האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

תהליך חישובי למיפוי מרקם קרקע ברזולוציה גבוהה בעזרת מדגם קרקע מצומצם וסקר חישה מקרוב

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

אסף ישראלי

הוגש לסנט האוניברסיטה העברית בירושלים

12/2025

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

תהליך חישובי למיפוי מרקם קרקע ברזולוציה גבוהה בעזרת מדגם קרקע מצומצם וסקר חישה מקרוב

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

אסף ישראלי

הוגש לסנט האוניברסיטה העברית בירושלים

12/2025

עבודה זו נעשתה בהדרכתם של פרופ' איגי (מיכאל) ליטאור ופרופ' עפר שיר.

תקציר

רקע

מרקם הקרקע משפיע על כושר תאחיזת המים בקרקע ובכך על הפרודוקטיביות בשדה חקלאי, לפיכך, אפיון מרחבי שלו הינו חיוני ליישום של השקיה מדייקת באזורים עם מחסור במים. מיפוי מרקם הקרקע מסתמך באופן מסורתי על דיגום סריג (Grid) צפוף או על הערכה סובייקטיבית של מומחה. סקר מוליכות חשמלית נדמית (ECa) הנמדדת באמצעות השראה אלקטרומגנטית מספק חלופה יעילה עם כיסוי מרחבי גבוה. עם זאת, תרגום ערכי ECa למפות מרקם מדויקות דורש תכנון מדגם מיטבי, קביעת ערכי סף לסיווג גודל גרגר המתאימים לשיטות השתברות לייזר (LD) ובחירת גישות מידול מתאימות.

השערות ומטרות המחקר

בעבודה זו פותח ונבחן בניסוי שדה אלגוריתם דגימה חדשני מבוסס חלוקה לאשכולות ושברונים (quantile-cluster - *QC*) למיפוי מרקם הקרקע בעזרת ECa, יחד עם התייחסות לשאלות הבאות: (1) האם דיגום מיטבי יכול לשפר את המתאמים המתקבלים בין רכיבי מרקם הקרקע ל-ECa כמו גם את דיוק החיזוי, לעומת דיגום סריג אחיד באותו גודל-מדגם? (2) במדידה מבוססת השתברות לייזר, מהו ערך הסף בין פרקציות גודל-הגרגר חרסית וסילט (2-8 מיקרומטר) המוביל לדיוק מיטבי בסיווג המרקם לפי ערכי ECa? ו-(3) האם רגרסיה על מדדי אנטרופיה רציפים מדויקת יותר מחיזוי ישיר של סוג המרקם (classification)? השערות המחקר הניחו כי תכנון מדגם המאזן בין כיסוי מרחבי וייצוג מלא של ספקטרום מרחב התדר (ECa) יניב קורלציות חזקות יותר (H1); סף חרסית\סילט אופטימלי יהיה שונה מהערך הנהוג של 2 מיקרומטר, עקב הבדלים בתהליכים הפיזיקליים המעורבים במדידה (H2); וכי רגרסיה תתאר את היחסים בין ECa לרכיבי מרקם הקרקע בצורה טובה יותר מאשר סיווג (H3).

מתודולוגיה

המחקר נערך בחלקת גידולי שדה בשטח של 65 דונם בעמק החולה, ישראל, המתאפיינת בקרקעות אלוביאליות באקלים צחיח-למחצה. סקר ECa אסף 20,800 מדידות באמצעות מכשיר *EM38* בארבע תצורות עומק, בתנאי תכולת רטיבות קרובה לקיבול שדה, במשך פחות משלוש שעות. האלגוריתם לעיצוב מדגם *QC* משלב פיזור מרחבי בחלוקה לאשכולות גיאוגרפיים באמצעות k-means, עם ריבוד של מרחב מאפייני ECa, בתוספת 10% נקודות אקראיות במרחק קרוב לנקודות דגימה קיימות. שלוש תכניות דגימה (n=17 כל אחת) נבחנו: (1) *QC*, האלגוריתם המוצע; (2) QC_{var} - גירסה של *QC* בתוספת אילוץ המקטין את מרחב החיפוש למיקום נקודות

הדיגום לאזורים בעלי אי-ודאות נמוכה; ו-(3) דיגום סריג רגיל (*Grid*) כשיטת ייחוס. 85 דוגמאות קרקע (51 דגימות מאופק הקרקע העליון בעומק 0-20 ס"מ, ו-34 דגימות תת-קרקע בעומק 40-60 ס"מ) נמדדו באמצעות מכשיר מבוסס השתברות לייזר (Mastersizer 3000) ל-101 קבוצות גודל הנעות בין 0.01 ל-3,500 מיקרומטר. האיכות של מודלים ללמידת מכונה בעזרת שיטת Random Forest הוערכה באופן שיטתי על פני 5,600 תצורות עם פרמטרים שונים: סיווג לעומת רגרסיה על מדד האנטרופיה *D*; ערך סף חרסית/סילט (2-8 מיקרומטר); מקדם שונות (המשפיע על גודל קבוצת האימון); ופרמטריזציות שונות של מודל הלמידה, תוך שימוש בנתוני אמת בלתי-תלויים (*n*=17) להערכת מידת הדיוק.

