Tel Hai College
Department of Water Sciences

Multiobjective Optimization of Soil Sampling Design Using Information Metrics

by Assaf Israeli

Supervisors
Michael (Iggy) Litaor
Ofer M. Shir

In partial fulfillment of the requirements for the degree of
Master of Science in Water Sciences

June 2020

Acknowledgements

I would like to express my gratitude to my teachers, fellows and friends who encouraged me to accomplish this work. Thanks to Prof. Iggy Litaor and Dr. Ofer Shir for bringing me up to the challenge and for an excellent learning experience. Special thanks to Prof. Michael T. M. Emmerich, to Dr. Michal Horowitz and to the hydro-geochemistry lab members at *MIGAL* that provided me the data and advice for this study, including Dr. Oren Reichmann, Nitzan Malachy, Shachar Cliff, Shalev Malul, Hadar Gilboa and Nimrod Meitav. Research was also supported by internal grants of *MIGAL* and by *COST* Action CA15140 (*ImAppNIO*).

Dedicated to all sentient beings in the universe, with hope that food production can be sustained with positive reverberations.

Abstract

Soil sampling design for assessment of soil nutrients content in an agricultural crop field can be aided by ancillary data to reduce sample-size – a main factor affecting the economic viability of a survey. Most soil sampling methods assume a given sample-size reflecting a predefined budget, without aiming to minimize it, thus not providing a practical solution for precision agriculture. In this study we formulate a bi-objective optimization task for soil-sample design that minimizes the conditioned Latin hypercube sampling criterion (cLHS) for a full representation of the soil information spectrum, while concurrently maximizing the distance between sampling locations for lower estimation error. We present a two-step approach for finding a minimal sample-size, by generating Pareto-optimal sampling schemes of varying sizes using an evolutionary multiobjective algorithm, then scrutinizing the solutions grouped by size with information-theory statistics to identify a satisfactory sample-size at which the rate of marginal information gain is declining. Individual schemes are then evaluated and ranked, as a decision support tool to aid in choosing a single design. We have applied the methods in field experiments, devising sampling-plans for different purposes. The extracted soil analysis data was used to verify management zones delineation, and can be used to calibrate a model of remote-sensed data to produce digital soil maps.

Contents

List of Figures

Chapter 1

Introduction

Since the early settlements of humankind, maintaining favourable conditions for plants growth has been a limiting factor on the scaling of human civilization. In the 19th century new perceptions of plant nutrition promoted the use of chemical fertilizers for crop production, followed by a widespread adoption of high-yield crop varieties and industrial agricultural practices in the mid 20th century, known as the *Green Revolution* [29]. Research-based application protocols prescribing optimal nutrients dosage for each crop have become the standard cultivation method. Although world population had more than doubled in the period between 1960 and 2010, the production of cereal crops tripled, with only a 30% increase in land area cultivated [48]. However, prevailing uniform application methods, coupled with the difficulty to estimate within-field spatial variation of available soil nutrients perpetuate a sub-optimal state, as some areas are fertilized below the recommended level, thus not reaching the maximal potential yield, while other areas are over-fertilized, causing a waste of resources and amplifying the environmental stress on downstream waterways due to leaching of excess nutrients.

Recent technological advancements facilitate the use of *variable rate applica-*

tion, which utilizes adaptive dosage according to prescription maps. Although visualizing dynamic soil processes in real time has been an elusive goal in soil science, it could nowdays be achieved by the convergence of several fields of science and technology, thus facilitate the implementation of site-specific management practices that promote soil conservation [34]. However, lacking a reliable diagnostic measure of soil nutrients content by pure means of remote sensing, Precision Agriculture (PA) relies on soil sampling and laboratory analyses to calibrate and validate remote-sensed data to produce digital soil maps.

Spatial coverage sampling strategies, such as grid sampling or stratified random sampling, provide a sound geographical distribution but require extensive sampling to generate effective spatial models [61]. The recommended range of soil samples to obtain a reliable spatial interpolation using geostatistical methods has been assessed to be at least 100 to 150 to provide six to ten estimates within the expected effective range [46]. Kerry et al. [32] advised that a sampling grid for spatial prediction should exhibit spacing no coarser than half the range of spatial dependence of the soil variable and ideally one third to two fifths of the range. In the absence of any prior information about the spatial scale of variation, a reconnaissance survey can provide a first approximation. If the property of interest appears related to ancillary data, such as those from remote and proximal sensing they could be used to approximate the spatial scale [32]. These guidelines set the bar high above economic viability in commercial agricultural systems.

An alternative approach for sample design, adapted from experimental design is the so-called *Conditioned Latin Hypercube Sampling* (cLHS) [43], which accounts also for a predefined feature space. cLHS aims to maximize the stratification of the feature space in the sample while preserving the data distribution, yet it may produce uneven coverage of the geographical space, thus reducing the overall estimation quality [25].

Hengl et al. [27] concluded that prediction accuracy in a sample may be improved by covering both the feature and the geographical spaces. Gao et al. [25] added a spatial measure to cLHS by aggregation into a single objective function, and achieved smaller mapping errors in comparison to strictly spatial or feature space methods. Lark [39] demonstrated the feasibility of multiobjective optimization of spatial sampling design using the AMOSA (Archived Multiobjective Simulated Annealing) algorithm [3] on a theoretical use-case, where the objective functions are the total distance travelled for sampling and the variance of the sample mean.

There is no single best sampling design for digital soil mapping, as the adequacy of a sampling scheme depends on the method used for mapping the soil. Sampling methods can be distinguished as *design* or *model* based approaches [10]. In design-based methods locations are selected by probability (random) sampling without using a known spatial model in estimation. This approach includes Geometric designs, such as *Spatial Coverage sampling* [52] and Adapted Experimental designs, of which cLHS is a primary example. In a model-based approach a stochastic model is used in estimation, for instance a linear regression or Kriging model. As the model already contains a random error term, probability sampling is not required in this approach, which opens up the possibility of optimized non-probability sampling [10]. Model-based sampling design requires prior information about the spatial variation (i.e. the variogram), as well as assumptions about the mean of the study variable and an explicit tolerable variance (error) range, therefore it is not suitable in many scenarios.

Following these studies, we elevate the perspective of multi-objective sample design optimization to the realm of evolutionary algorithms, and propose herein an experimental design-based approach for obtaining a sample design by solving a bi-objective optimization problem of concurrently maximizing the feature space stratification *and* the geographical distribution of the sampling points. In doing so, we aim to provide sampling schemes for various purposes that would cover of the entire area and portray the full information spectrum of the soil, supported by statistical analysis to assist in plan selection and scaling down the sample-size – a significant contributor to the operational costs.

Given the complex nature of soil attributes, geostatistical methods consider target variables as realizations of random fields [57]. These methods infer statistical parameters and calculate predictions based on partial observations of the random field realization. Assuming a normally distributed stochastic process with second-order stationarity (a constant *mean* and an *autocovariance* function dependent solely upon the distance between any two values), sampled data can be interpolated by the widely used Ordinary Kriging (OK) method [40], which provides a best linear unbiased prediction (BLUP) of values at un-sampled locations. OK requires a positive definite model of spatial variability, calculated as a function fitted to an experimental variogram of the sample data. The reliability of the model depends on the capacity of sample information to capture the variability of the observed phenomenon, affected primarily by the number of observations [61].

The use of ancillary data can reduce prediction errors associated with contraction in sample-size and facilitate cost-effective sampling [62]. In this context, *ancillary data* is considered as any source of spatial information with some relation to soil properties and in the form of a digital soil map. Importantly, high-resolution data – e.g., multi-spectral aerial imaging, proximal sensing or yield maps – are available at a rather low cost. Although the exact relation of these data to soil attributes may be unknown, a spatial variability model can be formed to guide soil sampling design [4] and density [33], as well as to improve the approximation accuracy as a covariate in co-Kriging [49].

The quality of a sampling scheme can be evaluated *a priori* by the uncertainty of prediction maps resulting from interpolation of known ancillary data values at sample locations, reflected by metrics such as Mean Ordinary Kriging Variance (MOKV), subject to the assumption that the attribute under study is a realization of a stationary Gaussian random function [28], or by the Root Mean Square Error (RMSE) between predicted and true values at all locations. We studied additional metrics to support agricultural soil-survey planning, focusing on practical questions concerning sample-sizing and composition, hence we introduce in this work a preferential index for ranking candidate sampling schemes according to expected MOKV and information-theoretic statistical measures, calculated by the Kullback-Leibler Divergence (D_{KL}), a quantifier of the similarity between two probability distributions, [37] and Akaike Information Criterion (AIC) [1], which estimates the relative quality of a statistical model for a given set of data.

The current study targets the following research question:

> Which model captures the effectiveness as well as the cost-efficiency of sampling-plans when accounting for both diversity and representation?

The proposed contributions of the current study are:

1. Modelling of objective functions quantifying sampling-plans designed for the efficient use of ancillary data;
2. Formulation of multiobjective optimization problems for optimizing sampling strategy;
3. Solving these optimization problems in an agricultural farm using real-field data.

The Thesis has the following structure: In Chapter 2 we provide a short introduction to Evolutionary Algorithms focusing on the Non-dominated Sorting Genetic Algorithm II (NSGA-II) used in this study. Chapter 3 outlines the geostatistical learning challenge – we describe the data processing steps, specify our notation, defining the objective functions and motivate the multi-objective optimization perspectives. Our practical observations on real-world case-studies are reported in Chapter 4 where we also discuss the attained solutions. Finally, we summarize our work and findings in Section 5, where we also draw possible directions for future work.

Chapter 2

Computational Methodology

Summary

In this chapter we present some historical milestones in optimization techniques, then introduce the general concepts of Evolutionary Algorithms (EAs), the Genetic Algorithm (GA) and Evolutionary Multiobjective optimization Algorithms (EMOAs); in addition, we describe NSGA-II, the particular EMOA used in this study.

2.1 Optimization: a brief history

Optimization, derived from Latin *optimum*, neuter singular of *optimus* (“best”, ”very good”), meaning the best or most favorable condition under specific sets of comparable circumstances.[1] In the context of applied mathematics, optimization can be defined as a quantitative and systematic methodology that represents a problem as consisting of three core elements: variables, objectives, and constraints. The objective functions are to be minimized (or maximized,

[1] Online etymology dictionary: `www.etymonline.com/word/optimum`

without loss of generality) by adjusting the decision variables, while satisfying the given constraints.[1] Formally, an optimizer $\vec{x}^*$ adheres to this general formulation:

$$\begin{aligned} \vec{x}^* := \text{arg}\min_{\vec{x}} \quad & f(x_1, x_2, ..., x_n), \\ \text{subject to} \quad & g_i(x_1, x_2, ..., x_n) \leq 0 \;\; \forall i = 1, \dots, m, \end{aligned} \tag{2.1}$$

where f is the objective function, x_k $(k = 1, \dots, n)$ are the decision variables, and g_i $(i = 1, \dots, m)$ are constraint functions.

Early Greek mathematicians solved optimization problems related to their geometrical studies. Circa 300 BC Euclid considered the minimal distance between two points, and proved that a square has the largest area among the rectangles with a given perimeter length [35]. The discovery of derivatives, attributed to Fermat and Lagrange among other scholars in the 17th century, established calculus-based formulae for identifying optima, whereas in the 18th century Newton and Gauss proposed iterative methods for moving towards an optimum, and Fourier proved that certain problems could be defined as a system of linear inequalities.

The next milestone took place in the 1930's with the conception of Linear Programming (LP) by Kantorovich and Koopmans, followed by the development of the *Simplex* method by Dantzig in 1947, an algorithm that became a state-of-the-art tool and its variants are still in use today. In the 1950's the unifying tool of linear and integer programming (where variables are only of integral nature) became available, and the area of Operations Research got intensive attention [53]. In parallel, a broad understanding of Convex Optimization [8] has been accomplished throughout the years. However, even with advance-

[1] Brief History of Optimization: `empowerops.com/en/blogs/2018/12/6/brief-history-of-optimization`

ments in computation capabilities, mathematical methods are inadequate for non-convex problems, especially with high-dimensional search spaces, such is the nature of many real-world problems.

