensitive area (buffer zone with radius of the original polygon's centroid to its farthest edge around a union of original locations and existing obfuscated polygons) from the primary obfuscated area generated by any point-based obfuscation methods, as an obfuscated area and then construct the obfuscated polygon. Furthermore, to obtain a high level of confidentiality, in polygon-based obfuscation methods the shape and size of the polygon were modified to reduce the risk of identification via shape or size of obfuscated location. The results showed

that the use of these methods guaranteed that the false-identification and non-unique obfuscation did not occur.

These five polygon-based obfuscation methods that rely on the distance from the main location without considering the underlying environmental conditions. Interaction between environmental and climate conditions with agricultural activities indicates the importance of use of combined agri-environmental data in data-driven approaches to achieve sustainable agriculture and mitigate climate change. Since spatial data are part of agri-environmental and climatic data and have an effective impact on the results of data-driven models, for some cases, it is essential to consider the preservation of environmental factors while obfuscating agricultural spatial data. The importance of contextual factors such as environmental and climatic conditions in agri-environmental research raised the subsequent question:

- 3) How well the proposed geometric-based obfuscation methods preserve these contextual factors?

In response to this research question, in Chapter 5, the PESOM polygon-based obfuscation method which is a context-aware and cluster-oriented method was developed to achieve absolute environmental and climate condition preservation. Clustering a study area based on contextual factors to generate obfuscation area and construct the obfuscated location is the primary purpose of this method. Choice of clustering algorithm highly depend on available and structure of data and purpose of research which can assist in choosing the right features.

The choice of obfuscation method is a challenge for data administrators or data providers. Although there are various evaluation metrics that can help to determine the suitable obfuscation method for a certain purpose, still the reliability of obfuscated data to solve real-world problems is a concern. Therefore, the question that was raised was:

- 4) Whether the research outcome using obfuscated data is the same or close to the research outcome when the original data is used?

To address the final research question a case study was conducted to showcase the useability of obfuscated data to solve real world problems and highlight the importance of the choice of obfuscation method in outcome and suitability of obfuscated data for a certain purpose (Chapter 6). Estimation of SOC using internal and external features using different data-driven modelling was explored. The importance of feature selection and data-driven modelling to

obtain significant SOC estimation was identified which then led to choosing the most proper obfuscation method for a certain purpose.

7.2. Contribution

After reviewing various point-based obfuscation methods in Chapter 2, several methods were selected in this research to examine the ability of point-based obfuscation method to deal with the polygon spatial data (Chapter 3). For the first time in literature the term non-unique obfuscation was introduced, and importance of false-identification and non-unique obfuscation was identified particularly for static polygon spatial data.

The high percentage of false-identification and non-unique obfuscation implies both vulnerable geoprivacy protection and low statistical accuracy was identified as the main drawbacks of using point-based obfuscation methods on polygon spatial data. Therefore, to enhance the performance of obfuscation process in both aspects of privacy and statistical accuracy for polygon data which shows the important need for new mechanisms led to the design and development of polygon-based methods with the assurance of preventing the occurrence of false-identification and non-unique obfuscation. Five novel polygon-based obfuscation methods including PN*Rand, PDonut-k, PDensity-k, PAHilb, and PDonut_AHilb methods were developed and implemented in Chapter 4 with consideration of polygon spatial data properties. In addition to avoiding these two factors, PESOM method which is the contextual polygon-based obfuscation method was developed to avoid contextual dissimilarity as well (Chapter 5).

To identify the best trade-off between confidentiality and spatial pattern preservation various obfuscation methods with different parameters should be examined which often is a tedious and time-consuming process. Donut method is one of the predominant approaches widely used in different disciplines. To overcome the exploring impact of different inner and outer radii in the donut method's outcome, qualitative approaches, including Donut-k and Density-k methods, were introduced to identify the optimal values of radii based on a combination of the Donut method and k-anonymity satisfaction, and subsequently, optimal k-anonymity determination to minimize the risk of identification and maximize spatial pattern preservation by generating the optimal obfuscation area with the highest level of k-anonymity (Chapter 3 section 3.4 and Chapter 4 section 4.3).

According to Chapter 3 and 4, obfuscation methods either point-based or polygon-based that provide smaller location distribution and correlation alteration, less MCE, global and local

divergence, and a higher percentage of lower points displacement perform better in terms of statistical accuracy and spatial pattern preservation, particularly in local and regional scales, for example, the density-based methods can be useful for local and regional research purpose. A high percentage of locations that satisfy higher K-anonymity indicates higher geoprivacy protection.

The results of Chapters 3 and 4 found that the density-based methods demonstrating the best performance in terms of spatial pattern preservation and geoprivacy protection. However, the performance of density-based methods highly depends on how the size of each obfuscation area for each individual location is obtained. The Density-k introduced in Chapter 3 (and polygon-based in Chapter 4) with generating optimal obfuscation size associated with optimal level of k-anonymity satisfaction was identified to provide a best trade-off between privacy protection and statistical accuracy.

Maintaining the geographical context within obfuscation methods is the central for some applications, particularly where geographic context is relevant to research outcomes. Two types of clustering preservation were examined in this research, spatial clustering preservation based on spatial coordinates of available data (Chapters 3 and 4) and clustering preservation based on external features (Chapters 5 and 6). Spatial clustering preservation examines the ability of obfuscation methods to preserve the correlation between spatial coordinates and internal features. The density-based methods such as Density-k and AHilb with lower MCE and less displacement demonstrate less correlation alteration than NRand methods, therefore providing higher spatial clustering preservation. Clustering preservation based on external features is an essential consideration for agri-environmental applications so environmental similarity obfuscation methods such as PESOM were developed with specific consideration of this factor. Absolute contextual clustering preservation, extremely high level of confidentiality, and no false-identification and non-unique obfuscation are evidence of the suitability of PESOM obfuscated data for agri-environmental and climate change research with the potential to share with a wide range of researchers and even for open access. This would increase the scope and capacity to address complex multidisciplinary research, reuse of data for different purposes and disciplines and minimized the recollection of such data (Tenopir et al., 2011). For example, the Central Statistical Office (CSO) can use this data to improve statistical analysis in various sections such as agri-environmental, ecosystem accounts-grasslands and croplands, climate-plant phenology, and more sections. On the other hand, PESOM method provides high percentage of large displacement and large TCA, which shows the large correlation alteration

between spatial coordinates and internal features that means in this case PESOM neglects spatial clustering preservation. Hence it is an inappropriate method for local and regional purposes. This indicates the importance of examining the impact of internal or external features on application outcomes and consequently choosing the appropriate obfuscation method.

In Chapter 6, section 6.3. three different scenarios were presented to illustrate this point. The results show that when the application outcome just depends on internal features, the outcome based on original data is the same as outcome after obfuscation, regardless of the type of obfuscation methods therefore, a higher level of geoprivacy protection can be considered. When the influence of external features is high, contextual cluster preservation is an important factor to be considered, hence contextual-based obfuscation methods such as PESOM are potential obfuscation methods that can be used.

In application or research with low or no external features impact, the obfuscation method with high geoprivacy preservation is the most appropriate obfuscation method. Hence, clarity on research priorities is essential to help determine the appropriate methodological approaches to obfuscate spatial data to achieve the optimal geoprivacy protection and spatial patterns and statistical accuracy.

Obfuscation can be real-time or offline processes, depending on the user's needs and the purpose of applications, for example, location-based services (LBS) applications and cellular phones require real-time obfuscation methods. The obfuscation approaches that were examined or designed and developed in this research are offline processes in which the data administrator or data provider's obligation is to provide obfuscated data prior to sharing data with users. Therefore, the data administrator's challenge is to explore different obfuscation methods with different parameters to find the most appropriate obfuscation method that complies with the need, purpose of research, influence of internal and external features on research outcomes, and level of data vulnerability which is the fourth aim of this research outlined in Chapter 1. However, a comparison of these obfuscated methods using various evaluation metrics as undertaken within this research (Chapters 3, 4 and 5) provides guidance which summarised in Figure 7.1 on how to choose the proper obfuscation method and. For example, spatial distribution preservation and low Gdi which indicates the preservation of location patterns globally are key factors for national or global studies. When internal features are the main interest of study TCA might be something to check. False-identification and non-unique

obfuscation are important for static spatial polygon data. Contextual clustering preservation is essential for study with a focus on external features.

