# Operational sampling designs for poorly accessible areas based on a multi-objective optimization method

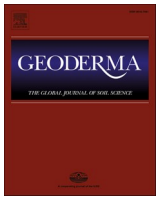
Maxime Dumont, Guilhem Brunel, Paul Tresson, Jérôme Nespoulous, Hassan Boukcim, Marc Ducousso, Stéphane Boivin, Olivier Taugourdeau, Bruno Tisseyre

Geoderma 445 (2024) 116888

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



# Operational sampling designs for poorly accessible areas based on a multi-objective optimization method

Maxime Dumont [a,b,*], Guilhem Brunel [b], Paul Tresson [a,c], Jérôme Nespoulous [a], Hassan Boukcim [a], Marc Ducousso [d], Stéphane Boivin [a], Olivier Taugourdeau [a,e], Bruno Tisseyre [b]

[a] VALORHIZ, 912 Rue de la Croix Verte, 34090 Montpellier, France
[b] ITAP, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France
[c] UMR AMAP, CIRAD, Montpellier, France
[d] AGAP, Univ Montpellier, CIRAD, INRAE
[e] Service Line Environnement, Egis Group, 34000 Montpellier, France

ARTICLE INFO

Handling Editor: Budiman Minasny

*Keywords:*
Soil sampling
cLHS
Field constraints
Pareto optimality
Digital Soil Mapping

ABSTRACT

Sampling for Digital Soil Mapping is an expensive and time-constrained operation. It is crucial to consider these limitations in practical situations, particularly when dealing with large-scale areas that are remote and poorly accessible. To address this issue, several authors have proposed methods based on cost constraints optimization to reduce the travel time between sampling sites. These methods focused on optimizing the access cost associated to each sample site, but have not explicitly addressed field work time required for the whole sampling campaign. Hence, an estimation of fieldwork time is of great interest to assists soil surveyors in efficiently planning and executing optimized field surveys. The goal of this study is to propose, implement and test a new method named Multi-Objective Operational Sampling (MOOS), to minimize sampling route time, while ensuring that sample representativeness of the area is maintained. It offers multiple optimal sampling designs, allowing practitioners to select the most suitable option based on their desired sample quality and available time resources. The proposed sampling method is derived from conditioned Latin Hypercube sampling (cLHS) that optimizes both total field work time (travel time and on-site sampling time) and sample representativeness of the study area (cLHS objective function). The use of a multi-objective optimization algorithm (NSGA II) provides a variety of optimal sampling designs with varying sample size. The sampling route time computation is based on an access cost map derived from remote sensing images and expert annotation data. A least-cost algorithm is used to create a time matrix allowing precise evaluation of the time required to connect each pair of sites and thus determine an optimal path. The proposed method has been implemented and tested on sampling for $pH_{H_2O}$ mapping within a 651 points kilometric grid in the northern part of Saudi Arabia, where soil analyses were conducted over a 1,069 $km^2$ area. MOOS method was compared to two other common approaches: classical cLHS and cLHS incorporating access cost. The performance of each method was assessed with the cross-validated RMSE and sampling route time in days. Results show that the MOOS method outperforms the two others in terms of sampling route time, especially with increasing sample size, gaining up to 1 day of work for the presented case study. It still ensures a relevant map accuracy and sample representativeness when compared to the two methods. This approach yields promising outcomes for field sampling in digital soil mapping. By simultaneously optimizing both sample representativeness and cost constraints, it holds potential as a valuable decision support tool for soil surveyors facing sampling designs in poorly accessible areas.

## 1. Introduction

Ensuring soil security is crucial to address contemporary challenges such as climate change, soil erosion, desertification or biodiversity loss (Koch et al., 2013). Thus, from local to global scale, mapping soil properties related to soil security (de Gruijter et al., 2016; Kidd et al., 2015b; Koch et al., 2015; Stockmann et al., 2015) is a key issue for a better management and monitoring of anthropized or natural terrestrial

* Corresponding author at: VALORHIZ, 912 Rue de la Croix Verte, 34090 Montpellier, France
*E-mail address:* maxime.dumont@valorhiz.com (M. Dumont).

https://doi.org/10.1016/j.geoderma.2024.116888
Received 31 January 2024; Received in revised form 9 April 2024; Accepted 12 April 2024
Available online 18 April 2024

ecosystems. Digital Soil Mapping (DSM) is one of the main leverage to contribute to soil security (Arrouays et al., 2021). It typically allows to predict soil properties at unseen locations by using (geo)statistical or machine learning (ML) models calibrated with a finite number of sampling sites. This relies on the relation of soil property to map with environmental covariates (geomorphological, soil, climatic data). McBratney et al. (2003) introduced the "scorpan" framework, describing the environmental covariates that are good proxies of soil forming factors.

Field sampling is a first and most critical step for any new DSM project. The representativeness of the collected data over the study area and the sample size significantly impact the quality of the resulting maps (De Gruijter et al., 2006). If no relevant environmental data is available, a sampling design can be representative of the soil property by spreading the sample sites all over the study area by using a simple random sampling (SRS) or a local pivotal method (Grafström et al., 2012). If environmental data that capture well the soil property to map are available, sampling design would rather be done by spreading the samples in the covariates space with methods such as conditioned Latin Hypercube Sampling (cLHS) (Minasny and McBratney, 2006) or k-means sampling (Brus, 2019). With the recent advent of high-resolution and widely available spatial data, the second option has become the most popular in the DSM community.

Sampling is often costly and is allocated a specific, limited timeframe within a mapping project. In poorly accessible areas such as deserts, sampling is even more constrained due to the lack of road networks and rugged terrain. Yet, arid and semi-arid deserts cover a third of the earth's land surface and are crucial as they host 20 % of the world's plant diversity and contribute significantly to carbon storage (Prăvălie, 2016). Thus, facilitate sampling in such constrained conditions is of great importance to increase knowledge about these environments. To solve this challenge, several authors introduced sampling methods optimizing operational constraints with different approaches. Cambule et al., (2013) proposed to sample in easily accessible areas with comparable environmental covariates to those of poorly-accessible areas. It aimed to create a model that would translate well to poorly-accessible areas. Clifford et al., (2014) adapted the cLHS method to maximize a geographical spread of sampling sites and minimize cost to reach each site (computed as the sum of distances of each site to reach the closest road). Sena et al., (2021) combined cLHS with a similarity method based on k-means to propose alternative sampling sites in case some sites are inaccessible. cLHS relies on a single criteria optimization algorithm (simulated annealing) which iteratively increases sample representativeness so that the sample presents a covariate distribution close to the distribution of covariates over the whole study area. This makes it prone to be adapted by adding an operational constraint objective to it. This approach was used by (Roudier et al., 2012) and is of paramount interest. These authors added a cost function to the cLHS algorithm to select sampling sites in the most easily accessible areas. A cost map describing the "ease of reach" (in arbitrary unit) of each point in the landscape was created by modeling the constraints with friction (i.e., difficulty to reach a point associated to each land type), slope and distance to near roads. At each iteration of the algorithm both representativeness and cost are optimized to obtain an optimal sampling design. However, as pointed out by Roudier et al. (2012), the proposed method calculates the cost of visiting each point independently, without accounting for the whole route. This does not allow to precisely estimate the cost of a sampling design as the route between each site and starting site highly affect the total cost. In the same manner, an optimal order of sites for sampling cannot be provided to support the practitioner in planning the sampling campaign. Additionally, as the cost is defined in arbitrary unit, the method cannot be used directly in assessing the operational time required to actually perform the sampling. Finally, it aims at optimizing both sample representativeness and operational constraints without considering it as a bi-criteria optimization problem. Indeed, there is not only one but several optimal solutions when dealing with multiple objectives. Not using the appropriate optimization method might lead to the selection of suboptimal solutions. Meanwhile, multi-objective optimization has already been used for sampling in DSM or Precision Agriculture for other aims (Israeli et al., 2019; Li et al., 2022), but never for optimizing operational constraints.

This article proposes a new method for creating optimized sampling designs adapted to poorly accessible areas. It relies on NSGA-II, a multi-objective optimization algorithm, which allows both a sample representativeness criterion of a study area and an operational criterion to be optimised. In this study, the sample representativeness criterion used will be the same as in cLHS. The operational criterion will be the total fieldwork time (expressed in days), as a sum of the time spent for sampling route and on-site sampling time. Sampling route time is created by taking into account the cost to reach each point sequentially (as proposed by Roudier et al., 2012) and expressing this cost as a speed (i.e., the time to cross each spatial unit of the study are) instead of a unit free "friction". On-site sampling time is determined by the practitioner, based on specificities of the study. This makes this operational criterion a relevant estimation of total fieldwork time, an indicator directly usable for decision-making. This method will be referred as Multi-Objective Operational Sampling (MOOS) in the rest of the document.

The objectives of this study were as follows: i) develop MOOS, a multicriteria optimization approach that integrates both the sample representativeness of available covariates and total fieldwork time, (ii) perform a comparative evaluation of the MOOS method with more common methods presented in the scientific literature to assess its effectiveness; (iii) apply and validate this approach in a real operational scenario to confirm the results and identify any potential limitations.