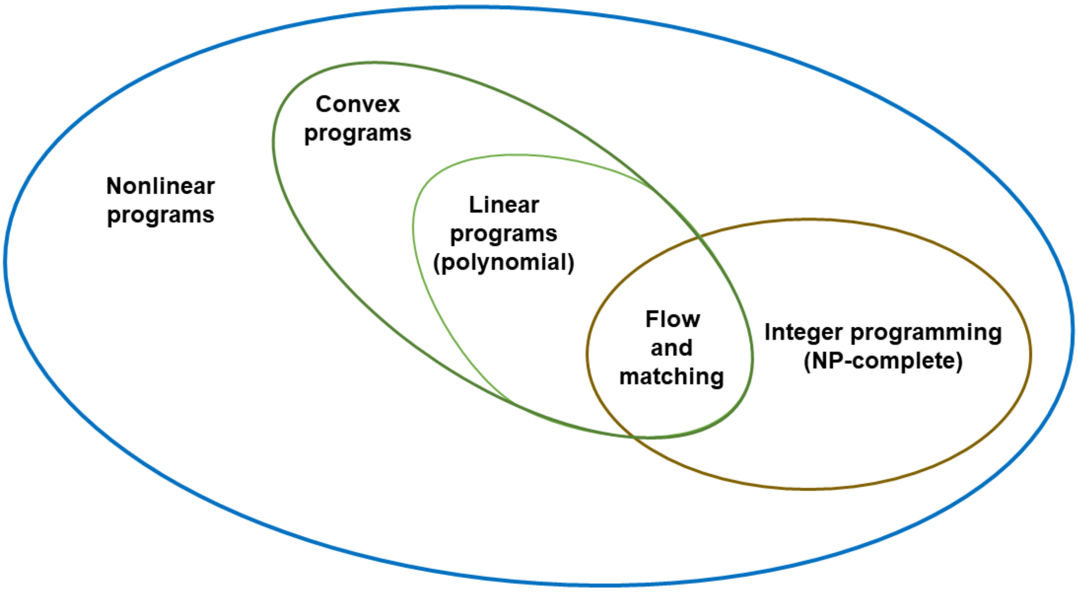
Nonlinear programs
Convex programs
Linear programs (polynomial)
Flow and matching
Integer programming (NP-complete)

Figure 2.1: A schematic taxonomy diagram of mathematical optimization branches [55].

2.2 Evolutionary Algorithms (EAs)

This shortcoming of mathematical optimization motivated the research on problem solving in Nature, leading to the foundation of the Evolutionary Computing field, a class of stochastic optimization methods that abstract the process of natural evolution by breeding a population of solutions. The scope of Evolutionary Algorithms (EAs) covers several meta-heuristics, the most prominent are: Genetic Programming (GP), introduced by Friedberg in 1958; Evolution Strategies (ES), developed by Rechenberg and Schwefel in the late 1960's; and

Genetic Algorithms (GA), proposed by Fogel et al. and Holland in the 1970's. EAs differ by their solution representation, selection mechanism and the mutation operator, making them appropriate to use in different applications. Although initially distinct, the lines between the above strands of research are becoming blurred, with representation and strategies being used interchangeably between the algorithms. As such, today it is common to use the term evolutionary algorithm to encompass all of the above approaches [9].

EAs share the concept of directing a population of solutions through iterative variation and selection to improve the fitness of their offspring in a certain environment (the modeled problem). By inclination towards better population members the solution is refined, gradually approaches the global optimum. The use of a population helps to achieve an implicit parallelism [15], which makes an EA computationally attractive for solving difficult problems.

The generalized EA, adapted from [2] is presented in Algorithm 1. The procedure begins with the formation of an initial population at random (or according to a predefined scheme). Then, a loop consisting of the steps evaluation (fitness assignment), selection, recombination, and/or mutation is executed a certain number of iterations. Each loop iteration is called a generation, and often a predefined maximum number of generations serves as the termination criterion of the loop, but also other conditions, e.g., stagnation in the population or existence of an individual with sufficient quality, may be used to stop the simulation. In *elitism* mode the parent population in each generation takes part in the environmental selection, thus ensuring the best solutions found so far are always propagated to the next generation, regulating a monotonically non-degrading performance. The best individuals in the final population represent the outcome of the EA [63].

Notation-wise, μ (mu) denotes the number of parents and λ (lambda) stands for the number of offspring, whereas a plus sign indicates the implementation of elitism, and a comma otherwise. Thus, an elitist-EA with a population of 10 parents and 10 offspring would be denoted as (10+10)-EA.

Algorithm 1 The Generalized Evolutionary Algorithm

```
$t \leftarrow 0$                {t: generation}
$P(t) \leftarrow init()$        {P: population}
$evaluate(P(t))$
**repeat**
  $P'(t) \leftarrow sexualSelection(P(t))$
  $P''(t) \leftarrow variation(P'(t))$
  $evaluate(P''(t))$
  **if** elitism **then**
    $Q(t) \leftarrow P(t)$
  **else**
    $Q(t) \leftarrow \varnothing$
  $P(t+1) \leftarrow environmentalSelection(P''(t) \bigcup Q(t))$
  $t \leftarrow t+1$
**until** stopping criterion is met
```

Next, we describe the GA as the prototype of the algorithm used to solve the problem at hand.

2.3 The Genetic Algorithm (GA)

Heredity, the passing of traits from parents to their offspring, is a key concept in the the theory of evolution and natural selection [14]. In genetics, a strong distinction is drawn between the *genotype* and the *phenotype*; the former contains genetic information, whereas the latter is the physical manifestation of that information [9]. As species reproduce, variation is introduced through recombination of the parents genes and occasional mutation, whereas survival of the fittest individuals is regulated by environmental selection acting upon the phenotype. The Genetic Algorithm (GA) simulates the evolutionary process by maintaining and iteratively improving a population of candidate solutions, each encoded as a binary sequence and subject to variation operators, while the selection mechanism in the form of fitness evaluation is applied to the corresponding phenotype. As the search evolves, the population includes fitter and fitter solutions, and eventually it converges, meaning that it is dominated by a single solution. Holland presented a proof of convergence (the schema theorem [30]) to the global optimum given an infinite population size with binary chromosome representation [36].

In canonical GAs, solution vectors are represented as binary strings. The mutation operator performs a bit-flip in each coordinate with a small probability $P_m \in [0, 1]$, whereas recombination is realized as a crossover of bit sequences with probability $P_c \in [0, 1]$. Selection operates in two stages: the sexual selection picks a couple for mating among the candidates, and the environmental selection assesses the performance of the new offspring, ensuring survival of the fittest.

Tournament selection is a commonly used mating selection strategy, in which n individuals are randomly picked from the population of size N without replacement, and the winner is the highest fitting individual. Calibration of $n \ll N$ controls the selection pressure, which is analogous to the convergence rate of the procedure. A key challenge in GA design is the balance between *exploration* of new regions of the search space, stimulated by mutation, and *exploitation* of the vicinity of already discovered fit solutions, encouraged by crossover and selection [9].

The canonical GA usually operates with an identical population size for parents and offspring and with constant control parameters (P_m, P_c). Although shown to be powerful problem solver that have been successfully applied to various real-world problems, obtaining the best results requires a careful design of the algorithm using any domain knowledge available.

Some limitations of GAs are inherent to all EAs: i) There is no guarantee that an optimal solution will be found in a finite time (or number of iterations); ii) progress towards better solutions may be intermittent rather than gradual, and iii) the algorithm relies on feedback in the form of fitness evaluations, which can be difficult or expensive for some problems [9].

Although the original implementation of the traditional GA considers genes as binary digits, the flexibility of the framework facilitates solution representation of any data structure, and in certain cases, e.g. of combinatorial decision variables, a pure phenotype representation is more straightforward (see Section 3.8).

2.4 Multiobjective Optimization

Real-world optimization scenarios often involve multiple, conflicting objectives within implicit feasibility constraints. For example, in the context of sample design, discrepancy is evident between spatial dispersion and feature space coverage (see Section 3.8), as improvement of one objective comes at the expense of the other. Therefore, the algorithmic goal alters from identifying the best solution to obtaining a range of good compromises amongst the objectives; the role of the decision-maker (DM) now becomes to select among this set of solutions.

Traditionally, multiobjective optimization problems are often solved using *scalarization* techniques [42], in which the weighted normalized objective functions are aggregated (or reformulated as constraints), and then a constrained single-objective problem is solved. This approach is called a-*priori* [36], since the weights express the DM's preference in advance. One can consider a multistart option, i.e., to run the EA several times with different aggregation parameters, and collect solutions in an archive [63]. This approach can possibly lead to a set of optimal solutions, but the complexity of the parameters' tuning requires some assumptions about the search landscape (i.e., estimation of the location of optimal solutions) [36], while there is an uncertainty regarding the revelation of the entire set of efficient solutions, which makes interpretability of the search space more difficult, thus not facilitating the DM's confidence that a final solution is the most preferred one or at least close to that [21].

Another classical multiobjective optimization strategy is to optimize only one objective, and constrain the other objective to be strictly lower than its value obtained in the previous optimization run (assuming minimization), resulting with a vector of optimal solutions per each run of a single-objective EA with an additional constraint [63]. This approach encounters the same limitations posed by scalarization.

Evolutionary Multiobjective Algorithms (EMOAs, also termed MOEA or EMO), originating in the 1990's [23], have become established as a separate subdiscipline combining the fields of evolutionary computation and classical multiple criteria decision making (MCDM) [63]. The growing popularity of EMOAs, which generalize the idea of single-objective EAs to a higher dimensional objective function space, is mainly attributed to the fact they do not require any derivative information and to their relatively simple implementation and flexibility – which altogether make them suitable for a broad range of applications [21]. As a by-product, EMOA-based solutions have the potential to assist in revealing important hidden knowledge about a problem – a matter which is difficult to achieve otherwise [15].

The concept of Pareto dominance, named after Vilfredo Pareto [44], is of fundamental importance for multiobjective optimization, as it allows to compare two objective vectors in a precise sense, according to the following formulation.

Given a multiobjective optimization problem with m objectives, let an objectives vector in $\mathbb{R}^m$ be denoted as

$$\vec{f}\left(\vec{x}\right) = \left(f_1\left(\vec{x}\right), f_2\left(\vec{x}\right), \ldots, f_m\left(\vec{x}\right)\right)^T,$$

and let all its coordinates assumed to be subject to *minimization*. A partial order is defined on the m-dimensional objective space, $\mathcal{F} = \vec{f}(\mathcal{X})$, by means of the *Pareto domination* concept: given any $\vec{f}^{(1)} \in \mathbb{R}^m$ and $\vec{f}^{(2)} \in \mathbb{R}^m$, it is stated that $\vec{f}^{(1)}$ weakly dominates $\vec{f}^{(2)}$, noted as $\vec{f}^{(1)} \preceq \vec{f}^{(2)}$, if and only if the following holds:

$$\forall i \in \{1, \ldots m\} : f_i^{(1)} \leq f_i^{(2)}. \tag{2.2}$$

We also consider the strict Pareto domination:

$$\vec{f}^{(1)} \prec \vec{f}^{(2)} \Longleftrightarrow \vec{f}^{(1)} \preceq \vec{f}^{(2)} \ \wedge \ \exists i \in \{1, \ldots, m\} : f_i^{(1)} < f_i^{(2)}. \tag{2.3}$$

We then state that $\vec{f}^{(1)}$ and $\vec{f}^{(2)}$ are *incomparable* or *indifferent*, noted as $\vec{f}^{(1)} || \vec{f}^{(2)}$, if and only if $\vec{f}^{(1)} \npreceq \vec{f}^{(2)} \wedge \vec{f}^{(2)} \npreceq \vec{f}^{(1)}$. For any non-empty compact subset of $\mathbb{R}^m$, say $\mathcal{F}$, there exists a non-empty set of minimal elements for the partial order $\preceq$ [19]. Non-dominated points are the set of minimal elements for $\preceq$:

$$\mathcal{F}_N = \left\{\vec{f} \in \mathcal{F} | \nexists \vec{f}' \in \mathcal{F} : \vec{f}' \prec \vec{f}\right\}. \tag{2.4}$$

The goal of multiobjective optimization is to obtain the *non-dominated set* for $\mathcal{F} = \vec{f}(\mathcal{X})$, entitled the *Efficient Frontier* (also known as *Pareto front*, see Figure 2.2), and its pre-image in $\mathcal{X}$, the so-called *Pareto set*.