The guideline represented in Figure 7.1 is based on the *NMP Online* dataset and specific distribution and characteristics. The guideline would have been more relevant and accurate if a wide range of different spatial data with different distribution and characteristic properties had been explored.

In data-driven modelling with spatial data in addition to the choice of features and predicting models, the choice of obfuscation methods is a crucial factor to achieve high accuracy. The results of the case study while pointing to the importance of the choice of features and predicting models to achieve a high level of prediction accuracy indicate the impact of the obfuscation method in prediction. The estimation of SOC with obfuscated data based on internal and external features using different predicting models was compared with the estimation of SOC using original data and the same features and predicting models, not the actual values of SOC. In view of the fact that the aim of this comparison was to obtain obfuscated data that can proceed the same as original data means that if the accuracy of the model using original data is low, the suitable obfuscation method is one that provides approximately the same low accuracy, not that obfuscation method provides close estimation to real value and high accuracy. The SOC estimation using PESOM obfuscated data for the same features in the same predicting model offers precise estimation to the SOC estimation using original data than Rand obfuscated data with the same condition. Therefore, the level of influence of external features on prediction outcome is an important factor to choose between contextual cluster-oriented methods and geometric-based obfuscation methods. This addresses the final research question in Chapter 6 which can help data administrators or data providers in the choice of proper obfuscation for a certain purpose, and address the final research question in Chapter 6. This could help to overcome the limitation of sharing data with inside organizations, and external research organizations or even make data publicly available to achieve the data-sharing benefits mentioned (Tenopir et al., 2011).

Computational time is another factor that should be considered in the choice of obfuscation method. In general polygon-based obfuscation methods involve more computation process and subsequently longer computational time than point-based obfuscation methods. N*Rand and Donut methods by themselves are not slow but the process to find the best parameters to obtain a more accurate trade-off between privacy protection and statistical

accuracy might take a long time as the method should be run for different parameters. This process is manual, and the choice of parameters needs some investigation about location distribution and underlying spatial patterns. Donut-k and Density-k are also time consuming process as they need to run several times to find inner and outer radii for different k-anonymity and finally find the optimal value of k and subsequently the size of donut shape which takes time but this is a computational process and generate the best possible trade-off between accuracy and privacy protection. These methods might take less time to complete than N*Rand and Donut method. AHilb and Donut-AHilb methods are slow processes, particularly polygon-based PAHilb and PDonut-AHilb methods. Although to reduce the processing time, in PAHilb and PDounal_Ahilb methods, the algorithms guarantee that each cell just visits once, still these methods are slow. Further research is recommended to explore different data structures such as B⁺-tree (a tree data structure where each tree node stores the number of leaf nodes in each of its subtrees) (Ghinita et al., 2007b) and Chord which is a ring structure (Stoica et al., 2001) to maintain the Hilbert sequence in order to improve the computational time.

Finally, taken altogether results indicate the improvement in both geoprivacy and statistical analysis aspects and provide a baseline for evaluation and comparison of different obfuscation methods which have general relevance for the wider GIScience community. The results using agricultural data in this research provide more confidence and reduce the concerns of data owners and administrators with regard to data confidentiality to make this data accessible for a wide range of researchers to improve agricultural sustainability for example, nutrient management planning and more informed decision making which is one of the most important factors to achieve sustainable agriculture and development.

The proposed polygon-based obfuscation methods in this research have the potential to be used in different disciplines with different characteristics and different spatial data types. Although, this needs more investigation to identify the advantages and disadvantages of these methods regards to level of geoprivacy protection, pattern preservation, distribution preservation, statistical accuracy, and computational processing time. Figure 7.1 can be used in identifying the most appropriate candidate methods for specific purposes. with consideration of further investigation to apply some modifications and adaptations to make them compatible with the specific situation.

PN*Rand	• Higher confidentiality level as the radius increases. • Lower analytical accuracy, as the radius increases. • Time consuming due to examining different value of radiuses.
PDonut-k	• Higher confidentiality level than PN*Rand PDensity-k • Different confidentiality level for different locations • Lower level of regional analytical accuracy than PDensity-k • Lower distribution and pattern preservation than PDensity-k • Time consuming due to finding optimal k-anonymity and corresponding inner and outer radii
PDensity-k density-based	• Fixed geoprivacy protection for all location • High spatial clustering preservation • High level of accuracy both in regional and national scale • High distribution and pattern preservation • Best trade-off between confidentiality and accuracy • Time consuming due to finding optimal k-anonymity and corresponding inner and outer radii
PAHilb density-based	• High spatial clustering preservation • High level of accuracy both in regional and national scale • High distribution and pattern preservation • Time consuming due to finding optimal k-anonymity and using Hilbert sequence. Slower than PDensity method.
PDonut_AHilb density-based	• High spatial clustering preservation • High level of analytical accuracy both in regional and national scale • High distribution and pattern preservation • Best trade-off between confidentiality and accuracy • Time consuming due to finding optimal k-anonymity and using Hilbert sequence. Slower than PDensity method.
PESOM	• Absolute contextual clustering preservation • Low spatial clustering preservation. • Very high confidentiality level • Low distribution and pattern preservation. • Not suitable for regional research • Faster than PDensity-k and PAHilb methods

Figure 7.1: Polygon-based obfuscation methods descriptions

7.3. Research Constraints

The main constraint encountered in this study was getting access to the required agricultural data. To comply with GDPR and the EU Code of Conduct on Agricultural Data Sharing mentioned in Chapter 1 and consistent with the main aim of this thesis which is data privacy protection we had to complete a time consuming process to obtain data access permission. This includes the following:

- 1) Data sharing agreement between data provider (Teagasc) and University College Cork (UCC)
- 2) Data Protection Impact Assessment (DPIA)
- 3) Ethical approval assessed by Social Research Ethics Committee (SREC)

Due to the unavailability of labelled data, the choice of machine learning models was limited to the unsupervised learning algorithm in both spatial clustering preservation addressed in Chapter 3 and contextual clustering preservation in Chapters 5 and

The lack of labelled data and dependent features in the given dataset makes it difficult to design the case study to show the use and level of confidence of using obfuscated data. Hence, for the case study presented in Chapter 6 the NSDB point spatial data was used.

7.4. Recommendation for further work

Privacy and particularly geoprivacy protection is a serious concern in today's modern world where the use of high-tech and data-driven technology is unavoidable. In this research, several geoprivacy techniques and evaluation metrics focused on agriculture data which contain polygon spatial data with specific properties were designed and developed to achieve the trade-off between statistical accuracy and geoprivacy preservation. To support a transition towards sustainable agriculture, obfuscated data transformations provide a pathway to make this data accessible to a wide range of organisations and stakeholders to boost the usability of potentially available agriculture data. However, it is necessary to address some limitations within this study and a number of possible directions for further investigation as follows:

- 1) Map visualization is an effective tool to clearly recognize underlying geographical patterns within different layers and gain knowledge. However, in polygon-based obfuscation methods, it is challenging to construct obfuscated polygons regarding shape, size, topology, and the relationship between the polygon (field parcel) with the surrounding area such as other field parcels, roads, rivers, urban areas, etc. In addition to considering these factors that make obfuscated locations unidentifiable through their appearance and decrease the probability of geoprivacy violation via topology and relationship with other features, improvement of map visualization is another important factor that should be considered in obfuscation development. In this research, the obfuscated polygon was constructed by a circle with radius equal to the distance of polygon's centre to its random edge. Randomly selecting a polygon shape from a predefined set of irregular polygons is another possible solution to construct the obfuscated polygon. This approach was implemented to generate mock data in Chapter 4. To generate more realistic map visualizations the random polygon generation technique can be used to construct arbitrary polygon shapes. However, a further study could assess how these randomly generated polygons fill the space and how can map topology could be improved.