To do so, MOOS was applied and assessed on a pH $H_2O$ mapping project taking place in an arid and poorly accessible area in Al Ula County (Medina Province, Saudi Arabia). To assess the potential of MOOS in this situation, it has been compared to the classical method cLHS and the adapted version as proposed by Roudier et al. (2012), and referred as $cLHS_{cost}$. The three approaches were compared on two different metrics, the pH map prediction error (RMSE) and route sampling time (in days), for various sampling size. Then, MOOS was evaluated within the same mapping project, simulating the approach that soil surveyors would take in planning an actual sampling campaign.

## 2. Material and methods

### 2.1. Case study and data

The study area covers 1,069 $km^2$ of the Al Ula region (Medina Province, Saudi Arabia). The area has an arid climate, with an average daily maximum temperature of 45 °C in summer and a minimum of 10 °C in winter. A rainfall gradient is observed over the area, with the western part receiving less than 30 mm/year and the south eastern part experiencing the highest rainfall at 170 mm/year (Nazzal et al., 2014).

The Al Ula region is qualified as poorly accessible since its geomorphology generates numerous route constraints (Fig. 1a.). Indeed, canyons with steep slopes predominate in the centre, as well as in the centre-east and the north-east parts, while a volcanic harrat plateau with steep borders covers the whole western part. The rest of the region is mostly covered by sand which can also affect the ease of access, depending on the presence of dunes or rocks. In plains, navigation can be constrained in stony colluviums due to the presence of large stones. Moreover, road network mainly connects the urban and rural zones located in the west and the centre of the region. Only one road links the north to the south on the eastern part of the region. For effective navigation, practitioners employ a four-wheel-drive vehicle on roads, tracks, or drivable sand. On all other land types for which driving is not possible, travel is made by walk. Note that due to terrain steepness in canyons and near harrat plateau, some zones are considered as totally inaccessible.

Soil samples were obtained from a kilometric grid covering the Al

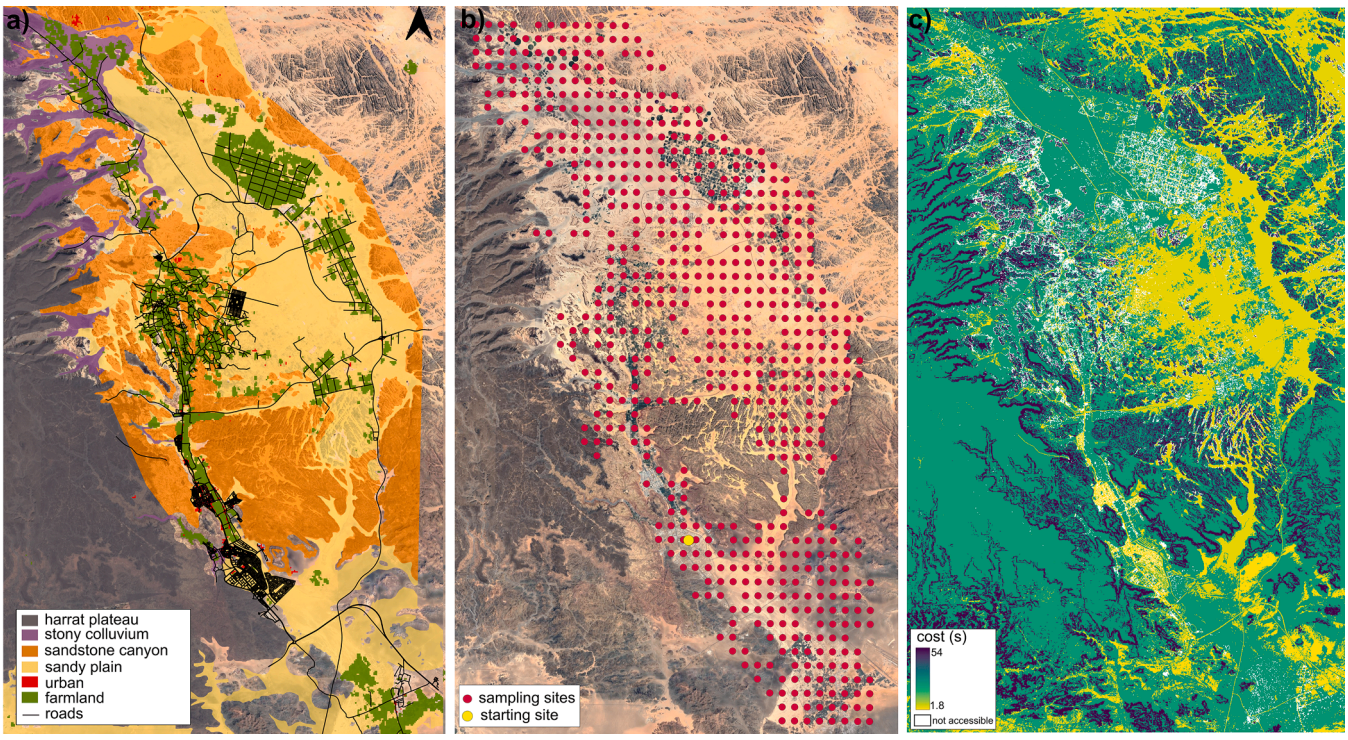



**Fig. 1.** General description of the study area with a) geomorphologic characteristics and land occupation, b) location of project's sampling sites and starting site and c) time cost map derived from geomorphic characteristics and land occupation.

Ula region. The sampling was done in two campaigns, during autumn 2019 and autumn 2020. Fig. 1.b. shows the soil sampling sites, note that the whole region was not sampled since it was not possible to reach inaccessible areas. The dataset includes 651 monitoring sites, each located at the center of a 1 x 1 km cell. Detailed descriptions of soil profiles, physico-chemical and microbial characteristics, site environment, location, vegetation, and land management were conducted for each site. The analyzed soil samples were a composite of 5 samples collected on a depth of 30–40 cm across a 1 x 1 m plot. For this study, among all available properties, only the pH $H_2O$ was considered. This is a commonly used indicator of soil chemical properties (Sparks et al., 2024). It plays a crucial role in several soil functions such as bioavailability of nutrients, physical structure or biological activity (Neina, 2019) and is strongly linked to former agricultural use of the land in arid regions (Maleki et al., 2021).

Soil pH was measured using a pH-meter (pH Meter Knick 766) in 1:5 of deionized water (i.e., pH $H_2O$). Overall, pH $H_2O$ of the zone was found to be highly alkaline compared to other arid lands deserts with values ranging from 7.5 to 9.5. This was explained by the volcanic plateau minerals migrating because of erosion.

One location (marked as a yellow point on Fig. 1.b.) was considered as the starting point from which a practitioner starts a sampling campaign. This was the actual accommodation of the practitioners; it is located in the Al Ula urban area with all amenities.

11 available environmental covariates were considered to describe soil pH $H_2O$ variability over the study area. Covariates were directly derived or computed from available spatial data: SRTM Digital Elevation model (30 m resolution), Sentinel-2 L1-C (10 m resolution) and Sentinel-1c-band (5 m resolution). Without any references on mapping pH in arid lands, covariates were selected following methodologies as proposed in the literature either for large-scale soil pH mapping (Lu et al., 2023) or for soil mapping in arid lands with optical remote sensing (Elhag, 2016; Taghizadeh-Mehrjardi et al., 2021). Studies aiming at estimating bare soil surface texture or soil moisture were also considered (Niang et al., 2014; Yang et al., 2019)

Topographic indexes were computed from the DEM with SAGA GIS software (https://saga-gis.sourceforge.io) (Conrad et al., 2015). Spectral indices were obtained from Sentinel 2 using "rasterio" Python package (Rasterio, access to geospatial raster data — rasterio documentation).

When possible, the acquisition date of remote sensing data was chosen in august 2019, just before the beginning of the actual sampling campaign.

### 2.2. Literature methods

The proposed method (MOOS) was compared to two other sampling methods, namely conditioned Latin Hypercube Sampling (cLHS) (Minasny and McBratney, 2006) and a variant of this latter, incorporating operational constraints ($cLHS_{cost}$) (Roudier et al., 2012)). Specifically designed for sampling in the presence of environmental covariates, cLHS uses an iterative optimization process based on simulated annealing. The aim of this iterative process is to optimize an objective function ensuring the representativeness of the sample across covariates. The cLHS objective function ($cLHS_{fObj}$) is the sum of three sub-parts called hereafter Oi,i = 1:3. $O_1$ assesses the resemblance of the sample distribution over numeric covariates to the overall data covariates distribution, $O_2$ performs a similar assessment for categorical covariates, and $O_3$ evaluates if the correlation of sampled covariates replicate those of the entire dataset. Compared to cLHS, $cLHS_{cost}$ includes an operational cost optimization. Prior to implementation, the cost associated to each potential sample site must be defined. According to Roudier et al. (2012), this was done by computing the distance of each potential sampling site to the closest road, taking into account the time cost map displayed in Fig. 1.c (the process to obtain this cost map is described later in 2.3.2). Finally, the cost of one sampling design equals to the sum of the cost of each individual site, as proposed by the author.