Unlike single criterion problems, a multiobjective problem has multiple,

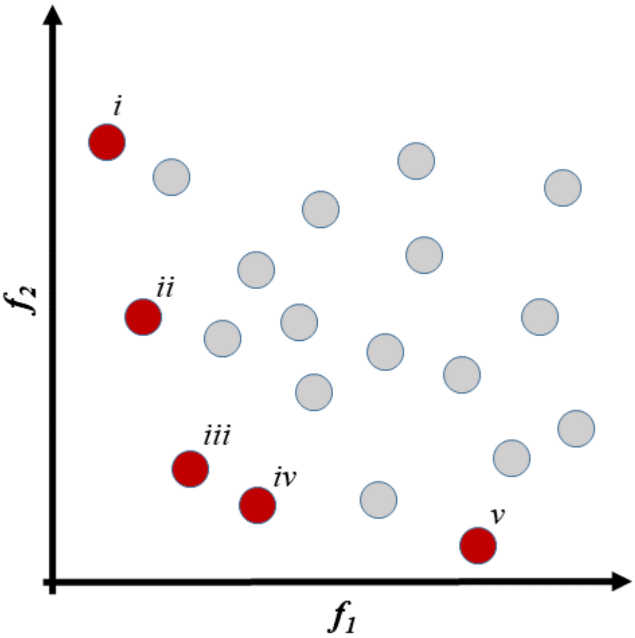

Figure 2.2: The Pareto frontier is depicted in red for a bi-objective problem of minimizing both f_1, f_2. Solution-points (i) and (v) constitute the ideal points for f_1 and f_2, respectively.

possibly infinite number of solutions, rendering the calculation of the entire Pareto set infeasible. Therefore, the target is to find an approximation of the Pareto set given available computational resources.

Similar to other *a posteriori* MCDM methods [15], most EMOAs are designed to gradually approach sets of Pareto optimal solutions that are well-distributed across the objective and/or the decision spaces – a multiobjective task by itself [63]. The distinction between different classes of EMOAs is mainly due to the paradigm used to define the selection operator, whereas the choice of the decision-space variation operators is most likely problem-dependent [21].

There is no guarantee that an EMOA will find any Pareto-optimal solution in a finite number of evaluations for an arbitrary problem. However, the preference of non-dominated and isolated solutions guarantees that population members iteratively progress towards the Pareto-optimal front [15].

The main contemporary paradigms for EMOAs' design are [21]:

I Pareto-based selection, which uses a two-level ranking scheme. The Pareto dominance relation governs the first ranking and contributions of points to diversity is the principle of the second level ranking, which applies to points that share the same rank in the first cycle;

II Indicator-based approaches which implicitly measure convergence *as well as* spread (for instance, the *hypervolume* or *R2* indicators) for the performance of an approximation set to guide the search; and

III Decomposition-based methods, in which the problem is decomposed into several subproblems, each one of them targeting different parts of the Pareto front, whereas each subproblem is assigned with different weighting of a scalarization method.

Elitism has shown to improve the performance of EMOAs [15], and is a common design feature. The implementation is not as straightforward as in single-objective EA, mainly due to the large number of possible elitist solutions, considering all non-dominated solutions discovered so far [36]. Elitism is implemented in EMOAs by either keeping elitist solutions in the population, or by storing elitist solutions in an external secondary list (archive) and reintroducing them to the population [11].

Real-world optimization problems are typically bound with constraints that must be satisfied. Single-objective GAs employ one of four constraints handling strategies: (i) discard infeasible solutions, (ii) penalize the fitness of infeasible solutions, (iii) if possible, customize variation operators to produce only feasible solutions, and (iv) repair infeasible solutions [36]. They are all applicable in the multiobjective case, whereas the single penalty function in (ii) is replaced by another mechanism, such as multi-dimensional penalty or the constraints-domination concept [16].

The resulting Pareto-optimal set is a portfolio of candidate solutions for further consideration by the DM, often involving non-technical, qualitative and experience-driven information. By narrowing down the choices and visualizing the trade-offs between the objectives, EMOAs have the potential to facilitate better decision making.

2.5 NSGA-II

NSGA-II (Non-dominated Sorting Genetic Algorithm II) [16] is one of the most popular EMOAs. Being a Pareto-based procedure, it employs an elitist strategy with a fast non-dominated sorting selection approach and an explicit diversity preserving mechanism [15].

NSGA-II follows the generic EA procedure (Algorithm 1), integrating a bilevel selection operator with Pareto non-domination preference as the primary criterion, and a density measure as a secondary criterion.

The *fast non-dominated sorting* procedure (Algorithm 2) reduces the computational complexity of Pareto order ranking (a.k.a. Pareto-sort) by iteratively identifying non-dominated solutions, then excluding them from subsequent pairwise comparisons. The *crowding distance* (Algorithm 3) is a measure of the objective space surrounding a point which is not occupied by any other solution in the population. It is calculated by estimating the perimeter of the cuboid formed by the nearest neighbors in the objective space as the vertices (Figure 2.3) [15]. This procedure is repeated for each objective function by sorting the population in an ascending order, then the value is calculated as the absolute normalized difference in the function values of two adjacent solutions, whereas the boundary solutions are assigned an infinite distance value, ensuring their inclusion. The overall crowding-distance value is calculated as the sum of individual distance values corresponding to each objective [16].

The crowded-comparison operator ($\preceq_n$) formulates these measures. Given the solution's non-domination rank (i_{rank}) and crowding distance ($i_{distance}$), a partial order $\preceq_n$ is defined as:

$$\begin{aligned} i \preceq_n j \quad if \quad & (i_{rank} < j_{rank}) \\ or \quad & ((i_{rank} == j_{rank}) \quad and \quad (i_{distance} < j_{distance})) \end{aligned}$$

such that non-domination rank is the primary criterion to favor a solution, whereas the distance measure is used as a tie-breaker between solutions of the same non-domination rank.

A fixed population size of N is maintained by NSGA-II. At each generation t, the offspring population Q_t is created by using the parent population P_t and problem-specific variation operators. Thereafter, the two populations are combined to form a new population $R_t = P_t \bigcup Q_t$ of size $2N$, which is then classified into different non-domination classes. Subsequently, the new population P_{t+1} is filled by points of different non-domination fronts, one at a time, starting with the first non-domination front (of class one), continues with points of the second non-domination front, and so on. Since $|R_t| = 2N$, not all points can be accommodated in N slots available for the new population, hence the points of the last front, which could not be fully accommodated, are sorted in a descending order of their *crowding distance* values and the top points of the ordered list are chosen, until the population size $|P_{t+1}| = N$.

The non-dominated sorting approach ensures that the best solutions so-far are kept during the search, effectively implementing elitism without using a

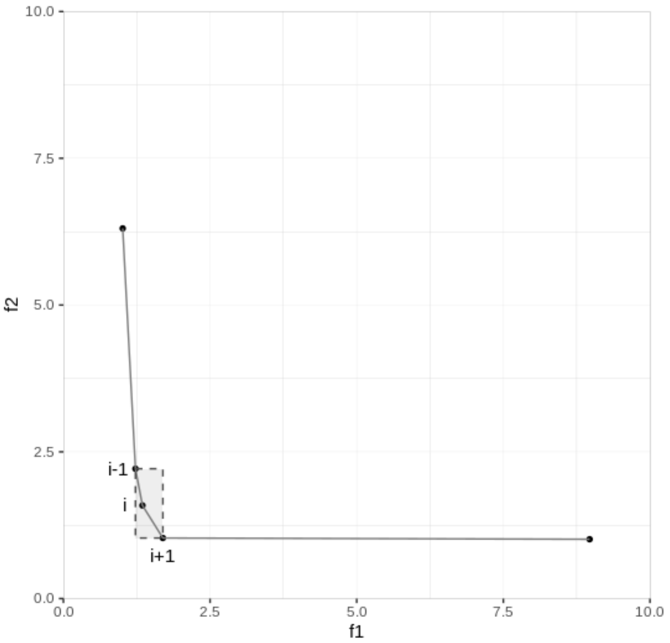

Figure 2.3: The crowding distance measure, calculated by the rectangle enveloping each point's nearest neighbours, outlined here for point *i* by the dashed line.

secondary external population [15].

An advantage of the crowding distance measure is the density calculation around a solution without requiring a user-defined parameter. When the combined parent and offspring population contains more than N non-dominated solutions, NSGA-II becomes a pure elitist GA where only non-dominated solutions participate in crossover and selection. This leads to a straightforward implementation, emphasizing population size as an important parameter since no external archive is used to store discovered non-dominated solutions [36].

The overall complexity of the algorithm is $O(MN^2)$ (where M is the number of objectives and N is the population size), governed by the non-dominated sorting part of the algorithm [16].

Algorithm 2 fast-non-dominated-sort - NSGA-II [16]

```
Input: P {current population}
for all $p \in P$ do
  $S_p = \varnothing$
  $n_p = 0$
  for all $q \in P$ do
    if $p \prec q$ then
      {if $p$ dominates $q$}
      $S_p = S_p \bigcup q$ {Add $q$ to the set of solutions dominated by $p$}
    else
      if $q \prec p$ then
        {Increment the domination counter of $p$}
        $n_p = n_p + 1$
  if $n_p == 0$ then
    {$p$ belongs to the first front}
    $p_{rank} = 1$
    $F_1 = F_1 \bigcup p$
$i = 1$ {Initialize the front counter}
while $F_i \neq \varnothing$ do
  $Q = \varnothing$ {Used to store the members of the next front}
  for all $p \in F_i$ do
    $n_q = |S_p|$
    for all $q \in S_p$ do
      $n_q = n_q - 1$
      if $n_q == 0$ then
        {$q$ belongs to the next front}
        $q_{rank} = i + 1$
        $Q = Q \bigcup q$
  $i = i + 1$
  $F_i = Q$
```

Algorithm 3 Crowding-distance-assignment - NSGA-II [16]

```
input: I {a non-dominated set}

$\ell = |\mathbf{I}|$ {number of solutions in I}

for each $i = 1, \ldots, \ell$ do
  $\mathbf{I}[i]_{distance} = 0$ {initialize distance}

for each objective $m$ do
  I=sort($\mathbf{I}, m$) {sort using each objective value}
  $\mathbf{I}[1]_{distance} = \mathbf{I}[\ell]_{distance} = \infty$ {so that boundary points are always selected}

  for $i = 2$ to $(\ell - 1)$ do
    {for all other points}
    $\mathbf{I}[i]_{distance} = \mathbf{I}[i]_{distance}$ + ($\mathbf{I}[i+1]_m$-$\mathbf{I}[i-1]_m$)/($f_m^{max} - f_m^{min}$)
    {normalized difference in objective space m between adjacent points}
```

Chapter 3

From Ancillary Data to Sampling-Plans

Summary

In this chapter we describe the utilization of an EMOA (Chapter 2) to obtain sampling plans with good coverage of both the geographic and the feature spaces, accounting for ancillary data. Information-theoretic statistical measures are then applied with the goal of identifying a minimal yet effective sampling plan for validating the partitioning of Site-Specific Management Units (SSMUs) in a 37 ha agricultural crop field.

3.1 Ancillary Data Collection

Ancillary data were recorded in the bare-soil field by the outset of the rainy season, encompassing geospatial measurements of apparent electrical conductivity (ECa) measured with an electromagnetic induction (EMI) sensor using a modified procedure outlined in [13], which consists of several sequential steps employing EM38-MK2 (Geonics Ltd. Ontario, Canada) in dual operational mode (EM$_{vertical}$ and EM$_{horizontal}$) at different depth response profiles (1.5 and 0.75 m, respectively) providing simultaneous measurements of both ECa and apparent magnetic susceptibility (MSa) values [18]. The sensor was connected to a differential global positioning system (GPS) with high planimetric accuracy. The precision and accuracy of ECa maps are affected by the swath width of the ECa sensing range. Based on [22], we used 12–15 m parallel swath widths across the field at an average speed of $3\ km\ h^{-1}$ with one per second data collection frequency which yielded data points of one in every 0.85 m.

Remote sensing multispectral data were acquired at resolution of 6x6 cm with a sensor (Parrot Sequoia, Paris, France) capturing four spectrum bands (green, red, red edge and near infra-red) mounted on an unmanned aerial system (Mavic Pro, DJI, Shenzhen, China). The fragmented data were processed to an orthomosaic map with *Pix4Dmapper* (Pix4D S.A., Prilly, Switzerland), and the Normalized Difference Vegetation Index (NDVI) calculated by the Red and Near Infra-Red (NIR) bands:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3.1}$$

3.2 Preprocessing

The outcome of ancillary data measurements are large datasets comprised of different formats and resolutions. Some preliminary steps are required to standardize the layers to common units and spatial resolution before further processing can take place.

Recorded points' data were first merged into a single table and stripped from comments, missing values and other non-informative rows or columns. Since the GPS receiver was anchored on the vehicle, while the EMI was dragged on a wagon 3 m behind, we shifted the values' location five table rows ahead to compensate for the constant gap. For computational speed this large dataset was averaged over every 20 successive readings. Normal distribution is a prerequisite for Ordinary Kriging (OK) interpolation [61], hence we ran exploratory analysis of the data histograms, with particular attention to deviance from Normal distribution, measured by the *skewness* metric, which equals zero for a symmetric distribution. Accordingly, the data was trimmed to 98% (removing the lower and upper 1% values) followed by log transformation, altogether shaping the distribution towards Normality.