In the proposed obfuscation methods in this research, the obfuscation process was implemented individually to reduce the risk of identification via topology, background knowledge, and the relationship between obfuscated polygons with adjacent polygons. These methods guarantee that no false-identification and non-unique obfuscation occurred which incurs no duplication and overlap between obfuscated polygons to obey the rules of no duplication and overlap of topology specified. To satisfy the rule of no gap between

adjacent polygons of topology further research is required to overcome this limitation. An alternative obfuscation process to satisfy the rule of no gap between adjacent polygons of topology can be group obfuscation, which means obfuscating, randomly selecting a set of adjacent polygons. Graph structure (using Graph theory) can be adapted to maintain the relationship between this group of adjacent polygons and then randomly changes the adjacent relationship in the graph structure to construct the topology of a set of adjacent obfuscated polygons. To develop and establish this idea and examine its efficiency in maintaining data confidentiality and statistical accuracy further research is recommended. Therefore, to obtain more realistic map visualization and topology between features while minimizing risk of disclosure of sensitive information via geographical data, further research is required.

- 2) Large displacement and subsequently low level of local and regional statistical accuracy, and a large TCA, which shows the large correlation alteration between spatial coordinates and internal features are the main drawbacks of the PESOM method due to generating unintegrated obfuscated areas which might contain several regions with long distances. Thus, randomly selected locations, as the obfuscated location might be in the furthest region. An integration of PESOM and existing obfuscation methods particularly density-based methods or determining the threshold distance could be used to reduce the percentage of large displacement and enhance the performance of PESOM method in terms of local statistical accuracy with perfect contextual clustering preservation and subsequently reduce the TCA. Moreover, to overcome this drawback and avoid having large and spread obfuscated areas, the nearest integrated obfuscation areas by considering a certain level of k-anonymity satisfaction can be considered as an obfuscated area. However, further research is recommended to explore the impact of these suggestions on outcomes and how to determine the parameters such as optimal level of k-anonymity, threshold distance, or density-based methods parameters.

The interaction between internal and external features might have an impact on performance of obfuscation methods in terms of both spatial and contextual clustering preservation. A further study with more focus on interaction and relationship between these features is therefore suggested. Furthermore, the performance of PESOM in terms of spatial clustering preservation, local statistical accuracy, and spatial pattern preservation is dependent on clustering parameters (type and the number of features and the number of clusters) and further research is recommended to examine different clustering, and feature

selection methodologies, and more case studies beyond SOC. Therefore, further investigation recommended in development of context-based algorithm to preserve geoprivacy, contextual correlation between agricultural data and external environmental data, and spatial clustering.

- 3) The proposed polygon-based obfuscation methods in this research are based on the avoidance of obfuscated locations being in a sensitive area. This sensitive area could be a union of all existing polygons in the dataset and obfuscated polygons or some region with irrelevant and unrealistic conditions. Spatial scale is an important factor that significantly impacts spatial analysis (Ghimire et al., 2010; Holloway & Miller, 2015; Miller & Hanham, 2011). Thus, spatial scale can influence the generated sensitive area and subsequently impact the generated obfuscated data. Using a much finer detailed study area map might generate different sensitive area than coarse study area map. For example, small lake or river might be in sensitive area using fine scale map but not be in sensitive area using coarse map. Further investigation is recommended to explore the effect of spatial scale on sensitive area and obfuscated data in terms of statistical accuracy, regional and national pattern preservation, and privacy protection.

Appendices

Appendix 1

Supplementary Information chapter 3: Pseudocode for obfuscation methods

Algorithm SI3.1: N*Rand methods, return obfuscation point of each point from dataset

```
1. input (Obf_area, pnt, N , method)    // for any point (pnt) in point-dataset , N: number of
2.                                     random point, obf_area was generated by Rand, Donut,
3.                                     Density, pinwheel, AHilb ad Donut-AHilb method
4. output(obf_point)
5. if method == Rand then
6.   Obf_point <- random point inside obf_area
7. else then
8.   NRand_point <- N random points inside obf_area
9.   If method == NRand then
10.    Obf-point <- farthest point of NRand_point from pnt
11.   If method == NMix then
12.    Obf-point <- middle point of NRand_point from pnt
13. Endif
```

AlgorithmSI3.2: Rand, Donut-k and Density-k method, return obfuscation area around each point of dataset

```
5. input (area, pnt, method, R)    // for each point (pnt) of point-dataset , R: radius threshold
6. output(obf_area)
7. if method == Rand then
8.   obf_area <- intersection of area and a buffer of radius R around point's coordinates
9. if method == Donut then
10.  obf-area <- intersection of area and donut shape with radiuses max(rmin) and max(rmax)
11.  calculated by find_raduis for all points
12. if method == Density then
13.  obf-area <- intersection of area and donut shape with radiuses rmin and rmax calculated
14.  by find_raduis
15. Return(onf_area)
```

AlgorithmSI3.3: gen_pinwheel, return pinwheel

```
1. input (alpha, th, R)    // for each alpha between [0°, 360°], th: pinwheel wings width,
2.                        //R: radius threshold
3. output (pinwheel shape)
4. return ((alpha mod th) /th * R)
```

AlgorithmSI3.4: Pinwheel method, return pinwheel shape obfuscation area

```
16. input (area, pnt, method, R)    // for each point (pnt) of point-dataset , R: radius threshold
17. output(obf_area)
18. pinw <- empty set
19. for each i in [0°, 360°] do
20.   p <- gen_pinwheel (i, th, R)
21.   calculate the respective Cartesian coordinates and add to pinw set
22. end for
23. if method == Rand then
24.   obf_area <- intersection of area and pinw set
25. if method == Donut then
26.   obf_area <- intersection of area and donut shape with radiuses max(rmin) and max(rmax)
27.   calculated by find_raduis for all points and pinw set
28. if method == Density then
29.   obf_area <- intersection of area and donut shape with radiuses rmin and rmax calculated by
        find_raduis and pinw set
30. return (obf_area)
```

AlgorithmSI3.5: AHilb_grid: return a data frame with HID, UID and number of point in each cell

```
// first step or offline step in AHilb method
31. input (area, level, point_dataset)    // level: resolution of grid and Hilbert curve
32. output(data frame)
33. H <- initialize a Hilbert curve with level resolution
34. grid <- rasterize the point-dataset with the same resolution of H
35. grid.H <- Create a new data structure contains ind, Hvalue and num
36.   // ind: cell index of grid, Hvalue: Hilbert curve order from H associated with ind,
37.   // num: number of points in cell associated with ind
38. return (grid.H)
```

AlgorithmSI3.6: AHilb method: return an obfuscation area for a point pnt

```
// second step or online step in AHilb method
39. input (pnt, grid.H, res, k)    // for each point (pnt) , grid.H: a data created by AHilb_grid,
40.                                // grid resolution, k: level of k-anonymity
41. output(obf_area)
42. next_ind <- find the index of cell that point located
43. total <- grid.H$df.Num[next_ind]    // number of point inside a cell which that point located
44. obf_area <- add this cell[next_ind] to obf_area
45. Nind <- {ind -1, ind +1, ind - res, ind + res}    // cell's neighbours indexes
46. while (total <= k) do
47.     // first consider the cell's four neighbours left, right, down and up with considering the grid
48.     // boundaries
49.     for (ind in Nind and total <= k) do
50.         total <- total + grid.H$df.Num[ind]
51.         obf_area[ <- add this cell[ind] to obf_area
52.     end for
53.     // follow by Hilbert value order
54.     // find the index of cell with following Hilbert value
55.     next_ind <- which(grid.H$df.Hvalue == grid.H$df.Hvalue[next_ind]+1)
56.     if (H$df.Num[next_ind] > 0 and total <= k) then
57.         total <- total + grid.H$df.Num[next_ind]
58.         obf_area[ <- add this cell[next_ind] to obf_area
59.     end if
60. end while
61. return (obf_area)
```