ממצאים עיקריים

המדגם לפי *QC* הגיע לדיוק של 76% בסיווג המרקם בתת-הקרקע, יותר מאשר דיגום סריג (65%) עם גודל-מדגם זהה. *QC* הניב גם מתאם (פירסון) חזק יותר בין ECa למרקם הקרקע בשני האופקים: R=0.74 עם חרסית עבור *QC* לעומת R=0.63 עבור *Grid* בשכבת הקרקע העליונה; ו-R=0.51 לעומת R=0.26 בתת-הקרקע, יתרון של 90%. QC_{var} שמר על דיוק דומה (71%) תוך הגבלת מיקומי הדגימה לאזורים בעלי רמת ודאות גבוהה. ניתוח התפלגות גודל גרגר ברזולוציה גבוהה הראה קורלציות חזקות ביותר ב-2.4 מיקרומטר (R=0.81) ו-143 מיקרומטר (R=-0.88), לא בגבולות המסורתיים של המחלקות חרסית, סילט וחול, דבר המצביע על כך שסילט דק (2-5 מיקרומטר) וחול דק (100-200 מיקרומטר) משפיעים באופן ניכר על ערכי ECa, ככל הנראה באמצעות השפעות של שטח פנים ומבנה הנקבוביות. הערכה שיטתית זיהתה את הערכים בטווח 6-8 מיקרומטר כערכי סף מיטביים עבור תת-הקרקע, והראתה שיפור של 13% במתאם (R=0.69 → 0.78) ושיפור של 29% במידת הדיוק (59% → 76%) בהשוואה לערך הסף המקובל של 2 מיקרומטר, דבר המשקף הבדלים בשיטות המדידה בין השתברות לייזר המסתמכת על קוטר שווה ערך כדורי לעומת שיטות מבוססות שיקוע המסתמכות על קוטר הידראולי של חלקיקים. המתאמים בתת-הקרקע היו חזקים יותר ב-64% מאשר בפני השטח, ניתן לייחס זאת להשפעות עיבוד מופחתות וסביבה יציבה יותר. חיזוי ישיר של סיווג מרקם-הקרקע הראה ביצועים טובים מעט יותר מאשר רגרסיה של מדד האנטרופיה D, עם דיוק של 76% לעומת 71%, בניגוד ל-H3, כנראה עקב שגיאה מצטברת מובנית בהמרה מערכי תכולת חרסית וחול ל-D ולהיפך.

השלכות ותרומות

מחקר זה מעלה שלוש תרומות עיקריות: (1) תהליך חישובי מאומת לדגימה שהשיג קורלציות חזקות יותר עד 90% מדיגום שריג רגיל באמצעות מיטוב של כמה מטרות; (2) מתודולוגיה מבוססת ראיות המצביעה על 6-8 מיקרומטר כערך סף מיטבי לחיזוי לפי ECa בשיטות של השתברות לייזר לעומת הערך המקובל של 2

מיקרומטר; ו-(3) הדגמה שיטתית שתכנון המדגם משפיע באופן משמעותי על דיוק מפות החיזוי, אפילו כאשר גודל המדגם זהה. רמת הדיוק שהושגה (76%) נמצאת בטווח הדיוק של סיווג מומחה אנושי (70-85%), ואילו האפליקציה האינטרנטית הייעודית שפותחה מאפשרת מיפוי מפורט באופן נגיש של צרכי השקיה ברחבי העולם. ניתוח ברזולוציה גבוהה זיהה פרקציות גודל גרגר מסוימות (2-5 מיקרומטר, 100-200 מיקרומטר) עם מתאם חזק ביותר ל-ECa, כבעלות פוטנציאל אפשרי לשיפור מודלים לחיזוי קיבולת המים הזמינים לצמח (AWC) וחלוקה לאזורי ניהול עבור פיתוח חקלאות בת קיימא.

Soil Sampling Design for ECa-Based Texture Mapping: A Novel Quantile-Cluster Algorithm with High-Resolution Particle Size Analysis

Thesis for the degree of Doctor of Philosophy

by

Assaf Israeli

Submitted to the Senate of the Hebrew University of Jerusalem

12/2025

Soil Sampling Design for ECa-Based Texture Mapping: A Novel Quantile-Cluster Algorithm with High-Resolution Particle Size Analysis

Thesis for the degree of Doctor of Philosophy

by

Assaf Israeli

Submitted to the Senate of the Hebrew University of Jerusalem

12/2025

This work was carried out under the supervision of Prof. Michael (Iggy) Litaor and Prof. Ofer M. Shir.