The treated tables, containing 3,164 $ECa_{vertical}$ and 3,415 $ECa_{horizontal}$ readings, were converted into spatial objects [47] with a UTM coordinates reference system (CRS) assigned. This data structure enables the calculation of a variogram spatial model that portrays the dependence between points values as a function of their pairwise distances, and serves as an input for Kriging interpolation. OK is a geostatistical method to predict values at unsampled locations by computing a weighted average of the known values in the neighborhood of each point (whereas weights are given by the variogram model) [40]. In this study, the target for prediction was an empty grid covering the field area with a cell size of a practical 1x1 m resolution. Field perimeter was hand-drawn as a polygon in QGIS [59] and saved as a shapefile, providing a mask for clipping the resulting estimation maps to the field boundaries. The final step involves normalization of the values to the range [0,1]. Correlation tests of EMI data showed high collinearity (Pearson's $R > 0.97$) of the four layers, a phenomenon that may disqualify the use of three of them. However, in the absence of more meaningful ancillary data layers we opted to use them nonetheless.

The NDVI raster layer was scaled-down to 1x1 m resolution by *resampling*, then clipped to field perimeter and normalized to [0,1].

Finally, a unified normalized table of ancillary data was constructed from the raster layers, consisting of five value columns with geo-referenced values.

3.3 Site-Specific Management Units

Cluster analysis of ancillary data layers is conducted to delineate Site-Specific Management Units (SSMUs) using the fuzzy ***c***-means clustering procedure [26; 45]. A suite of *cluster validity indices* is used to assess the number of SSMUs: Partition Coefficient (PC) [5], Partition Entropy (PE) [6], Fukuyama-Sugeno index (FS) [24] and the Calinski-Harabasz Criterion (CHC) [58]. The partitioning results of multiple fuzzy ***c***-means executions, with the number of clusters k incremented within a considerable range, are evaluated by each of these indices. By visualizing index values for all possible number of clusters, one can determine the optimal number of clusters in the data, indicated by local extrema (maxima for PC and CHC, minima for PE and FS) [50].

The resulting partitions are quite irregular in shape and size, which may be difficult to cultivate without advanced variable rate application central pivot and it is completely unsuitable to current drip irrigation technology. Hence, the fuzzy ***c***-means clusters have undergone smoothing using a median filter (Figure 3.1), as suggested by Córdoba et al. [12], so some spatial accuracy was lost in favor of easier management. Multiple passes of an edge-detection filter (*boundaries* function in `raster` R-package) defines the feasible space by excluding a buffer from the field perimeter and from the SSMUs boundaries.

3.4 Notation

The unified table of ancillary data layers was used to search for an optimal sample design according to the following notation.

Let N denote the number of feasible sampling sites, each associated with spatial (x, y)-coordinates, $\{(x_i, y_i)\}_{i=1}^{N}$ and ancillary data layers which assumed to be acquired in dimensionality k, represented by vectors in $\mathbb{R}^k$ at each of the N sites.

Let $\mathcal{A}$ denote the corresponding $N \times k$-dimensional ancillary data matrix, and importantly, let it define the *feature space*. At the same time, every sampling-point is associated with spatial (x, y)-coordinates, $\{(x_i, y_i)\}_{i=1}^{N}$, subscribing to the so-called *geographical space*, which is defined by the field's boundaries and

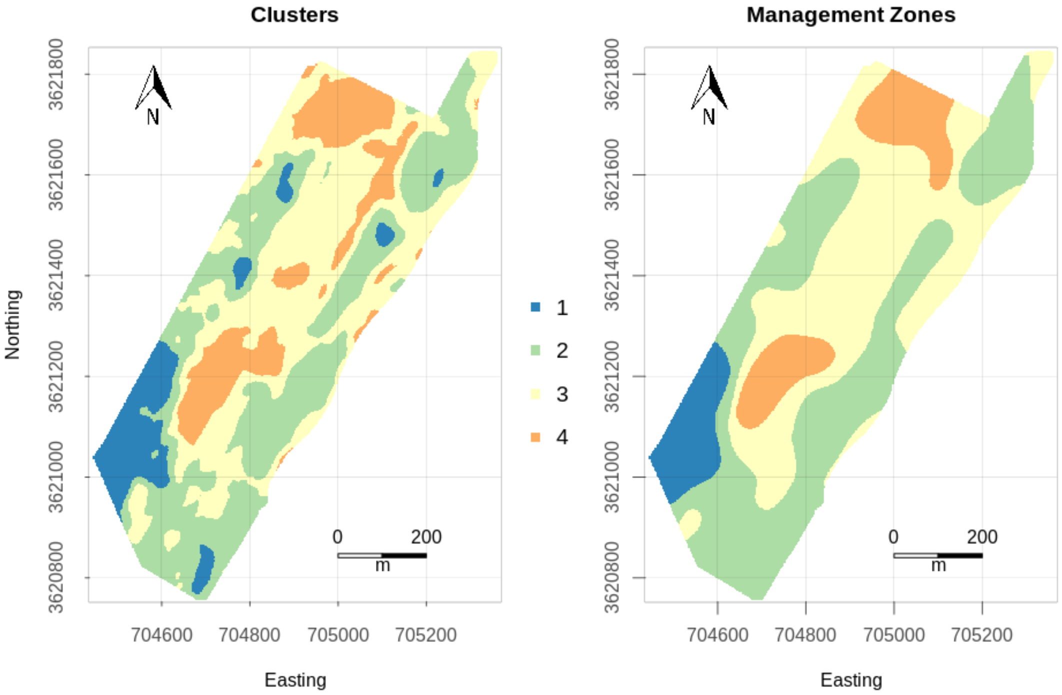

Figure 3.1: Clusters resulting from fuzzy c-means clustering (left) and smooth management zones (SSMUs) after successive passes of median filter (right).

operational constraints. These N pairs of coordinates constitute the geographical data matrix $\mathcal{G}$:

$$\mathcal{G} = \begin{pmatrix} x_1 \; y1 \\ x_2 \; y2 \\ \vdots \\ x_N \; y_N \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,k} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,k} \\ & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \cdots & \alpha_{N,k} \end{pmatrix} \tag{3.2}$$

Pairwise distance calculations use the Euclidean metric, denoted by $d_{i,j}^{(\mathcal{G})}$ for every $i, j \in \{1, \ldots, N\}$.

In our notation $\mathcal{Z}$ represents the *augmented* data matrix of dimension $N \times (k+2)$, encompassing the N sites' ancillary data and their geographical coordinates:

$$\mathcal{Z} = \left(\mathcal{G} | \mathcal{A} \right). \tag{3.3}$$

The ultimate target is to form a sampling-plan by locating $\mathbf{n} \ll \mathbf{N}$ sites, whose ancillary data vectors best represent the feature space's distribution, and at the same time, are spatially disperse concerning the geographical space. At this point, it is assumed that the user provides a value of n and the discussion on setting this value is postponed to Section 3.9.

Formally, a candidate sampling-plan p is a mapping π indicating the subset selection of the n indices. Importantly, a candidate sampling-plan p is associated with the following components:

I An ancillary data matrix of dimension $n \times k$, denoted by $\mathbf{A}^{(p)}$, whose rows constitute a subset of $\mathcal{A}$'s rows adhering to the mapping π

$$\mathbf{A}^{(p)} = \begin{pmatrix} \alpha_{\pi(1),1} & \alpha_{\pi(1),2} & \cdots & \alpha_{\pi(1),k} \\ & \vdots & \ddots & \vdots \\ \alpha_{\pi(n),1} & \alpha_{\pi(n),2} & \cdots & \alpha_{\pi(n),k} \end{pmatrix} \tag{3.4}$$

II A geographical data matrix $\mathbf{G}^{(p)}$, defined in an equivalent manner

$$\mathbf{G}^{(p)} = \begin{pmatrix} x_{\pi(1)} \; y_{\pi(1)} \\ \vdots \\ x_{\pi(n)} \; y_{\pi(n)} \end{pmatrix} \tag{3.5}$$

III An augmented matrix $\mathbf{Z}^{(p)}$:

$$\mathbf{Z}^{(p)} = \left(\mathbf{G}^{(p)} | \mathbf{A}^{(p)}\right) . \tag{3.6}$$

Thus, the sample planning process can be translated into obtaining the subset selection mapping π as a combinatorial optimisation problem in the domain $\mathbf{Z}^{(p)}$.

3.5 Conditioned Latin Hypercube Sampling

Minasny and McBratney [43] devised a method for obtaining optimal sampling design in the presence of ancillary data, namely cLHS. This method solves sampling design as a single-objective optimisation task. It aims to maximally stratify the multivariate distribution of ancillary data layers by forming a Latin hypercube of their *quantiles*, while preserving the structure of their statistical correlation, for a representation of the full information spectrum. Importantly, this method accounts only for the feature space.

Given the ancillary data matrix $\mathbf{A}$, the idea is to compute n statistical quantiles per each of its k ancillary data column, then aim to place a single member of each quantile in the sample. Given a candidate sampling-plan p of n sites, defined by a mapping π, let η hold histogram information of $\mathbf{A}^{(p)}$ with respect to $\mathbf{Z}$ in the following manner: given the i^{th}-quantile of $\mathbf{A}$'s j^{th} column, $q_j^{(i)}$, the element $\eta^{(p)}\left[q_j^{(i)} \leq \alpha_{\pi(i),j} < q_j^{(i+1)}\right]$ is the number of occurrences of $\alpha_{\pi(i),j}$ with values in the corresponding quantile. Accordingly, the first evaluation criterion within cLHS is defined by

$$\psi_1\left(\mathbf{A}^{(p)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left|\eta\left[q_j^{(i)} \leq \alpha_{\pi(i),j} < q_j^{(i+1)}\right] - 1\right| . \tag{3.7}$$

Also, let $\mathbf{C}^{(\mathbf{A})}$ and $\mathbf{C}^{(\mathbf{A}^{(p)})}$ denote the *correlation matrices* of $\mathbf{A}$ and $\mathbf{A}^{(p)}$, respectively, both $k \times k$-dimensional. The second evaluation criterion is the following:

$$\psi_2\left(\mathbf{A}^{(p)}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\left|\mathbf{C}_{i,j}^{(\mathbf{A})} - \mathbf{C}_{i,j}^{(\mathbf{A}^{(p)})}\right| . \tag{3.8}$$

For categorical data such as soil classification the sub-objective is to match the probability distribution for each of the classes. With strictly continuous ancillary data cLHS defines an objective function for evaluating p's quality as a weighted sum of the two criteria ($\omega_1, \omega_2 > 0$):

$$f_{\text{cLHS}}\left(p\right) = \omega_1 \cdot \psi_1\left(\mathbf{A}^{(p)}\right) + \omega_2 \cdot \psi_2\left(\mathbf{A}^{(p)}\right) \longrightarrow \min . \tag{3.9}$$

As recommended for general purpose [43] we set the weights $\omega_1 = \omega_2 = 1$. To achieve dimensionless-scaling, the function f_{cLHS} is normalized (i.e., divided by number of sampling points n times ancillary data layers k).

In terms of problem-solving, cLHS operates with a dedicated variation operator, which swaps a random site within the subset mapping π of the candidate sampling-plan p with one of the elements in its complement π^C, to obtain $\tilde{\pi}$ (defining $\tilde{p}$):

$$\{p, \pi\} \rightsquigarrow \{p', \pi'\} \text{ such that } \delta\left(\pi, \pi'\right) = 1, \tag{3.10}$$

where δ counts the differing subsets' attributes.

Overall, cLHS culminates at a perfect stratification of the feature space, yet it does not account for the geographic dispersion of the sampling points, thus results in many impractical solutions. It was suggested to run cLHS for a subset of the points, followed by a space-filling algorithm for the remaining points [62]. In practice, we have found that this procedure still produces inefficient solutions, and suggest as a remedy to augment the cLHS objective function with a spatial dispersion objective function, as described in the following subsection.