For both methods, the number of iterations was set to 1,000. In this study, the Python package "clhs" ("cLHS: Conditioned Latin Hypercube Sampling — clhs 1.0.0 documentation," n.d.) was used as is for classical cLHS, and modified to integrate the cost functionality, aligning with the existing R package (Roudier, n.d.).

### 2.3. Multi-objective operational sampling (MOOS)

#### 2.3.1. General methodology

This section provides a general description of MOOS. Details about each component of the method are further described in the following sections.

The general principle of MOOS relies on the iterative optimization of sampling designs regarding two important criteria: sample representativeness of covariates and total time spent to perform the sampling campaign (referred as total fieldwork time). It is assumed that all relevant covariates are used for sampling, and capture well the variability of the parameter to map. Following the same idea of cLHS and $cLHS_{cost,}$ the initialization starts by creating random sampling designs. Then, those sampling designs are iteratively modified. At each iteration, the new sampling design is kept if it is more optimal than the precedent one regarding both criteria simultaneously.

Total fieldwork time comprises the time spent to reach all sampling sites, starting by a starting site; and the time spent for soil sample collection at each sampling site. These two different components of total fieldwork time are respectively referred as sampling route time and on-site sampling time.

Sampling route time requires two first steps before being computed: i) creating a time cost map (i.e., a map of the study area with each pixel corresponding to the time required to cross it), ii) creating the least cost paths between each potential sampling sites and store it in a matrix. During the optimization, the matrix is used at each iteration to determine the optimal order of sampling sites and therefore have an estimate of sampling route time.

Compared to $cLHS_{cost}$, mapping the "ease of access" with a time instead of a "friction" and considering the whole route to reach each sampling sites allows to estimate the actual time spent by a practitioner. The on-site sampling time was estimated by the practitioners.

The sample representativeness criterion was the same as the one used for cLHS and $cLHS_{cost}$ methods.

Finally, contrarily to cLHS and $cLHS_{cost}$, the optimization was not performed with simulated annealing, but with a multi-objective optimization process, described in 2.3.4. When optimizing several criteria there might be not only one optimal solution, but several. The multi-objective optimization is adapted to solve a problem with more than one criterion. For the case of field sampling, it allows to select a set of sampling designs that are optimal regarding both sample representativeness and fieldwork time.

#### 2.3.2. Total fieldwork time estimation

Total fieldwork time (in days) comprises sampling route time and on-site sampling time. Prior to optimization, creating the data for sampling route time computation was done in two steps. First, a time cost map was created by gathering all data that gives information about possible route constraints (i.e., land occupation, administrative data, slope). These data were discretized in order to assign to each pixel, the time required to cross it. Secondly, the least cost paths between each pair of potential sampling sites were computed using the classical path search algorithm "A-star" (Hart et al., 1968). This allowed to estimate the time an operator would spend to go from one site to another, avoiding most constrained areas. Route time was then stored in a matrix which was used during the optimization step. A cell of the matrix of coordinates {i,j} contained the time to go from the $i^{th}$ to the $j^{th}$ sampling site.

During the optimization, the sampling route time associated to a specific sampling design was computed by finding the optimal order and associated time to reach each sampling site, beginning from starting site. This is a common problem known as Travelling Salesman Problem (TSP) (Hoffman et al., 2013). This was solved using the "fast_tsp" python package. Note that this method allowed to retrieve recommended optimized path for a sampling design.

On-site sampling time was added to sampling route time to obtain total fieldwork time for a specific sampling design. It was computed as the estimated time for sampling at one site (determined by practitioners) multiplied by the sample size.

For this study, road network, slope (expressed in degrees) and land occupation (buildings, farmlands, types of sand) data were used to design the time cost map. Pixels of size 30 m were considered as it allowed to keep enough information while reducing computation time. Zones with slope higher than 45° or buildings were considered as not accessible (white zones in Fig. 1.c). Accessible zones were given a speed value according to (Table 1) the type of land occupation, presence of roads/tracks and slope as detailed in Table 2. Sampling sites with pH $H_2O$ and environmental covariates values will be used as potential sampling sites to assess the sampling methods that are presented in the following sections.

Least cost paths were computed for all 651 potential sampling sites and starting site. On-site sampling time was estimated to 40 min by practitioners who conducted the actual sampling campaign in the Al Ula region.

Then this actual time spent for sampling (expressed in seconds) is converted in whole days of work. According to the classical French working time of 35 h per week that was applied to practitioners, a day of fieldwork was set to 7 hours.

#### 2.3.3. Sample representativeness objective

Representativeness criterion used in the MOOS was the same as the objective function of the cLHS already described in 2.2. Here, the aim was to minimize $cLHS_{fObj}$, as a smaller value corresponded to a relative better sample representativeness of covariates. As stressed in Israeli et al. (2019), this criteria is sensitive to the sample size, it was therefore normalized (i.e. divided by the sample size) in order to be able to compare the representativeness of sampling designs with different sample size. This criterion was computed for all eleven covariates presented in 2.1.

#### 2.3.4. Multi-criteria optimization

The multi-objective algorithm NSGA-II (Deb et al., 2002) was used to select optimal sampling designs that minimize both sample representativeness of covariates and total fieldwork time. It is based on a genetic algorithm. Each sampling design is an "individual" which has one "chromosome", a vector describing all its sampling sites. This vector contains X Boolean elements, X being the number of potential sampling sites. "True" value at the $i^{th}$ element of the vector indicates that the $i^{th}$

**Table 1**
Detail of environmental covariates used to predict soil pH $H_2O$.

| Environmental data | Resolution | Description | Reference |
|---|---|---|---|
| *SRTM Global DEM* | 30 m | | |
| *Altitude (m)* | | *Original data* | |
| *LS factor* | | *Length-slope factor* | (Module LS Factor, SAGA GIS) |
| *TWI* | | *Topographic Wetness Index* | (Module Wetness Index, SAGA GIS) |
| *Slope (degrees)* | | *Gradient of altitude* | *SAGA slope function* |
| *Sentinel-2 L1C* | *10 m* | | |
| *Red, Green, Blue, NIR* | | *Original data* | |
| *Salinity index 1* | | $SI = \sqrt{B \times R}$ | (Khan et al., 2005) |
| *Normalized Difference Salinity index* | | $NDSI = \frac{R - NIR}{R + NIR}$ | (Major et al., 1990) |
| *Normalized Difference Vegetation index* | | $NDVI = \frac{NIR - R}{NIR + R}$ | (Tucker, 1979) |
| *Brightness index* | | $BI = \sqrt{R + NIR^2}$ | (Khan et al., 2005) |
| *Sentinel-1C- band* | *5 m* | | |
| *C- band vertical polarization* | | *Original data* | |
| *C- band horizontal polarization* | | *Original data* | |

**Table 2**
Land occupation types of study area and associated transportation, estimated speed (km/h) and corresponding pixel value.

| Land occupation type | Transport | Speed (km/h) | Pixel Value |
|---|---|---|---|
| Large roads | Car | 100 | 1.08 |
| Small roads | Car | 60 | 1.8 |
| Drivable sand / tracks | Car | 30 | 3.6 |
| Not drivable sand | Walk | 5 | 21.6 |
| Farmland | Walk | 5 | 21.6 |
| Steep slope (30° to 45°) | Walk | 2 | 54 |
| Urban | Impossible | X | noData |
| Very steep slopes (>45°) | Impossible | X | noData |

sampling sites is contained in the sampling design. "False" value at the $i^{th}$ element indicates the contrary.

The optimization process used for MOOS is described Fig. 2. First, N different sampling designs with various sample size are randomly created (initial population) for the first iteration (generation 0). Then, for each sampling design, both $cLHS_{fObj}$ and total fieldwork time are computed. Sampling designs were then ranked according to two consecutive methods:

a) Pareto optimality, where a solution is considered more optimal to another if both criteria are simultaneously better (Luc, 2008). This process results in the creation of a first Pareto front, representing the set of optimal solutions where no solution can be improved in one objective without compromising performance in another. Then, other Pareto fronts that are less optimal are progressively discovered. Therefore, 1 to N different Pareto fronts can be identified and ranked from the most optimal to the less optimal. This rank is associated to each sampling design.
b) the crowding distance value which provides an estimate of the density of solutions (sampling design) surrounding the solution under consideration. Within a Pareto front, if a sampling design is surrounded by only a few other sampling designs, it will be given a high crowding distance. For instance, a sampling design on the borders of the front has the highest crowding distance, set to $+\infty$. Sampling designs are ranked from higher to smaller crowding distance value within each front. Keeping sampling designs with highest crowding distance ensures to explore a large variety of optimal solutions during optimization and not to focus on few local optima. An example of rank for N = 10 is shown in Fig. 2.