Abstract

Background and Rationale

Soil texture governs water retention and agricultural productivity, making accurate field-scale characterization essential for precision irrigation in water-scarce regions. Traditional texture mapping relies on dense grid sampling or subjective expert assessment. Apparent electrical conductivity (ECa) surveys using electromagnetic induction provide cost-effective alternatives with high spatial coverage. The common USDA standard, which is based on Stokes' Law sedimentation, prescribes a clay/silt cutoff at 2 μm, while laser diffraction (LD) based methods produce clay content values different from those of sedimentation-based methods. Nevertheless, translating ECa to accurate texture maps requires planning sampling design, determining appropriate particle size classification thresholds for LD-based methods, and selecting proper modeling approaches.

Research Objectives and Hypotheses

This thesis developed and validated a novel quantile-cluster (*QC*) sampling algorithm for ECa-based texture mapping, addressing the following questions: (1) Can heuristic sampling design improve ECa-texture correlations and prediction accuracy when compared to traditional *Grid* sampling? (2) What clay/silt cutoff (2-8 μm) best fits laser diffraction-based texture classification for prediction with ECa? (3) Does regression on continuous entropy indices outperform direct texture classification? It was hypothesized that strategic sampling design balancing spatial coverage and feature-space representation would yield stronger correlations (H1), optimal cutoff would differ from conventional 2 μm due to measurement physics differences (H2), and that regression would capture ECa-texture relations more effectively than classification (H3).

Methodology

Research was conducted in a 6.5 ha irrigated crop field (Hula Valley, Israel), characterized by alluvial soils in a semi-arid climate. An ECa survey using *EM38* collected 20,800 measurements across four depth configurations under near-field-capacity conditions in less than three hours. The *QC* algorithm combines geographic *k*-means clustering for spatial dispersion with quantile stratification across ECa feature space, augmented with 10% close-pair random points. Three sampling designs ($n = 17$ each) were implemented: (1) *QC*, (2) variance-filtered QC_{var} with samples restricted to low-uncertainty zones, and (3) regular *Grid* sampling as reference. Soil samples (85 total, divided to 51 surface 0-20 cm and 34 subsoil 40-60 cm) were analyzed via laser diffraction into 101 size classes ranging 0.01-3,500 μm. Random Forest models were systematically trained and tested across 5,600 configurations varying response variable (classification vs. regression on *D* entropy index); clay/silt cutoff (2-8 μm); variance filter (controlling training-set size); and internal model parameterizations, using cross-design validation

for independent assessment.

Principal Findings

QC achieved 76% texture classification accuracy in the subsoil, outperforming *Grid* sampling (65%) with identical sample size. *QC* also yielded stronger ECa-texture correlations across horizons: $R = 0.74$ with clay for *QC* vs. $R = 0.63$ for *Grid* in the topsoil; and $R = 0.51$ vs. $R = 0.26$ in the subsoil, a 90% advantage. QC_{var} maintained accuracy (71%) while restricting sampling locations to high-confidence zones. High-resolution particle size distribution analysis revealed peak correlations at 2.4 µm ($R = 0.81$) and 143 µm ($R = -0.88$), not at traditional boundaries, indicating fine silt (2-5 µm) and fine sand (100-200 µm) disproportionately influence ECa, likely through surface area and pore architecture effects. Systematic evaluation identified 6-8 µm as best fitting cutoff for subsoil, showing 13% correlation improvement ($R = 0.69 \rightarrow 0.78$) and 29% accuracy improvement (59% $\rightarrow$ 76%) when compared to conventional 2 µm, reflecting measurement physics differences between laser diffraction using spherical equivalent diameter vs. sedimentation that rely on hydraulic diameter. Subsoil correlations were 64% stronger than surface, attributable to reduced management interference and greater stability. Direct classification performed slightly better than regression (76% vs. 71%), contrary to H3, possibly due to the inherent cumulative error in conversion steps from PSD to *D* and vice versa.

Implications and Contributions

This research provides three primary contributions: (1) a validated sampling framework achieving up to 90% stronger correlations than traditional *Grid* sampling through addressing multiple targets (geographic and feature-space coverage, short-range variability), (2) evidence-based methodology establishing 6-8 µm as optimal for LD-based PSD prediction with ECa vs. conventional 2 µm, and (3) systematic demonstration that sampling design substantially impacts prediction accuracy even with identical sample sizes. The 76% accuracy attained is on par with expert classification (70-85%) while being scalable. Our publicly available implementation, as an open-source web application, enables cost-effective precision irrigation mapping worldwide. High-resolution analyses identified specific particle fractions (2-5 µm, 100-200 µm) driving ECa response, which can improve soil water holding capacity modeling and management zone delineation for sustainable agricultural intensification.

Publications Included in this Thesis

Submitted:

Israeli, A., Shir, O. M. and Litaor, M. I. (2025). **Algorithmically-Guided Soil Texture Mapping Using Small Sample Coupled with Proximal Soil Survey** [Manuscript submitted for publication]. Soil and Water Sciences, The Robert H Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem.

Other Publications During Candidature

Published:

Shulman, D., Israeli, A., Botnaro, Y., Margalit, O., Tamir, O., Naschitz, S., Gamrasni, D., Shir, O. M. and Dattner, I. (2024). **Physics-Guided Inverse Regression for Crop Quality Assessment**. *Journal of Agricultural, Biological and Environmental Statistics*, 2024.