3.6 max-min Diversity

The max-min diversity is one of the simplest notions for promoting dispersion. Despite the simplicity of this diversity indicator, finding maximally diverse subsets is an $\mathcal{NP}$-hard problem [38]. Here, it aims to maximize the minimal pairwise (geographical) distances among all sampling points:

$$f_{d^{(\mathbf{G})}_{\min}}\left(p\right) = \min_{\pi(i),\pi(j)} \left\{ d^{(\mathbf{G})}_{\pi(i),\pi(j)} \right\} \longrightarrow \max \qquad i, j \in 1, \ldots, n,\ i \neq j. \tag{3.11}$$

3.7 Bi-Objective Formulation

In the context of this study, we are interested in investigating the competition between spatial dispersion and feature space coverage. Accordingly, given the model functions presented in Sections 3.5 and 3.6, we formulate a bi-objective optimisation task. For the sake of compatibility, we compute the multiplicative inverse of f_2, so that all objectives are subject to minimization:

$$\boxed{\begin{aligned} &\textbf{[P0]} \\ &f_1 := f_{\text{cLHS}}\left(p\right) \longrightarrow \min \\ &f_2 := 1/f_{d^{(\mathcal{G})}_{\min}}\left(p\right) \longrightarrow \min \end{aligned}} \tag{3.12}$$

3.8 Algorithmic Approach

To solve the bi-objective task **P0** (Eq. 3.12) we used the renowned NSGA-II algorithm [16], utilizing the `ecr` R-package [7]. We set the parental and offspring population sizes both to $\mu = \lambda = 10$, and the maximally available iterations to 50,000. A simple mutation operator tailored to the current domain swaps a random point in the sample set $\mathbf{Z}^{(p)}$ with a random member of its complementary set $(\mathbf{Z}^{(p)})^C$, as outlined by Algorithm 4.

The vector of objective functions is evaluated using **P0** (Eq. 3.12). The cLHS fitness function, inspired by the `clhs` R-package [51], first segments each ancillary data variable into n iso-probable quantiles (strata) according to its CDF (Cumulative Distribution Function) and calculates the correlation matrix $\mathbf{C}^{(\mathcal{A})}$.

The optimisation procedure starts by initializing a population of μ individuals, each constituting a candidate sampling-plan comprising n random points. It then iterates over a serial execution of the variation and the selection operations and terminates at the predefined maximal number of objective

Algorithm 4 The utilized mutation operator. Upon receiving a sampling-plan π as input, $\mathcal{P}^C$ represents the subset of unsampled sites. The operator swaps a single site $\pi(i_{\text{rmv}})$ with a uniformly random unsampled site i_{add}. The modified mapping π' is returned as output.

```
Input: $\pi, N$ {set of sampling points, overall number of points}

$n \leftarrow$ length($\pi$)
$\mathcal{P}^C \leftarrow \{1, \ldots, N\} \setminus \{\pi(1), \ldots, \pi(n)\}$
$i_{\text{rmv}} \leftarrow$ uniformly randomly from $\{1, \ldots, n\}$
$i_{\text{add}} \leftarrow$ uniformly randomly from $\mathcal{P}^C$
$\pi' = (\pi(1), \ldots, \pi(i_{\text{rmv}}) \rightsquigarrow i_{\text{add}}, \ldots, \pi(n))$

Return: $\pi'$
```

function evaluations. Technically, through CPU parallelization, 30 independent optimisation runs are concurrently executed to ensure sufficient replications. Results of the last generation of all runs are Pareto-sorted to exclude the dominated solutions, producing a portfolio of Pareto-optimal sampling-plans to consider.

3.9 Sample-Size Identification

The choice of the sample-size n reflects the economic concept of *marginal profit*, as each additional sampling-point presumably improves prediction accuracy at the cost of increased operational expenses. We seek to quantify the information gain by increments of sample-size, in a workflow of multiple optimization tasks (Section 3.8) with n varying within a pragmatic budget range, in a modus operandi that resembles cluster validity indices (Section 3.3). Several representative solutions are selected from the approximated Pareto frontier of each sample-size for evaluation by the following information-theoretic metrics:

1. **AIC**, which provides an approximation of a statistical model´s relative

predictive accuracy, as measured by out-of-sample deviance [41]. First, a linear model is fitted with a single ancillary data layer as a independent variable $\mathbf{Y}$ and all other p layers $\mathbf{X}$ as dependent variables:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{3.13}$$

Assuming $\widehat{L}$ is the maximum value of the likelihood function of $\mathbf{Y}$, and p is the number of estimated parameters in the model, then AIC value of the model is calculated by

$$\text{AIC} = 2p - 2\ln(\widehat{L}) \tag{3.14}$$

These two steps are repeated for each ancillary data layer, and the mean AIC value is returned.

2. **MOKV**, calculated as the mean variance resulting from ordinary Kriging interpolation of ancillary data values at sampling locations onto the entire field (using `gstat` R-package). As this operation is time-consuming, we opted to run it only for ECa$_V$ and ECa$_H$.
3. $\mathbf{D_{KL}}$, derived from the ratio of ancillary data distributions in a sample to those of the entire field, given by

$$D_{KL}\left(\mathbf{A} \mid\mid \mathbf{A}^{(p)}\right) = -\sum \mathbf{Pr}(\mathbf{A}) \log\left(\frac{\mathbf{Pr}(\mathbf{A}^{(p)})}{\mathbf{Pr}(\mathbf{A})}\right). \tag{3.15}$$

To calculate KL divergence, $\mathbf{Pr}(\mathbf{A})$ is computed as the CDF of ancillary data layers of the entire field, and $\mathbf{Pr}(\mathbf{A}^{(p)})$ is the CDF of ancillary data at locations given by a candidate sample, then $\mathbf{D_{KL}}$ is computed via the *KL.plugin* function in `entropy` R-package.

These three quantifiers are negatively-proportional to a model´s quality, thus a descending performance is expected as the sample-size grows. The results are averaged for each sample-size, hence by observing the charts one can gain insight into the relation between the number of points and the latent quality of a sample. We are seeking a satisfactory sample-size that could be manifested as a local minima or a knee-point, at which the rate of improvement decreases.

3.10 Scheme Selection Criteria

Once an applicable sample-size n is found (Section 3.9), decision-making needs to take place post-optimization, for selecting a singular sampling plan among the candidate solutions. To this end, the same information quantifiers (AIC, MOKV and D_{KL}) are used to evaluate individual solutions. The performance on each metric is ordered by ranking (1 for the best solution, 2 for the next and so on). The ranks of each sample are aggregated to form a cumulative performance rank that supports decision-making, which is likely to account for additional practical and subjective aspects.

Chapter 4

Experimental Results

Summary

In this chapter we report on our empirical findings in two real-field settings. The utilized units abbreviations read: [ha] for Hectare, [m] for meters.

4.1 Sampling for validation of SSMU delineation

As a part of a PA research focusing on fertilizer management, we have tested the aforementioned methods to devise a soil-survey plan for a 37 ha plot in Newe Ya'ar Research Center, located in Jezreel valley, northern Israel.

Ancillary data measurements were collected in the field right after winter wheat was sown. EMI data was recorded on late November, 2018, following a rain storm when soil moisture content was expected to be close to field capacity, yielding four proximal sensing layers (ECa$_V$, ECa$_H$,MSa$_V$ and MSa$_H$). NDVI layer was constructed from data of a multi-spectral imaging campaign in the field on mid December, 2018.

The pre-processing stage followed the procedure described in Section 3.2, involving ECa and MSa data compaction, log transformation and OK interpolation to a grid at a resolution of 1×1 m, followed by a crop to field boundaries and normalization.

The optimal number of clusters was determined by the suite of validity tests described in Section 3.3, which assess the average compactness and separation of the partitions generated by the algorithm. Although the PE and PC indices did not show any local extrema (not shown), the FS and CHC tests agreed on four management zones as an optimal number of clusters (Figure 4.1).

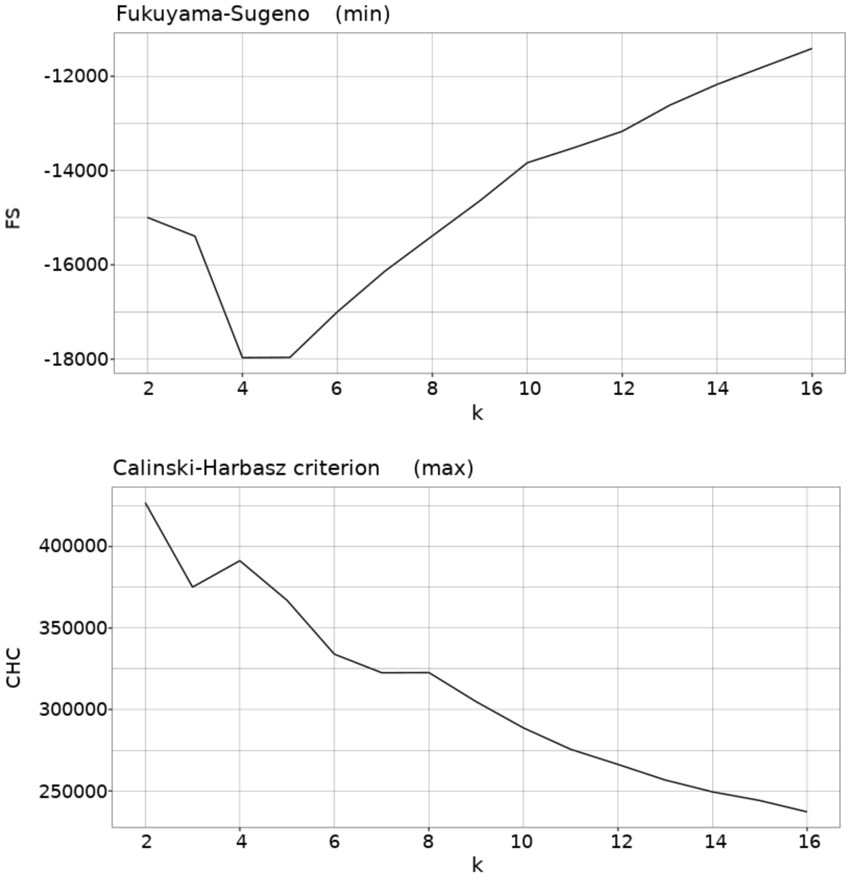

Figure 4.1: The fuzzy c-means validation indices used to determine the optimal number of SSMUs of the study area, where k represents the number of clusters.

The field was then divided into four SSMUs using Fuzzy-c-means clustering of the matrix $\mathcal{A}$, which were subsequently smoothed with a median filter to reduce zone fragmentation [54]. The feasible search space was defined by exclusion of a 14 m buffer from field boundaries, and 7 m buffer from SSMUs´ boundaries.

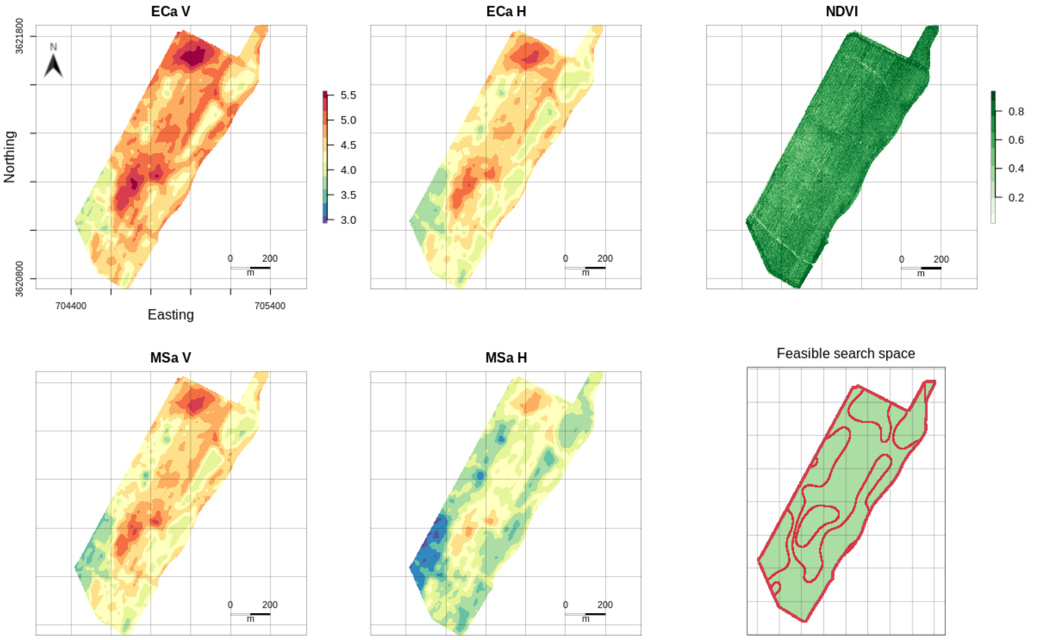

Figure 4.2: Ancillary data layers: ECa and MSa in vertical and horizontal modes, NDVI, and feasible search area (green).

Figure 4.2 provides a summary of ancillary data layers and the search space, where red areas represent high ECa and MSa values, green areas represent mid-range values while blue areas exhibit the lowest ECa and MSa values, probably due to coarse texture and low soil moisture content. It seems that the NDVI data add little to the clustering power because of relatively low spatial variations in the NDVI signal of bare soil.

