Optimal sampling using Conditioned Latin Hypercube for digital soil mapping: An approach using Bhattacharyya distance

Adnan Khan a,*, Matt Aitkenhead b, Craig R. Stark c, M. Ehsan Jorat a

a School of Applied Science, Abertay University, Dundee, UK
b James Hutton Institute, Aberdeen, UK
c School of Design and Informatics, Abertay University, Dundee, UK

ARTICLE INFO

Handling Editor: B. Minasny

Keywords:
Digital soil mapping (DSM), Conditioned Latin hypercube (CLH) sampling, Soil sampling, Optimum sample size, Bhattacharyya distance

ABSTRACT

Soil properties are important because they determine the soil’s suitability for different types of plant growth, ecosystems and biota functioning. Soil properties influence nutrient cycling, carbon sequestration and soil management. Digital Soil Mapping (DSM) is a procedure to map soil properties. Soil sampling for DSM is a foundational step in building prediction accuracy and essential for incorporating variability in terms of environmental covariates (ancillary variables). Conditioned Latin Hypercube (CLH) sampling is a method for generating a sample of points from a multivariate distribution that has been conditioned on one or more covariates. It is an extension of Latin Hypercube sampling, which is a popular technique for generating samples from a multivariate distribution in a way that ensures that each dimension is sampled uniformly. CLH sampling carries the benefit of selecting sampling locations covering the feature space and forming a Hypercube of the original sample. However, determining the optimum sample size is crucial in soil survey exercises constrained by budget and time limits. For this purpose, a study was carried out on Scotland’s Finzean Estate (44.8 km²) location. A dataset of 21 independent features (16 continuous and five categorical) and 17,932 sampling locations was created using Digital Elevation Model (DEM) derivatives, soil classes map, land cover map, peat depth map and parent material map to further generate sub-samples and compare the generated sub-samples with the original population. Two hundred CLH sampling datasets were extracted from the original population (17,932 data points) with different sizes (5, 10, 15, 20, ..., 100) and each size was given 10 repetitions e.g. (5, 1, 5, 2, ..., 5, 10). The sample datasets were analysed by comparing the mean, standard deviation, boxplot and estimates of the probability density function (pdf) for all the 16 continuous independent features. All the mentioned comparisons suggest that the impact of increasing sample size on the distribution of covariates can be observed up to a certain point, beyond which further increases in sample size may not yield noticeable differences. Bhattacharyya distance, a statistical measurement that quantifies the similarity between two probability distributions, was calculated between every quantitative and qualitative element of respective sampling size and for the original population. In contrast, as the CLH sample dataset size increased, the Bhattacharyya distance value decreased and became constant. The optimum number of samples based on the study was determined for the spatial extent of the Finzean Estate in Scotland and a range of 25–50 CLH samples was suggested based on the study. This work, therefore, achieved both reductions in sampling location numbers compared to classical approaches and identification of the precise location of these sample location to achieve optimal DSM.

1. Introduction

Soil is a vital part of the ecosystem and is responsible for maintaining the life cycle on this planet. Soil is directly or indirectly responsible for 95% of the food available around the world (FAO, 2015). Although soil supports the growth of plants and hosts thousands of biological organisms, it has been under tremendous pressure due to increased population, urbanisation, climate change, and lack of awareness and resources for appropriate management and conservation.

Soil properties are essential soil health indicators which provide information about the condition of the soil and need to be known before any decision-making (Maharjan et al., 2020; Aitkenhead et al., 2016a).
Farmers, land managers and policymakers need a quick quantitative knowledge of soil health indicators before any decision-making. Quantitative data of soil properties can be used to address the problem of soil mismanagement, soil erosion, land use change, food security, and water security. (Adhikari et al., 2014, 2015). However, to derive soil property quantitative data, access to laboratory-based soil analysis and skills (e.g., soil sampling, soil testing and soil analysis) are required (Aitkenhead et al., 2016b).

Soil conventional maps with quantified parameters have been the most familiar source of information for farmers for decades and traditionally farmers and land managers have used these maps for land management. Traditional conventional maps have limitations such as restricted coverage, detail, precision, and outdated information. They are also costly to produce. Digital Soil Mapping (DSM) overcomes these limitations by providing more detailed, accurate, and up-to-date information on soil properties over a large area, cost-effectively and quickly (Biswas and Zhang, 2018). McBratney et al. (2003) formulated a function based on Jenny’s 1941 ‘crop’ model that formalized DSM. The function is a quantitative empirical equation that estimates either soil class or soil property at a given point in space and time using seven environment covariates, namely: soil (S), climate (C), organisms (O), relief (R), parent material (P), age (A), and spatial location (N) often referred as SCORPAN (McBratney et al., 2003). DSM maps the target soil property using environmental covariates, associating weights to each environment covariate. Advancements in technology, including remote sensing devices, computational devices, and spatial data analysis tools, have significantly enhanced the process of collecting, processing, and interpreting large volumes of data necessary for DSM, facilitating the creation of high-resolution Digital Elevation Models (DEM) and their derivatives, land cover maps, soil types, and geology maps. However, we would like to emphasize that these high-resolution datasets are not readily available, but are the outcome of extensive work and application of sophisticated models such as DSM. Additionally, the improvement in remote sensing devices, computational devices, and spatial data analysis tools have made DSM a powerful approach for estimating soil health indicators.