Based on this ranking, half of sampling designs are kept. These latter are considered as "parents" from which other sampling designs will be created (offspring). In our case, only "mutations", a random modification of "individual's chromosome" (random permutation of n sampling sites) were used to create new sampling designs. A mutation often generates a sampling design with a different size as a permutation may result to add or remove several sampling sites. Contrarily to the classical use of NSGA-II, no crossover (combination of 2 sampling designs to

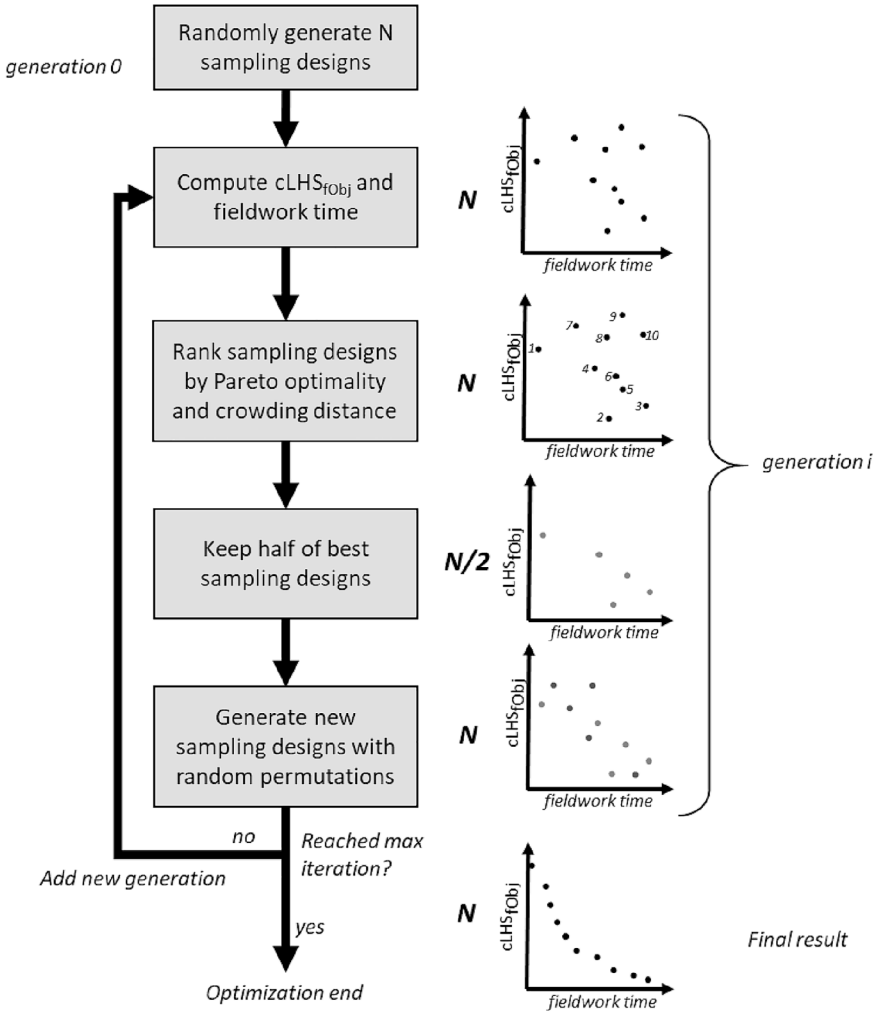



**Fig. 2.** MOOS optimization process and example of sampling designs obtained at each stage of a generation and at the end of the optimization. N or N/2 indicate the current size of the set of sampling designs. Each point on a plot corresponds to one sampling design.

create a new one) were used as it could generate sampling designs with a large sample size, thus a long fieldwork time, that would have high chances not to be kept during optimization. Once the offspring are created a new population is available, forming the next generation. Thus, generation after generation, most optimal individuals are kept. The optimization ends when the maximum number of generations is reached. At the end of the optimization, it is assumed that the most optimal sampling designs are obtained.

The method was chosen as it is widely used in the optimization literature and easy to implement and customize with the "Pymoo" python package (Blank and Deb, 2020).

In the presented optimization process, the sample size of generated sampling designs can evolve along iterations. However, practitioners may have operational constraints on the size of sampling design. Therefore, MOOS also allows to limit the range of possible sample size by discarding all sampling designs that do not go within this range and replacing it by sampling design with the right size. Setting a range of possible sample size as presented here is the classical use intended of MOOS. It fits an operational for which no precise assumption is made to chose one optimal sample size.

Note that to compare MOOS, cLHS and $cLHS_{cost}$ on the same basis, the possible sample size to explore with MOOS was fixed. This is due to cLHS and $cLHS_{cost}$ only allowing the optimization to be performed with a fixed sample size. This was achieved by generating an initial population with various sampling designs of same sample size and doing mutations that can only randomly replace n sites with n other sites (n being a random number).

#### 2.3.5. Multi-criteria decision making

As the optimization leads to a set of various optimal sampling designs, it can be challenging to choose the best possible solution considering operational context. Solutions are often not uniformly distributed

```
Split the dataset fully randomly into 5 subsets of equal size;
for k = 1 to 5 do
    Define the k-th subset as the validation subset. Merge the remaining subsets and define
     it as training subset for sampling;
    for sampleSize in (10,25,50,100,200) do
        Run MOOS with a population size of 100 and a sample size of s on training subset;
        Retrieve sampling route time obtained with each sampling design;
        Train Random Forest model with each obtained sampling design and compute
         RMSE on validation subset;
        for repetition =1 to 100 do
            Run cLHS and cLHScost with sample size s on K-1 subset;
            Compute sampling route time of each sampling design;
            Train Random Forest model with obtained sample and compute RMSE on
             validation subset;
        end
    end
end
Plot fieldwork time and RMSE for every method and varying sample size;
```

along a Pareto front, and different methods were proposed (Li et al., 2020) to identify relevant solutions within the front. Among these methods, the best trade-off between objectives metric (Rachmawati and Srinivasan, 2009) was preferred in this study because its practical use is intuitive and it can easily be implemented by a practitioner with little knowledge in optimization methods. The best trade-off metric relies on "knees" identification within the Pareto front. Das (1999) defined the "knee" with the following statement: "*It is noticed from practical experience that given the trade-off curve or surface for a particular multicriteria problem, the user or designer usually picks a point 'in the middle' of the surface. Often this is also the point where the Pareto surface "bulges out the most*"". Indeed, solutions located at "knees" of the front are of particular interest. Compared to their neighboring solutions, they allow for a large improvement of one of the objectives with only a small degradation of the other. Translated to MOOS, this would mean a sampling design which allows to gain substantially on representativeness with only a small increase in total fieldwork time; and/or a substantially shorter fieldwork time with a similar representativeness.

This was implemented using "Pymoo" package (Blank and Deb, 2020) on the results of optimization obtained with MOOS.

### 2.4. Evaluation methodology

MOOS is evaluated first by comparing it to cLHS and $cLHS_{cost}$ to assess its efficiency over more conventional approaches. Secondly, the evaluation is based on a realistic use-case study, to evaluate the approach's effectiveness as a support for decision-making.

The three methods MOOS, cLHS and $cLHS_{cost}$ were compared by randomly splitting the dataset into 5 subsets of equal size to perform a cross-validation. Iteratively, each subset was used as a validation set, and the 4 other subsets were merged and used as training set. For each method, samples of size of 10, 25, 50, 100 and 200 were created from the training set and used to train a Random Forest model (Breiman, 2001) to predict pH $H_2O$ with the eleven presented environmental covariates as predictors. The RMSE of each model on the validation subset was then computed. In the meantime, sampling route time of each sampling design obtained with MOOS was determined. Since cLHS and cLHScost do not inherently optimize this criterion, sampling route time was computed (with the same method described in 2.3.2) after each run. For each subset, and each sample size, cLHS and cLHScost were run a hundred times and MOOS once with a population size of 100.

The detailed procedure is described in the following:

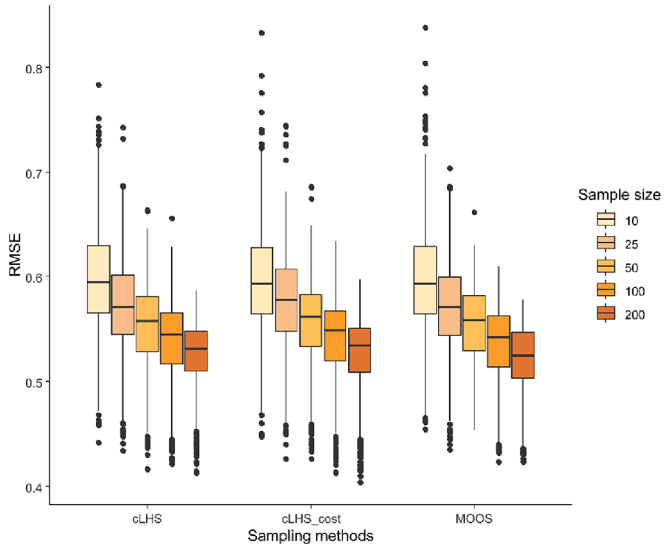



**Fig. 3.** Cross-validated RMSE of pH $H_2O$ obtained with varying sample size with 3 different sampling methods: cLHS, $cLHS_{cost}$ and MOOS. For every sample size, a pairwise Dunn test, conducted at a 95% confidence level revealed no statistically significant differences in medians among the three methods.