AlgorithmSI3.7: Donut_AHilb method: return an obfuscation area for a point pnt

```
// second step or online step in AHilb method
62. input (pnt, grid.H, res, k1, k2)    // for each point (pnt), grid.H: a data created by AHilb_grid,
63.                                // res: grid resolution, k1 and k2: level of k-anonymity
64. output(obf_area)
65. obf_area1 <- AHilb(pnt, gid.H, res, k1)
66. obf_area2 <- AHilb(pnt, grid.H, res, k2)
67. obf_area <- difference between obf_area2 and obf_area1
68. return (obf_area)
```

Appendix 2

Supplementary Information Chapter 4:

Pseudocode for Polygon-based PAHilb obfuscation methods

Algorithm SI4.1: PAHilb method: return an obfuscation polygon

```

1. input (input_data,k,H,pnt, existing_obf_pnt)    //pnt: point in input_data, H: a grided data contains Hilbert
   //value created by AHilb_grid, k: level of k-anonymity, existing_obf_pnt: existing obfuscated centroid
2. output(obf_poly)                               // obfuscated polygon
3. pre_visit_cell <- c()                           // the cell already visited
4. org_poly <- readWKB(plot_shape[pnt])           // read original polygon by its centroid
5. rMOP <- farthest_edge                           //distance of centroid of a polygon and its farthest edge
6. identify current_cell                           // cell that the contains the centroid of org_poly
7. aobf_area <- current_cell                        //at start point-based obfuscation area is the cell contains centroid of polygon
8. sensitive <- org_poly                           //at start sensitive area is the original polygon
   //return point-based obfuscation area and sensitive area related to the current cell and its 8 adjacent cells
   // might contain part of polygon that their centroid is inside the current cell, number of polygons and list
   // of visited cells
9. call PAHilb_adjacent
10. extended_sensitive <- buffer(sensitive, rMOP)   //buffering the sensitive area by rMOP
   // Generate polygon-based obfuscation area by removing extended sensitive area from point-based
   // obfuscation area
11. poly_obf_area <- difference (obf_area, extended_sensitive)
12. while (k-anonymity is not satisfied) do        // use poly_num to check k-anonymity satisfaction
   // consider the cell's four neighbours left, right, down and up with considering the cell boundaries
   // and k-anonymity satisfaction
13.   for (current cell's four neighbours) do
14.     call PAHilb_adjacent
15.     extended_sensitive <- buffer(sensitive, rMOP) //buffering the sensitive area by rMOP
16.     // Generate polygon-based obfuscation area by removing extended sensitive area from point-based
     // obfuscation area
17.     poly_obf_area <- difference (obf_area, extended_sensitive)
18.   endif
19. end for
   // follow by Hilbert value order
20. Current_cell <- cell[next Hilbert value order]
21. if ( current_cell is not in pre_visit_cell & k-anonymity is not satisfied) then
   call PAHilb_adjacent
   extended_sensitive <- buffer(sensitive, rMOP)    //buffering the sensitive area by rMOP
22.   poly_obf_area <- difference (obf_area, extended_sensitive)
23. end if
24. end while
25. obf_point <- random point inside poly_obf_area
   //generate MOP of obfuscated point with radius of its rMOP
26. MOP <- buffer(existing_obf_pnt, rMOP)
27. sensitive <- add this MOP to sensitive area
28. end for
29. Obf_poly <- Generate obf_polygon by using polygon generation algorithm
30. return(obf_polygon)

```

Algorithm SI4.2: *PAHilb_adjacent*: return point-based obfuscation area and sensitive area related to current cell and its 8 adjacent cells

```
1.  Input (H, pnt, obf_area, sensitive, num_poly, pre_visit_cell , existing_obf_pnt, current_cell)
    //pnt: point in input_data, H: a grided data contains Hilbert value created by AHilb_grid,
    // obf_area: previous point-based obfuscation area, sensitive: previous sensitive area,
    //num_poly: previous number of polygons, pre_visit_cell: cells that previously was considered,
    //existing_obf_pnt: existing obfuscated centroid
2.  Output (obf_area, sensitive, num_poly, pre_visit_cell)    //new obfuscation area, sensitive area, number of
                                                                //polygons and pre_visit_cell

3.  for (each polygon inside current cell) do
4.      sensitive <- Add this polygon to sensitive area
5.      num_poly <- num_poly + 1                                //increased number of polygons
6.      while (part of polygon is inside the one of its 8 adjacent cells with considering the cell boundaries) do
7.          if (adjacent cell is not in pre_visit_cell) do
8.              Pre_visit_cell <- add adjacent cell to pre_visit_cell    // to reduce the computational time
9.              Obf_area <- Add this adjacent cell to obfuscation area
10.             for (each polygon in the adjacent cell) do
11.                 Sensitive <- Add each new polygon to sensitive area
12.                 num_poly <- num_poly + 1                            //increased number of polygons
13.             endfor
14.         endif
15.     while end
16. end for
17. return(obf_area, sensitive, num_poly, pre_visit_cell)
```

Appendix 3

Data is not available due to the privacy restrictions of the Nutrient Management Plan Online. Mock data and codes (R) to implement the analysis are provided to demonstrate how the codes work. Instructions on how to use mock data with the codes are provided. Results are presented from this data to ensure reproducible research where possible.

Please see electronic link <https://zenodo.org/record/6496256>

References:

- Akrami, N., Ziarati, K., & Dev, S. (2022). Graph-based local climate classification in Iran. *International Journal of Climatology*, 42(3), 1337–1353. <https://doi.org/10.1002/joc.7306>
- Ardagna, C. A., Cremonini, M., Damiani, E., De Capitani Di Vimercati, S., & Samarati, P. (2002). *Location Privacy Protection Through Obfuscation-Based Techniques*. <https://pdfs.semanticscholar.org/a016/888695fab5f979c8499c92aca643782f3152.pdf>
- Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999a). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5), 497–525. [https://doi.org/10.1002/\(sici\)1097-0258\(19990315\)18:5<497::aid-sim45>3.0.co;2-%23](https://doi.org/10.1002/(sici)1097-0258(19990315)18:5<497::aid-sim45>3.0.co;2-%23)
- Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999b). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5), 497–525. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990315\)18:5<497](https://doi.org/10.1002/(SICI)1097-0258(19990315)18:5<497)
- Auer, T., & Held, M. (1996). RPG-Heuristics for the Generation of Random Polygons. *Proc. 8th Canada Conf. Comput. Geom. Ottawa, ...*, 1–12. <http://ukpmc.ac.uk/abstract/CIT/154185%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.5609&rep=rep1&type=pdf>
- Bethere, L., Sennikovs, J., & Bethers, U. (2017). Climate indices for the Baltic states from principal component analysis. *Earth System Dynamics*, 8(4), 951–962. <https://doi.org/10.5194/esd-8-951-2017>
- Bora, G. C., Nowatzki, J. F., & Roberts, D. C. (2012). Energy savings by adopting precision agriculture in rural USA. *Energy, Sustainability and Society*, 2(1), 1–5. <https://doi.org/10.1186/2192-0567-2-22>
- Bouma, J., Montanarella, L., & Evanylo, G. (2019). The challenge for the soil science community to contribute to the implementation of the UN Sustainable Development Goals. *Soil Use and Management*, 35(4), 538–546. <https://doi.org/10.1111/sum.12518>
- Breinholt, G., & Schierz, C. (1998). Algorithm 781: Generating Hilbert's Space-Filling Curve by Recursion. *ACM Transactions on Mathematical Software*, 24(2), 184–189. <https://doi.org/10.1145/290200.290219>
- Brum-Bastos, V., Łoś, M., Long, J. A., Nelson, T., & Demšar, U. (2021). Context-aware movement analysis in ecology: a systematic review. In *International Journal of Geographical Information Science*. <https://doi.org/10.1080/13658816.2021.1962528>
- Carlis, J., & Bruso, K. (2012). RSQRT: An heuristic for estimating the number of clusters to report. *Electronic Commerce Research and Applications*, 11(2), 152–158. <https://doi.org/10.1016/j.elerap.2011.12.006>
- Cassa, C. A., Grannis, S. J., Overhage, J. M., & Mandl, K. D. (2006). Application of Information Technology j A Context-sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection. *J Am Med Inform Assoc*, 13, 160–165. <https://doi.org/10.1197/jamia.M1920>
- Claridge, J., Diggle, P., McCann, C. M., Mulcahy, G., Flynn, R., McNair, J., Strain, S.,