Soil sampling is the initial step in DSM and is essential for every experimental analysis. Soil sampling is always desired to capture spatial variability in land use, topography, soil type, and vegetation (referred to as environmental covariates). Sampling design defines the model’s foundation by providing appropriate inputs to predict the model output and providing sampling points with good spatial coverage which eventually affects the model’s accuracy (Kidd et al., 2015). The sample’s design and size play a crucial role in model accuracy (De Gruijter et al., 2010) and also decide the cost and time of the project. Researchers pointed out soil sampling in DSM as the most expensive step of the study (Minasny and McBratney, 2006; Biswas and Zhang, 2018; Brus, 2019; Yang et al., 2020). The number of sampling points is always restricted by time and financial resources, therefore, it is often impractical to acquire a substantial number of sample locations to analyze the properties of soil at a site location (Webster et al., 2012).

Two challenges arise in selecting the sampling: (1) covering the variability in terms of environmental covariates and geographical space so that the machine learning model can be provided with variability; and, (2) keeping the size of sampling locations to a minimum while sampling the location of interest to save money and time. Addressing the first challenge, various designs are presented in the literature which cover the variability in terms of geographical space or feature space. Designs such as grid sampling (Jonard et al., 2013; Püikki et al., 2013), transect sampling (Zhou et al., 2016), spatial coverage sampling, (Kempen et al., 2015), simple random sampling (Evans and Hartemink, 2014), stratified random sampling, and CLH sampling (Minasny and McBratney, 2006) have been used in the past for sampling purposes. There are also studies highlighting the comparisons between different sampling designs. Minasny and McBratney (2006) compared simple random, stratified, spatial, and CLHs for a study area of 11 km². Using three continuous (CTI, NDVI, slope) and one categorical variable (land use) they showed in their study that the probability distribution at every interval of the variable was represented in the CLH sample while the random and stratified random sampling techniques under- and over-sampled some areas of NDVI and land use distribution. Ramirez-Lopez et al. (2014) compared Kennard Stone sampling, Fuzzy k-means sampling, and CLH sampling to map soil organic carbon and found the superiority of the CLH sampling algorithm over the other two. Adhikari et al. (2017) investigated grid, random and CLH sampling designs over different areas to map organic carbon content and showed that CLH sampling showed the highest value of the coefficient of determination. Zhang et al. (2022) compared four different sampling designs, namely grid sampling, stratified sampling, grid random sampling and CLH sampling and concluded that grid sampling and CLH sampling resulted in better-predicted values with grid sampling providing better spatial coverage and CLH sampling providing better feature space.

Several other studies also reported the sampling designs which outperformed CLH sampling. For instance, Wadoux and Brus (2021) indicated that CLH sampling performed worse than other sampling designs when utilizing a Random Forest model, especially for larger sample sizes. Similarly, Ma et al. (2020) compared CLH sampling with Feature Space Coverage Sampling (FSCS) and found that FSCS outperformed CLHS in terms of overall accuracy and consistency among different sample sets. However, it’s imperative to note that CLH sampling has been shown to be effective in numerous other studies, particularly in ensuring representative and stratified sampling that efficiently captures the variability of environmental covariates. The performance of CLH sampling can be dependent on specific data characteristics and modeling approaches. Despite certain limitations, CLH sampling’s ability to address the complex nature of soil properties through its stratified sampling and optimization criteria, while preserving multivariate correlation and original distributions, makes it especially suited for studies that are constrained by budget and require an efficient sampling strategy.

Therefore as the literature suggests, the first challenge of covering the variability in terms of feature space (environmental covariates) is satisfied by the CLH sampling, the most widely used sampling technique in the presence of ancillary variables. Minasny and McBratney (2006) proposed an efficient soil sampling method in terms of covering feature space based on Latin Hypercube Sampling (LHS). LHS is a method for drawing samples from multiple variables, considering their distributions. A stratified sampling procedure divides the variables into equally probable strata and extracts a single sample from each stratum. It works like a Latin Square where each row and column (corresponding to variable strata) contains a single sample. A generalisation of the Latin Square in higher dimensions is referred to as a Latin Hypercube.

LHS is not suitable for DSM as it may generate combinations of features that are not actually present in the data. According to a study by Minasny and McBratney (2006), only four out of ten sampling points selected using LHS had geographical relevance, while the remaining six did not match any actual locations. CLH sampling addresses this issue by repeatedly searching through the covariate distribution until the most ideal combination is chosen, resulting in a more accurately populated hypercube. Additionally, CLH sampling allows for more optimization criteria to be added while preserving multivariate correlation and the original distribution, making it an ideal choice for budget-constrained personnel looking to map soil properties.

However, the second challenge of selecting an appropriate number of sampling locations is a crucial aspect of DSM and is often based on expert knowledge, which may introduce the potential for human error. In recent years, studies have used the CLH sampling algorithm to select sampling points for various areas. However, as Table 1 illustrates, there is no clear relationship between sample size and the study area, as the sample size in DSM is determined by a combination of factors, including geographical space and variability in environmental covariates. For example, Brungard and Boettinger (2010) found that a sample size of
Table 1