To complete the comparison a pairwise Dunn test evaluating the statistical significance ($p < 0.05$) of the differences in median sampling route time and RMSE among the three methods for each sample size was conducted. The random forest algorithm was implemented with *RandomForestRegressor* function of "skcikit-learn" python package (sklearn.ensemble.RandomForestRegressor, scikit-learn) with default settings.

Secondly, MOOS was qualitatively assessed on the same sampling task running with a population size of 100 for 2000 generations. To better simulate a real use-case, the possible sample size to explore during optimization was set between 60 and 120. This range was considered to make sure the resulting Pareto front represented a large variety of sampling designs that a practitioner can choose from. The 'knee' method introduced in 2.3.5 was applied on the obtained Pareto front to identify the sampling design with the best trade-off between objectives. The best sampling design was then mapped to illustrate its potential to plan a sampling campaign.

## 3. Results

### 3.1. Sampling methods comparison

#### 3.1.1. RMSE

Fig. 3 displays the cross-validated RMSE of pH $H_2O$ obtained with the three methods with varying sample size. RMSE values range from 0.8 to 0.4 points of pH $H_2O$, with a decrease in both median and variability as the sample size increases. This figure also shows that the method used does not impact the RMSE. No significant difference of RMSE values is observed among the three methods.

#### 3.1.2. Sampling route time

Fig. 4 shows a comparison of sampling route time obtained across varying sample size for the three sampling methods considered in the experiment. To better display differences between methods, only the sampling route time (i.e., time spent to travel from the start point to reach each sampling site) was included. The on-site sampling time was not included as it does not change for a fixed sample size. In this case, each run of MOOS was done with a fixed sample size for comparing with other methods.

Complementing Fig. 4, Table 3 shows the result of Dunn test, evaluating the statistical significance ($p < 0.05$) of the differences in median sampling route times among the three methods for each sample size. Methods with different letter have a statistically significant difference in median route time.Sampling route time (expressed in days) ranges from less than one day for 10 sampling sites up to 5 days with 200 sampling sites. Whatever the sample size considered; the sampling route time observed with the MOOS method is always lower. The difference in sampling route time between MOOS and other methods increases with the number of sampling sites. Indeed, with 10 sampling sites, MOOS results in a 0.4 days median gain compared to $cLHS_{cost}$ and cLHS. With 200 sampling sites, the difference of median time increases up to 1 day compared to $cLHS_{cost}$ and 1.4 days compared to cLHS. Table 3 confirms that MOOS results in significantly smaller sampling route time than the two other methods.

### 3.2. Operational use of MOOS

Fig. 5.a is a scatter plot showing the 17 most optimal sampling designs obtained with one run of MOOS, with a flexible sample size set for optimization. Each point is therefore a specific sampling design characterised by its sample representativeness ($cLHS_{fObj}$) and its total fieldwork time (expressed in days). Contrarily to Fig. 4, the total fieldwork time (i.e. sampling route time + on-site sampling time) is displayed, as it

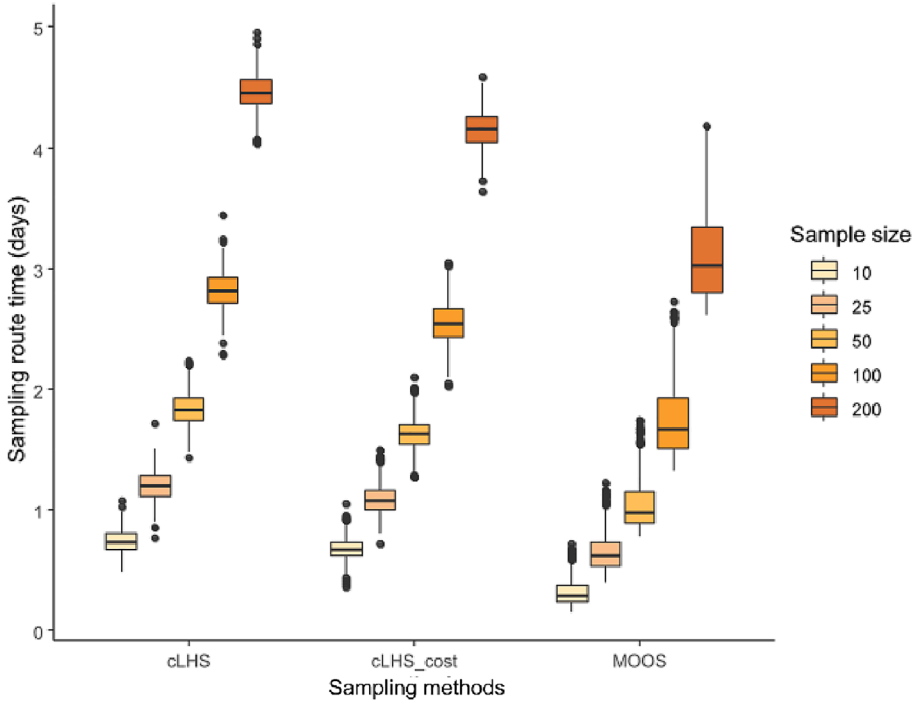



**Fig. 4.** Sampling route time (expressed in days) needed to reach each sampling site considering 5 different sample sizes and 3 different sampling methods: cLHS, cLHScost and MOOS.

**Table 3**
Dunn test results for differences in median sampling route time (time spent to travel to each sampling site) obtained with the 3 methods and 5 different sample sizes. Common letters indicate non-significant differences ($p < 0.05$) for each sample size.

| | Sample size | | | | |
|---|---|---|---|---|---|
| | 10 | 25 | 50 | 100 | 200 |
| cLHS | *a* | *a* | *a* | *a* | *a* |
| $cLHS_{cost}$ | *a* | *a* | *b* | *b* | *a* |
| MOOS | *b* | *b* | *c* | *c* | *b* |

is the final criterion that serves decision-making. All 17 sampling designs form a Pareto front, which means they dominate bi-criteria wise other sampling designs obtained along MOOS run. The number next to each sampling design is an ID to help referring to a specific sampling design in the following of this article. The objective here is to describe the kind of results a soil surveyor can expect when preparing a sampling campaign with MOOS. This Pareto front represents the set of possible solutions a soil surveyor will have to choose for field sampling.

Fig. 5.a highlights sampling designs with $cLHS_{fObj}$ values ranging between 5 and 4.4 and total fieldwork time ranging from 7 to 11.5 days. As expected, sampling designs with smaller sample size (top left of the plot) have the highest $cLHS_{fObj}$ values (less representative), but lead to the shortest total fieldwork time. On the other hand, sampling designs with larger sample size (bottom right corner) have the best representativeness and longest total fieldwork time. Due to the randomness of the optimization and the complex relation between spatial distribution of the samples, representativeness and associated route; the repartition of the sampling designs along the Pareto front is non-uniform. Indeed, the front contains gaps, and plateaux where one of the objectives is almost constant while the other substantially changes. To precise, there are three $cLHS_{fObj}$ plateaux: from sampling design n° 16 to 15, n° 3 to 1 and 8 to 17. On the other hand, there are two total fieldwork time plateaux: from sampling design n° 15 to 3 and n°1 to 8. This is particularly interesting because points connecting plateaux and forming a "knee" on the front, such as sample designs n°15, 1 and 17, are those presenting the best trade-off between two objectives compared to their surrounding solutions. For example, the sampling design n°1 (with the highest trade-off, indicated by a black circle) compared to n°2 and 14, allows to gain substantially on one of the objectives while the other only slightly changes. For almost no change in total fieldwork time, using solution 1 over solution 2 allows to gain 0.05 points of $cLHS_{fObj}$. Similarly, using solution 1 over solution 14 allows to gain 1.5 days with only a slight decrease in $cLHS_{fObj}$. Thus, in a real case, if there are no other constraints, solution n°1 can be considered as the best option to choose.

The spatial organization of sampling sites of sampling design n° 1 and the resulting recommended path are shown in Fig. 5.b. The sampling design n°1 includes 75 sampling sites, it has a $cLHS_{fObj}$ value of 4.6 and an estimated field work time of 9.1 days. As the estimated time spent on each site to sample is of 40 min, the total time spent on the field comprises 7.1 days (75 sites x 40 min) for on-site sampling time and 2 days for sampling route time. This map shows that sampling sites are mostly concentrated in easily reachable areas. Indeed, when compared with time cost map (Fig. 1.c), and characteristics of the study zone (Fig. 1.a.), recommended path passes through the most accessible zones (roads, sandy plains, near urban areas) and avoids canyons or steep slopes. Additionally, sampling sites are spatially spread all over the study area, yet slightly clustered, forming lines close to roads in certain zones.

## 4. Discussion

### 4.1. Using MOOS reduces operational costs

The proposed method (MOOS) aimed at optimizing both the sample representativeness ($cLHS_{fOb}$) and the total fieldwork time, which is in our case the most important part of operational costs.