Shir, O.M., Yazmir, B., Israeli, A. and Gamrasni, D. (2022). **Algorithmically-guided postharvest by experimental combinatorial optimization.** *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO-2022*, New York, NY, USA, ACM Press (2022), pp. 2027–2035

Shir, O.M., Israeli, A., Caftory, A., Zepko, G. and Bloch, I. (2022). **Algorithmically-guided discovery of viral epitopes via linguistic parsing: Problem formulation and solving by soft computing** *Applied Soft Computing*, 129, 109509, 2022

Acknowledgments

The completion of this thesis would not have been possible without the support and advice of many people, to whom I am deeply grateful. I wish to express my sincere appreciation to my advisors Iggy and Ofer, whose patient guidance, insightful feedback, and willingness to both challenge and encourage me shaped this work at every stage. I am also grateful to the members of my thesis committee for the time and thoughtful criticism they generously offered. I gratefully acknowledge MIGAL Galilee Research Institute for providing the financial support and facilities for this research, and thank my fellow students and colleagues at the Hydrogeochemistry Lab in MIGAL for making this work possible. My deepest gratitude goes to my dear friends and family, for the love and support that made us get through these challenging times. Finally, I would like to pay tribute to the many innocent lives lost during the war, and offer my prayer for a lasting peace.

Contents

List of Figures

List of Tables

List of abbreviations

Abbreviations

AWC	Available Water Capacity
cLHS	conditioned Latin Hypercube Sampling
CCC	Concordance correlation coefficient
CEC	Cation Exchange Capacity
D_{KL}	Kullback-Leibler divergence
DSM	Digital Soil Mapping
EMI	Electromagnetic Induction
ECa	apparent Electrical Conductivity
IDW	Inverse Distance Weighting
K_{sat}	Saturated hydraulic conductivity
LD	Laser Diffraction
LOI	Lost On Ignition
OM	Organic Matter
PTF	Pedotransfer function
PSD	Particle Size Distribution
QC	Quantile-Cluster
QC_{var}	Variance constrained QC
RF	Random Forest
SD	Standard Deviation
VRI	Variable-rate irrigation

Introduction

1.1 Soil texture and precision irrigation

Agriculture accounts for approximately 70% of global freshwater withdrawals (Koncagül, Connor, and Abete 2024), making it the largest consumer of freshwater resources worldwide. The growing global demand for agricultural produce underscores the significance of water usage efficiency, particularly in water-scarce regions (Nikolaou et al. 2020). Soil texture is the main factor that controls the amount of water available for plants. However, arable lands often show significant within-field variation in soil texture, with textural differences occurring at scales ranging from several meters to hundreds of meters within a single field. This variability has significant implications for optimal irrigation planning, considering that the topsoil acts as a reservoir, enabling plants water uptake on demand. The practice of precision irrigation aims to supply the optimal water amount required for each plant (or management zone) at each growth stage and season, through monitoring and control systems that apply strategies which rely on site-specific soil texture data, plant properties, weather conditions, topography and real-time soil moisture data (Violino et al. 2023). Variable-rate irrigation (VRI) can be implemented through multiple technologies, including zone-controlled sprinklers, center pivots with individual sprinkler control, or precision drip irrigation systems with zone valves. These systems can optimize the profitability of water usage by matching application rates to spatial variability in soil water-holding capacity, whereas recent technological developments, such as autonomous emitters (Al-agele et al. 2021) and low-pressure emitters (Ghodgaonkar et al. 2025), promise to make precision irrigation more accessible and affordable for diverse farm scales. Beyond water conservation, improved water management contributes to the sustainability of ecosystems and biodiversity; by reducing over-irrigation, precision approaches minimize nutrient leaching and agricultural runoff, thus mitigating associated water pollution and habitat degradation (Ray and Majumder 2024). Field studies have demonstrated that precision irrigation can reduce water consumption by 10-60% compared to uniform application while maintaining or improving crop yields (Gundim et al. 2023; Lakhiar et al. 2024).

Available water capacity (AWC) – defined as the difference between field capacity (-33 kPa matric potential) and permanent wilting point (-1500 kPa) – quantifies the amount of water accessible for plant uptake that can be retained within a soil profile (Veihmeyer and Hendrickson 1928; Rai, Singh, and Upadhyay 2017). While the concept is fundamental to irrigation planning, its definition varies somewhat across disciplines, with some approaches considering depth-weighted profiles while others focus on specific soil horizons. AWC is primarily determined by soil texture, which is classified based

on particle size distribution (PSD) (Arya and Paris 1981; Amemiya 1965), as texture controls pore size distribution and water retention characteristics. Secondary factors include bulk density (affecting total porosity); organic matter (OM) content, which typically enhance water retention by 1-2% per 1% OM increase (Bhadha et al. 2017); and clay mineral composition (with 2:1 clays like smectites retaining more water than 1:1 clays like kaolinite) (Bruand and Tessier 2000).