<table>
<thead>
<tr>
<th>Study area (km²)</th>
<th>Metric</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 160 CEC, pH</td>
<td>SP</td>
<td>Wang et al., 2022</td>
</tr>
<tr>
<td>45 42 Salinity</td>
<td>PV</td>
<td>Khongnawang et al. (2022)</td>
</tr>
<tr>
<td>53.9 126 SOC</td>
<td>R²</td>
<td>Maleki et al., 2022</td>
</tr>
<tr>
<td>1140 210 SOC</td>
<td>R² / RMSE</td>
<td>Aqdam et al., 2021</td>
</tr>
<tr>
<td>46 150 PSF and AWC</td>
<td>R²</td>
<td>Chakan et al., 2019</td>
</tr>
<tr>
<td>6 100 SOC</td>
<td>CV</td>
<td>Rosemary et al., 2017</td>
</tr>
<tr>
<td>0.94 58 pH</td>
<td>Score</td>
<td>Taghizadeh-Mehrdadi et al. (2014)</td>
</tr>
<tr>
<td>207 217 Soil class</td>
<td>Kappa</td>
<td>Brungard et al. (2015)</td>
</tr>
<tr>
<td>190 103 Soil class</td>
<td>Kappa</td>
<td>Brungard et al. (2015)</td>
</tr>
<tr>
<td>300 300 Soil class</td>
<td>Kappa</td>
<td>Brungard et al. (2015)</td>
</tr>
<tr>
<td>296 57 Soil class</td>
<td>Kappa</td>
<td>Brungard et al. (2015)</td>
</tr>
<tr>
<td>10 70 SOC</td>
<td>R² / RMSE</td>
<td>Lacoste et al. (2014)</td>
</tr>
<tr>
<td>850 99 Soil class</td>
<td>Accuracy</td>
<td>Rad et al. (2014)</td>
</tr>
<tr>
<td>110 206 Soil suitability map</td>
<td>Accuracy</td>
<td>Van Zijl et al. (2014)</td>
</tr>
</tbody>
</table>

200–300 was optimal for a 300 km² area with an original population of approximately 30,000 sampling points, and Van zijd et al. (2014) concluded that the total number of soil samples to map a soil property or class depends on the heterogeneity of the site rather than its size. Additionally, Somarathna et al. (2017) found that increasing sample size improves the prediction accuracy of carbon, but at a decreasing rate.

In this study, the Bhattacharyya distance serves as the metric for determining the optimal sample size, as it quantifies the similarity between two probability distributions, making it ideal for comparing sample distributions with the original population. This geometric measure is symmetric and stable, particularly when the distributions have little overlap, and effectively concentrates on the overlap between distributions. On the contrary, Kullback-Leibler (KL) divergence, previously used in studies like Malone et al. (2019) and Saurette et al. (2023), is information-theoretic, and is more sensitive to differences in the tails of distributions, but lacks symmetry and can be numerically unstable. Additionally, Jensen-Shannon divergence (JSD) extends KL divergence by averaging the KL divergences between two distributions and their average, yielding a symmetric measure. However, akin to KL divergence, JDS is still information-theoretic and may face numerical instability for distributions with minimal overlap. Another metric, normalized variance used in Saurette et al. (2023), examines the dispersion in datasets by considering the variance of one distribution in relation to the variance of the other. While this metric is valuable for assessing distribution variability, it does not offer a direct gauge of similarity or overlap between distributions.

The choice of using the Bhattacharyya distance over metrics, like KL divergence and Jensen Shannon divergence because it does not focus on the differences, between two distributions but also takes into account their overlap or similarity. While both KL divergence and Jensen Shannon divergence compare probability density functions (PDFs), the main goal is to highlight a metric that gives a representation of the distribution. The objectives of this study necessitate a metric that not only is stable and geometric but also adeptly captures the similarity between distributions. The Bhattacharyya distance meets these criteria, and hence is chosen for its capability to accurately and robustly represent the key attributes of the population distribution in the sample set, which is essential for determining the optimal sample size in DSM.

Determining the optimum sample size in CLH sampling is a crucial aspect of DSM, particularly when time and money are limited. The goal is to reduce the sample size while still preserving the variability in feature and geographical space, and having enough samples to represent landscape attributes. This study proposes a methodology for selecting the optimum number of CLH samples for a specific site, with the Finzean Estate in Scotland being used as a case study.

In this research, the 'optimal sample size' is defined, as the point where it strikes a balance between representing the features and achieving optimal model performance. Precisely it refers to the sample size that minimizes the Bhattacharyya distance between the sample and the population and represents the variability and characteristics of the original dataset in the feature space. At the time this sample size aims to ensure that metrics such as, R², RMSE and others that indicate model performance are optimized. In addition, the study extends its applicability by including the analysis on DSM mapping of Organic Matter content percentage (LOI) and pH levels at the Estate based on the derived optimal sample size. This study aims to provide a generalizable approach that can be applied to other sites, regardless of their location. By using a range of sample sizes and comparing them to the original population using the Bhattacharyya distance, mean, standard deviation, estimates of the probability density function (pdf), and boxplots, this methodology allows for the determination of the optimal sample size for a given site with the case study of DSM to show its relevance. This study contributes to the field of DSM by providing a method that can be used to improve the efficiency and accuracy of soil mapping, which has implications for sustainable land management and food production worldwide.


2. Methodology

A rectangular grid of regular geographic data points was generated at a spacing of 50 m producing 48,411 observations or location points over the area of Finzean Estate. The observations lying outside of the Finzean Estate were disregarded. A final dataset (original population) of the remaining 17,932 records and 21 legacy features were used for sampling. The dataset was processed through an open-source QGIS software and then through the CHLs module.

It’s important to understand that the quantity of these covariates, in this case, 21, significantly impacts determining the optimal sample size. More covariates could have increased the model’s complexity and the computations required, potentially introducing issues of collinearity. However, it could also have provided a more comprehensive understanding of the data’s variability, leading to a more accurate model, provided that multicollinearity and overfitting issues were effectively managed.