Results showed that employing MOOS for sampling in poorly accessible areas significantly reduces fieldwork time compared to cLHS and $cLHS_{cost}$, particularly with large sample size, without impacting sample representativeness and resulting mapping accuracy. This was expected as MOOS aims specifically to optimize the total fieldwork time criterion unlike cLHS and $cLHS_{cost}$. For a sound comparison, the three methods were compared on the same operational criterion: total

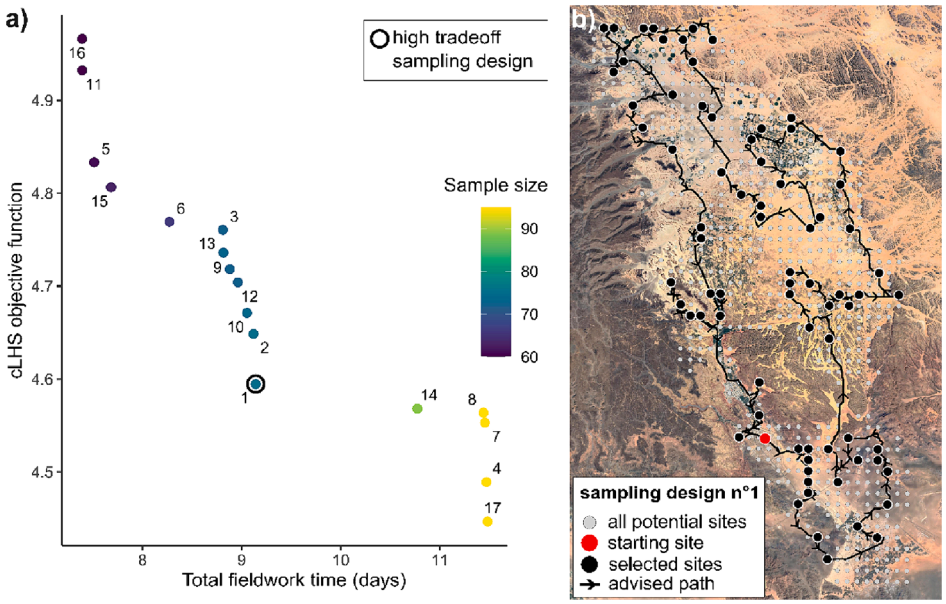



**Fig. 5.** Operational use case of MOOS. a) Pareto-optimal set of points obtained with MOOS with sample size varying from 60 to 100, and selected sampling design (sampling design n°1) with best trade-off between objectives; (the number associated to each sampling route is an id. facilitating identification), b) selected sampling design and resulting sampling sites with recommended path from one site to another.

fieldwork time. For the record, for cLHS and $cLHS_{cost}$, which do not incorporate natively this new cost criterion, this latter was computed once the optimization done. Indeed, in a practical application, these methods do not provide any optimal site order or recommended path to optimally explore all the sampling sites. This would certainly lead to even longer sampling route time than depicted in Fig. 3 for cLHS and $cLHS_{cost}$.

MOOS maximizes the sample size that can be achieved in a limited time. Therefore, compared to cLHS and cLHScost, it allows to create more accurate maps with the same total fieldwork time. This makes MOOS a relevant alternative to existing sampling methods. To our knowledge, no other sampling method allows to estimate and optimize fieldwork time like MOOS does. As a result, it is expected to also outperform other methods that consider operational constraints without using a representativeness criterion like $cLHS_{fObj}$ (Cambule et al., 2013; Kidd et al., 2015a).

### 4.2. MOOS as a decision support to design realistic sampling designs suited to field experts needs

Regarding operational application of MOOS, results show that it aligns with the practices of soil surveyors when organizing sampling surveys over large areas for the following reasons.

First, total fieldwork time criterion gives a relevant estimation of the time required to perform the sampling, noting that this information was not available with other methods. This certainly supports the practitioner in its objective to minimize the risk of unexpected time spent on the field. Secondly, the method provides not only one best solution, but a set of sampling designs that allows the practitioner to choose the sampling design that is most appropriate for a specific use. Due to the selection based on Pareto optimality and "crowding-distance", NSGA-II explores preferentially, zones of the front with few sampling designs, like its outer edges. Therefore, it tends to select a wide range of optimal sampling designs. Overall, this provides valuable information on the time needed for a large diversity of sampling designs with different sizes and/or different representativeness.

Regarding the example chosen for Fig. 5, the set of sampling designs showed that the method succeeds in providing a variety of possible solutions which differ in representativeness and total fieldwork time. The shape of the Pareto front with "knees" helps choosing among all sampling designs, the ones with the best trade-off between both criteria. Regarding the spatial repartition shown in Fig. 5.b, the sampling sites spread all over the study area may indicate how well the two objectives are balanced. Indeed, this is a visual confirmation that a diverse set of sites will be explored during sampling. Moreover, the recommended path resulting from the application of MOOS avoids inaccessible zones (i.e., zones marked as white in Fig. 1.c.). This way, the practitioner can verify that the risk of encountering sites that are inaccessible during the fieldwork remains relatively low.

In the case presented, it is assumed that the project is not constrained by any time limit, the aim is to choose the sampling designs that has the best trade-off between both criteria. Considering a case where the number of total fieldwork time is strictly constrained, one may prefer to choose the solution with the best sample representativeness with a total fieldwork time as close as possible to the limit. Regarding example presented Fig. 4.a., the consideration of a 11 days limit for the survey would have, for example, led to choose the solution n°14.

As NSGA-II algorithm is highly versatile, its parameters can be easily adapted. For example, by setting specific constraints on objectives, sample size, specific initialization or stopping criterion or type of random permutation to create new sampling designs. For instance, the practitioner can choose to constraint the number of total fieldwork time so that solutions exceeding the fixed threshold are discarded during the optimization.

However, practitioners should choose cautiously the sample size to ensure that the resulting model still yields a sufficient quality of prediction. Several authors investigated this issue, indicating that the optimal sample size surely depends on covariates resolution, type of variable of interest, spatial extent of the area, sampling method and model used (Bouasria et al., 2023; Loiseau et al., 2021; Saurette et al., 2022; Schmidinger et al., 2024). Thus, it is recommended to choose a minimum sampling size which follows the recommendations of these authors that is adapted to each specific DSM study. Choosing an optimal sample size could be further aided by using other metrics. Stumpf et al., (2016) identified optimal sample size by identifying the 'knee' of the difference of sample variance and global variance for different sample sizes. In the same way, Saurette et al., (2024) demonstrated that the Jensen-Shannon divergence metric is useful to determine an optimal

sample size as it reflects well the final model performance (providing all relevant covariates are used), while being robust to the spatial extent of the area.

In this specific case study, $cLHS_{fObj}$ and sampling cost expressed in days were used as optimization criteria. Yet, practitioners have the opportunity to explore alternative representativeness criteria that may align more closely with the mapping method and availability of covariates. As proposed by Brus (2019), if relevant covariates are missing or no covariates are available, a representativeness criteria based on space coverage sampling should be preferred. On the contrary, if all relevant covariates are available and the mapping is done by kriging with external drift, a model-based sampling would be recommended. As long as it is possible to find or design a relevant criterion allowing different sampling designs to be compared, it can be used with MOOS.

MOOS could be further improved by using a representativeness criterion specifically tailored for decision-making. Such a criterion ought to assist a practitioner in selecting a sampling design that ensures adequate prediction accuracy. For instance, rather than stipulating a maximum of 10 days of fieldwork days, a practitioner could specify a desired accuracy level of at least 70 % for the map. As previously indicated, the Jensen-Shannon divergence, as introduced by Saurette et al., (2024), presents a viable initial alternative. Nonetheless, additional research involving diverse soil parameters, models, and spatial extents is essential for verification of this metric for an operational use.

In the same way, other types of operational costs metrics could be used to better satisfy specific requirements. For instance, Brus et al. (1999) exhaustively described costs associated with sampling and used a monetary (US$) metric to include the fieldwork time as well as equipment and laboratory costs. Additionally, the NSGA-II optimization algorithm used in MOOS, enables the optimization of more than two criteria, providing an opportunity to incorporate additional criteria. To maximize sample representativeness, a criteria ensuring good spatial coverage could be added, such as in Israeli et al. (2019) which optimized jointly $cLHS_{fObj}$ and a "spatial dispersion objective".

### 4.3. Limits and perspectives

In any Digital Soil Mapping project, having quality environmental covariates that capture well the soil parameter's variability is crucial (McBratney et al., 2003). Remote and poorly accessible areas are often underexplored and poorly characterized by previous studies. Therefore, sampling in these areas often imply a lack of relevant environmental data or legacy data. This can be solved by numerous means and has already been extensively covered in the literature (Hartemink et al., 2008). This is an issue shared by all sampling methods based on ancillary data such as response surface sampling (Lesch et al., 1995), Kennard-Stone sampling (Kennard and Stone, 1969), feature space coverage sampling using k-means (Brus, 2019), among others. This aspect is recalled here as an important limitation, but not specific to the approach presented in this study.

