Original Study • DOI: 10.15212/ijafr-2025-0011 IJAFR • 64(1) • 2025 • 64-80

Irish Journal of Agricultural and Food Research

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

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

[1]Department of Geography, University College Cork, Cork, Ireland
[2]Teagasc, Crop, Environment and Land Use Programme, Johnstown Castle, Co Wexford, Ireland
[3]Environmental Research Institute, University College Cork, Cork, Ireland

Abstract

Geoprivacy protection is a significant concern when sharing agricultural data with geographical information. Open and transparent access to agricultural data is essential to optimise its use, as well as integration with other data sources, such as environmental and climate data to support agri-environmental research. However, the choice of obfuscation method is crucial for achieving maximum statistical accuracy and geoprivacy protection in shared agricultural data. The primary objective of this research is to evaluate the effectiveness of obfuscated data in solving real-world problems. Random forest and multilinear regression models were used to predict soil organic carbon (SOC) based on internal (soil-related) and external (climate) features under three scenarios. The model estimations of SOC using original data were compared to model estimations of SOC using two datasets derived from obfuscation techniques, the environmental similarity obfuscation method (ESOM) and Rand. ESOM is a novel cluster-oriented obfuscation method that considers the importance of the contribution of contextual factors (external/climate conditions) in agri-environmental research, which guarantees that obfuscated data are relocated to areas of the same external conditions (in this instance climate) as the original data. Rand is a simple, yet widely used, obfuscation method that transfers the original location to a random location inside an obfuscation area. The results demonstrate that ESOM SOC estimation is more accurate than that obtained using the Rand method, despite the low modelled impact of climate conditions on the SOC estimation. In addition to the importance of feature selection and prediction models in achieving high accuracy and enhancing model performance, the choice of obfuscation method is another crucial factor when the use of original data is restricted. Therefore, while integrating internal and external features can improve results, it is essential to carefully assess how these features impact accuracy when selecting an obfuscation method.

Keywords

Agriculture • environment • geoprivacy • obfuscation • spatial analysis

Introduction

Sustainable agriculture is an essential component of various international initiatives, including the 2030 agenda of the UN Sustainable Development Goals (SDGs), particularly SDG2-Zero Hunger (Bouma *et al.*, 2019). The objective of sustainable agriculture is to optimise the use of natural and environmental resources to increase agriculture productivity to satisfy food security while maintaining environmental integrity. However, agriculture significantly impacts the environment, affecting soil and water quality, biodiversity and greenhouse gas (GHG) emissions. Agricultural activities contribute to approximately 10% of the European Union (EU)'s GHG emissions (McEldowney, 2020; EEA, 2024) with sources such as methane (CH_4), nitrous oxide (N_2O) and carbon dioxide (CO_2). Methane is produced by various livestock digestion processes and decomposition of manure, and 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). There is subsequently a pressing need for solutions to GHG emissions in the agricultural sector that maintain productivity and mitigate environmental degradation (Qayyum *et al.*, 2023).

Agricultural land use and soil can be a sink or a source of anthropogenic emissions, with land management having the potential to either reduce or increase emissions. Although agricultural activities contribute to CO_2 emissions, effective land management practices can mitigate these emissions by increasing carbon storage in both trees and soils, thereby helping to reduce atmospheric CO_2 concentrations (Xu *et al.*,

†Corresponding author: P. Nowbakht

E-mail: 111222849@umail.ucc.ie

2011; McEldowney, 2020). For example, hedgerows are an important element in agricultural landscapes that can enhance terrestrial carbon sinks; however, this may be counteracted by intensive management and hedgerow removals (Black *et al.*, 2023). 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 (Lal, 2004; Schulte *et al.*, 2016). While the majority of SOC is available for mineralisation, research has identified more stable SOC associated with micro-aggregates and silt, plus clay fractions stored in subsoils and can act as a long-term carbon store (Torres-Sallan *et al.*, 2017). From a management perspective, this requires consideration of the trade-off between productivity and carbon emissions due to drainage (O'Sullivan *et al.*, 2015; Torres-Sallan *et al.*, 2017). 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 emissions, particularly CO_2, it is recommended to adopt practices that enhance carbon storage in trees and soils. These practices include growing diverse crops, rotating crop types and implementing mixed land use strategies such as agroforestry (McEldowney, 2020). Increasing carbon content of soil and carbon sequestration improves soil fertility and 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 strategies to collect and assess the high spatial and vertical variability of SOC are time-consuming and expensive (Smith *et al.*, 2020).

Environmental and climate data coupled with statistical and/or machine learning approaches can be used to generate models that represent the relationship between SOC with management and environmental conditions, such as soil type and climate. These models can then be used to estimate the SOC in locations that have no direct measurement, supporting our understanding of the factors that have a positive impact on SOC and government policies surrounding carbon sequestration. Multilinear regression (MLR) and machine learning methods, such as random forest (RF) models, are widely used for predicting crop suitability, crop yield production and SOC based on environmental data, including climate, soil, water and fertilisation information from various sources such as soil sampling, weather stations, remote sensing and satellite data (Jeong *et al.*, 2016; Ganesan *et al.*, 2021; Sothe *et al.*, 2022). For example, Broeg *et al.* (2023) utilised machine learning to predict soil properties from covariates derived from traditional soil mapping, elevation, land use and remote sensing, finding that such models could be transferred to areas that had no direct soil mapping undertaken in southern Germany.

Therefore, to achieve sustainable agriculture and effective land management, adopting an innovative digital farming approach that utilises such data and modelling techniques is essential. These approaches can boost productivity, optimise fertiliser and pesticide use and minimise environmental impacts while preserving natural resources (Ferris, 2017; Holloway *et al.*, 2018; Laborde & Piñeiro, 2018; Hrustek, 2020). The advancement of digital farming generates vast amounts of agricultural data including geographical data. Spatial data analysis techniques allow researchers to extract valuable insights, discover spatial patterns and explore interactions among diverse phenomena. For example, the interaction of environmental conditions on SOC consequently can improve decision-making in agricultural systems.

Despite its potential, spatial agricultural data can be used to identify the fields and farms from which the data originate and can lead to the disclosure of private information such as size of farm, type of crop, herd size, fertiliser use and financial information. Therefore, geoprivacy protection is a significant concern when sharing agricultural data. Obfuscation is one of the predominant techniques developed to alter the location data to reduce risk of disclosure while maintaining statistical accuracy and spatial patterns (Armstrong *et al.*, 1999; Wiseman *et al.*, 2019; Swanlund *et al.*, 2020; Wang & Kwan, 2020; Wei *et al.*, 2024). The challenge is to identify the appropriate obfuscation method for a certain purpose, while considering the trade-off between location confidentiality and spatial pattern preservation (Wang *et al.*, 2022).

Anonymisation and randomisation are the two predominant obfuscation techniques that are used to protect geoprivacy. Anonymisation exploits the *k*-anonymity concept that generates an obfuscation area that contains at least $k - 1$ other locations to ensure the original location is not identifiable, thereby reducing the risk of identification by $1/k$ (Gruteser & Grunwald, 2003). Randomisation transfers the original location to a random (or specified, i.e., furthest) location among *N* random locations in a predefined obfuscation area (Wightman *et al.*, 2011; Murad *et al.*, 2014; Zandbergen, 2014; Seidl *et al.*, 2018). The predominant obfuscation methods designed and developed to date are geometric based, relying on distance from the original location without considering the underlying environmental condition. Due to the influence of agriculture on the environment and conversely environment and climate conditions on agriculture, recent research has begun to explore generating obfuscation techniques based on environmental similarity, which preserve geoprivacy while maintaining the environmental and climate conditions of the original data (Nowbakht *et al.*, 2024). While promising,

these methods remain largely untested in agri-environmental research, meaning there remains a pressing need for further research to consider how obfuscation techniques that consider the environmental context of a study area can be used on real-world problems.

Obfuscation alters the location of data which can change the distribution, topology and agri-environmental relationship of features (e.g., SOC and climate). Different obfuscation methods affect distinct aspects of spatial data, and while various evaluation metrics have been applied to evaluate these (e.g., privacy protection, distribution preservation), the question of whether obfuscated data can be used in agri-environmental models remains open. This is important, as if obfuscated data can be reliably used to solve real-world problems in agri-environmental research, it should substantially advance digital farming and sustainable agriculture. The use of obfuscated data in statistical models that incorporate explicit geographic locations and spatial data is novel, meaning it is also unknown how internal features (i.e., farm attributes) and external features (i.e., climate conditions) impact the choice of selecting an appropriate obfuscation method, advancing current guidelines in geographic information science (Wang *et al.*, 2022; Lorestani *et al.*, 2024).