On the other hand, fewer covariates would have simplified the model and reduced computational intensity. But, this might come at the cost of missing important explanatory variables contributing to the data’s variability, which could lead to a less accurate or oversimplified model. Therefore, regardless of the number, the relevance and meaningful contribution of the selected covariates to the model are crucial.

An open-source Python CLH sampling module written by Erica Wagoner Wagoner, E., & Zheng, Z. (2019), was produced in October 2019 based on Minasny and McBratney (2006). The module takes input parameters such as the predictors (original population), and the number of samples to be drawn, the number of observations to be drawn for each
sample (num_sample), the maximum number of iterations (In our study, 5,000 iterations were found to effectively minimize the objective function for all sample datasets, providing consistent results across datasets), along with other parameters which in this study were kept as default (see appendix). The module was executed in Google Collaboratory to extract 200 samples of different sizes from the original dataset (Fig. 1). A nested loop was used over a CLHs function to iterate within the range of 5 to 100 with a step size of 5 (5, 10, ..., 100); another loop was built, repeating over the same step size ten times (5_1, 5_2, ..., 5_10). Below, we provide the stepwise pseudocode representation for this operation:

1. Initialize an empty list named samples for storing the output from each iteration.
2. Begin an outer loop, iterating over a series of numbers ranging from 5 to 105, with increments of 5. Let’s denote each number in this series as num_sample.
3. For each num_sample value, initiate an inner loop that repeats 10 times.
4. Within each iteration of the inner loop, perform CLH sampling on population dataset, specifying num_sample as the sample size and setting the maximum number of iterations as 5000.
5. Append the resulting sampled data to the samples list.
6. Repeat steps 3–5 for each num_sample value in the outer loop.

After all iterations are completed, samples list will contain the output from each CLHs operation. All the density curves for continuous covariates (16 continuous variables) were approximated using kernel density estimation, which is the standard method to generate pdfs of a random variable in a non-parametric way (Silverman, 1986; Scott, 1992). The Python module gaussian_kde available in the scipy Python library was used to produce density curves in this study (Virtanen et al., 2020).

2.1. Location of the study

The study area, Finzean Estate, is a geographic region located in Birse, Aberdeenshire, Scotland. The area encompasses a total of 44.8 km² and is depicted in Fig. 2b. The Estate comprises eight farms, a farm shop, holiday cottages, a school, a village hall, and a church, and is inhabited by an estimated population of 400 individuals. The area was chosen for this study due to its diverse landscape and variability in soil classes, land cover and capability (Fig. 3). The area is characterized by diverse soil types, with the most prevalent soil type being Mineral Podzols (43.16%), Peat (17.24%), Immature Soil (12.5%), Brown soils (10.6%), Peaty podzols (9.69%), Alluvial soil (5.5%), and Mineral gleys (Hydromorphic soils) (1.24%). The Estate experiences an average annual precipitation of 992 mm, with a standard deviation of 129 mm. The mean annual temperature of the Estate was 10.22 °C, with a standard deviation of 2.26 °C. The warmest month being July, with an average temperature of 15.44 °C, and the coldest month is January, with an average temperature of 5.00 °C (Met Office, 2023).

To the north of the Estate, a narrow hilltop of low hills separates the Estate from the neighbouring Ballogie Estate. The southern and western part of the Finzean Estate is covered by granite hills, with the highest...
Fig. 2. A) Scotland boundary from UK census data b) band1 dataset (2 m) for the location of Finzean Estate (Table 2). c) Panoramic photograph of Finzean landscape showing different land cover.

Fig. 3. Soil class, land cover and land capability maps for the Finzean Estate.
being Peter Hill (617 m). This area is mostly covered with heather moor and peatland. The eastern valley of the Finzean Estate has a small patchwork of woodlands, scattered houses, small settlements and farmlands. Fig. 2 highlights the diverse landscape with various land cover, land capability, and soil classes within the Estate. The land capability map shows that most of the area in the estate can produce a narrow range of crops, with the rest of the area suitable for improved grasslands and rough grazing.

2.2. Environmental covariates and data sources in the study

The study utilized the SCORPAN model, which utilizes environmental covariates, to predict soil attributes and classes within the spatial extent of the Finzean Estate. A variety of continuous and categorical ancillary variables were utilized, including topography, soil classes, land capability, and land cover features, as described in Table 2. To generate ancillary variables were utilized, including topography, soil classes, land capability, and land cover features, as described in Table 2. To generate these features, the study used the Ordinance Survey Terrain 5 (OST 5) dataset, which provides a high-resolution digital elevation model (DEM) of the terrain in the United Kingdom. The dataset has a cell size of 5 m and is based on a combination of different data sources, including Light Detection and Ranging (LiDAR), Radio Detection and Ranging (RADAR), and stereo aerial imagery, to provide accurate and reliable information on the terrain’s elevation, and other topographical features. This dataset was used to calculate terrain derivatives, such as slope, aspect, roughness, the Multiresolution Valley Bottom Flatness (MRVBF) index, and the Terrain Roughness Index (TRI). Additionally, the Topographic Position Index (TPI) and general curvature were calculated using the OST 5 dataset.