AWC represents a critical parameter in irrigation scheduling and crop water balance models, with studies identifying it as a primary source of uncertainty in crop model predictions (Cousin et al. 2022). This uncertainty propagates through modeling chains, potentially leading to suboptimal irrigation recommendations when AWC is poorly characterized. Recent advances in spatial prediction of AWC and soil texture have employed both proximal and remote sensing approaches, with variable success rates. Proximal sensing methods, particularly electromagnetic induction surveys of bare soil, have demonstrated promise for enhancing spatial estimation of soil texture (Heil and Schmidhalter 2017; Gozdowski et al. 2015; Domsch and Giebel 2004), though calibration challenges remain (Lück et al. 2009). Remote sensing approaches using satellite imagery have shown mixed results (Mgohele et al. 2024; Chen et al. 2022; Dharumarajan and Hegde 2022; Blaschek et al. 2019), with prediction accuracy varying substantially with soil depth, land cover conditions, and regional soil characteristics. Accumulated evidence suggests that fusion of remote and proximal sensor data with environmental covariates can improve model performance over single-sensor approaches (Schmidinger, Barkov, et al. 2024; Rodrigues et al. 2024; Kalumba et al. 2022), though optimal sensor combinations and integration strategies remain active areas of research.

Practical precision irrigation requires soil texture maps with sufficient details for management zone delineation – typically a spatial resolution of 1-10 meters to capture within-field variability – while maintaining prediction accuracy adequate for hydraulic property estimation. Integrating these considerations, reliable high-resolution soil texture maps can be transformed into spatial AWC estimates via pedotransfer functions (PTFs) – empirical or physically based equations linking easily measured soil properties to hydraulic parameters (Wösten, Pachepsky, and Rawls 2001). Such AWC maps provide the spatial baseline for effective variable-rate irrigation scheduling (Cousin et al. 2022; Xing and Wang 2024), enabling water application rates to be matched to local soil water-holding capacity.

1.2 Sampling methods for digital soil mapping

The design of soil sampling schemes for digital soil mapping (DSM) aims to capture the spatial variation of one or more target soil properties. To that end, a genuine soil-sample would cover both the geographical space and the feature-space (the postulated spectrum of the property of interest), which might include spectral reflectance values, terrain attributes, or climatic variables (Brus and Heuvelink

2007). Since soil sampling is a sparse, costly process, an exhaustive coverage of an entire crop field is impractical. Therefore, ancillary data – high-resolution datasets which are relatively easily available, correlated to some degree with the target variable (e.g., digital elevation models, remotely sensed imagery, or electromagnetic induction surveys) – can be used as an a priori expression of the target property's spatial variability. Notwithstanding partial correlation with the target variable, adequate preprocessing permits a sampling design to incorporate such information to determine sampling locations in a way that achieves those targets.

In this manner, model-based soil sampling design strategies, such as spatial coverage or variance-based methods, leverage a spatial model to capture the variability of soil properties, thus providing an informed selection of sampling points. In contrast, design-based sampling methods are predicated on probability sampling (e.g., simple random sampling, stratified random sampling) and are more appropriate for map validation, since it offers unbiased estimates of map quality indices (Brus 2019; Biswas and Zhang 2018). Hybrid approaches that combine model-based efficiency with design-based validation strategies are becoming increasingly common in practice (e.g. Brus, Kempen, and Heuvelink 2011). Sampling designs are typically evaluated using multiple criteria, including spatial coverage, feature-space coverage, and prediction accuracy (Zhang et al. 2022; Wadoux, Brus, and Heuvelink 2019). The trade-off between sample minimization and information maximization constitutes a fundamental optimization problem rooted in Information Theory—specifically, the efficient allocation of sampling resources to maximize mutual information or minimize expected entropy relative to inference goals (Saurette et al. 2024).

The density of soil observations is the main determinant of mapping accuracy (Loiseau et al. 2021). It has been shown that at least 100 sampling points are required to produce a reliable method-of-moments estimate of the variogram, sufficient for kriging interpolation, regardless of the study area size (Webster and Oliver 1992), although a reliable residual maximum likelihood (REML) estimate of the variogram could be attained with fewer than 50 sampling units (Kerry and Oliver 2007). However, these preconditions were established primarily for geostatistical interpolation methods and may not apply equally to all DSM approaches. Recent advances demonstrate that ML-based prediction models can substantially reduce sampling requirements compared to traditional interpolation approaches. For instance, Schmidinger et al. (2024) achieved moderate model quality (concordance correlation coefficient, CCC > 0.65) with only 10 samples. Moreover, greedy algorithms (i.e., iterative selection of locally optimal sites) can simplify complex optimization procedures for sampling design while maintaining computational efficiency. This intuition is supported by a well-established result in Combinatorial Optimization. Sampling-design objectives such as entropy, mutual information, and variance-reduction criteria over Gaussian process models are *submodular* set functions (Krause, Singh, and Guestrin 2008), that is, they exhibit diminishing returns as additional sites are added. For maximization of a monotone submodular function under a cardinality constraint, the greedy

algorithm runs in polynomial time and yields a solution within a factor of $(1 - 1/e) \approx 0.63$ of the global optimum (Nemhauser, Wolsey, and Fisher 1978). Greedy site selection therefore offers a principled and computationally tractable alternative to exhaustive or metaheuristic search, which scale poorly with the number of candidate locations – a point of practical importance for the design of soil-sampling campaigns.