Subsequently, the main aim of this research is to explore the usability of obfuscated data in addressing a real-world agricultural problem. For this purpose, we estimated SOC based on different environmental features that include both internal soil features (e.g., nitrogen (N), potassium (K), phosphorus (P) and potential of hydrogen [pH]) and external features (e.g., climate) in Ireland. This research also adds to the burgeoning SOC literature in Ireland and beyond. We used both the original data (i.e., unobfuscated data) and then compared modelled estimates with the obfuscated data. To generate the obfuscated data, we implemented two techniques. The first is a widely used geometric-based method and the second is a new method that uses clustering to maintain environmental similarity. The rationale behind this case study is such that if the spatial relationships can be replicated, then in principle it should allow high-risk data to be obfuscated and shared in order to support research purposes.

Materials and methods

Study area and data

The National Soil Database (NSDB) contains information on soil geochemistry for Ireland including major nutrients, as well as major and trace elements (Fay *et al.*, 2007). The project was conducted during 2003–2005, with the aim of compiling soil geochemistry information that included national scale maps (Fay *et al.*, 2007). To measure SOC concentration, and other geochemistry features including N, P and K for the near-surface soils (0–10 cm), the sampling strategy was adapted to collect 1,015 samples across Ireland, excluding the south-eastern region which was previously sampled in 1995. The samples were collected from predetermined sites from the national grid system with a grid resolution of 10 × 10 km. Two samples were taken from each 100 km^2 grid, one on the intersection and another from the centre of the grid. This sampling strategy ensures coverage of the four corners and the centre of each grid, providing a representative sample of different spatial areas within the grid. This approach helps capture the spatial variability of soil properties and can detect patterns that may exist due to various environmental factors or agricultural practices. At sites, a 20 × 20-m grid was created, with the sampling position at the centre with cores taken at 5 m intervals to a depth of 10 cm resulting in 25 cores to form a composite sample (Fay *et al.*, 2007). The NSDB can be used to demonstrate the effect of point-based obfuscation methods on SOC estimation. After pre-processing, 931 points were extracted from the NSDB to use in this case study.

Climate data, including precipitation and minimum and maximum temperatures from December 2003 to November 2004, were obtained from Wingler *et al.* (2021) at a pre-processed 5 km resolution from the Met Éireann Reanalysis (MÉRA) dataset. MÉRA offers a 38-year high-resolution regional climate reanalysis for Ireland from 1981 to 2019 (Gleeson *et al.*, 2017; Gleeson & Whelan, 2020). These data were used to calculate seasonal variables for winter (December 2003–February 2004), spring (March 2004–May 2004), summer (June 2004–August 2004) and autumn (September 2004–November 2004). Seasonal climate data were standardised before further processing.

Methods

The NSDB is a point-feature spatial dataset. Therefore, we employed two point-based obfuscation methods to investigate the usability of obfuscated data generated by different techniques for specific purposes. We selected two obfuscation methods for comparison:

(1) **Environmental similarity obfuscation method (ESOM):** This is a cluster-oriented method that ensures absolute environmental clustering preservation by generating obfuscation locations that maintain the original data's environmental characteristics (Nowbakht *et al.*, 2024).

(2) **Rand method:** This geometric-based approach provides a lower level of environmental clustering preservation, with >90% environmental misclassified error (MCE) according to Nowbakht *et al.* (2024).

These methods were chosen to highlight the differences in their performance, especially in solving real-world problems where the preservation of environmental clustering is crucial. The ESOM represents the best-case scenario with excellent

clustering preservation, while the Rand method illustrates the challenges of maintaining environmental integrity during obfuscation. Rand represents the worst-case scenario for cluster preservation, but it is one of the most widely used obfuscation approaches, providing an industry benchmark for comparison.

To compare SOC estimations using original data with those using obfuscated data, an independent-sample *t*-test was employed. This statistical method determines whether there is a significant difference between the means of two independent groups. By applying the independent-sample *t*-test, we can assess whether the observed differences between the SOC estimations are likely due to random variation or if they reflect a true difference in the populations being compared.

Rand method

Rand transforms the original point *p* to a random point inside a circular obfuscation area with the centre *p* and radius *r*. The obfuscated point can be located within *r*, either the farthest point from *p* or representative of the average point within *N* random points (or *N* random points from specific set of points). The selection of the location for the random obfuscated point depends on the size of the obfuscation area, which is dependent on spatial distribution, population density, attribute sensitivity and level of obfuscation (Wightman *et al.*, 2011; Murad *et al.*, 2014). 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) (Figure 1A and B).

Environmental similarity obfuscation method (ESOM)

The ESOM is a point-based adaptation of the polygon-based environmental similarity obfuscation method (PESOM) (introduced by Nowbakht *et al.*, 2024). ESOM was developed using environmental and climate data to generate obfuscation areas that are located in regions with similar environmental and climate conditions. The technique clusters the study area based on the 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 partition the data based on their similarity. Temporal and spatial variability of climate data can then be used to cluster the study area based on the similarity of their climate conditions. 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.

It is crucial that obfuscated locations are not generated within sensitive areas. Sensitive areas are defined as the union of all existing locations (points) and obfuscated locations within the primary obfuscation area. According to Nowbakht *et al.* (2023), a new obfuscated area can be constructed by excluding the sensitive area from the primary obfuscation area. This approach helps prevent false identification and ensures unique obfuscation. Within this predefined obfuscation area, the obfuscated location can be chosen randomly or determined as the farthest/middle point among *N* random locations (Figures 1C and 2, step 3).

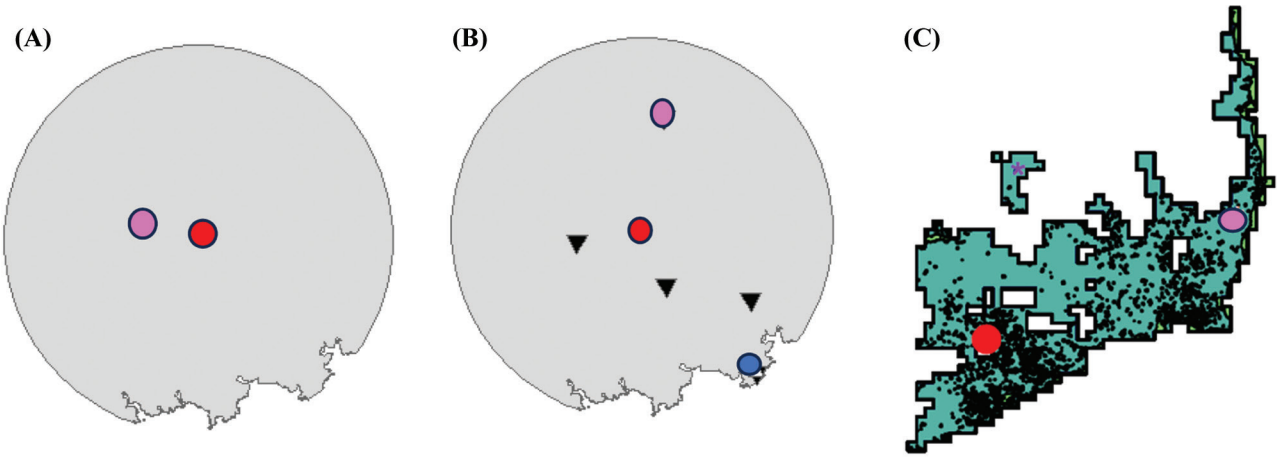


Figure 1. Examples of an obfuscated point generated by different methods (red point: original point). (A) Simple Rand, purple point: obfuscated point. (B) Rand (farthest blue point/middle purple point): obfuscated point. (C) Environmental similarity obfuscation method (ESOM), purple point: obfuscated point.

To apply unsupervised clustering, two key issues must be considered: (1) the number of features and (2) the number of clusters. Nowbakht *et al.* (2024) extensively detailed the process of determining optimal feature selection and the appropriate number of clusters. Consequently, ESOM is implemented in multiple stages, as illustrated in Figure 2.