- Welsh, M., Baylis, M., & Williams, D. J. L. (2012). *Fasciola hepatica* is associated with the failure to detect bovine tuberculosis in dairy cattle. *Nature Communications*, 3(May), 853–858. <https://doi.org/10.1038/ncomms1840>
- Colizzi, L., Caivano, D., Ardito, C., Desolda, G., Castrignanò, A., Matera, M., Khosla, R., Moshou, D., Hou, K. M., Pinet, F., Chanet, J. P., Hui, G., & Shi, H. (2020). Introduction to agricultural IoT. In *Agricultural Internet of Things and Decision Support for Precision Smart Farming*. <https://doi.org/10.1016/B978-0-12-818373-1.00001-9>
- Curtis, A., Mills, J. W., Agustin, L., & Cockburn, M. (2011). Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems*, 35(1), 57–64. <https://doi.org/10.1016/j.compenvurbsys.2010.08.002>
- Dailey, D., & Whitfield, D. (2008). Constructing random polygons. *SIGITE'08: Proceedings of the 9th ACM SIG-Information Technology Education Conference. Cincinnati OH USA, January 2008*, 119–124. <https://doi.org/10.1145/1414558.1414592>
- Damiani, M. L., Bertino, E., & Silvestri, C. (2009). Protecting location privacy against spatial inferences: The PROBE approach. *SPRINGL 2009 - Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS, January 2009*, 32–41. <https://doi.org/10.1145/1667502.1667511>
- Dong, Q., Chen, X., Dong, S., & Zhang, J. (2021). Classification of pavement climatic regions through unsupervised and supervised machine learnings. *Journal of Infrastructure Preservation and Resilience*, 2(1). <https://doi.org/10.1186/s43065-021-00020-7>
- Duckham, M., & Kulik, L. (2005). *A Formal Model of Obfuscation and Negotiation for Location Privacy*. <http://www.geosensor.net/papers/duckham05.PERVASIVE.pdf>
- EEA. (2023). *EEA greenhouse gases - data viewer*. European Environment Agency. <https://www.eea.europa.eu/data-and-maps/data/data-viewers/greenhouse-gases-viewer>
- Elkhodr, M., Shahrestani, S., & Cheung, H. (2014). *A Semantic Obfuscation Technique for the Internet of Things*. 448–453.
- FAO, IFAD, UNICEF, WFP, & WHO. (2022). *The State of Food Security and Nutrition in the World 2022. Repurposing food and agricultural policies to make healthy diets more affordable*. <https://doi.org/10.4060/cc0639en>
- Fay, D., Mcgrath, D., Zhang, C., Carrigg, C., Flaherty, V. O., Kramers, G., Carton, O. T., & Grennan, E. (2007). *NATIONAL SOILS DATABASE END OF PROJECT REPORT RMIS 5192 Authors Johnstown Castle Research Centre Wexford July 2007*. 1, 1–24. <http://www.teagasc.ie/research/reports/environment/5192/eopr-5192.pdf>
- Ferris, J. L. (2017). Data Privacy and Protection in the Agriculture Industry: Is Federal Regulation Necessary? *Minnesota Journal of Law, Science & Technology*, 18(1), 307–342. <https://scholarship.law.umn.edu/mjlst/vol18/iss1/6>
- Fielke, S. J., Garrard, R., Jakku, E., Fleming, A., Wiseman, L., & Taylor, B. M. (2019). Conceptualising the DAIS: Implications of the ‘Digitalisation of Agricultural Innovation Systems’ on technology and policy at multiple levels. In *NJAS - Wageningen Journal of Life Sciences* (Vols. 90–91). <https://doi.org/10.1016/j.njas.2019.04.002>
- Fujimori, S., Hasegawa, T., Krey, V., Riahi, K., Bertram, C., Bodirsky, B. L., Bosetti, V., Callen, J., Després, J., Doelman, J., Drouet, L., Emmerling, J., Frank, S., Fricko, O.,

- Havlik, P., Humpenöder, F., Koopman, J. F. L., van Meijl, H., Ochi, Y., ... van Vuuren, D. (2019). A multi-model assessment of food security implications of climate change mitigation. *Nature Sustainability*, 2(5), 386–396. <https://doi.org/10.1038/s41893-019-0286-2>
- Ganesan, M., Andavar, S., & Raj, R. S. P. (2021). Prediction of Land Suitability for Crop Cultivation Using Classification Techniques. *Brazilian Archives of Biology and Technology*, 64. <https://doi.org/10.1590/1678-4324-2021200483>
- Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), 189–206. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>
- Ghimire, B., Rogan, J., & Miller, J. (2010). *Contextual land-cover classification_ incorporating spatial dependence in land-cover classification _ Enhanced Reader.pdf*.
- Ghinita, G., Kalnis, P., & Khoshgozaran, A. (2008). Private Queries In Location Based Services Anonymizers Are Not Necessary Categories And Subject Descriptors. *Proc. of the 2008 ACM SIGMOD ...*, 8(ii), 1–12. <https://doi.org/10.1145/1376616.1376631>
- Ghinita, G., Kalnis, P., & Skiadopoulos, S. (2007a). MOBIHIDE: A mobile peer-to-peer system for anonymous location-based queries. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4605 LNCS(August), 221–238. https://doi.org/10.1007/978-3-540-73540-3_13
- Ghinita, G., Kalnis, P., & Skiadopoulos, S. (2007b). PRIVE: Anonymous location-based queries in distributed mobile systems. *16th International World Wide Web Conference, WWW2007, June 2014*, 371–380. <https://doi.org/10.1145/1242572.1242623>
- Ghinita, G., Zhao, K., Papadias, D., & Kalnis, P. (2010). A reciprocal framework for spatial K-anonymity. *Information Systems*, 35(3), 299–314. <https://doi.org/10.1016/j.is.2009.10.001>
- Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Climate Smart Agriculture: A Synthesis of Empirical Evidence of Food Security and Mitigation Benefits from Improved Cropland Management. *Science*, 327, 812–818. <http://www.ncbi.nlm.nih.gov/pubmed/20110467>
- Gruteser, M., & 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*, 31–42. <https://doi.org/10.1145/1066116.1189037>
- Hamfelt, A., Karlsson, M., Thierfelder, T., & Valkovsky, V. (2011). Beyond K-means: Clusters identification for GIS. In *Popovich, V., Claramunt, C., Devogele, T., Schrenk, M., Korolenko, K. (eds) Information Fusion and Geographic Information Systems. Lecture Notes in Geoinformation and Cartography()*, vol 5. Springer, Berlin, Heidelberg. (Issue June 2014). <https://doi.org/10.1007/978-3-642-19766-6-8>
- Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, 172(9), 1062–1069. <https://doi.org/10.1093/aje/kwq248>