The adapted NSGA-II was executed in 30 parallel runs with the number of points varying in our practical budget range $n \in \{10, 12, \ldots, 48, 50\}$ to solve **P0** (Eq. 3.12), in the configuration described in Section 3.8. The *Hypervolume Indi-*

cator (HVI) is a benchmark for the size of the subspace dominated by the evolving Pareto frontier, bound from above by an arbitrary reference point [20]. The progress of the HVI during 30 parallel runs for sampling-plans with $n = 26$ points is depicted in Figure 4.3, exhibiting convergence. The corresponding 300 solution are displayed in Figure 4.4 as points over the objective space with their associated ranks, and the cumulative Pareto frontier approximation as well. By scrutinizing three Pareto-optimal solutions of each sample-size n with respect to the information-theoretic quantifiers described in Section 3.9, we identified a certain sample-size, $n^* = 22$, beyond which model improvement is decaying, manifested as local minima in Figure 4.5. Upon selecting this sample-size, additional two sets of solutions were generated to address an operational constraint requiring at least 3 points per SSMU, repeating the optimization task twice with 22 points.

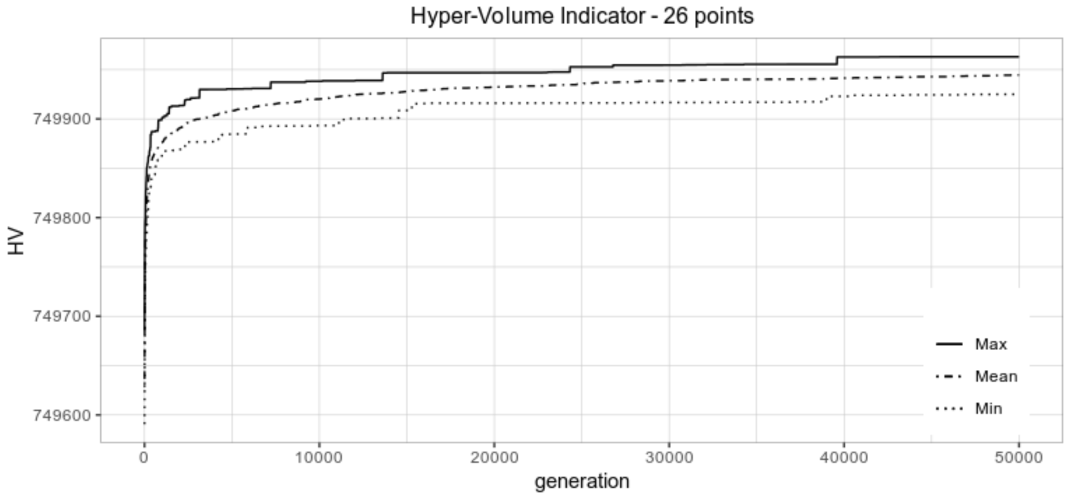

Figure 4.3: Statistical summary of the *Hypervolume Indicator*'s evolution along 30 runs, each featuring 50,000 generations, with $n = 26$ points.

To guide the choice of a particular plan, the Pareto-optimal solutions for $n = 22$ (Figure 4.6) were evaluated by the measures $MOKV$ and D_{KL} and ordered by rank accordingly. Evidently, aggregated ranks (Figure 4.7) suggest

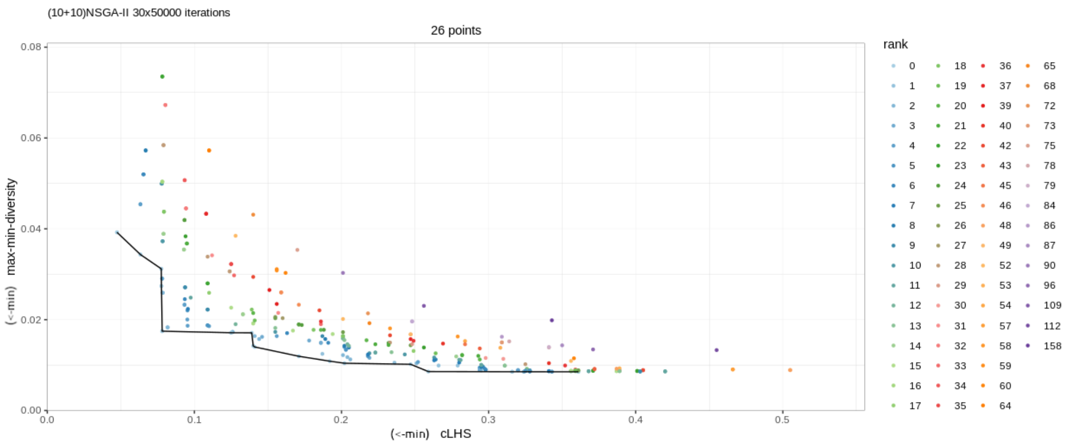

Figure 4.4: All solution-points, their rank and the Pareto front (rank 0, black line) obtained by optimization of cLHS and max-min-diversity with $n = 26$ points.

Sample 13 as the best candidate plan. Inspection of the plan revealed that it was not well spatially distributed. As Figure 4.6 suggest, this may be attributed to its location on one extreme of the frontier. The selection process proceeded to the next candidate, namely *Sample 21* (Figure 4.8) which met the requirements, and thus was selected as a *blueprint*.

Aftermath Provided with this algorithmically generated sampling-plan, the agricultural field has been precisely sampled according to its prescription, followed by laboratory analysis of more than 30 soil physico-chemical attributes in 4 depths. The soil analysis results were used to assess the validity of the SSMU delineation by a binary tree classification with soil properties as independent variables, and SSMU assignment as dependent variable. The results showed classification was accurate in 90.9% (20/22) of the cases, referring the top-soil layer (0-30 cm).

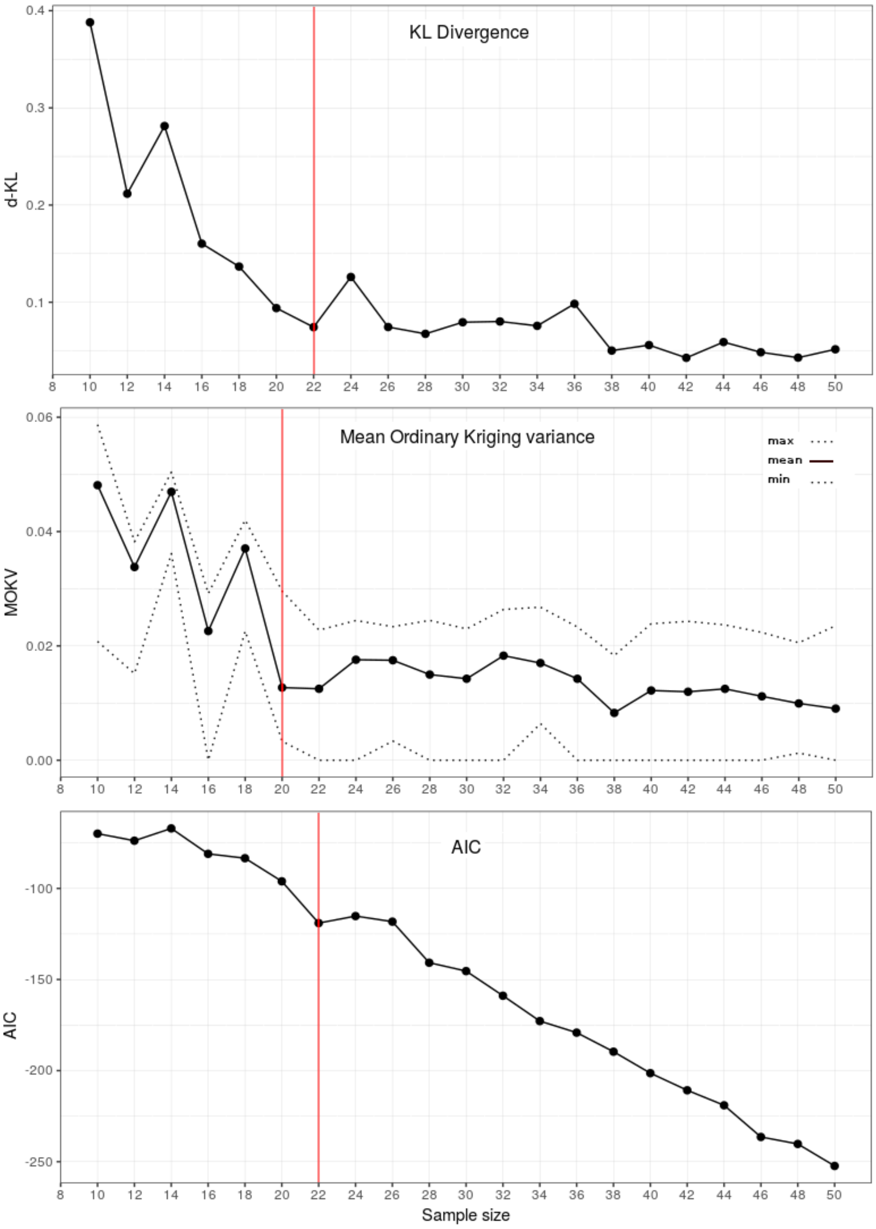

Figure 4.5: Statistical measures by sampling size. A knee-point is apparent (marked by red line) in all measures at 22 (D-KL, AIC), and 20 (MOKV), indicating a recommended sample-size of 22 points.

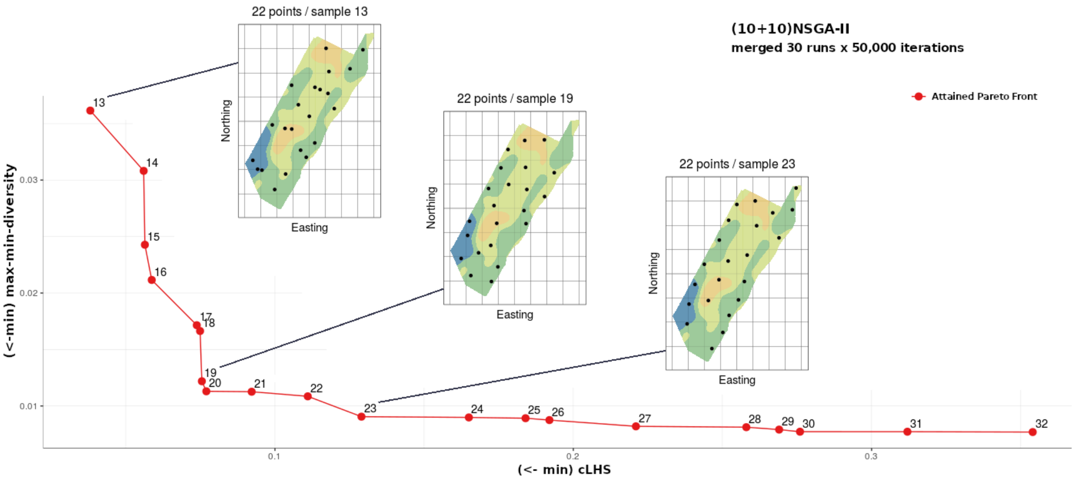

Figure 4.6: Attained Pareto front for samples with 22 points and some candidate solutions with their respective position.

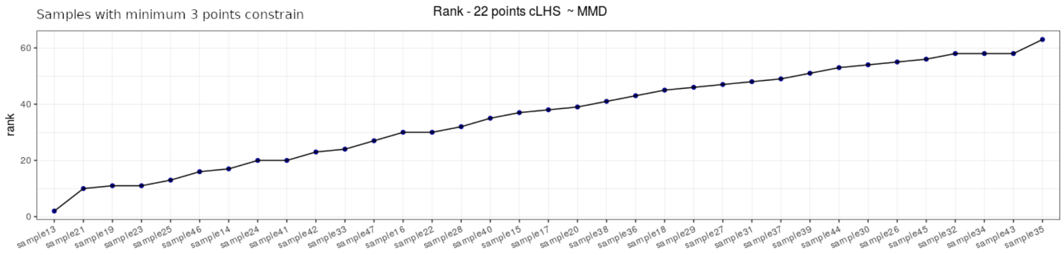

Figure 4.7: Ranking of sampling plans with 22 points by cumulative performance of statistical measures (D-KL, MOKV).