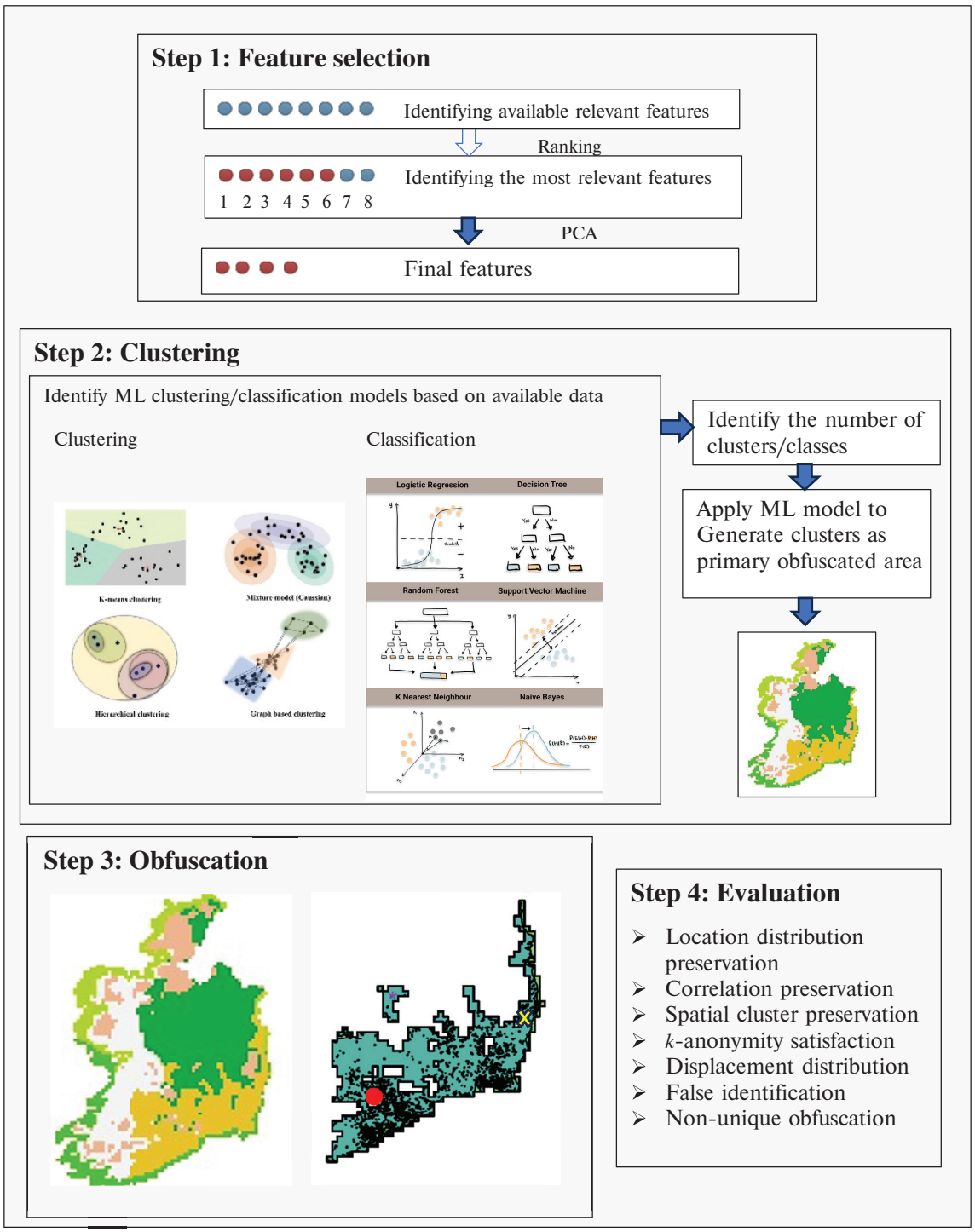


Figure 2. Environmental similarity obfuscation method (ESOM) procedure. ML = machine learning; PCA = principal component analysis.

Number of features

Spatially derived climate data offer a wide range of variables, including precipitation and temperature, captured at various temporal resolutions such as hourly, daily, monthly and annually (Dong *et al.*, 2021). However, we used seasonal climate data in the clustering algorithm to avoid overfitting of monthly data or over-simplification of annual data following recommendations (Bethere *et al.*, 2017; Akrami *et al.*, 2022). Three features, total precipitation, minimum temperature and maximum temperature, were considered for each of the four seasons (spring, summer, autumn and winter), resulting in a total of 12 features used in the K-means clustering algorithm. This seasonal approach ensures a balanced representation of climatic variability while maintaining computational efficiency.

In clustering, multicollinearity can be problematic as colinear features (highly correlated) may result in some features receiving a higher weight. Principal component analysis (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 (PCs), which is a linear combination of the 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, and thus reduce multicollinearity (Jollife & Cadima, 2016; Gwelo, 2019).

Number of clusters

K-means clustering is an algorithm that divides an unlabelled dataset into distinct groups, using the Euclidean distance metric to measure similarity, as recommended by Singh *et al.* (2013). To estimate the number of clusters, we applied the RSQRT function (Eq. 1), following the recommendation of Carlis & Bruso (2012), which refers to a recursive square root of the number of observations or features. The choice between the number of observations and features depends on the scale of the data. When all features are numeric, the number of observations is used; otherwise, the number of features is considered.

$$\mathrm{RSQRT}(x) = \{x^{2^{-r}};\ r = 1,2,\ldots, r_{\max};\ x^{2^{-r_{\max}}} < 2.25\} \qquad (1)$$

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

Predictive machine learning techniques (MLR and RF)

SOC was predicted based on soil-related features contained within NSDB and climate features obtained from MÉRA, including precipitation, minimum temperature and maximum temperature from December 2003 to November 2004. To estimate and predict SOC based on the NSDB, we included in the explanatory variables soil-related features (herein we refer to these as internal features, as they are internal to the NSDB) and climate features (herein we refer to these as external features, as they are external to the NSDB). We explored the SOC–environmental relationships across three scenarios. In scenario 1, we included just internal features. In scenario 2, we included both the internal and external features. In scenario 3, we included only the external features.

Multilinear regression 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. Multilinear regression 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.

Random forest is a supervised machine learning technique used in both classification and regression problems. In RF, the bootstrapping technique is used to construct a large number of decision trees as a forest on randomly selected observations and features of the training dataset. This learning process leads to improvement in the predictive accuracy of the model. The outcome is the majority vote of the forest for classification techniques and the average outcome for regression techniques.

Results

Parameter determination of obfuscation process

For the Rand method, an obfuscation area was defined as a circle with a 50,000-m radius around each original point, and the obfuscated point was determined as the farthest among five randomly selected points within this area (Nowbakht *et al.*, 2022).

In step 1 (feature selection) of implementing the ESOM procedure, the PCA resulted in us selecting the top seven PCs for step 2. In step 2 (clustering), due to the unavailability of labelled data, the choice of machine learning clustering models was limited to unsupervised learning algorithms; K-means clustering was co-opted to cluster the study area using the seven climate features obtained in step 1. The overall study area consisted of 3,440 5-km grid cells, meaning that there are 3,440 observations of each feature. RSQRT function was applied on 3,440 observations with seven numerical features (PC1, PC2, …, PC7) to obtain the potential number of clusters for K-means clustering. RSQRT (3,440) = (58.6, 7.6, 2.7, 1.6). Therefore, the largest number of clusters, 59, was selected to cluster the study area and generate an optimal obfuscation area. This resulted in the final step of the ESOM procedure, step 3 (obfuscation), whereby a random point within the

obfuscation area defined in the previous step was selected as the obfuscated location.