- Holloway, P., Kudenko, D., & Bell, J. R. (2018). Dynamic selection of environmental variables to improve the prediction of aphid phenology: A machine learning approach. *Ecological Indicators*, 88, 512–521. <https://doi.org/10.1016/j.ecolind.2017.10.032>
- Holloway, P., & Miller, J. A. (2015). Exploring spatial scale, autocorrelation and nonstationarity of bird species richness patterns. *ISPRS International Journal of Geo-Information*, 4(2), 783–798. <https://doi.org/10.3390/IJGI4020783>
- Houfah-Khoufah, W., & Touya, G. (2021). Geographically Masking Addresses to Study COVID-19 Clusters. *Cartography and Geographic Information Science*. <https://doi.org/10.1080/15230406.2021.1977709>
- Hrustek, L. (2020). Sustainability Driven by Agriculture through Digital Transformation. *Sustainability*, 12(20), 8596. <https://doi.org/10.3390/su12208596>
- Hughes, G. P. (1968). Comments “on the Mean Accuracy of Statistical Pattern Recognizers.” *IEEE Transactions on Information Theory*, IT-15(3), 420–423. <https://doi.org/10.1109/TIT.1969.1054314>
- Hwang, R. H., Hsueh, Y. L., & Chung, H. W. (2014). A novel time-obfuscated algorithm for trajectory privacy protection. *IEEE Transactions on Services Computing*, 7(2), 126–139. <https://doi.org/10.1109/TSC.2013.55>
- Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, 11(6), 1–15. <https://doi.org/10.1371/journal.pone.0156571>
- Jouanjean, M.-A., Casalini, F., Wiseman, L., & Gray, E. (2020). Issues around data governance in the digital transformation of agriculture: The farmers’ perspective. *OECD Food, Agriculture and Fisheries Papers*, 146. <http://dx.doi.org/10.1787/53ecf2ab-en>
- Kachore, V. A., Lakshmi, J., & Nandy, S. K. (2015). Location Obfuscation for Location Data Privacy. *Proceedings - 2015 IEEE World Congress on Services, SERVICES 2015*, 213–220. <https://doi.org/10.1109/SERVICES.2015.39>
- Kalnis, P., Ghinita, G., Mouratidis, K., & Papadias, D. (2007). Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering*, 19(12), 1719–1733. <https://doi.org/10.1109/TKDE.2007.190662>
- Kern, M., & Holz, R. (2013). Anonymity: A Formalization of Privacy, l-Diversity Michael. *Seminar Future Internet SS2013 Lehrstuhl, Technische Universität München, August*, 17–23. <https://doi.org/10.2313/NET-2013-08-1>
- Khoshgozaran, A., Shahabi, C., & Shirani-Mehr, H. (2011). Location privacy: Going beyond K-anonymity, cloaking and anonymizers. *Knowledge and Information Systems*, 26(3), 435–465. <https://doi.org/10.1007/s10115-010-0286-z>
- Kim, J., & Kwan, M.-P. (2021). An Examination of People’s Privacy Concerns, Perceptions of Social Benefits, and Acceptance of COVID-19 Mitigation Measures That Harness Location Information: A Comparative Study of the U.S. and South Korea. *ISPRS International Journal of Geo-Information*, 10(1), 25. <https://doi.org/10.3390/ijgi10010025>
- Kounadi, O., Lampoltshammer, T. J., Leitner, M., & Heistracher, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic*

- Information Science*, 40(2), 140–153. <https://doi.org/10.1080/15230406.2013.777138>
- Kounadi, O., & Leitner, M. (2015). Spatial Information Divergence: Using Global and Local Indices to Compare Geographical Masks Applied to Crime Data. *Transactions in GIS*. <https://doi.org/10.1111/tgis.12125>
- Kounadi, O., & Resch, B. (2018). *A Geoprivacy by Design Guideline for Research Campaigns That Use Participatory Sensing Data*. <https://doi.org/10.1177/1556264618759877>
- Kwan, M.-P., Casas, I., & Schmitz, B. (2004). Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2), 15–28. <https://doi.org/10.3138/x204-4223-57mk-8273>
- Kwan, M.-P., & Schwanen, T. (2018). Context and Uncertainty in Geography and GIScience: Advances in Theory, Method, and Practice. In *Annals of the American Association of Geographers* (Vol. 108, Issue 6, pp. 1473–1475). <https://doi.org/10.1080/24694452.2018.1457910>
- Laborde, D., & Piñeiro, V. (2018). Monitoring agricultural productivity for sustainable production and R&D planning. *Economics*, 12(1), 1–14. <https://doi.org/10.5018/economics-ejournal.ja.2018-66>
- Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, 304(5677), 1623–1627. <https://doi.org/10.1126/science.1097396>
- Leitner, M., & Curtis, A. (2004). Cartographic Guidelines for Geographically Masking the Locations of Confidential Point Data. *Cartographic Perspectives*, 49, 22–39. <https://doi.org/10.14714/cp49.439>
- Li, J., Nie, M., Powell, J. R., Bissett, A., & Pendall, E. (2020). Soil physico-chemical properties are critical for predicting carbon storage and nutrient availability across Australia. *Environmental Research Letters*, 15(9). <https://doi.org/10.1088/1748-9326/ab9f7e>
- Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016). Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling and Software*, 84, 494–504. <https://doi.org/10.1016/j.envsoft.2016.07.017>
- Long, J., Robertson, C., & Nelson, T. (2018). Stampr: Spatial-temporal analysis of moving polygons in R. *Journal of Statistical Software*, 84(April). <https://doi.org/10.18637/jss.v084.c01>
- Longley, P., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic Information Science and Systems*. In *John Wiley & Sons*. <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118676955.html>
- Lovell, M., & Foy, M. A. (2018). General Data Protection Regulation May 2018 (GDPR). *Bone & Joint* 360, 7(4), 41–42. <https://doi.org/10.1302/2048-0105.74.360622>
- Lu, Y., Yorke, C., & Zhan, F. B. (2012). Considering Risk Locations When Defining Perturbation Zones for Geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47(3), 168–178. <https://doi.org/10.3138/carto.47.3.1112>

- Lynch, J., Cain, M., Frame, D., & Pierrehumbert, R. (2021). Agriculture's Contribution to Climate Change and Role in Mitigation Is Distinct From Predominantly Fossil CO₂-Emitting Sectors. *Frontiers in Sustainable Food Systems*, 4. <https://doi.org/10.3389/fsufs.2020.518039>
- Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). ℓ -Diversity: Privacy beyond k-anonymity. *Proceedings - International Conference on Data Engineering*, 2006, 24. <https://doi.org/10.1109/ICDE.2006.1>
- Maciąg, P. S. (2017). A survey on data mining methods for clustering complex spatiotemporal data. *Communications in Computer and Information Science*, 716, 115–126. https://doi.org/10.1007/978-3-319-58274-0_10
- Mafie, G. K. (2022). The Impact of Climate Change on Agricultural Productivity in Tanzania. In *International Economic Journal* (Vol. 36, Issue 1). <https://doi.org/10.1080/10168737.2021.2010229>
- Manley, G. (2010). Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto International*, 25(6), 443–452. <https://doi.org/10.1080/10106049.2010.496496>.
- Manuel, P., & Masseno, D. (2021). *EU Code of Conduct on Agricultural Data Sharing by Contractual Arrangement*. November. https://fefac.eu/wp-content/uploads/2020/07/eu_code_of_conduct_on_agricultural_data_sharing-1.pdf
- Mao, Y., Liu, Y., Zhuo, L., Wang, W., Li, M., Feng, B., & Wu, P. (2021). Quantitative evaluation of spatial scale effects on regional water footprint in crop production. *Resources, Conservation and Recycling*, 173(May), 105709. <https://doi.org/10.1016/j.resconrec.2021.105709>
- Martin, M. E., & Schuurman, N. (2020). *Social Media Big Data Acquisition and Analysis for Qualitative GIScience: Challenges and Opportunities*. <https://doi.org/10.1080/24694452.2019.1696664>
- McEldowney, J. (2020). EU agricultural policy and climate change. *European Parliamentary Research Service*, December 2019. [https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651922/EPRS_BRI\(2020\)651922_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651922/EPRS_BRI(2020)651922_EN.pdf)
- McLafferty, S. (2004). The Socialization of GIS. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2), 51–53. <https://doi.org/10.3138/F333-6V74-815U-4631>
- Mechiche-Alami, A., & Abdi, A. M. (2020). Agricultural productivity in relation to climate and cropland management in West Africa. *Scientific Reports*, 10(1), 3393. <https://doi.org/10.1038/s41598-020-59943-y>
- Merlos, F. A., & Hijmans, R. J. (2020). The scale dependency of spatial crop species diversity and its relation to temporal diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 117(42), 26176–26182. <https://doi.org/10.1073/pnas.2011702117>
- Miller, J. A., & Hanham, R. Q. (2011). Spatial nonstationarity and the scale of species-environment relationships in the Mojave Desert, California, USA. In *International Journal of Geographical Information Science* (Vol. 25, Issue 3, pp. 423–438).