Despite cLHS being a widely used method in the DSM community, several authors recently showed that using cLHS has no particular interest over a Simple Random Sampling when mapping with Kriging with external drift or Random forest, two of the most used models in DSM (Wadoux et al., 2019; Wadoux and Brus, 2021). In the context of MOOS, it can be argued that using cLHS objective function is still interesting as it is combined with fieldwork time criterion. Indeed, the representativeness and fieldwork time criteria can be seen as two competing objectives. The interest of MOOS relies in finding a compromise between both, which is not possible with SRS that do not use a specific criterion.

Initial tests with MOOS were undertaken as a preliminary observation to further validate the proposed approach. These first results indicated that the total fieldwork time was slightly overestimated. Mainly because practitioners spent less time for on-site sampling than expected. This illustrates one of the main limitations of MOOS: the estimation of total fieldwork time relies heavily on the available data source and assumptions made to create the time cost map and determine on-site sampling time. The time cost map reliability depends on the quality of the data source to map land characteristics and derive associated access constraints. It also depends on the assumptions made to define each type of environment and the travel speed associated with each of them. This requires an extensive knowledge of the study zone. Moreover, on-site sampling time requires a good knowledge of the specific sampling task, to estimate accurately the time spent on each sampling site. Not respecting these conditions may lead to a high uncertainty in total fieldwork time, which can significantly affect the quality of final result, especially with a large sample size. Indeed, providing practitioners results with an associated uncertainty is crucial for a more realistic estimation of fieldwork time. It could help them to better manage risks and prepare a sampling campaign with awareness. To solve this issue, uncertainty can be incorporated into the algorithm process thanks to fuzzy-logic or Bayesian reasoning, as already explored in the optimization literature for different applications (Bahri et al., 2018; Laumanns and Ocenasek, 2002). Choosing among a Pareto front under uncertainty seems to be still an open question for the current research. However, a practitioner using such method for field sampling might choose sampling designs that have the best trade-off between objectives as well as uncertainty to reduce risks.

To better estimate total fieldwork time, another solution may be to dynamically update the time cost map, or on-site sampling time while sampling. For example, this would mean recording the actual speed (or range of speed) for a specific land occupation while doing the sampling, and updating accordingly the routes and chosen sampling sites to optimize fieldwork time. Although the dynamic recommendation of new sampling sites was already investigated for field sampling (Ma et al., 2020; Zhao et al., 2019), no author explored dynamic rerouting. One should investigate solutions solving dynamic routing problems that proved to improve field work parameters (Seyyedhasani and Dvorak, 2018). Whether achieved with an on-board or remote computation unit, this would raise new technical challenges to solve to fit operational constraints.

To summarize, while employing MOOS necessitate an initial investment in data curation (time cost map creation), knowledge of the study area and environmental data of quality, it serves as a valuable decision support tool for soil surveyors. MOOS facilitates the design of more efficient sampling campaigns, leading to a reduction in fieldwork time.

## 5. Conclusion

This study introduced a methodological approach aiming at optimizing sampling designs for poorly accessible areas using a multi-objective optimization technique. Its primary goal was to minimize sampling route time between sampling sites and on-site sampling time, while ensuring a high level of sample representativeness that captures the variability of the study area. The developed method, named Multi-Objective Operational Sampling (MOOS), was designed to simultaneously optimize operational costs and a sample representativeness. When compared to common approaches of the literature, MOOS presents very similar results in map error but with significant sampling route time reduction. MOOS especially outperforms other solutions in terms of route time with increasing sample size. The potential of MOOS for decision support in choosing the best possible sampling route was confirmed through its application in a real-life case study performed over a large study in Saudi Arabia. In summary, this study proposed a new operational method that helps soil surveyors efficiently plan and conduct optimized field surveys. It offers various optimal sampling designs, giving practitioners the flexibility to choose the best option based on the required sample quality and operational constraints.

## CRediT authorship contribution statement

**Maxime Dumont:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Guilhem Brunel:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **Paul Tresson:** Validation, Investigation, Data curation. **Jérôme Nespoulous:** Writing – review & editing, Validation, Supervision, Resources. **Hassan Boukcim:** Supervision, Resources, Project administration, Funding acquisition. **Marc Ducousso:** Resources, Project administration, Funding acquisition, Data curation. **Stéphane Boivin:** Writing – review & editing, Validation, Resources, Investigation, Data curation. **Olivier Taugourdeau:** Validation, Supervision, Methodology, Investigation, Conceptualization. **Bruno Tisseyre:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Conceptualization.

## 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

The authors do not have permission to share data.

## Acknowledgments

This work was supported by the Valorhiz company and the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004.

*Use of Generative AI statement.*

Generative AI was used during the redaction of this article for syntax and language correctness purposes only.

## References

Arrouays, D., Mulder, V.L., Richer-de-Forges, A.C., 2021. Soil mapping, digital soil mapping and soil monitoring over large areas and the dimensions of soil security – A review. Soil Security 5, 100018. https://doi.org/10.1016/j.soisec.2021.100018.

Bahri, O., Talbi, E.-G., Ben Amor, N., 2018. A generic fuzzy approach for multi-objective optimization under uncertainty. Swarm Evol. Comput. 40, 166–183. https://doi.org/10.1016/j.swevo.2018.02.002.

Blank, J., Deb, K., 2020. Pymoo: Multi-Objective Optimization in Python. IEEE Access 8, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567.

Bouasria, A., Bouslihim, Y., Gupta, S., Taghizadeh-Mehrjardi, R., Hengl, T., 2023. Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. Eco. Inform. 102294 https://doi.org/10.1016/j.ecoinf.2023.102294.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464–480. https://doi.org/10.1016/j.geoderma.2018.07.036.

Brus, D.J., Spätjens, L.E.E.M., de Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89, 129–148. https://doi.org/10.1016/S0016-7061(98)00123-2.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353. https://doi.org/10.1016/j.geoderma.2012.08.020.

Clifford, D., Payne, J.E., Pringle, M.J., Searle, R., Butler, N., 2014. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput. Geosci. 67, 62–68. https://doi.org/10.1016/j.cageo.2014.03.005.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.

Das, I., 1999. On characterizing the ?knee? of the Pareto curve based on Normal-Boundary Intersection. Structural Optimization 18, 107–115. https://doi.org/10.1007/BF01195985.

De Gruijter, J.J., Bierkens, M.F.P., Brus, D.J., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33161-1.

de Gruijter, J.J., McBratney, A.B., Minasny, B., Wheeler, I., Malone, B.P., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130. https://doi.org/10.1016/j.geoderma.2015.11.010.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Computat. 6, 182–197. https://doi.org/10.1109/4235.996017.

Elhag, M., 2016. Evaluation of Different Soil Salinity Mapping Using Remote Sensing Techniques in Arid Ecosystems, Saudi Arabia. J. Sens. 2016, e7596175.

Grafström, A., Lundström, N.L.P., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x.

Hart, P.E., Nilsson, N.J., Raphael, B., 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Trans. Syst. Sci. Cybernet. 4, 100–107. https://doi.org/10.1109/TSSC.1968.300136.

Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M. de L., 2008. Digital Soil Mapping with Limited Data. Springer Science & Business Media.

Hoffman, K.L., Padberg, M., Rinaldi, G., 2013. Traveling Salesman Problem. In: Gass, S.I., Fu, M.C. (Eds.), Encyclopedia of Operations Research and Management Science. Springer, US, Boston, MA, pp. 1573–1578. https://doi.org/10.1007/978-1-4419-1153-7_1068.

Israeli, A., Emmerich, M., Litaor, M. (Iggy), Shir, O.M., 2019. Statistical learning in soil sampling design aided by pareto optimization, in: Proceedings of the Genetic and Evolutionary Computation Conference. Presented at the GECCO '19: Genetic and Evolutionary Computation Conference, ACM, Prague Czech Republic, pp. 1198–1205. Doi: 10.1145/3321707.3321809.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. Technometrics 11, 137–148. https://doi.org/10.2307/1266770.

Khan, N.M., Rastoskuev, V.V., Sato, Y., Shiozawa, S., 2005. Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. Agricultural Water Management, Special Issue on Land and Water Use: Environmental Management Tools and Practices 77, 96–109. Doi: 10.1016/j.agwat.2004.09.038.

Kidd, D., Malone, B., McBratney, A., Minasny, B., Webb, M., 2015a. Operational sampling challenges to digital soil mapping in Tasmania, Australia. Geoderma Reg. 4, 1–10. https://doi.org/10.1016/j.geodrs.2014.11.002.

Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A., 2015b. Digital soil assessment of agricultural suitability, versatility and capital in Tasmania, Australia. Geoderma Reg. 6, 7–21. https://doi.org/10.1016/j.geodrs.2015.08.005.

Koch, A., McBratney, A., Adams, M., Field, D., Hill, R., Crawford, J., Minasny, B., Lal, R., Abbott, L., O'Donnell, A., Angers, D., Baldock, J., Barbier, E., Binkley, D., Parton, W., Wall, D.H., Bird, M., Bouma, J., Chenu, C., Flora, C.B., Goulding, K., Grunwald, S., Hempel, J., Jastrow, J., Lehmann, J., Lorenz, K., Morgan, C.L., Rice, C.W., Whitehead, D., Young, I., Zimmermann, M., 2013. Soil Security: Solving the Global Soil Crisis. Global Pol. 4, 434–441. https://doi.org/10.1111/1758-5899.12096.