Estimating SOC–environmental relationships using the original data

Using the MLR approach, SOC had a statistically significant strong positive correlation with N (r = 0.89, $p \leq 0.05$), a moderate negative correlation with pH (r = −0.53, $p \leq 0.05$), a weak positive correlation with P (r = 0.25, $p \leq 0.05$) and a very weak non-significant correlation with K (r = −0.05, p = 0.09) (Figure 3A). The results of the correlation analysis also show a statistically significant moderate correlation between SOC and two components of climate features PC2 (r = −0.4, $p \leq 0.05$) and PC1 (r = 0.21, $p \leq 0.05$) and a very weak non-significant correlation with the rest (Figure 3B). These results suggest that the MLR model can be used to predict SOC

(A)

	pH	P	K	N	SOC
pH	1	0.1	0.1	−0.38	−0.53
P		1	0.36	0.27	0.25
K			1	0.07	−0.05
N				1	0.89
SOC					1

(B)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	SOC
PC1	1	−0.4	0.34	0.01	0	−0.04	0.08	0.21
PC2		1	0.21	−0.04	−0.09	0.03	0.08	−0.4
PC3			1	−0.02	−0.04	0	0.36	−0.06
PC4				1	−0.04	0.11	−0.05	0
PC5					1	0.09	0.22	−0.01
PC6						1	0.06	0.05
PC7							1	0.03
SOC								1

(C)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	SOC
PC1	1	−0.4	0.34	0.01	0	−0.03	0.08	0.21
PC2		1	0.21	−0.04	−0.09	0.03	0.08	−0.4
PC3			1	−0.02	−0.04	0	0.36	−0.06
PC4				1	−0.04	0.11	−0.05	0
PC5					1	0.09	0.22	−0.01
PC6						1	0.06	0.05
PC7							1	0.03
SOC								1

(D)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	SOC
PC1	1	−0.41	0.33	−0.04	0.1	0.03	−0.11	0.08
PC2		1	0.24	0.21	−0.09	0.13	0.17	−0.3
PC3			1	0.17	−0.13	−0.1	−0.06	−0.09
PC4				1	0.22	−0.13	−0.13	0.14
PC5					1	0.28	−0.08	−0.03
PC6						1	0.05	−0.15
PC7							1	−0.12
SOC								1

Figure 3. Correlation matrix between soil organic carbon (SOC) in % from the original National Soil Database (NSDB) dataset and (A) soil-related features including nitrogen (N), potassium (K), phosphorus (P) and potential of Hydrogen (pH). Correlation matrix between soil organic carbon (SOC) in % and (B) climate data represented as seven principal components (PCs) from the original NSDB dataset, and from the obfuscated NSDB data generated by (C) environmental similarity obfuscation method (ESOM) and (D) Rand method.

based on soil nutrient properties (e.g., N, P, K) and pH levels, along with climate features.

Estimating SOC–environmental relationships using the obfuscated data

A comparison of Figure 3B–D demonstrated that the correlation between SOC and climate features remained unchanged for ESOM obfuscated data, while the Rand method altered this correlation. This suggests that SOC predictions using Rand-generated obfuscated data might be less accurate than those using ESOM obfuscated data.

Both methods demonstrated a high level of privacy protection, 90% of obfuscated locations satisfied 20-anonymity. This means for 90% of the obfuscated locations, the chance of identification is 1/20, indicating that the obfuscated location is in a region with 20 other obfuscated data points. The ESOM offered a higher level of privacy protection for 75% of obfuscated locations, with the risk of identification decreasing to 1/50 (the obfuscation location is in a region with 50 other locations). Comparatively, the Rand method only provided this level of protection just for 45% of obfuscated locations (Figure 4A).

Displacement refers to the distance of the obfuscated location from the original location. The percentage of displacements for certain values is shown in Figure 4B. This indicates that all obfuscated locations generated by Rand methods have <50 km displacement while 60% of the obfuscated points generated by ESOM are >50 km, with >70% of locations between 100 km and 400 km displacement. Larger location displacement reduces the statistical accuracy.

ESOM provided perfect environmental and climate cluster preservation with zero MCE. This is due to the nature of the method that guarantees the obfuscated location is in the same environmental and climate condition as the original location. The MCE for the Rand method based on the same climate clustering is reported to be 97%, which means 97% of obfuscated locations are in different climatic conditions from the original locations (Supplementary Figure S1).

The SOC estimation results across three scenarios revealed the following:

(1) The MLR model, which estimates SOC using soil-related features with an adjusted R^2 of 0.85, shows that SOC estimation remains unaffected by obfuscation methods since it relies solely on internal features (Supplementary Table S1).

(2) Random forest models trained on both internal and external features explained over 91% of SOC variance, with the RF3 model showing high accuracy and minimal mean difference in predictions between original and obfuscated data, though ESOM obfuscated data provided slightly better alignment with the 1:1 line compared to Rand obfuscated data (Table 1, Figure 5D and E).

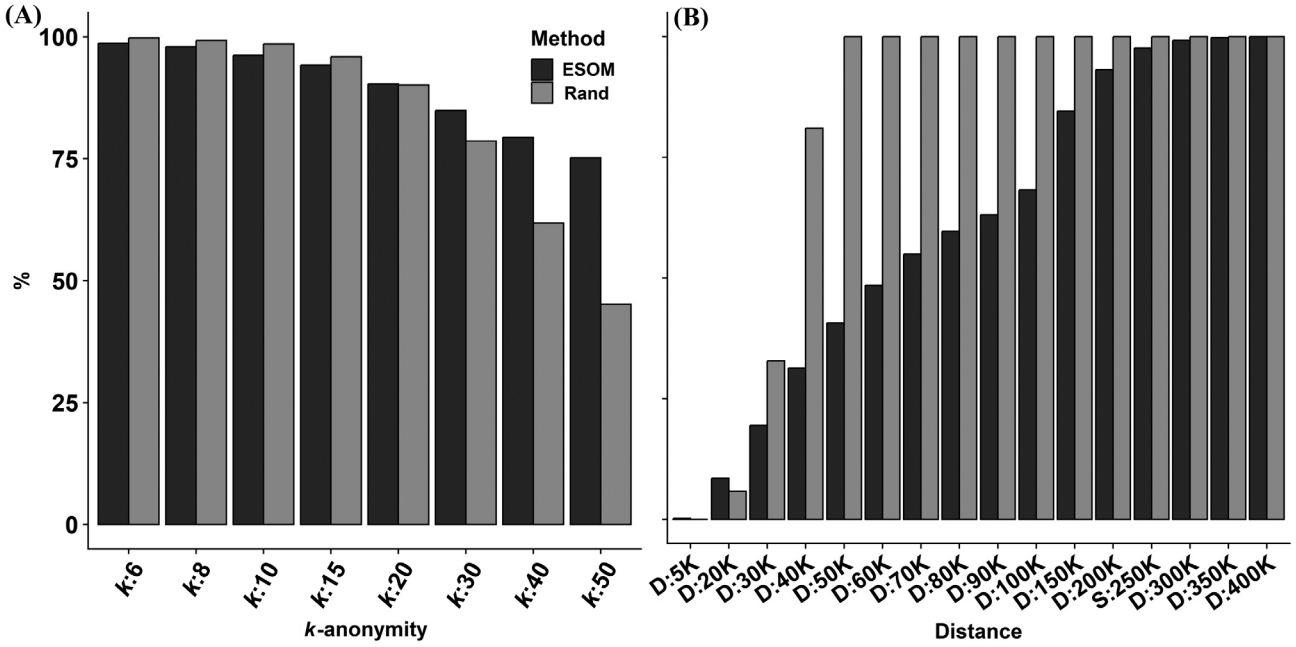


Figure 4. Comparison of performance of environmental similarity obfuscation method (ESOM) and Rand method. (A) Percentage of obfuscated locations that satisfies *k*-anonymity for different values of *k*. (B) Percentage of obfuscated location with displacement less than different values.

Table 1: Comparison of various RF models' performance in predicting SOC for the original NSDB dataset (first three rows), ESOM obfuscated data and Rand obfuscated data using internal features including N, K, P and pH and external features

Model	% Variance explained	Root mean squared residuals of train dataset	Root mean squared error (RMSE) for test dataset	All internal features (N, pH, P, K) and external features
RF1 (original)	91.13	4.42	3.34	PC1, PC2, PC3, PC4, PC5, PC6, PC7
RF2 (original)	91.03	4.45	3.24	
RF3 (original)	91.43	4.34	3.26	PC2, PC1
RF (ESOM)	91.42	4.34	3.28	PC2, PC1
RF (Rand)	91.29	4.38	3.44	PC2, PC4

ESOM = environmental similarity obfuscation method; K = potassium; N = nitrogen; NSDB = National Soil Database; P = phosphorus; PC = principal component; pH = potential of hydrogen; RF = random forest; SOC = soil organic carbon.

(3) The MLR model using climate data showed low explanatory power, but ESOM obfuscated data maintained accuracy identical to the original data, whereas Rand obfuscated data led to decreased predictive performance, demonstrating that ESOM provides better obfuscation for SOC prediction (Supplementary Table S2).