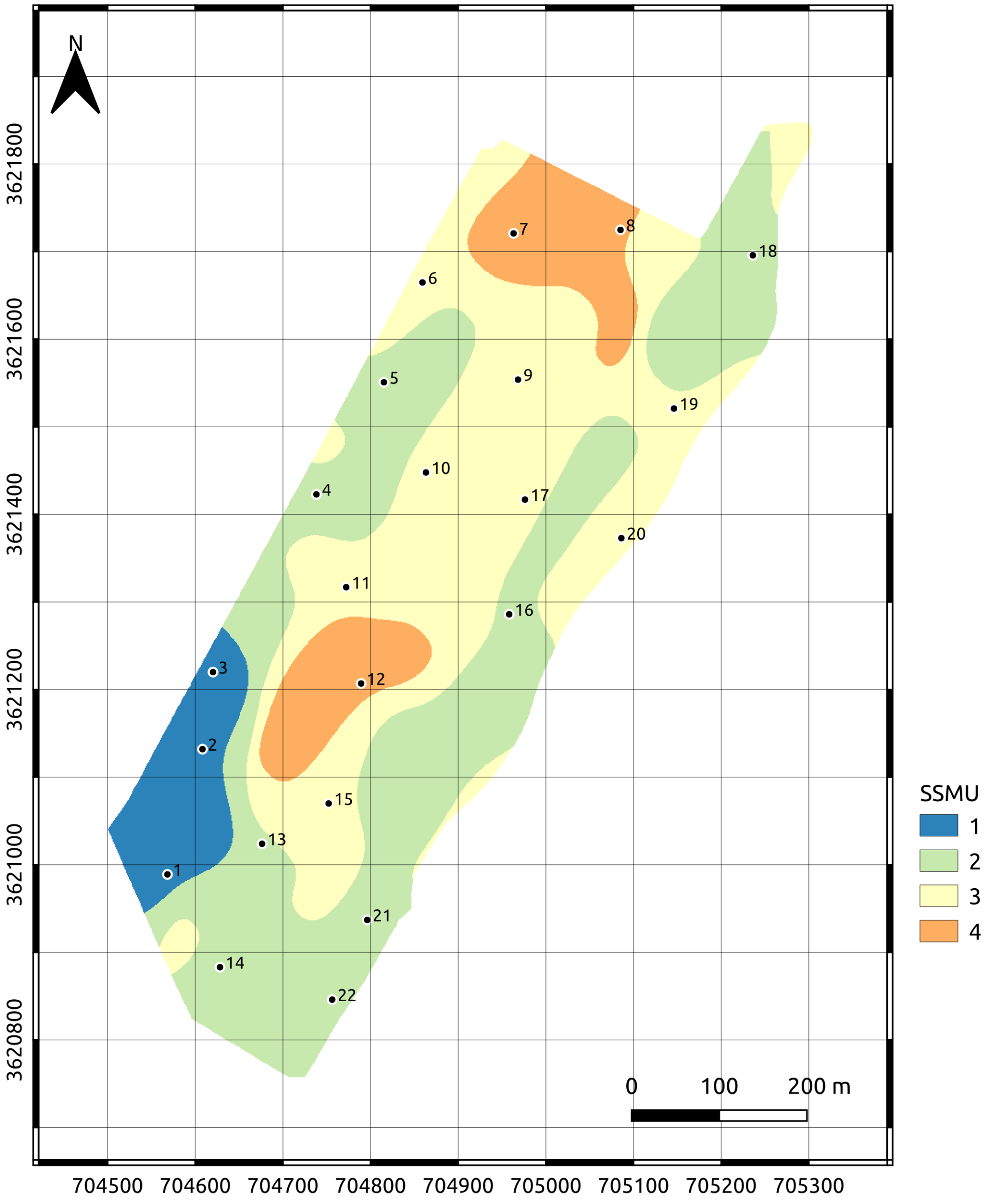

Figure 4.8: The selected soil sampling-plan in the 37 ha field, superimposed on management units.

4.2 Sampling for Calibration of Thermal Infra-Red Data

Another field experiment considered only a single ancillary data layer of Thermal Infra-Red (TIR) measurements in a 10 ha sub-plot of the 37 ha field in Newe Ya'ar (Section 4.1), with the aim of fitting a model between the TIR values and the soil texture (i.e., fractions of sand, clay and silt), so a spatial prediction of these sparsely measured attributes can be done based on the dense TIR data.

The TIR image values were normalized within $[0, 1]$. The feasible sampling area (Figure 4.9) was defined by exclusion of 7 m buffer from field boundaries and omission of the areas covered by vegetation, defined by Green-Red Vegetation Index (GRVI) data, derived from an aerial RGB image.

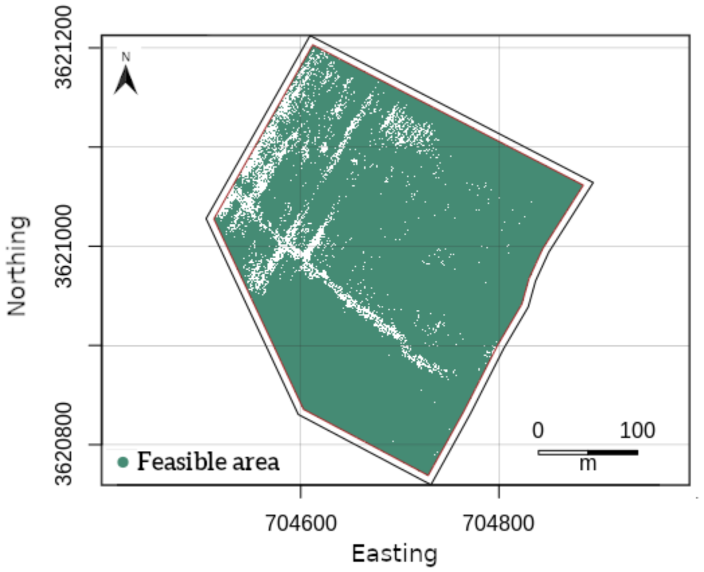

Figure 4.9: The feasible search space in the 10 ha plot.

A sample-size of 12 location was pre-defined by the expert commissioning the survey, according to the requirements and applicability constrains. This dictated a single run of 30 parallel optimization tasks (described in Section 3.8), resulting with 300 solutions that were later screened out of the dominated solutions, portraying the efficient frontier (Figure 4.10). The *Hypervolume Indicator* progress (Figure 4.11) exhibits a convergence profile after about 2000 iterations, suggesting that these results could be obtained with less computation.

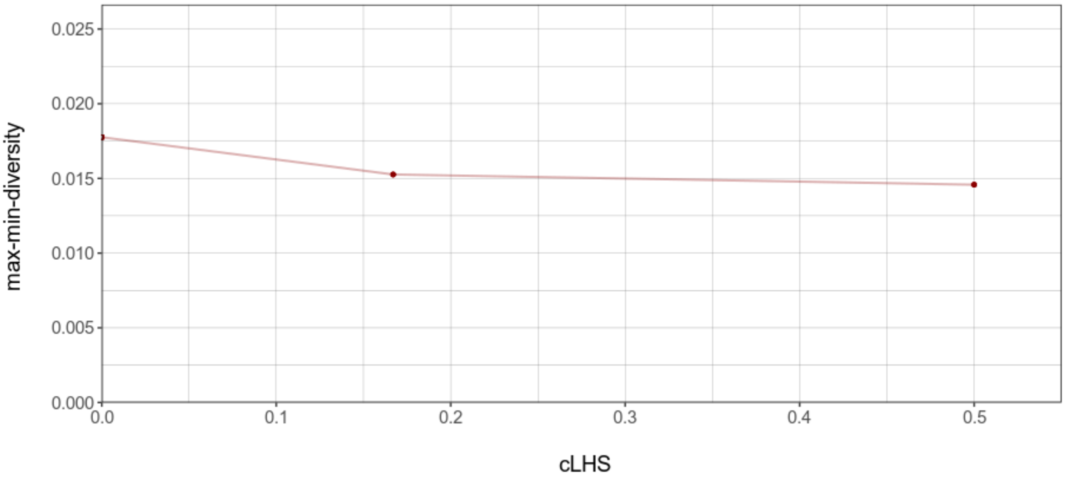

Figure 4.10: Solutions points on the attained Pareto frontier (line) for schemes with 12 sampling locations in the 10 ha plot. Both function are subject to minimization.

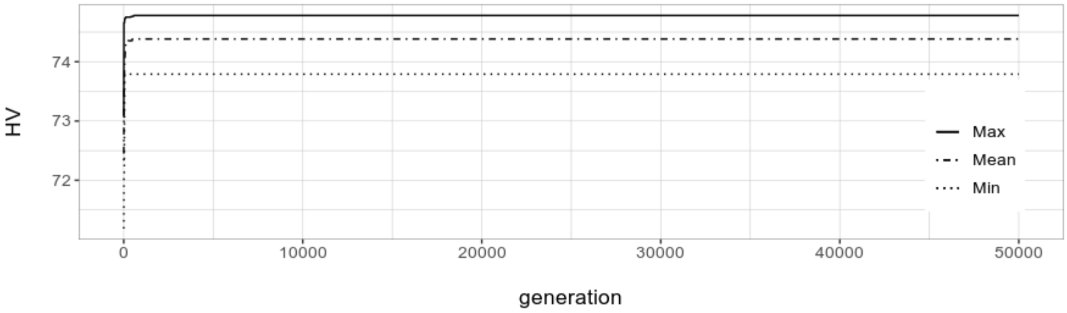

Figure 4.11: Statistical summary of the *Hypervolume Indicator*'s evolution along the run of 50,000 iterations (max, mean, min).

As the aim of this survey is to identify correlations between variables, we selected only solutions with perfect stratification (i.e. cLHS score of 0), such that the entire information spectrum is represented in the sample. This narrowed down the selection to 9 optimal sampling schemes, which were evaluated by the surveyor, accounting for considerations such as spatial dispersion and vegetation growth since the last thermal imaging campaign. Scheme #2 (Figure 4.12) was selected and implemented for sampling.

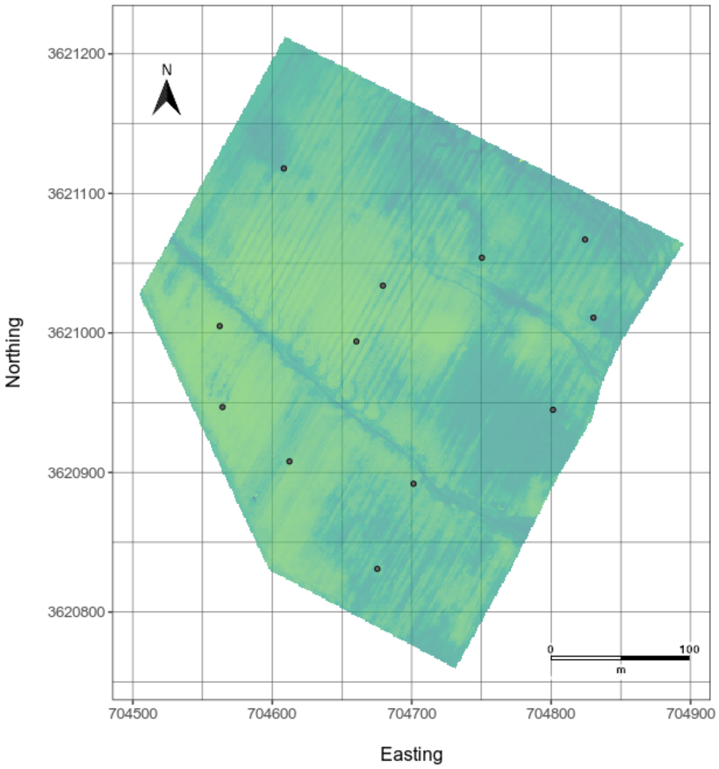

Figure 4.12: The selected sampling-plan in the 10 ha plot.

Chapter 5

Discussion

5.1 Summary and Conclusions

In this study we have demonstrated that multiobjective optimization with simultaneous targets of geographic dispersion and feature space stratification is a suitable approach for soil sampling design. In the use-cases considered here, application of EMOA in a real farm with cLHS and max-min-diversity as objective functions produced many feasible solutions, found according to the survey's purpose on one edge of the approximated Pareto frontier when stratification is beneficial for calibration, or at the knee-point area – offering an apt compromise between the objectives.

A-priori evaluation of a sampling-plan quality is a key for an informed process of sampling scheme selection. Several information-theoretic quantifiers based on available ancillary data have been presented herein, alongside their application to qualify actual sampling schemes and to optimize the sample-size n, providing a decision-support tool for soil-survey planning.

Clearly, the proposed procedure can be improved, notably by introducing a more sophisticated variation operator which preserves a perfect Latin hypercube state, by devising additional information criteria for model evaluation, by revising the selection process of candidate solutions to be included and by adding a termination criterion on HVI stagnation.

The versatility of the proposed approach makes it suitable to use with different objective functions and for different sampling scenarios. A comparative analysis with other sampling methods is beyond the scope of this study, although it could assess whether the relative complexity of this procedure is worthwhile.