<https://doi.org/10.1080/13658816.2010.518147>

- Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Comptes Rendus - Mécanique*, 347(11), 845–855. <https://doi.org/10.1016/j.crme.2019.11.009>
- Montazeri, Z., Houmansadr, A., & Pishro-Nik, H. (2016). Defining perfect location privacy using anonymization. *2016 50th Annual Conference on Information Systems and Sciences, CISS 2016*, 204–209. <https://doi.org/10.1109/CISS.2016.7460502>
- Moosmann, L., SIEMONS, A., FALLASCH, F., SCHNEIDER, L., URRUTIA, C., WISSNER, N., & OPPELT, D. (2021). The COP26 Climate Change Conference. Status of climate negotiations and issues at stake. *Study for the Committee on the Environment, Public Health and Food Safety, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, Luxembourg, 2021, October*, 1–114.
- Murad, A., Hilton, B., Horan, T., & Tangenberg, J. (2014). Protecting Patient Geo-Privacy Via A Triangular Displacement Geo-Masking Method. *GeoPrivacy '14*. <https://doi.org/https://doi.org/10.1145/2675682.2676399>
- Murphy, P., Crosson, P., O'Brien, D., & Schulte, R. P. (2013). The Carbon Navigator: a decision support tool to reduce greenhouse gas emissions from livestock production systems. *Animal : An International Journal of Animal Bioscience*, 7 Suppl 2, 427–436. <https://doi.org/10.1017/S1751731113000906>
- Murphy, P., Lalor, S. T. J., Mehan, S., Plunkett, M., & 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>.
- Nainggolan, R., Perangin-Angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, 1361(1). <https://doi.org/10.1088/1742-6596/1361/1/012015>
- Nowbakht, P., O'Sullivan, L., Cawkwell, F., Wall, D. P., & Holloway, P. (2022). A comparison of obfuscation methods used for privacy protection: Exploring the challenges of polygon data in agricultural research. *Transactions in GIS*, 26(2), 949–979. <https://doi.org/10.1111/tgis.12892>
- Nowbakht, P., O'Sullivan, L., Wall, D. P., & Holloway, P. (2023). Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data. *Transactions in GIS*, 27(Feb 2023), 84–104. <https://doi.org/10.1111/tgis.13009>
- O'Mahony, J., & Cerro, T. (2021). Modelling the Distribution of the Red Macroalgae *Asparagopsis* to Support Sustainable Aquaculture Development. *AgriEngineering*, 251–265. <https://doi.org/10.3390/agriengineering3020017>
- Öman, S. (2004). Implementing data protection in law. *Scandinavian Studies in Law*, 389.
- Ord, J. K., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 27(4), 286–306. <https://doi.org/10.1111/j.1538-4632.1995.tb00912.x>
- Palmisano, A. (2018). Poisson and Binomial Distribution. *The Encyclopedia of*

- Archaeological Sciences*, June, 1–4. <https://doi.org/10.1002/9781119188230.saseas0467>
- Parameswaran, R., & D. Blough. (2005). A robust data obfuscation approach for privacy preservation of clustered data. In *Workshop Proceedings of the 2005 IEEE International Conference on Data Mining*, 18–25.
<https://blough.ece.gatech.edu/research/papers/psdm05.pdf>
- Peacock, J. A. (1983). *Two-dimensional goodness-of-fit testing in astronomy*.
<https://academic.oup.com/mnras/article/202/3/615/967854>
- Peng, W., Ma, N. L., Zhang, D., Zhou, Q., Yue, X., Khoo, S. C., Yang, H., Guan, R., Chen, H., Zhang, X., Wang, Y., Wei, Z., Suo, C., Peng, Y., Yang, Y., Lam, S. S., & Sonne, C. (2020). A review of historical and recent locust outbreaks: Links to global warming, food security and mitigation strategies. *Environmental Research*, 191(July), 110046.
<https://doi.org/10.1016/j.envres.2020.110046>
- Pingel, T. (2018). The Raster Data Model. *Geographic Information Science & Technology Body of Knowledge*, 2018(Q3). <https://doi.org/10.22224/gistbok/2018.3.11>
- Powlson, D. S., Whitmore, A. P., & Goulding, K. W. T. (2011). Soil carbon sequestration to mitigate climate change: A critical re-examination to identify the true and the false. *European Journal of Soil Science*, 62(1), 42–55. <https://doi.org/10.1111/j.1365-2389.2010.01342.x>
- R Development Core Team. (2020). R Development Core Team, R: a language and environment for statistical computing. In *R: A Language and Environmental for Statistical Computing: Vol. ver. 3.5.1*.
- Rajendran, K., Jayabalan, M., & Ehsan Rana, M. (2017). A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. *IJCSNS International Journal of Computer Science and Network Security*, 17(12), 172–177.
http://paper.ijcsns.org/07_book/201712/20171225.pdf
- Ray, D. K., West, P. C., Clark, M., Gerber, J. S., Prishchepov, A. V, & Chatterjee, S. (2019). S1 Text Additional data and statistical analysis information Climate change already affects global food production. *PLoS ONE*, 14(5).
https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.01/
- Samarati, P., & Sweeney, L. (1998). Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression. *Proc of the IEEE Symposium on Research in Security and Privacy*, 384–393.
<https://doi.org/http://dx.doi.org/10.1145/1150402.1150499>
- Scheider, S., Wang, J., Mol, M., Schmitz, O., & Karssenber, D. (2020). Obfuscating spatial point tracks with simulated crowding. *International Journal of Geographical Information Science*, 34(7), 1398–1427.
<https://doi.org/10.1080/13658816.2020.1712402>
- Schulte, R. P. O., O’Sullivan, L., Coyle, C., Farrelly, N., Gutzler, C., Lanigan, G., Torres-Sallan, G., & Creamer, R. E. (2016). Exploring Climate-Smart Land Management for Atlantic Europe. *Agricultural & Environmental Letters*, 1(1), 160029.
<https://doi.org/10.2134/ael2016.07.0029>
- Seidl, D. E. (2014). Obfuscation of Spatial Survey Data in Carinthia for Privacy Preservation. *Marshall Plan Scholarship*. San Diego, CA: San Diego State University.

- Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and False Identification Risk in Geomasking Techniques. *Geographical Analysis*, 50(3), 280–297. <https://doi.org/10.1111/gean.12144>
- Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, 63, 253–263. <https://doi.org/10.1016/j.apgeog.2015.07.001>
- Simo, I., Schulte, R., O’Sullivan, L., & Creamer, R. (2019). Digging deeper: Understanding the contribution of subsoil carbon for climate mitigation, a case study of Ireland. *Environmental Science and Policy*, 98(March), 61–69. <https://doi.org/10.1016/j.envsci.2019.05.004>
- Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, 67(10), 13–17. <https://doi.org/10.5120/11430-6785>
- Smith, P., Soussana, J. F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., & Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, 26(1), 219–241. <https://doi.org/10.1111/gcb.14815>
- Soomro, T. R. (2015). Chapter 4: GIS Enabling Smart Agriculture. In *Smart Agriculture: An Approach Towards Better Agriculture Management*. OMICS Group eBook (Issue August). OMICS International. <https://doi.org/10.4172/978-1-63278-023-2-024>
- Sothe, C., Gonsamo, A., Arabian, J., & Snider, J. (2022). Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. *Geoderma*, 405, 115402. <https://doi.org/10.1016/j.geoderma.2021.115402>
- Stinchcomb, D. (2004). Procedures for Geomasking to Protect Patient Confidentiality. *ESRI International Health GIS Conference October 19, 2004 in Washington, DC*, 1–17.
- Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., & Balakrishnan, H. (2001). Chord: a scalable peer-to-peer lookup protocol for internet applications. *IEEE/ACM Transactions on Networking*, 11(1), 17–32. <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1180543>
- Sui, D. (2014). E-Research and GeoComputation. In *GeoComputation* (Issue 2nd Edition, pp. 262–279). <https://doi.org/10.1201/b17091-17>
- Swanlund, D., & Schuurman, N. (2016). Mechanism Matters: Data Production for Geosurveillance. *Annals of the American Association of Geographers*, 106(5), 1063–1078. <https://doi.org/10.1080/24694452.2016.1188680>
- Swanlund, D., Schuurman, N., Zandbergen, P., & Brussoni, M. (2020). Street masking: A network-based geographic mask for easily protecting geoprivacy. *International Journal of Health Geographics*, 19(1), 1–11. <https://doi.org/10.1186/s12942-020-00219-z>
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6), 1–21. <https://doi.org/10.1371/journal.pone.0021101>
- Thornton, P., & Lipper, L. (2014). How Does Climate Change Alter Agricultural Strategies

- to Support Food Security? *SSRN Electronic Journal*, April.
<https://doi.org/10.2139/ssrn.2423763>
- Tompson, L., Johnson, S., Ashby, M., Perkins, C., & Edwards, P. (2015). UK open source crime data: Accuracy and possibilities for research. *Cartography and Geographic Information Science*, 42(2), 97–111. <https://doi.org/10.1080/15230406.2014.972456>
- Top, M., & Kesk, M. (2019). Effectiveness of GNSS-based tractor auto steering systems in crop spraying. *Mustafa Kemal University Journal of Agricultural Sciences*, 4(24), 78–90.
- Um, J. H., Kim, H. D., & 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*, 1093–1098.
<https://doi.org/10.1109/SocialCom.2010.162>
- Van Der Burg, S., Wiseman, · Leanne, & Krkeljas, J. (2021). Trust in farm data sharing: reflections on the EU code of conduct for agricultural data sharing. *Ethics and Information Technology*, 23(2). <https://doi.org/10.1007/s10676-020-09543-1>
- Vu, K., Zheng, R., & Gao, J. (2012). Efficient algorithms for K-anonymous location privacy in participatory sensing. *Proceedings - IEEE INFOCOM, October 2014*, 2399–2407.
<https://doi.org/10.1109/INFCOM.2012.6195629>
- Wall, D., & Plunkett, M. (2020). *Major and Micro Nutrient Advice For Productive Agricultural Crops* (Issue 5). Teagasc, Environment Research Centre,.
<http://hdl.handle.net/11019/2475>
- Wang, J., Kim, J., & Kwan, M. P. (2022). An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data. *Cartography and Geographic Information Science*, 49(5), 385–406.
<https://doi.org/10.1080/15230406.2022.2056510>
- Wang, J., & Kwan, M. P. (2020). Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. *International Journal of Health Geographics*, 19(1), 1–14. <https://doi.org/10.1186/s12942-020-00201-9>
- Wang, Y., Tao, F., Chen, Y., & Yin, L. (2022). Interactive impacts of climate change and agricultural management on soil organic carbon sequestration potential of cropland in China over the coming decades. *Science of the Total Environment*, 817, 153018.
<https://doi.org/10.1016/j.scitotenv.2022.153018>
- Wightman, P., Coronell, W., Jabba, D., Jimeno, M., & Labrador, M. (2011). Evaluation of location obfuscation techniques for privacy in location based information systems. *2011 IEEE Latin-American Conference on Communications, LATINCOM 2011 - Conference Proceedings*. <https://doi.org/10.1109/LatinCOM.2011.6107399>
- Wightman, P., Zurbaran, M., & Santander, A. (2014). High variability geographical obfuscation for location privacy. *Proceedings - International Carnahan Conference on Security Technology, March 2015*. <https://doi.org/10.1109/CCST.2013.6922079>
- Wightman, P., Zurbarán, M., Zurek, E., Salazar, A., Jabba, D., & Jimeno, M. (2013). *θ-Rand : Random Noise-based Location Obfuscation Based on Circle Sectors*. 100–104.
<https://doi.org/10.1109/ISIEA.2013.6738976>

- Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS - Wageningen Journal of Life Sciences*, 90–91(December 2018), 100301. <https://doi.org/10.1016/j.njas.2019.04.007>
- Xu, X., Liu, W., Zhang, C., & Kiely, G. (2011). Estimation of soil organic carbon stock and its spatial distribution in the Republic of Ireland. *Soil Use and Management*, 27(2), 156–162. <https://doi.org/10.1111/j.1475-2743.2011.00342.x>
- Xue, M., Kalnis, P., & Pung, H. K. (2009). Location Diversity: Enhanced Privacy Protection in Location Based Services. *Springer*, 5561(4), 3. <https://doi.org/10.1007/978-3-642-01721-6>
- Yildirim, V., & Yomralioglu, T. (2004). An address-based geospatial application. *FIG Working Week, Athens, Greece*, 1–10. https://www.fig.net/resources/proceedings/fig_proceedings/athens/papers/ts20/TS20_6_Yildirim_Yomralioglu.pdf
- Yuanhui, X. (2016). *R package*. <https://cran.r-project.org/package=Peacock.test>
- Zandbergen, P. A. (2014). Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine*, 2014, 1–14. <https://doi.org/10.1155/2014/567049>
- Zang, H., & Bolot, J. (2011). *Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study* (Issue 11). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.44&rep=rep1&type=pdf>
- Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2015). The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1), 22–34. <https://doi.org/10.1080/15230406.2015.1095655>
- Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J. L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., ... Asseng, S. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences of the United States of America*, 114(35), 9326–9331. <https://doi.org/10.1073/pnas.1701762114>
- Zhong, G., & Hengartner, U. (2008). Toward a Distributed K-Anonymity Protocol for Location Privacy. *Association for Computing Machinery, WORKSHOP*, 33–38. <https://doi.org/https://doi.org/10.1145/1456403.1456410>
- Zurbaran, M. A., Salazar, A., Brovelli, M. A., & Wightman, P. M. (2020). An Evaluation Framework for Assessing the Impact of Location Privacy on Geospatial Analysis. *IEEE Access*, 8, 158224–158236. <https://doi.org/10.1109/access.2020.3019631>
- Zurbarán, M., Ávila, K., Wightman, P., & Fernández, M. (2014). *Near-Rand : Noise-based Location Obfuscation Based on Random Neighboring Points*. <https://doi.org/10.1109/LATINCOM.2014.7041890>
- Zurbarán, M., González, L., Wightman Rojas, P., Labrador, M., Revisión, A. DE, & Article, R. (2014). *A Survey on Privacy in Location-Based Services*. 32, 314–343. <http://www.scielo.org.co/pdf/inde/v32n2/v32n2a08.pdf>
- Zurbarán, M., Wightman, P., Brovelli, M., Oxoli, D., Iliffe, M., Jimeno, M., & Salazar, A.

(2018). N-Rand-K: Minimizing the impact of location obfuscation in spatial analysis. *Transactions in GIS*, 22(5), 1257–1274. <https://doi.org/10.1111/tgis.12462>