Scenario 1: using internal features

To estimate and predict SOC based on soil-related features (internal features including N, K, P and pH) and climate features (external, seven PCs), a stepwise selection method was used to construct the multiple linear regression model. The adjusted R^2 value of 0.85 shows a high level of explanatory power. 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 ($p \leq 0.05$). Results from the correlation analysis and Figure 3A and B indicate a statistically significant strong-to-weak association between soil-related features and SOC ($p \leq 0.05$). Equation 2 can be used to estimate the SOC which shows that for 1% increase in pH and K the SOC decreases by −4.4 and −0.03, respectively. The coefficients for P and N are 0.26 and 15.52, respectively.

$$\widehat{\mathrm{SOC}} = 23.45 - 4.41\,\mathrm{pH} + 0.26\,\mathrm{P} - 0.03\,\mathrm{K} + 15.52\,\mathrm{N} \quad (2)$$

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 the type of obfuscation method.

Scenario 2: using internal and external features

Several 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 >91% of variation of SOC (Table 1). The RF3 model used six of the most important features, including both internal features (N, K, P, pH) and external features (PC1 and PC2). This model reported the best results with 91.43% of the variance explained with a root mean squared error (RMSE) of 3.26 based on the test dataset (Table 1). The marginal difference between RF3 and other models (e.g., RF2 model with an RMSE of 3.24 but slightly lower explained variance of 91.03%) highlights that the inclusion of external features only slightly improves the overall performance, but the RF3 model with slight improvement is used to demonstrate the comparison of different obfuscation methods when external features are included. The comparison allows us to evaluate how external features contribute to model performance and provides a basis for assessing different obfuscation techniques. This slight enhancement serves to demonstrate the minor but measurable influence of external features on predicting SOC, emphasising their potential role without drastically altering the predictive power of the model. The descriptive statistics of the observed and predicted SOC of RF3 model for the test dataset are similar, although the minimum predicted value is higher, and the maximum predicted value is lower than the observed value (Table 2). Comparison of the observed and predicted SOC values using the RF3 model shows deviations from the 1:1 line (true model). For SOC values <20%, the predicted values closely match the observed values, aligning well with the 1:1 line. However, for SOC values between 20% and 40%, the model tends to over-predict, while for SOC values >40%, it tends to under-predict (Figure 5A).

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 ($p = 0.44 > 0.05$) with the mean difference in SOC values of 0.18 (Supplementary Table S3).

Table 1 summarises the performance of the RF model when applied to ESOM and Rand obfuscated data using the same training dataset. Since the most important features for

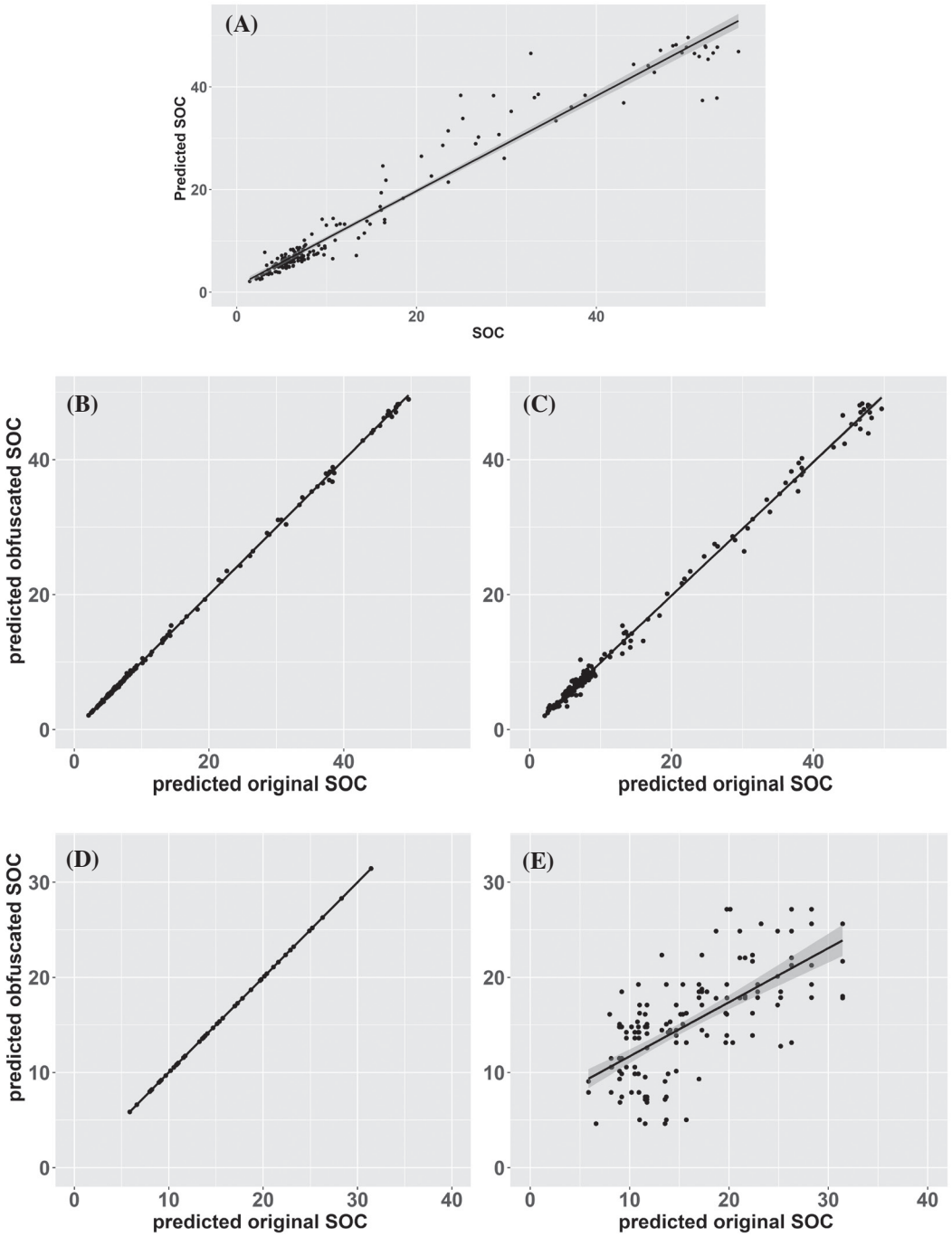


Figure 5. (A) Random forest (RF3) model performance for the test dataset. Scatterplot of observed values of soil organic carbon (SOC) versus predicted values in %. RF model performance. Scatterplot of predicted values of SOC in % using the original test dataset versus predicted SOC values in % using the (B) environmental similarity obfuscation method (ESOM) obfuscated dataset and (C) Rand obfuscated dataset. Multilinear regression (MLR) model performance based on external features (climate features). Scatterplot of predicted values of SOC in % using the original dataset versus predicted SOC values in % using the (D) ESOM obfuscated dataset with a root mean squared error (RMSE) = 0 and (E) Rand obfuscated dataset with an RMSE = 4.

Table 2: Comparison of the descriptive statistics of observed SOC from the original test dataset, SOC predictions from the original dataset using random forest RF3 and MLR models and SOC predictions from the obfuscated datasets generated by ESOM and Rand obfuscation methods using RF3 and MLR models

Model	Data	Min	Max	Mean	s.d.	Sample size
Original	Observe SOC	1.44	55.8	13.7	14.7	186
RF3 model	Predict SOC (original)	2.10	49.62	13.88	13.96	186
	Predict SOC (ESOM)	2.11	49.92	13.90	13.93	186
	Predict SOC (Rand)	2.05	48.29	13.8	13.84	186
MLR model	Predict SOC (original)	5.86	31.43	14.85	6.16	186
	Predict SOC (ESOM)	5.86	31.43	14.85	6.16	186
	Predict SOC (Rand)	4.63	27.16	14.46	5.27	186

ESOM = environmental similarity obfuscation method; MLR = multilinear regression; RF = random forest; SOC = soil organic carbon.

estimating SOC are soil-related internal features, and climate features have a minimal impact on RF model performance, no substantial difference is expected between the RF model's performance using ESOM or Rand obfuscated data.

An independent-sample *t*-test compared the predicted SOC from the original data with those from ESOM and Rand obfuscated datasets. The mean difference between predicted SOC from the original data and ESOM obfuscated data was −0.02 (p = 0.33), smaller than the mean difference observed with Rand obfuscated data, which was 0.09 (p = 0.20). The results showed no significant difference between the predicted SOC from the original and either obfuscated dataset (Supplementary Table S4). This indicates that both obfuscated datasets can predict SOC as accurately as the original data.