Koch, A., Chappell, A., Eyres, M., Scott, E., 2015. Monitor Soil Degradation or Triage for Soil Security? An Australian Challenge. Sustainability 7, 4870–4892. https://doi.org/10.3390/su7054870.

Laumanns, M., Ocenasek, J., 2002. Bayesian Optimization Algorithms for Multi-objective Optimization, in: Guervós, J.J.M., Adamidis, P., Beyer, H.-G., Schwefel, H.-P., Fernández-Villacañas, J.-L. (Eds.), Parallel Problem Solving from Nature — PPSN VII, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 298–307. https://doi.org/10.1007/3-540-45712-7_29.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial Prediction of Soil Salinity Using Electromagnetic Induction Techniques: 1. Statistical Prediction Models: A Comparison of Multiple Linear Regression and Cokriging. Water Resour. Res. 31, 373–386. https://doi.org/10.1029/94WR02179.

Li, X., Gao, B., Pan, Y., Bai, Z., Gao, Y., Dong, S., Li, S., 2022. Multi-objective optimization sampling based on Pareto optimality for soil mapping. Geoderma 425, 116069. https://doi.org/10.1016/j.geoderma.2022.116069.

Li, W., Wang, R., Zhang, T., Ming, M., Li, K., 2020. Reinvestigation of evolutionary many-objective optimization: Focus on the Pareto knee front. Inf. Sci. 522, 193–213. https://doi.org/10.1016/j.ins.2020.03.007.

Loiseau, T., Arrouays, D., Richer-de-Forges, A.C., Lagacherie, P., Ducommun, C., Minasny, B., 2021. Density of soil observations in digital soil mapping: A study in the Mayenne region, France. Geoderma Regl. 24, e00358.

Lu, Q., Tian, S., Wei, L., 2023. Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning. Sci. Total Environ. 856, 159171 https://doi.org/10.1016/j.scitotenv.2022.159171.

Luc, D.T., 2008. Pareto Optimality, in: Chinchuluun, A., Pardalos, P.M., Migdalas, A., Pitsoulis, L. (Eds.), Pareto Optimality, Game Theory And Equilibria, Springer Optimization and Its Applications. Springer, New York, NY, pp. 481–515. Doi: 10.1007/978-0-387-77247-9_18.

Ma, T., Wei, T., Qin, C.-Z., Zhu, A.-X., Qi, F., Liu, J., Zhao, F., Pan, H., 2020. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. Earth Sci Inform 13, 39–53. https://doi.org/10.1007/s12145-019-00407-x.

Major, D.J., Baret, F., Guyot, G., 1990. A ratio vegetation index adjusted for soil brightness. Int. J. Remote Sens. 11, 727–740. https://doi.org/10.1080/01431169008955053.

Maleki, S., Karimi, A., Zeraatpisheh, M., Poozeshi, R., Feizi, H., 2021. Long-term cultivation effects on soil properties variations in different landforms in an arid region of eastern Iran. Catena 206, 105465. https://doi.org/10.1016/j.catena.2021.105465.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009.
Module LS Factor / SAGA-GIS Module Library Documentation (v2.2.0) [WWW Document], n.d. URL https://saga-gis.sourceforge.io/saga_tool_doc/2.2.0/ta_hydrology_22.html (accessed 1.24.24).
Module SAGA Wetness Index / SAGA-GIS Module Library Documentation (v2.2.3) [WWW Document], n.d. URL https://saga-gis.sourceforge.io/saga_tool_doc/2.2.3/ta_hydrology_15.html (accessed 1.24.24).
Nazzal, Y., Ahmed, I., Al-Arifi, N.S.N., Ghrefat, H., Zaidi, F.K., El-Waheidi, M.M., Batayneh, A., Zumlot, T., 2014. A pragmatic approach to study the groundwater quality suitability for domestic and agricultural usage, Saq aquifer, northwest of Saudi Arabia. Environ. Monit. Assess. 186, 4655–4667. https://doi.org/10.1007/s10661-014-3728-3.
Neina, D., 2019. The Role of Soil pH in Plant Nutrition and Soil Remediation. Appl. Environ. Soil Sci. 2019, e5794869.
Niang, M.A., Nolin, M.C., Jégo, G., Perron, I., 2014. Digital Mapping of Soil Texture Using RADARSAT-2 Polarimetric Synthetic Aperture Radar Data. Soil Sci. Soc. Am. J. 78, 673–684. https://doi.org/10.2136/sssaj2013.07.0307.
Prăvălie, R., 2016. Drylands extent and environmental issues. A Global Approach. Earth-Science Reviews 161, 259–278. https://doi.org/10.1016/j.earscirev.2016.08.003.
Rachmawati, L., Srinivasan, D., 2009. Multiobjective Evolutionary Algorithm With Controllable Focus on the Knees of the Pareto Front. IEEE Trans. Evol. Comput. 13, 810–824. https://doi.org/10.1109/TEVC.2009.2017515.
Rasterio: access to geospatial raster data — rasterio documentation [WWW Document], n.d. URL https://rasterio.readthedocs.io/en/stable/# (accessed 3.26.24).
Roudier, P., Beaudette, A.E.H.& D.E., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints: Pierre Roudier & Allan E. Hewitt Dylan E. Beaudette, in: Digital Soil Assessments and Beyond. CRC Press, Boca Raton, USA.
Saurette, D.D., Berg, A.A., Laamrani, A., Heck, R.J., Gillespie, A.W., Voroney, P., Biswas, A., 2022. Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon. Geoderma 425, 116054. https://doi.org/10.1016/j.geoderma.2022.116054.
Saurette, D.D., Heck, R.J., Gillespie, A.W., Berg, A.A., Biswas, A., 2024. Sample Size Optimization for Digital Soil Mapping: An Empirical Example. Land 13, 365. https://doi.org/10.3390/land13030365.
Schmidinger, J., Schröter, I., Bönecke, E., Gebbers, R., Ruehlmann, J., Kramer, E., Mulder, V.L., Heuvelink, G.B.M., Vogel, S., 2024. Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming. Precis. Agric. https://doi.org/10.1007/s11119-024-10122-3.
Sena, N.C., Veloso, G.V., Lopes, A.O., Francelino, M.R., Fernandes-Filho, E.I., Senra, E.O., Silva Filho, L.A. da, Condé, V.F., Silva, D.L. de A., Araújo, R.W. de, 2021. Soil sampling strategy in areas of difficult acess using the cLHS method. Geoderma Regl. 24, e00354. Doi: 10.1016/j.geodrs.2020.e00354.
Seyyedhasani, H., Dvorak, J.S., 2018. Dynamic rerouting of a fleet of vehicles in agricultural operations through a Dynamic Multiple Depot Vehicle Routing Problem representation. Biosyst. Eng. 171, 63–77. https://doi.org/10.1016/j.biosystemseng.2018.04.003.
sklearn.ensemble.RandomForestRegressor [WWW Document], n.d. . scikit-learn. URL https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed 1.26.24).
Sparks, D.L., Singh, B., Siebecker, M.G., 2024. Environmental soil chemistry, Third, edition. ed. Elsevier, Academic Press, Amsterdam.
Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S.Y., Rawlins, B.G., Field, D.J., 2015. Global soil organic carbon assessment. Glob. Food Sec. 6, 9–16. https://doi.org/10.1016/j.gfs.2015.07.001.
Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., Scholten, T., 2016. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. J. Plant Nutr. Soil Sci. 179, 499–509. https://doi.org/10.1002/jpln.201500313.
Taghizadeh-Mehrjardi, R., Schmidt, K., Toomanian, N., Heung, B., Behrens, T., Mosavi, A., S. Band, S., Amirian-Chakan, A., Fathabadi, A., Scholten, T., 2021. Improving the spatial prediction of soil salinity in arid regions using wavelet transformation and support vector regression models. Geoderma 383, 114793. Doi: 10.1016/j.geoderma.2020.114793.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.
Wadoux, A.-M.-J.-C., Brus, D.J., 2021. How to compare sampling designs for mapping? Eur. J. Soil Sci. 72, 35–46. https://doi.org/10.1111/ejss.12962.
Wadoux, A.-M.-J.-C., Brus, D.J., Heuvelink, G.B.M., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913. https://doi.org/10.1016/j.geoderma.2019.113913.
Yang, L., Feng, X., Liu, F., Liu, J., Sun, X., 2019. Potential of soil moisture estimation using C-band polarimetric SAR data in arid regions. Int. J. Remote Sens. 40, 2138–2150. https://doi.org/10.1080/01431161.2018.1516320.
Zhao, F.-H., Qin, C.-Z., Wei, T.-F., Ma, T.-W., Qi, F., Liu, J.-Z., Zhu, A.-X., 2019. Dynamic Recommendation of Substitute Locations for Inaccessible Soil Samples during Field Sampling Campaign. ISPRS Int. J. Geo Inf. 8, 127. https://doi.org/10.3390/ijgi8030127.