Peat, an organic soil that covers approx. 25% of Scotland’s land cover, is characterized by properties like carbon availability, soil quality, and water storage capacity, which play a significant role in determining the overall attributes of the soil (Aitkenhead, 2017). To account for this, the study utilized a peat depth map of Scotland, which is available through the Natural Asset Register Data Portal (hutton.ac.uk). Additionally, soil series maps, produced by the James Hutton Institute (JHI) were used as environmental covariates to show the composition of 7 dominant classes in the region. These maps were used to understand the composition of soil series, in the study area.

Although the Land Cover Survey of 1988 (LCS88) map may have higher accuracy compared to the current Land Cover Map (LCM2019), it may not be suitable for sole use due to its age. Land cover can change rapidly over time, and using an outdated map may not accurately represent the current land cover conditions. Therefore, to ensure inclusion of recent changes in land cover, it is important to use a current land cover map such as the LCM2019 in conjunction with the LCS88 map. The aerial photography used in the study was originally obtained at a resolution of 25 cm under the license of JHI from GetMapping. However, for the purposes of the study, the resolution was adjusted to 5 m, to reduce the size of the image files and ensuring all the datasets used have same resolution (5 m) to facilitate the analysis. All the independent features were processed using open-source Quantum Geographic Information System (QGIS) software, along with plugins such as SAGA (SAGA User Group Association, 2011) and GRASS.

2.3. Bhattacharyya distance

In this study, the Bhattacharyya distance was used to analyze the relationship between the distribution of the original population and the extracted samples of different sizes. The Bhattacharyya distance ($\omega_b$) is a measure of similarity between two distributions which takes two distributions as input parameters. The Python function written by Williamson (2018) was used to calculate the distance for the continuous variables which is based on the Bhattacharyya coefficient ($\theta$), whereas for the categorical variables the script was written in python (See appendix). The distance was calculated for all covariate and sample sizes of CLH samples as:

$$\omega_b(a, b) = -\ln(\theta(a, b))$$

Where $\omega_b$ is the Bhattacharyya distance between two probability distributions, $a$ and $b$, and $\theta$ is the Bhattacharyya coefficient. Bhattacharyya coefficient $\theta$ can be evaluated for both discrete ($\theta_d$) and continuous ($\theta_c$) covariates as:

$$\theta_d(a, b) = \sum_{x} \sqrt{a(x)b(x)}$$

$$\theta_c(a, b) = \int \sqrt{a(x)b(x)} \, dx$$

Moreover, for the analysis the normalisation of all environmental covariates was done so they can be analysed on the same scale. The formulae used for the normalisation was:

$$B(x) = \frac{\omega(x)}{\omega_{max}}$$

Table 2

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Soil forming factor/Type</th>
<th>Mean</th>
<th>Range</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>r/Q</td>
<td>273.483</td>
<td>87.88 – 618 m</td>
<td>OS Terrain 5</td>
</tr>
<tr>
<td>Slope</td>
<td>r/Q</td>
<td>8.4628</td>
<td>0 – 43.699</td>
<td>From DEM</td>
</tr>
<tr>
<td>Flow accumulation</td>
<td>r/Q</td>
<td>1.37455</td>
<td>26 – 22900</td>
<td>From DEM</td>
</tr>
<tr>
<td>TPI</td>
<td>r/Q</td>
<td>0.0004</td>
<td>1.6037 – 1.23</td>
<td>From DEM</td>
</tr>
<tr>
<td>Roughness</td>
<td>r/Q</td>
<td>1.95</td>
<td>0 – 13.16</td>
<td>From DEM</td>
</tr>
<tr>
<td>Aspect</td>
<td>r/Q</td>
<td>160</td>
<td>0 – 359.99</td>
<td>From DEM</td>
</tr>
<tr>
<td>MBVF</td>
<td>r/Q</td>
<td>0.6865</td>
<td>0 – 5.88</td>
<td>From DEM</td>
</tr>
<tr>
<td>MRTF</td>
<td>r/Q</td>
<td>0.3418</td>
<td>0 – 4.97</td>
<td>From DEM</td>
</tr>
<tr>
<td>TRI</td>
<td>r/Q</td>
<td>1.87</td>
<td>0 – 11.63</td>
<td>From DEM</td>
</tr>
<tr>
<td>Peat depth (cm)</td>
<td>r/Q</td>
<td>28.55</td>
<td>0 – 158</td>
<td>Dataset - Natural Asset Register Data Portal (hutton.ac.uk) – dataset produced by M Aitkenhead</td>
</tr>
<tr>
<td>Curvature</td>
<td>r/Q</td>
<td>– 0.00095</td>
<td>– 0.188 – 0.200</td>
<td>From DEM</td>
</tr>
<tr>
<td>Land capability</td>
<td>o/C</td>
<td>5 classes</td>
<td>JHI LCA map (National scale land capability for agriculture</td>
<td>Scotland’s soils (environment.gov.scot))</td>
</tr>
<tr>
<td>Land cover</td>
<td>o/C</td>
<td>12 classes</td>
<td>LCM2019 and LCS88</td>
<td></td>
</tr>
<tr>
<td>Soil series</td>
<td>o/C</td>
<td>7 classes</td>
<td>JHI soil map of Scotland (National soil map of Scotland</td>
<td>Scotland’s soils (environment.gov.scot))</td>
</tr>
<tr>
<td>Major soil groups in Scottish soil classification system</td>
<td>o/C</td>
<td>13 classes</td>
<td>JHI soil map of Scotland (National soil map of Scotland</td>
<td>Scotland’s soils (environment.gov.scot))</td>
</tr>
<tr>
<td>Band 1 (2 m)</td>
<td>o/D</td>
<td>76.44</td>
<td>8 – 242</td>
<td>GetMapping (Getmapping</td>
</tr>
<tr>
<td>Band 2</td>
<td>o/D</td>
<td>72.54</td>
<td>10 – 240</td>
<td>GetMapping</td>
</tr>
<tr>
<td>Band 3</td>
<td>o/D</td>
<td>65.71</td>
<td>0 – 232</td>
<td>GetMapping</td>
</tr>
</tbody>
</table>