To advance these developments while addressing the persistent challenge of efficient sample-size determination, this work presents a novel quantile-cluster (QC) algorithm – a fast, model-based soil sampling design procedure (detailed in Section 2.2) that integrates a screening phase using information metrics for optimal sample-size selection. Hereafter, the term "optimal" is used throughout to indicate the best solution found by the algorithm, not a mathematically guaranteed global optimum.

1.3 ECa for soil sampling design

Electrical conductivity of saturated paste (EC_e) is an established lab analysis indicator for several soil properties, including water content, salts (Ding et al. 2020), ionic strength, cation exchange capacity (CEC) and organic matter content (Martinez et al. 2009). While EC_e requires laboratory analysis of extracted soil samples, apparent Electrical Conductivity (ECa) offers a rapid, non-destructive alternative for field-scale in-situ measurement of soil resistivity, and is being increasingly studied as a proxy for soil texture mapping for non-saline soils (Cousin et al. 2022; Pace et al. 2024). ECa values (mS/m) are derived from an electromagnetic signal induced in the soil matrix via a non-intrusive process, obtained with dedicated measuring equipment, such as *EM38* (Geonics Ltd., Mississauga, ON, Canada). The received signal is affected by various soil conditions, such as water content (Evett and Parkin 2005), soil texture (Stępień et al. 2015; Kelley et al. 2017), salinity (Corwin and Lesch 2013; Jiang et al. 2019), organic matter (Saxton and Rawls 2006) and nutrients' content, as well as measurement-related factors like temperature (Friedman 2005), elevation above the ground and travel speed. Therefore, isolating a single-factor effect from the measured ECa signal is a complex task (Clay 2001; Heiniger, McBride, and Clay 2003), limiting its widespread adoption for quantitative soil property prediction, particularly in heterogeneous agricultural landscapes. In addition, ECa values are altered by pedogenic, edaphic (porosity, topography, hydrology), meteorological, biological and anthropogenic circumstances (Corwin and Scudiero 2019).

The multitude of confounding factors, some of which are unmeasurable, implies that the correlation between ECa and soil properties is site-specific. In several studies, the R^2 of observed site-specific relationships between ECa and AWC or particle size distribution (PSD) ranged from 0.2 to 0.8 (Cousin et al. 2022; Heil and Schmidhalter 2017), with stronger correlations typically observed in areas with greater textural variability and with minimal salinity effects. Therefore, field-scale ECa maps can be used as ancillary data to characterize spatial variability (Michael-Mertens, Pätzold, and Welp

2008), as well as for the delineation of management zones (Fortes, Millán, Prieto, et al. 2015) and soil sampling design, but for the time being, ECa surveys cannot directly quantify AWC without site-specific calibration, nor can they provide the detailed PSD data necessary for soil hydraulic modeling. Thus, strategic soil sampling remains essential for calibration and validation of ECa-based spatial predictions (Stępień et al. 2015). The *ECa-directed* sampling method used by De Feudis et al. (2025) prescribes delineation of n zones by ECa survey values and sampling from the center of each zone. While this zone-based approach ensures representation across ECa ranges, it may not optimally balance feature space coverage with geographic dispersion, particularly when zones are spatially fragmented or when sample size constraints are stringent.

Building upon the concept of *ECa-directed* sampling, our proposed *QC* algorithm advances this approach through dual optimization. The algorithm explicitly balances feature-space coverage (sampling across the entire ECa spectrum) with geographic dispersion via spatial constraints. Unlike feature-space zone-centroid approaches which may cluster samples in spatially contiguous areas, *QC* aims to sample the whole spectrum while maximizing the geographic dispersion with a minimal set of points, as described in Section 2.2.

Critical to the success of ECa-based sampling is the timing of field measurements. To isolate the influence of soil texture on ECa measurements while minimizing confounding effects, surveys should be conducted when soil is at or near field capacity (Corwin and Lesch 2013). At this moisture state, textural differences in water retention are maximized – sandy soils drain more rapidly than clayey soils – creating ECa patterns that primarily reflect AWC variability. Surveying during active drainage (2-3 days following significant precipitation or irrigation) captures this transitional period when spatial patterns in ECa most strongly correlate with textural controls on water retention. This timing strategy effectively uses soil physics principles to enhance the signal-to-noise ratio for texture mapping. In this study, ECa was selected as ancillary variable because it: (1) directly responds to water retention properties derived from texture; (2) provides high spatial resolution data (meter-scale) at field scale; (3) requires minimal preprocessing compared to remote sensing data; (4) can be rapidly collected at relatively low cost; (5) offers flexible operation, as the survey can be done by hand or by towing, even above a crop rather than bare soil; and (6) exhibits spatial patterns that remain relatively stable over time (Gonçalves et al. 2025).