To further evaluate the differences between the two obfuscation methods, a *t*-test comparing ESOM and Rand obfuscated data was conducted. The results (mean difference = 0.11, p = 0.14) indicate no statistically significant difference between the two methods (Supplementary Table S5). Although the results of this study highlight the low impact of external features on SOC prediction, this scenario was presented to demonstrate the impact of external features on the choice of obfuscation methods. Assuming there is a statistically significant difference between ESOM and Rand, the ESOM would be the better choice as it provides a balance between privacy protection and predictive accuracy.

The descriptive statistics of the predicted SOC from RF model using obfuscated data generated by ESOM are the same as the descriptive statistics of the predicted SOC using the original dataset. Some overestimation is observed when the RF model is implemented on Rand-generated obfuscated data (Table 2). Comparison of the predicted SOC using the RF model with original data and the RF model with ESOM and Rand obfuscated data shows that while both obfuscated datasets generate SOC predictions similar to the original, the predictions from ESOM obfuscated data align perfectly with the 1:1 line. In contrast, SOC predictions from Rand obfuscated data display minimal deviations from the 1:1 line (Figure 5B and C).

Scenario 3: using external features

The MLR model can be constructed just using climate data; Figure 3B shows the moderate-to-weak correlation between SOC and PC1 and PC2. The results of the MLR model to estimate the SOC based on climate data indicate that the adjusted R^2 value was 0.15, showing a low level of explanatory power. PC2 is the only component that shows a statistically significant impact on SOC ($p \leq 0.05$), with an F-statistic value of 131.8. The proposed LR model can be expressed as shown in Equation 3.

$$\widehat{\mathrm{SOC}} = 14.8 - 2.9\ \mathrm{PC2} \quad (3)$$

The ESOM obfuscation model preserves the correlation between SOC and climate features (external features) (Figure 3B and C). Therefore, the MLR model using ESOM obfuscated data was the same as the MLR using original data and can be represented in Equation 2 too. The MLR model for Rand obfuscated data can be modelled in Equation 4.

$$\widehat{\mathrm{SOC}} = 14 - 2.5\ \mathrm{PC2} + 4.3\ \mathrm{PC4} \quad (4)$$

The descriptive statistics of the predicted SOC from the MLR model using the original dataset and ESOM 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 2).

Implementation of MLR model on Rand obfuscated data to predict SOC over-predict and under-predict the minimum and maximum SOC to 4.63% and 27.16%, respectively (Table 2) and is different from the descriptive statistic of the predicted SOC from original data. The summary of implementation

of MLR model on ESOM and Rand obfuscated data for the same training dataset is shown in Supplementary Table S2. The results indicate that the MLR model on ESOM obfuscated data is exactly the same as the MLR model on original data and the MLR model on Rand obfuscated data reduced the explanatory level 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.

The ESOM provides perfectly obfuscated data to predict SOC based on external features as the predicted SOC is exactly the same as the predicted SOC from original data with almost no deviation from the 1:1 line (Figure 5D). The RMSE is very close to zero, indicating a very small error. The SOC prediction from the original data and that from Rand obfuscated data are clearly different and a large deviation from the 1:1 line confirms the difference (Figure 5E) and an RMSE of 4 is representative of a larger error.

Results from the independent-sample *t*-tests indicate no significant differences between the predicted SOC from the original data and both ESOM obfuscated data (mean difference = 0, $p = 0.2 > 0.05$) and Rand obfuscated data (mean difference = 0.4, $p = 0.26 > 0.05$) (Supplementary Table S6). Similarly, the *t*-test comparing SOC predictions from ESOM and Rand shows no significant difference (mean difference = 0.4, $p = 0.26 > 0.05$) (Supplementary Table S7) which are identical to those observed between Rand obfuscated data and the original data. This finding confirms that ESOM obfuscation produces predictions closely aligned with the original data. While there is no statistical difference between SOC predictions from ESOM and Rand, the fact that ESOM results in a mean difference of zero relative to the original data suggests that it aligns more closely with the original predictions. Therefore, ESOM is a more reliable choice for preserving predictive accuracy while providing the necessary privacy protection.

Discussion

The aim of this research was to determine the effect of different generated obfuscated data to identify the environmental and climate factors that influence SOC. In this case study, two obfuscation methods were adopted on NSDB data to predict SOC based on internal (soil-related) and external (climate) features and explore the importance of the choice of obfuscation method. The results highlighted the importance of feature selection, prediction models and obfuscation methods for achieving high accuracy and model performance when original data use is restricted. While integrating internal and external features can enhance outcomes, their impact on accuracy must be carefully evaluated when choosing an obfuscation method.

A comparison of SOC estimation across three scenarios highlights the critical role of feature selection in enhancing the accuracy of statistical and machine learning models. This investigation confirms that soil-related features with higher correlations to SOC are the most crucial features for accurately estimating SOC, particularly N (Eq. 2), consistent with global research (e.g., Xu *et al.*, 2021). While external features slightly impact the SOC estimation, the combination of internal and external features improved model performance, albeit not to a statistically significant extent (Figure 5B and C and Table 1). The choice of prediction model is another crucial factor in achieving more accurate results. This choice significantly depends on the type of available data, specific needs, facilities and equipment. Further research is recommended to explore how different prediction models impact the accuracy of SOC estimation.

Another critical factor arises when the use of original data is restricted: selecting appropriate obfuscation methods to ensure accurate estimation. The results of this case study indicate that despite the low influence of external features on the estimation of SOC, using ESOM obfuscated data for the same features and model provides more accurate estimation than Rand obfuscated data, even though the difference is not statistically significant (Figure 5B–E). This case study was designed to demonstrate the impact of external features on the choice of obfuscation method, where the significance of differences between obfuscation methods was not the primary focus.

It is important to note that model estimates based on obfuscated data are more accurate when they closely match the estimates derived from the original data, not necessarily the observed values. Therefore, the goal is to generate obfuscated data that perform similarly to the original data. This means that if the accuracy of the model using the original data is low, the appropriate obfuscation method should yield similarly low accuracy, rather than trying to achieve high accuracy or close estimates to the real values.

Internal features, which remained unchanged during the obfuscation process, were the most important factors for predicting SOC. The MLR model, utilising internal features to predict SOC, provides a high level of accuracy and easy interpretation. Since the obfuscation process does not affect the correlation between internal features and SOC, the estimation of SOC based on internal features remains consistent regardless of the obfuscation method used. This stability allows us to select the obfuscation method that offers the highest geoprivacy protection which best addresses farmers' concerns. By safeguarding sensitive

data without compromising prediction accuracy, stakeholders can confidently support farmers in nutrient management to enhance profits and assist governments in adopting informed agricultural policies for sustainable land management (Xu *et al.*, 2011; Forkuor *et al.*, 2017). External features were reported to have slight impact on SOC. The results indicate that the obfuscated data generated by a cluster-oriented method produced more substantial SOC estimation than a geometric-based obfuscation method to estimate SOC using original data. Several studies have confirmed the important role of climate factors in predicting SOC levels, highlighting their impact across various scales (Adhikari *et al.*, 2014; Akpa *et al.*, 2016; Sothe *et al.*, 2022; Wu *et al.*, 2022).

We suggest that the influence of internal and external features on statistical models derived in agri-environmental research and a consideration of the priority of privacy protection are components that should be considered when choosing the most appropriate obfuscation methods, which builds upon the *de facto* standard which is to select models that only consider privacy protection (Zurbarán *et al.*, 2018). Our results highlight that clustering preservation is the most important factor when environmental data have a major influence on statistical analysis. This indicates that cluster-oriented methods would be the best choice for data that need to be used in agri-environmental research, while density-based methods such as Density and AHilb (Nowbakht *et al.*, 2022) are more appropriate when the distribution preservation is the main focus of study.

In this study, unsupervised machine learning and climate data were used to cluster the study area. The selection of features and clustering techniques depends on data availability, labelled data, research interests and specific problems. For instance, clustering can be based on soil type, land use or proximity to water, demonstrating the flexibility of ESOM for various case studies. Moreover, different machine learning and deep learning techniques, such as convolutional neural networks (CNNs) and graph neural networks (GNNs), are widely used in clustering spatial data (Richter, 2018; Kaczmarek, 2023; Kaczmarek *et al.*, 2023; Houfaf-Khoufaf *et al.*, 2024). Kopczewska (2022) provides an overview of clustering techniques, highlighting that when clustering of locations and values simultaneously is needed, K-means clustering is the preferred method. Due to the paucity of research exploring the development of environmental similarity in obfuscation techniques, we opted to select only one method within this case study; however, the choice of data that points are clustered on (i.e., climate, soil, land use) and the clustering techniques that are used (e.g., K-means, DBSCAN) may impact the results. Therefore, further research is needed to build upon our findings, although we posit that selection of data and methodologies must be grounded in the context of the study under investigation.