$r$ = relief, $p$ = parent material, $o$ = organisms, $Q$ = quantitative, $C$ = categorical, $D$ = discrete.
Normalised value \( \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \)

2.4. Case study

Following the determination of the optimal sample size using CLH sampling, the results proceeded to fit the machine-learning models to the real-world dataset. This dataset comprised soil properties, specifically pH and LOI (%), from the Finzean Estate in Scotland. The goal of this application was to demonstrate the practical utility of our methodology and to evaluate its performance in a real-world context.

To conduct the DSM, we utilized three different machine learning models: Multi Linear Regression (MLR), CART (Classification and Regression Trees) Decision Tree, and Random Forest. These models were chosen for their unique strengths and their ability to provide diverse insights into the dataset. Multi Linear Regression (MLR) was chosen for its simplicity and interpretability. It is a straightforward model that assumes a linear relationship between the independent and dependent variables, making it easy to understand and interpret. It also provides a baseline against which to compare the performance of the other, more complex models.

The Decision Tree model, on the other hand, was selected for its ability to handle non-linear relationships and interactions between variables. It is a flexible model that can capture complex patterns in the data, and its hierarchical structure makes it relatively interpretable compared to other complex models.

Fig. 4. Comparative analysis of sample means with error bars (10 repetitions) and true means across different variables and sample sizes.
Finally, the Random Forest model was chosen for its robustness and accuracy. As an ensemble method that combines multiple decision trees, it is less prone to overfitting and typically provides more accurate predictions than a single decision tree. It also has the ability to handle large datasets with many variables, making it particularly suitable for our dataset.

The process of model fitting involved training and validating each model using k-fold cross-validation using 6 folds. This technique involves partitioning the original sample into a training set to train the model, and a validation set to validate it. This rigorous method ensures model robustness by validating it against different portions of the dataset.

The performance of each model was assessed using several metrics. The R-squared metric, also known as the coefficient of determination, was used to measure the proportion of the variance in the dependent variable that is predictable from the independent variables. The Mean Absolute Error (MAE) was used to measure the average magnitude of the errors in a set of predictions, without considering their direction. The Root Mean Squared Error (RMSE) was used to measure the differences between values predicted by the model and the values actually observed. This comprehensive methodology allowed us to delve deep into the dataset and extract valuable insights. The results of this application are presented in the following section.

![Graphs showing standard deviation and error bars for different variables and sample sizes.](image)

**Fig. 5.** Comparative analysis of sample standard deviation with error bars (10 repetitions) and true standard deviation across different variables and sample sizes.
3. Results

Fig. 4 illustrates the relationship between sample size and the mean of various continuous covariates. The mean was calculated for each covariate using samples of increasing size, with the original sample (17,932) serving as the true line in red. The blue line refers to the error bar representing the variation of mean across the 10 repetitions of the same sample size. As the sample size decreases, it can be observed that the mean of smaller samples deviates more significantly from the original mean. This is especially evident in the case of the flow accumulation covariate, where a sample size of 15 results in a 34% change in mean relative to the original, for 40 it was 7%, while for a sample size of 100 results in only a 9% change. Similarly, for the covariate slope, a sample size of 5 results in a 21% change in mean relative to the original, for 40 it was 2.5% while for the sample size of 100 results in only a 1.5% change. The plots show that increasing the sample size leads to convergence towards the original mean for most of the covariates, reducing variability in the mean and improving its accuracy. However, it should be noted that for some of the covariates such as TPI, the plots may not appear to converge due to the narrow range of values for the covariate, making it difficult to observe changes in the mean as sample size increases.

An analysis of the estimates of probability distribution function of various continuous variables was conducted using sample sizes (10, 20, 30, ..., 100). The density distribution of each covariate was plotted using the seaborn library in Python, as depicted in Fig. 6. The results indicate

In Fig. 5, the relationship between sample size and the standard deviation of various covariates is presented. The standard deviation was calculated for each covariate using samples of increasing size. The results show that the change in standard deviation is relatively low for smaller sample sizes but becomes more significant as the sample size increases. For example, for the covariate elevation, the fractional change relative to the original population standard deviation is 12% for a sample size of 5, 80% for a sample size of 10, 73% for a sample size of 40, 69% for a sample size of 70, and 56% for a sample size of 100. Similarly, for the covariate easting, the fractional change is 14% for a sample size of 5, 70% for a sample size of 10, 68% for a sample size of 40, 67% for a sample size of 70, and 58% for a sample size of 100. The results also show that there is no significant effect on the standard deviation when increasing the size of samples. This is because, in most of the covariates, the standard deviation distribution remains similar as the sample size increases. A large drop is seen when the sample size is increased from 5 to 10. However, further increasing the sample size results in smaller changes in the standard deviation.