1.4 High-resolution PSD analysis: Soil texture as a continuous variable

Accurate characterization of particle size distribution (PSD) is fundamental to predicting soil hydraulic properties. Traditional methods for measuring PSD, such as the pipette (Olmstead, Alexander, and Middleton 1930) and hydrometer (Bouyoucos 1962) methods, which are based on sedimentation rate

according to Stokes' law, commonly classify PSD into three classes ($I = 3$): *clay* for particles smaller than 2 µm, *silt* for particles of size 2-50 µm, and *sand* for particles in the range 50-2,000 µm, whereas larger particles are excluded (Gavlak et al. 2003). The resulting trio of values describes the relative content (%) of each class, although the ranges of classes differ worldwide (Hirotatsu, Yusuke, and Toshiyuki 2015; García-Gaines and Frankenstein 2015).

The USDA (2017) system classifies soils into 12 textural classes (e.g., *sandy loam*, *clay loam*, *silty clay*) based on the relative proportions of sand, silt, and clay. From these categorical classifications, it is possible to derive soil's field capacity (FC) and wilting point (WP) – the upper and lower bounds for irrigation planning – and then calculate AWC (expressed as volumetric water content or mm/m soil depth) as their difference:

$$AWC(mm/m) = FC - WP \tag{1.1}$$

Modern PSD analysis tools, such as the *Mastersizer 3000* (Malvern Panalytical Ltd., UK) which utilizes laser diffraction (LD) technology, measure PSD at a resolution of 101 classes within the range of 0.01-3,500 µm, which enables high dimensional correlation analysis with ancillary data. Laser diffraction determines particle size by measuring the angular variation in intensity of light scattered by particles suspended in a dispersant. Larger particles scatter light at small angles, while smaller particles scatter at larger angles, enabling simultaneous measurement across the full size range in minutes rather than hours, providing more reproducible results compared to traditional methods. Nevertheless, LD-based methods are sensitive to preparation procedures that include OM removal by hydrogen peroxide (H_2O_2; Allen and Thornley 2004), and are prone to underestimation of the clay fraction (Eshel et al. 2004), as opposed to traditional methods' bias for overestimating the clay fraction due to breakdown of aggregates and dispersion artifacts (Moreno-Maroto and Alonso-Azcárate 2022). Consequently, numerous clay/silt threshold values were suggested for LD-based measurements, including 3.9 µm (Svensson, Messing, and Barron 2022), 5.8 µm (Makó et al. 2017), 6 µm (Crouvi et al. 2018) and up to 9 µm (Fisher et al. 2017).

This discrepancy primarily stems from the different interpretation of particles' sphericity, and varies by soil type (Gorączko and Topoliński 2020), hence, thresholds should be adjusted when comparing measurement techniques. Given this controversy and its strong impact, we decided to explore this parametric choice in this study. In practice, various clay/silt boundaries from 2 to 8 µm were systematically evaluated by correlation with ECa measurements and soil texture prediction accuracy. While high-resolution PSD provides up to 101 data points per sample, with each data point representing the volumetric percentage of particles within a specific size bin, this dimensionality poses challenges for spatial modeling and interpretation. Information-theoretic measures offer parsimonious descriptors that capture essential characteristics of the distribution while reducing dimensionality – as being described in what follows.

Any discrete distribution can be described using entropy measures, which assess the degree of

uniformity in the occurrences of different classes. Martin et al. (2001) used the *Shannon H* index (Eq. 1.2) to quantify the entropy of PSD and have shown its relation to the homogeneity of particle size distributions. Suppose $\{A_i\}$ denotes a partition of the probability space into a finite number of non-overlapping subsets, and $\mu(A_i)$ is the probability of the occurrence of a set A_i, then the entropy ($H_{Shannon}$) of the partition is defined as follows:

$$H(\mathcal{A}) = -\sum_{i=1}^{n} \mu(A_i) \log \mu(A_i) \tag{1.2}$$

whereas higher $H_{Shannon}$ values indicate more uniform distributions across classes and lower values indicate dominance by particular size fractions. For the three-class system, $H_{Shannon}$ ranges from 0 (all particles in one class) to $\log_2(3) \approx 1.58$ (equal distribution across classes).

The common PSD division prescribes disproportionate class ranges, in which *sand* (1,950 µm) is ~41 times wider than *silt* (48 µm) and ~975 times wider than *clay* (2 µm), hence a uniform distribution of particles across all sizes would appear to have 97.6% sand content simply due to range definitions. This range bias implies that the $H_{Shannon}$ quantifier alone cannot distinguish between physically meaningful concentration in specific size ranges versus artifacts of class width definitions.