Next, we propose a possible direction of future research.

5.2 Future Work: Solow-Polasky Diversity

We want to consider the so-called Solow-Polasky Diversity [56; 60] as another dispersion measure. It has been proposed in the field of biodiversity conservation as a statistical measure for the diversity of a population of individuals, given by a set of vectors in a metric space. Given pairwise distances between sites i and j, d_{ij} (either within the geographical or the feature space), let $\Psi := (\psi_{ij}) \in \mathbb{R}^{n \times n}$ be constructed with matrix elements $\psi_{ij} = \exp(-\gamma \cdot d_{ij})$. Then, the Solow-Polasky Diversity is defined as:

$$D_{SP} = \vec{1}^T \Psi^{-1} \vec{1}, \tag{5.1}$$

with $\vec{1}$ denoting a vector of n ones, i.e., the summation is over all the elements of Ψ^{-1}; γ is a domain-specific *normalization factor*. The Solow-Polasky Diversity strives to quantify the number of existing species within a given population. It obtains its minimum at 1, meaning that the community consists of only one species, and its maximum at n (the number of points), meaning that every aspect is a unique species. Thus, the larger this scalar, the more diverse the sample is.

Importantly, in the current study, the Solow-Polasky Diversity can be applied both to the feature space as well as to the geographical space. Given a sampling plan p defined by a mapping π, two measures can be computed, using either $\left\{d^{(\mathcal{G})}_{\pi(i),\pi(j)}\right\}$ or $\left\{d^{(\mathcal{A})}_{\pi(i),\pi(j)}\right\}$, denoted as

$$D^{(\mathcal{G})}_{SP}(p), \quad D^{(\mathcal{A})}_{SP}(p),$$

respectively. We then formulate another bi-objective optimization problem:

$$\boxed{\begin{array}{l} \textbf{[P1]} \\ f_3 := 1/D^{(\mathcal{A})}_{SP}(p) \longrightarrow \min \\ f_4 := 1/D^{(\mathcal{G})}_{SP}(p) \longrightarrow \min \end{array}} \tag{5.2}$$

Preliminary calculations indicate that this is a promising direction. At the same time, the objective functions exhibit sensitivity to the defining normalization factor γ, which requires further investigation. Another interesting approach would be to look into low discrepancy sampling methods [17], which provide the promise of small approximation errors when combined with regression models, but at the same time are computationally challenging.

Parts of this thesis were published in a conference proceedings [31]. We intend to advance this study further through ongoing research.

References

[1] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 6 (December 1974), 716–723. (5)

[2] Bäck, T. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, Inc. New York, NY, USA, 1996. (10)

[3] Bandyopadhyay, S., Saha, S., Maulik, U., and Deb, K. A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation 12* (2008), 269–283. (3)

[4] Barca, E., Castrignanó, A., Buttafuoco, G., De Benedetto, D., and Passarella, G. Integration of electromagnetic induction sensor data in soil sampling scheme optimization using simulated annealing. *Environ Monit Assess 187* (2015). (5)

[5] Bezdek, J. C. Cluster validity with fuzzy sets. *Journal of Cybernetics 3*, 3 (1973), 58–73. (26)

[6] Bezdek, J. C. Mathematical models for systematics and taxonomy. (26)

[7] Bossek, J. ECR 2.0: A modular framework for evolutionary computation in R. In *Proceedings of the Genetic and Evolutionary Computation Conference*

Companion (New York, NY, USA, 2017), GECCO '17, ACM, pp. 1187–1193. (32)

[8] BOYD, S., AND VANDENBERGHE, L. *Convex Optimization*. Cambridge University Press, New York, 2004. (8)

[9] BRABAZON, A., O'NEILL, M., AND MCGARRAGHY, S. *Natural Computing Algorithms*, 1st ed. Springer Publishing Company, Incorporated, 2015. (10, 11, 12, and 13)

[10] BRUS, D. J. Sampling for digital soil mapping: A tutorial supported by r scripts. *Geoderma 338* (2019), 464–480. (3)

[11] COELLO, C. A. C., LAMONT, G. B., AND VELDHUIZEN, D. A. V. *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Springer-Verlag, Berlin, Heidelberg, 2006. (17)

[12] CÓRDOBA, M., BRUNO, C., COSTA, J., PERALTA, N. R., AND BALZARINI, M. Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems Engineering 143* (03 2016), 95–107. (26)

[13] CORWIN, D., AND LESCH, S. Characterizing soil spatial variability with apparent soil electrical conductivity: I. survey protocols. *Computers and Electronics in Agriculture 46*, 1 (2005), 103–133. Applications of Apparent Soil Electrical Conductivity in Precision Agriculture. (23)

[14] DARWIN, C. *The Origin of Species: By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859. (11)

[15] Deb, K. Multi-objective optimisation using evolutionary algorithms: An introduction. In *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing* (2011). (10, 14, 16, 17, 18, and 20)

[16] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Trans. Evol. Comp 6*, 2 (Apr. 2002), 182–197. (18, 19, 20, 21, 22, and 32)

[17] Doerr, C., Gnewuch, M., and Wahlström, M. *Calculation of Discrepancy Measures and Applications*. Springer International Publishing, Cham, 2014, pp. 621–678. (49)

[18] Doolittle, J. A., and Brevik, E. C. The use of electromagnetic induction techniques in soils studies. (24)

[19] Ehrgott, M. *Multicriteria Optimization*, second ed. Springer, Berlin, 2005. (15)

[20] Emmerich, M., Beume, N., and Naujoks, B. An EMO algorithm using the hypervolume measure as selection criterion. In *Evolutionary Multi-Criterion Optimization* (Berlin, Heidelberg, 2005), C. A. Coello Coello, A. Hernández Aguirre, and E. Zitzler, Eds., Springer Berlin Heidelberg, pp. 62–76. (39)

[21] Emmerich, M., and Deutz, A. H. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing 17*, 3 (Sep 2018), 585–609. (14, 16, and 17)

[22] Farahani, H., and Flynn, R. Map quality and zone delineation as affected by width of parallel swaths of mobile agricultural sensors. *Biosystems Engineering 96* (02 2007), 151–159. (24)

[23] FONSECA, C. M., AND F., P. J. Genetic algorithms for multiobjective optimization: Formulationdiscussion and generalization. In *Proceedings of the 5th International Conference on Genetic Algorithms* (San Francisco, CA, USA, 1993), Morgan Kaufmann Publishers Inc., pp. 416–423. (14)

[24] FUKUYAMA, Y., S. M. A new method of choosing the number of clusters for the fuzzy c-mean method. *Proc. 5th Fuzzy Syst. Symp., 1989* (1989), 247–250. (26)

[25] GAO, B., PAN, Y., CHEN, Z., WU, F., REN, X., AND HU, M. A spatial conditioned latin hypercube sampling method for mapping using ancillary data. *Transactions in GIS 20*, 5 (2016), 735–754. (3)

[26] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013. (26)

[27] HENGL, T., ROSSITER, D., AND STEIN, A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research 41 (2003) 8 41* (01 2003). (3)

[28] HEUVELINK, G. B. M., AND PEBESMA, E. J. Is the ordinary Kriging variance a proper measure of interpolation error? (5)

[29] HILLEL, D. Soil fertility and plant nutrition. In *Soil in the Environment*, D. Hillel, Ed. Academic Press, San Diego, 2008, pp. 151 – 162. (1)

[30] HOLLAND, J. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975. (12)

[31] ISRAELI, A., EMMERICH, M., LITAOR, M. I., AND SHIR, O. M. Statistical learning in soil sampling design aided by pareto optimization. In *Proceed-*

ings of the Genetic and Evolutionary Computation Conference (New York, NY, USA, 2019), GECCO '19, ACM, pp. 1198–1205. (49)

[32] KERRY, R., OLIVER, M., AND FROGBROOK, Z. *Sampling in Precision Agriculture*. 07 2010, pp. 35–63. (2)

[33] KERRY, R., AND OLIVER, M. A. Variograms of ancillary data to aid sampling for soil surveys. *Precision Agriculture 4*, 3 (Sep 2003), 261–278. (5)

[34] KHOSLA, R. Science breakthroughs 2030: Transforming food and agriculture research. *CSA News 63*, 9 (2018), 14–15. (2)

[35] KOCHENDERFER, M., AND WHEELER, T. *Algorithms for Optimization*. The MIT Press. MIT Press, 2019. (8)

[36] KONAK, A., COIT, D., AND SMITH, A. E. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering and System Safety 91*, 9 (9 2006), 992–1007. (12, 13, 14, 17, and 20)

[37] KULLBACK, S., AND LEIBLER, R. A. On information and sufficiency. *Ann. Math. Statist. 22*, 1 (03 1951), 79–86. (5)

[38] KUO, C., GLOVER, F., AND DHIR, K. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences 24*, 6 (1993), 1171–1185. (31)

[39] LARK, R. M. Multi-objective optimization of spatial sampling. *Spatial Statistics 18* (2016), 412 – 430. (3)

[40] MATHERON, G. *Les Variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson, Paris, 1965. (4, 25)

[41] McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall / CRC Press, 2016, p. 189. (34)

[42] Miettinen, K. *Nonlinear multiobjective optimization*, vol. 12. Springer, Berlin, 2012. (13)

[43] Minasny, B., and McBratney, A. B. A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences 32*, 9 (2006), 1378 – 1388. (2, 29, and 30)

[44] Mornati, F. Pareto optimality in the work of pareto. *Revue europenne des sciences sociales [En ligne] 51*, 2 (2017). (14)

[45] Odeh, I. O. A., Chittleborough, D. J., and McBratney, A. B. Soil pattern recognition with fuzzy-c-means: Application to classification and soil-landform interrelationships. *Soil Science Society of America Journal 56* (1992), 505–516. (26)

[46] Oliver, M., and Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA 113* (02 2014), 5669. (2)

[47] Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences 30* (2004), 683–691. (25)

[48] Pingali, P. Green revolution: Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences of the United States of America 109* (07 2012), 12302–8. (1)

[49] Reyes, J., Wendroth, O., Matocha, C., Zhu, J., Ren, W., and Karathanasis, A. D. Reliably mapping clay content coregionalized

with electrical conductivity. *Soil Science Society of America Journal 82* (05 2018). (5)

[50] Rezaee, M. R., Lelieveldt, B. P. F., and Reiber, J. H. C. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters 19* (1998), 237–246. (26)

[51] Roudier, P., Beaudette, D., and Hewitt, A. *A conditioned Latin hypercube sampling algorithm incorporating operational constraints*. CRC Press, 2012, pp. 227–232. (32)

[52] Royle, J. A., and Nychka, D. An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences 24*, 5 (1998), 479 – 488. (3)

[53] Schrijver, A. On the history of combinatorial optimization (till 1960). *Handbooks in Operations Research and Management Science 12* (08 2001). (8)

[54] Scudiero, E., Teatini, P., Manoli, G., Braga, F., Skaggs, T. H., and Morari, F. Workflow to establish time-specific zones in precision agriculture by spatiotemporal integration of plant and soil sensing data. *Agronomy 8*, 11 (2018). (38)

[55] Shir, O. M. Introductory mathematical programming for EC. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (New York, NY, USA, 2018), GECCO '18, ACM, pp. 539–552. (vi, 9)

[56] Solow, A., and Polasky, S. Measuring biological diversity. *Environmental and Ecological Statistics 1* (1994), 95–103. (48)

[57] Stein, M. L. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York, 2012. (4)

[58] T., C., AND HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics 3*, 1 (1974), 1–27. (26)

[59] TEAM, Q. D. Qgis geographic information system. open source geospatial foundation project., 2019. (25)

[60] ULRICH, T., BADER, J., AND THIELE, L. Defining and Optimizing Indicator-Based Diversity Measures in Multiobjective Search. In *Parallel Problem Solving from Nature - PPSN XI, 11th International Conference, Kraków, Poland, September 11-15, 2010, Proceedings, Part I* (2010), vol. 6238 of *Lecture Notes in Computer Science*, Springer, pp. 707–717. (48)

[61] WEBSTER, R., AND OLIVER, M. A. *Geostatistics for Environmental Scientists, Second Edition*. John Wiley and Sons Ltd., Chichester, England, 2007. (2, 4, and 25)

[62] ZHAO, Y., XU, X., TIAN, K., HUANG, B., AND HAI, N. Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. *Environmental Earth Sciences 75* (12 2016). (4, 31)

[63] ZITZLER, E., LAUMANNS, M., AND B., S. A tutorial on evolutionary multiobjective optimization. In *Metaheuristics for Multiobjective Optimisation* (2004), Springer Berlin Heidelberg, pp. 3–37. (10, 14, and 16)