An analysis of the estimates of probability distribution function of various continuous variables was conducted using sample sizes (10, 20, 30, ..., 100). The density distribution of each covariate was plotted using the seaborn library in Python, as depicted in Fig. 6. The results indicate...
that while the small sample sizes may not have been able to fully capture the entirety of the continuous variables, they were still able to accurately represent the original distribution. However, it can be observed from the figure that the distribution curve for the lower sample sizes deviated from the original distribution, while the distribution curves for larger sample sizes were more similar and overlapped each other. The best approximation of the original distribution is different for different covariates but largely lies in the range of 40–50. The probability distribution is dependent on the sample size and the best approximation of the original distribution varies for each covariate, depending on the shape and range of the distribution.

The boxplots in Fig. 7 provide insight into the distribution of continuous environmental covariates at different sample sizes (5, 10, 15, …, 100). Boxplots are a graphical representation that provides a summary of the distribution of numerical data, including the spread, central tendency, and skewness. The ends of the boxplot, known as whiskers, indicate the minimum and maximum values of the distribution. The lower end of the box represents the first quartile (Q1), which indicates that 25% of the data is below that point, while the upper end represents the third quartile, which shows that 75% of the data is below that point. The middle line within the box, known as the median (Q2), divides the data into two halves. By examining the spread, central tendency, and skewness of the data, as depicted by the position of the median, quartiles, and whiskers, it is evident that increasing the sample size can improve the approximation of the original distribution to a certain extent. For example, as the sample size increases, the boxplot becomes more tightly grouped around the median, indicating a decrease in the spread of the data. However, beyond a certain point, there is no significant change in the distribution, as evident by the stability of the quartiles and median. This suggests that increasing the sample size beyond this point does not provide additional benefits in approximating the original distribution.

The Bhattacharyya distance, a measure of similarity between two probability distributions, was calculated between the original distribution of each covariate and the distribution of each sample size. Cut-off points or elbow points were calculated using the first derivative of the Bhattacharyya distance. The cut-off points mark the juncture of a significant change in the rate of Bhattacharyya distance similar to a bend in a curve, cut-off points assisted to determine the sample sizes that capture the critical change shown in the form of a heat map in Fig. 8. Heatmap was used to visually analyze how elbow points varied with respect to different variables and thresholds. Various thresholds from 0.04 to 0.004 were examined to systematically understand the scale at which the change in the distance becomes substantial. The threshold of 0.004 was ultimately selected for further analysis as the aim was to capture even the smallest changes across the majority of the variables. Fig. 9 was plotted showing the line graph of the average Bhattacharyya distance for 10 repetitions with respect to sample sizes with a threshold of 0.004 as a reference point. In the case of the covariate slope, it can be observed that the Bhattacharyya distance decreases as the sample size increases. For example, for a sample size of 5; the distance is 0.21; for a sample size of 10; it decreases to 0.14; for a sample size of 15 it was 0.07; for a sample size of 25 it decreases to 0.05; for a sample size of 40 it was 0.04; for a sample size of 80 it was 0.03; and for a sample size of 100; it was 0.02. The graph shows that the distance is higher for smaller sample sizes and decreases as the sample size increases.

It can be observed from the Fig. 8 that the cut-off point for different variables at the threshold value of 0.004 was different. However, it is noteworthy variables like Easting and Land Cover (1988) required the highest sample size which was 50 and MRRTF required the lowest at 25. Based on these findings, further analysis involving the creation of machine learning models was carried out using a dataset 50 sampling location shown in Fig. 10.

3.1. Performance of machine learning model

Our application of the methodology to the real-world dataset yielded insightful results. We fitted the DSM models to the optimal sample size (50 samples) for the soil properties of pH and Organic matter content (%) (LOI), and assessed their performance using the R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) metrics. The digital soil maps produced with the help of best performing model (random forest) are shown in Fig. 11.

The performance of the three machine learning models varied. The
results, as shown in Table 3, indicate that the Random Forest model consistently outperformed the MLR and Decision Tree models in predicting both pH and Loss-On-Ignition (LOI).

For LOI, the Random Forest model achieved an R-squared value of 0.69, indicating that it was able to explain 69% of the variance in the LOI values. The MAE and RMSE values were 0.14 and 0.18 respectively, indicating a high level of accuracy in the predictions.

For pH, the Random Forest model achieved an R-squared value of 0.58, indicating that it was able to explain 58% of the variance in the pH values. The MAE and RMSE values were 0.12 and 0.15 respectively, also indicating a high level of accuracy in the predictions.

To address the optimality of the selected sample size, the performance of the Random Forest model was further tested across different sample sizes, ranging from 5 to 50 (step size of 5). For each sample size, we conducted 10 repetitions and then computed the average of the metrics. Fig. 12 shows the variation of R-squared, MAE, and RMSE with the increasing sample size. Notably, the R-squared value displayed an increasing trend and reached its peak at the sample size of 50, confirming the appropriateness of our selected sample size from the covariate analysis.

These results demonstrate the effectiveness of our methodology in a real-world context. The optimal sample size selected using cLHS was able to support accurate predictions of soil properties using DSM models, particularly the Random Forest model. This underscores the potential of our methodology for practical applications in the field of soil science and beyond.

3.2. Discussion

The accuracy of a soil prediction model is dependent on the coverage of the geographic (Easting and Northing) and feature space by the sampling points, and the type of prediction model used. However, selecting the appropriate sample size and creating the prediction model requires values of dependent covariates, such as soil properties and...
classes, which are obtained from laboratory measurements and legacy datasets before the sampling stage. This can be a limitation when conducting a soil survey, it’s worth noting that there are instances (Malone et al., 2011) in the literature where DSM maps have been crafted effectively without relying solely on these datasets.