Conclusion

The results of this case study show the importance of feature selection, prediction models and obfuscation methods when the use of original data is prohibited. The combination of internal and external features improved the results of SOC estimation (Figure 5A), with our results highlighting the importance of internal features, particularly N, but also a weaker, but still significant role of climate, building on the burgeoning literature on the subject. Moreover, when selecting an obfuscation method, it is essential to consider the degree of these features' impact on the privacy protection. In applications where external features have little to no impact, the obfuscation method with high geoprivacy preservation is the most appropriate obfuscation method. However, our results suggest that if obfuscated data are to be used in agri-environmental research with external features utilised in models (e.g., climate, soil, land use), then obfuscation techniques must include some environmental similarity, otherwise the models will perform poorly. The ESOM successfully resulted in lower environmental displacement and maintained the modelled relationships when compared to the unobfuscated data, suggesting that this approach can be widely adopted by researchers and practitioners. Clarity on research priorities is essential to help determine the appropriate methodological approaches to obfuscate spatial data to achieve the optimal geoprivacy protection, spatial pattern preservation and statistical accuracy.

Acknowledgements

The first author was funded by Teagasc, Walsh Scholarship Scheme, a joint project between Teagasc and UCC (Walsh Scholarships Ref Number 2018034).

Conflicts of interest

The authors have no conflict of interest to declare in relation to this work.

Data availability statement

Data are not available due to the privacy restrictions of the NSDB dataset.

References

Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B. and Greve, M.H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. *PLoS One* **9**: e105519.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E. and Amapu, I.Y. 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma* **271**: 202–215.

Akrami, N., Ziarati, K. and Dev, S. 2022. Graph-based local climate classification in Iran. *International Journal of Climatology* **42**: 1337–1353.

Armstrong, M.P., Rushton, G. and Zimmerman, D.L. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* **18**: 497–525.

Bethere, L., Sennikovs, J. and Bethers, U. 2017. Climate indices for the Baltic states from principal component analysis. *Earth System Dynamics* **8**: 951–962.

Black, K., Lanigan, G., Ward, M., Kavanagh, I., hUallacháin, D. and Sullivan, L.O. 2023. Biomass carbon stocks and stock changes in managed hedgerows. *Science of the Total Environment* **871**: 162073.

Bouma, J., Montanarella, L. and 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**: 538–546.

Broeg, T., Blaschek, M., Seitz, S., Taghizadeh-Mehrjardi, R., Zepp, S. and Scholten, T. 2023. Transferability of covariates to predict soil organic carbon in cropland soils. *Remote Sensing* **15**: 876.

Carlis, J. and Bruso, K. 2012. RSQRT: an heuristic for estimating the number of clusters to report. *Electronic Commerce Research and Applications* **11**: 152–158.

Dong, Q., Chen, X., Dong, S. and Zhang, J. 2021. Classification of pavement climatic regions through unsupervised and supervised machine learnings. *Journal of Infrastructure Preservation and Resilience* **2**: 5.

EEA. 2024. "EEA Greenhouse Gases – Data Viewer". European Environment Agency. Available online: https://www.eea.europa.eu/data-and-maps/data/data-viewers/greenhouse-gases-viewer.

Fay, D., Mcgrath, D., Zhang, C., Carrigg, C., Flaherty, V.O., Kramers, G., Carton, O.T. and Grennan, E. 2007. "National Soils Database – End of Project Report. RMIS 5192". Teagasc. Available online: https://t-stor.teagasc.ie/handle/11019/897, 1–24 pages.

Ferris, J.L. 2017. Data privacy and protection in the agriculture industry: is federal regulation necessary? *Minnesota Journal of Law, Science & Technology* **18**: 307–342.

Forkuor, G., Hounkpatin, O.K.L., Welp, G. and Thiel, M. 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PLoS One* **12**, e0170478.

Ganesan, M., Andavar, S. and Raj, R.S.P. 2021. Prediction of land suitability for crop cultivation using classification techniques. *Brazilian Archives of Biology and Technology* **64**, e21200483.

Gleeson, E. and Whelan, E. 2020. "Met Éireann's Contribution to Package D6.2 of the JPI Climate INDECIS Climate Indices Project". Met Éireann Technical Note No. 67. Available online: http://hdl.handle.net/2262/91470.

Gleeson, E., Whelan, E. and Hanley, J. 2017. Met Éireann high resolution reanalysis for Ireland. *Advances in Science and Research* **14**: 49–61.

Gruteser, M. and Grunwald, D. 2003. Anonymous usage of location-based services through spatial and temporal cloaking. *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, MobiSys 2003*, pages 31–42.

Gwelo, A.S. 2019. Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics* **4**: 79–91.

Holloway, P., Kudenko, D. and 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.

Houfaf-Khoufaf, W., Touya, G. and Le Guilcher, A. 2024. Geographically masking addresses to study COVID-19 clusters. *Cartography and Geographic Information Science* **51**: 242–256.

Hrustek, L. 2020. Sustainability driven by agriculture through digital transformation. *Sustainability* **12**: 8596.

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. and Kim, S.H. 2016. Random forests for global and regional crop yield predictions. *PLoS One* **11**: 1–15.

Jollife, I.T. and Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical and Engineering Sciences* **374**: 20150202.

Kaczmarek, I. 2023. Spatial objects classification using machine learning and spatial walk algorithm. *Open Geosciences* **15**, 20220542.

Kaczmarek, I., Iwaniak, A. and Świetlicka, A. 2023. Classification of spatial objects with the use of graph neural networks. *ISPRS International Journal of Geo-Information* **12**: 83.

Kopczewska, K. 2022. Spatial machine learning: new opportunities for regional science. *Annals of Regional Science* **68**: 713–755.

Laborde, D. and Piñeiro, V. 2018. Monitoring agricultural productivity for sustainable production and R&D planning. *Economics* **12**: 1–14.

Lal, R. 2004. Soil carbon sequestration impacts on global climate change and food security. *Science* **304**: 1623–1627.

Lorestani, M.A., Ranbaduge, T. and Rakotoarivelo, T. 2024. Privacy risk in GeoData: a survey.

Lynch, J., Cain, M., Frame, D. and 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**: 518039.

McEldowney, J. 2020. "Briefing: EU Agricultural Policy and Climate Change. European Parliamentary Research Service". PE 651.922 (Issue May 2020). Available online: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651922/EPRS_BRI(2020)651922_EN.pdf.

Murad, A., Hilton, B., Horan, T. and Tangenberg, J. 2014. Protecting patient geo-privacy via a triangular displacement geo-masking method. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Privacy in Geographic Information Collection and Analysis (GeoPrivacy'14)*, pages 1–9.

Nowbakht, P., O'Sullivan, L., Cawkwell, F., Wall, D.P. and 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.

Nowbakht, P., O'Sullivan, L., Wall, D.P. and Holloway, P. 2023. Implementation of novel polygon-based obfuscation methods to improve privacy of agricultural data. *Transactions in GIS*, **27**: 84–104.

Nowbakht, P., O'Sullivan, L., Wall, D.P. and Holloway, P. 2024. Maintaining environmental context and geoprivacy protection in agriculture. *Information Processing in Agriculture* (In press). Available online: https://www.sciencedirect.com/science/article/pii/S2214317324000623.

O'Sullivan, L., Creamer, R.E., Fealy, R., Lanigan, G., Simo, I., Fenton, O., Carfrae, J. and Schulte, R.P.O. 2015. Functional Land Management for managing soil functions: a case-study of the trade-off between primary productivity and carbon storage in response to the intervention of drainage systems in Ireland. *Land Use Policy* **47**: 42–54.

Qayyum, M., Zhang, Y., Wang, M., Yu, Y., Li, S., Ahmad, W., Maodaa, S.N., Sayed, S.R.M. and Gan, J. 2023. Advancements in technology and innovation for sustainable agriculture: understanding and mitigating greenhouse gas emissions from agricultural soils. *Journal of Environmental Management* **347**: 119147.

Richter, W. 2018. The verified neighbor approach to geoprivacy: an improved method for geographic masking. *Journal of Exposure Science and Environmental Epidemiology* **28**: 109–118.