Previous research by Brungard and Boettinger (2010) studied the optimum sample size using only soil-forming factors. They compared the distribution of these factors in the sample to the population distribution and chose the smallest sample size that closely represented the original distribution as the optimum size. However, the method does not use any
Table 3
Performance metrics of different machine learning models for predicting various soil properties.

<table>
<thead>
<tr>
<th></th>
<th>R2</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOI</td>
<td>0.36</td>
<td>1.54</td>
<td>2.07</td>
</tr>
<tr>
<td>MLR</td>
<td>0.27</td>
<td>2.22</td>
<td>2.78</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.49</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.58</td>
<td>0.12</td>
<td>0.15</td>
</tr>
</tbody>
</table>

quantitative parameters to measure the closeness of the sample distribution to the original distribution.

The current study aims to investigate the effect of increasing the sample size on the mean, distribution, standard deviation of continuous covariates and Bhattacharyya distance of both continuous and categorical covariates. Results show that as the sample size increases, the means of the continuous covariates become closer to the original mean, and the change in standard deviation is less for smaller sample sizes but increases as the sample size increases. Additionally, the study found that the best approximation of the original distribution is different for different covariates but lies in the range of 25–50. Overall, the study suggests that increasing the sample size can benefit the approximation of the original distribution to a certain extent, but there is a small effect on the distribution beyond that point.

This study also emphasises how crucial it is to comprehend the fundamental properties of the data and how they could influence the choice of the ideal sample size. For instance, the study discovered that the ideal sample size differed depending on the particular covariate being examined, which raises the possibility that not all data and research problems can be addressed using a general method.

The work of Stumpf et al. (2016) indicates that the optimization of sampling methods needs to strike a balance between effective estimate precision and the cost and difficulty of data collection. This philosophy is fundamental in our study as well, emphasizing the necessity of having an efficient and cost-effective CLH sampling design. However, their study used variance comparison between sample and population to determine the ideal sample size whereas this study utilizes the Bhattacharyya distance to gauge the similarity between the sample and population.

The study agrees with Malone et al. (2019) which emphasis on identifying an appropriate sample size when using CLH sampling in field sampling surveys. To achieve this, Malone et al. (2019) introduced KL divergence metrics as comparisons tools and found larger samples improved representation to a certain degree.

In addition, Saurette et al. (2023) investigated these metrics within the framework of converting data into pdfs with an emphasis, on binning techniques. They discovered that the metrics were significantly influenced by the number of bins chosen while the covariates did not have an impact. This highlights the significance of selecting a binning strategy to ensure an accurate representation of the data. Our study differs in the approach as it employs kernel density estimation (KDE), a non-parametric method for estimating the probability density function of a variable. KDE has its strengths, as it does not necessitate a predetermined number of bins, which could offer a truer representation of the underlying distribution. This also discards the arbitrary choice of bin size and location reducing the chance of bias. In contrast working with categorical features was different, as categorical data is inherently naturally discretized with each category displaying a bin, this approach able to effectively interpret the categorical data in its most natural form.

The study investigates the importance of finding a balance between the cost and effort involved in data collection and the need for accuracy in soil prediction models. Although a larger sample size may improve accuracy, it requires more time and resources for data collection and processing. Therefore, it is essential to carefully consider the trade-offs between accuracy and sample size when determining the optimal sample size for a particular research question (Somarathna et al., 2017). Overall, the approach described in this study to determine the optimal sample size using the CLH sampling and the Bhattacharyya distance can be applied to other fields beyond soil science, such as ecology, environmental science, and geology, where similar challenges of sampling and predicting spatial data exist. The study has simplified the application of this approach by sharing the code used in the analysis, which can help advance our understanding of the complex relationships between spatial data and environmental processes.

The primary limitations of our study stem from the focus on marginal distributions and our exclusive reliance on a single sampling design. Our approach, while providing detailed insights into individual variables, doesn’t take into account potential interdependencies and interactions between multiple variables that multivariate distributions could reveal. This may influence the accuracy of the DSM and the sampling distribution. In addition, the use of a single sampling design could potentially bias the outcomes due to the specific characteristics of this design. A comparison with different sampling designs might have enhanced the robustness of our findings. Future research should aim to address these limitations to enhance the depth and applicability of our results.

3.3. Conclusion

This study introduces a methodology to ascertain the optimal sample size and location reducing the chance of bias. In contrast working with categorical features was different, as categorical data is inherently naturally discretized with each category displaying a bin, this approach able to effectively interpret the categorical data in its most natural form.
size for capturing feature space in Digital Soil Mapping (DSM) through Conditioned Latin Hypercube (CLH) sampling. The analysis revolves around a case study conducted over an area of 44.8 km² in Scotland. Herein, ‘optimal sample size’ signifies the size where the sampled set best captures the variability and characteristics inherent to the entire dataset. An analysis was conducted on a set of 200 samples, each of varying sizes and distributions, which were derived from the original dataset.

In summary, the study indicates that increasing sample size can improve the accuracy of the original distribution, but only up to a certain extent. The mean and standard deviation of the covariates were found to change as the sample size increased. The study also suggests that a sample size of 50 is sufficient to provide a good approximation of the original distribution. Additionally, small sample sizes can still provide a representative distribution despite not covering all soil classes, land cover, and land use.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was conducted with financial support from the Research-led Innovation Nodes for Contemporary Society (RLINCS) programme of Abertay University, Scottish Government’s RESAS strategic programme and the Norman Fraser Design Trust. We acknowledge support from farmers at Finzean Estate for access to the farms.

References