Schulte, R.P.O., O'Sullivan, L., Coyle, C., Farrelly, N., Gutzler, C., Lanigan, G., Torres-Sallan, G. and Creamer, R.E. 2016. Exploring climate-smart land management for Atlantic Europe. *Agricultural & Environmental Letters* **1**: 160029.

Seidl, D.E., Jankowski, P. and Clarke, K.C. 2018. Privacy and false identification risk in geomasking techniques. *Geographical Analysis* **50**: 280–297.

Simo, I., Schulte, R., O'Sullivan, L. and Creamer, R. 2019. Digging deeper: understanding the contribution of subsoil carbon for climate mitigation, a case study of Ireland. *Environmental Science and Policy* **98**: 61–69.

Singh, A., Yadav, A. and Rana, A. 2013. K-means with three different distance metrics. *International Journal of Computer Applications* **67**: 13–17.

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. and 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**: 219–241.

Sothe, C., Gonsamo, A., Arabian, J. and Snider, J. 2022. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. *Geoderma* **405**: 115402.

Swanlund, D., Schuurman, N., Zandbergen, P. and Brussoni, M. 2020. Street masking: a network-based geographic mask for easily protecting geoprivacy. *International Journal of Health Geographics* **19**: 1–11.

Torres-Sallan, G., Schulte, R.P.O., Lanigan, G.J., Byrne, K.A., Reidy, B., Simó, I., Six, J. and Creamer, R.E. 2017. Clay illuviation provides a long-term sink for C sequestration in subsoils. *Scientific Reports* **7**: 45635.

Wang, J. and 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**: 1–14.

Wang, J., Kim, J. and 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**: 385–406.

Wei, Z., Li, R., Jiang, K., Luo, Q., Sun, Z., Han, T., Bi, Z. and Wang, Y. 2024. PCS-ADS: privacy computing system for agricultural data security. *Proceedings of the 2024 8th International Conference on Control Engineering and Artificial Intelligence*, pages 278–283.

Wightman, P., Coronell, W., Jabba, D., Jimeno, M. and Labrador, M. 2011. Evaluation of location obfuscation techniques for privacy in location based information systems. *Proceedings of the 2011 IEEE Latin-American Conference on Communications (LATINCOM 2011)*, pages 1–6.

Wingler, A., Cawkwell, F., Holloway, P., Misra, G., de la Torre Cerro, R.S.C. and Sweeney, C. 2021. PhenoClimate: impact of climate change on phenology in Ireland. Environmental Protection Agency.

Wiseman, L., Sanderson, J., Zhang, A. and 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**: 100301.

Wu, X., Yuan, Z., Li, D., Liao, Y. and Huang, C. 2022. Contributions of climate and soil properties to geographic variations of soil organic matter across the East Asian monsoon region. *SSRN Electronic Journal* **234**: 105845.

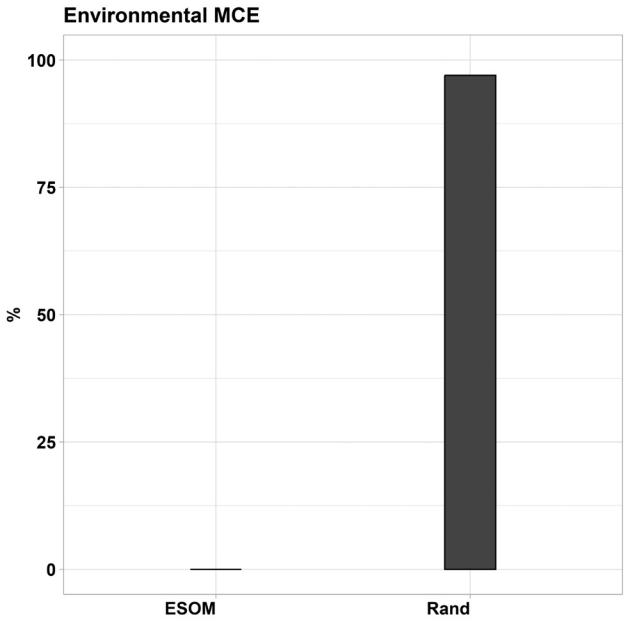
Xu, C., Xu, X., Ju, C., Chen, H.Y.H., Wilsey, B.J., Luo, Y. and Fan, W. 2021. Long-term, amplified responses of soil organic carbon to nitrogen addition worldwide. *Global Change Biology* **27**: 1170–1180.

Xu, X., Liu, W., Zhang, C. and Kiely, G. 2011. Estimation of soil organic carbon stock and its spatial distribution in the Republic of Ireland. *Soil Use and Management* **27**: 156–162.

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

Zurbarán, M., Wightman, P., Brovelli, M., Oxoli, D., Iliffe, M., Jimeno, M. and Salazar, A. 2018. NRand-K: minimizing the impact of location obfuscation in spatial analysis. *Transactions in GIS* **22**: 1257–1274.

Supplementary materials



Supplementary Figure S1. Comparison of environmental cluster preservation of environmental similarity obfuscation method (ESOM) and Rand method based on environmental misclassification error (MCE).

Supplementary Table S3: Independent-sample *t*-test results for random forest (RF3) predicted SOC of original dataset with observed SOC of original dataset

Model/data	*t*-Value	Confidence interval	Mean difference	*P*-value
RF3	0.77	−0.29, 0.66	0.18	0.44

RF = random forest; SOC = soil organic carbon.

Supplementary Table S4: Independent-sample *t*-test results for predicted SOC of original dataset with predicted SOC of ESOM and Rand obfuscated datasets

Model/data	*t*-Value	Confidence interval	Mean difference	*P*-value
RF (ESOM)	−0.99	−0.06, 0.02	−0.02	0.33
RF (Rand)	1.28	−0.04, 0.22	0.086	0.20

ESOM = environmental similarity obfuscation method; RF = random forest; SOC = soil organic carbon.

Supplementary Table S1: Comparison of MLR models' performance in predicting SOC for the original NSDB dataset, ESOM obfuscated data and Rand obfuscated data using just internal features including N, K, P and pH

Model/data	Adjusted R^2	F-statistic	Root mean squared error (RMSE)	*P*-value
MLR (original)	0.85	1,036	0.11	<0.001
MLR (ESOM)	0.85	1,036	0.11	<0.001
MLR (Rand)	0.85	1,036	0.11	<0.001

ESOM = environmental similarity obfuscation method; K = potassium; MLR = multilinear regression; N = nitrogen; NSDB = National Soil Database; P = phosphorus; pH = potential of hydrogen; RF = random forest; SOC = soil organic carbon.

Supplementary Table S2: Comparison of MLR models' performance in predicting SOC for the original NSDB dataset, ESOM obfuscated data and Rand obfuscated data using just external features

Model/data	Adjusted R^2	F-statistic	Root mean squared error (RMSE)	External features	*P*-value
MLR (original)	0.15	131.8	13.7	PC2	<0.001
MLR (ESOM)	0.15	131.8	13.7	PC2	<0.001
MLR (Rand)	0.12	51.47	13.95	PC2, PC4	<0.001

ESOM = environmental similarity obfuscation method; MLR = multi linear regression; PC = principal component; SOC = soil organic carbon.

Supplementary Table S5: Independent-sample *t*-test results for RF model predicted SOC of ESOM obfuscated dataset with predicted SOC of Rand obfuscated dataset

Model/data	*t*-Value	Confidence interval	Mean difference	*P*-value
RF	1.49	−0.03, 0.25	0.11	0.14

ESOM = environmental similarity obfuscation method; RF = random forest; SOC = soil organic carbon.

Supplementary Table S6: Independent-sample *t*-test results for LR model predicted SOC from original dataset with predicted SOC from ESOM and Rand obfuscated datasets

Model/data	*t*-Value	Confidence interval	Mean difference	*P*-value
LR (ESOM)	0.14	0, 0	0	0.2
LR (Rand)	1.12	−0.30, 1.08	0.40	0.26

ESOM = environmental similarity obfuscation method; LR = linear regression; SOC = soil organic carbon.

Supplementary Table S7: Independent-sample *t*-test results for LR model predicted SOC of ESOM obfuscated dataset with predicted SOC of Rand obfuscated dataset

Model/data	*t*-Value	Confidence interval	Mean difference	*P*-value
LR	1.12	−0.30, 1.08	0.40	0.26

ESOM = environmental similarity obfuscation method; LR = linear regression; SOC = soil organic carbon.