To compensate for this bias, Martin et al. (2004) proposed another measure – the balanced entropy (D) index (Eq. 1.3), which weighs both the probability of a particular segment (P_i) and its range (r_i), and, unlike $H_{Shannon}$, conveniently yields values in the range [0,1]. The case $D = 0$ represents a homogeneous soil with all particles concentrated in a single size class. Conversely, $D = 1$ occurs when $P_i = r_i$, corresponding to a heterogeneous distribution where mass is distributed proportionally to the width of each size class. D is formally described by the following equation:

$$D(\mathcal{A}) = \frac{H(\mathcal{A})}{H(\mathcal{A}) + d(P_i \| r_i)} \tag{1.3}$$

where H is the $H_{Shannon}$ index of the distribution, and $d(P_i||r_i) = \Sigma_i P_i \log(P_i/r_i)$ is the relative entropy, or *KL-divergence* (Kullback and Leibler 1951), which quantifies how much the actual particle size distribution (P) differs from a distribution proportional to class widths (r), and acts as a balancing component here (see also Eq. 2.2). As expected, the D index demonstrates high correlation to the sand fraction of the soil, due to its wide range compared with clay and silt (Martín, Rey, and Taguas 2004). D, as an expression of soil heterogeneity, has proven to be suitable for estimation of saturated hydraulic conductivity (K_{sat}) (García-Gutiérrez, Pachepsky, and Martín 2018), achieving $R^2 > 0.9$, and for terroir classification in vineyards (Cámara, Lázaro-López, and Gómez-Miguel 2016), linking the entropy index with lithological groups with R^2 up to 0.86.

For precision irrigation applications requiring spatially explicit soil texture predictions, treating PSD as a continuous variable enables more nuanced spatial predictions aligned with the continuous nature of soil hydraulic properties. One hypothesis of this study is that entropy indices derived from high-resolution (~100 classes) PSD measurements capture sufficient information about the continuous

distribution to enable accurate regression-based spatial prediction. Unlike categorical texture classes that impose arbitrary boundaries, entropy measures, as continuous variables, are naturally suited to modeling relationships with continuously distributed ancillary data such as ECa. This approach forms the basis for spatial prediction which uses ECa as a continuous predictor for entropy-based PSD characterization.

1.5 Spatial Prediction Using Machine Learning: Regression and Classification

Spatial prediction of soil properties from ancillary data constitutes a classic digital soil mapping challenge. While geostatistical methods like kriging are widely used, ML approaches offer advantages for modeling complex, non-linear relationships between soil properties and environmental covariates. Among ML methods, Random Forest (RF) has emerged as a simple yet particularly robust for soil mapping applications (Dharumarajan and Hegde 2022). RF implements an ensemble of decision trees with random input selection (bootstrap sampling, feature subsampling) and out-of-bag error estimate, relying on the Law of Large Numbers to avoid overfit (Breiman 2001). Given a set of input features, it uses a training subset to fit either a classifier (for categorical values) or a regression model (for continuous values), then evaluates the accuracy or error with a test subset. RF handles non-linearity with robustness to outliers and has demonstrated strong performance across diverse soil mapping applications. In a comprehensive literature review, Mgohele et al. (2024) found RF to be the most widely used model for predicting soil texture from satellite imagery, with a median overall accuracy of 0.49 for PSD classes (reflecting the challenging nature of satellite-based texture prediction). When predicting continuous PSD fractions, RF has shown moderately to strong predictive power: Chagas et al. (2016) explained 63% and 56% of the spatial variability of sand and clay fractions, respectively, while He et al. (2024) achieved R^2 values of 0.73 for clay fraction when using environmental variables as predictors.

One of the main challenges in training a model on spatial data is the scarcity and high cost of ground-truth data points – i.e., PSD data in this case. Therefore, in this study, relying on the spatial autocorrelation assumption regarding soil properties, the response variable data is augmented by interpolating ground-truth measurements (Dos Santos et al. 2025), then filtering by a variance-based criterion to ensure the training dataset includes only locations with low prediction uncertainty (see Section 2.4.2 for details). Kriging interpolation and Random Forest complement each other well, each addressing different aspects of spatial complexity in soils. While *kriging* exploits spatial autocorrelation, minimizes the estimation variance under the assumption of stationarity and has built-in uncertainty quantification (Webster and Oliver 2007), *Random Forest* handles non-linear relationships and easily processes multiple predictors – continuous or categorical without a stationarity assumption

(Hengl et al. 2018), making it suitable for heterogeneous terrain.

Hereby, we examine the direct prediction of soil texture classification from ECa values, compared by accuracy to a regression prediction of the continuous variable D, which involves a double transformation: first from high-resolution PSD to D for model training, then from predicted D values back to texture classes for map production. Multiple runs were evaluated with different parameters, including training-set size, clay/silt cutoff threshold (2-8 µm), seed initialization and number of RF trees, to assess model stability and optimize prediction performance.

1.6 Study Objectives

This study describes an efficient procedure for producing a high-resolution (1×1 meter) soil texture map, which can be converted into an AWC map using established pedotransfer functions. Such maps enable the delineation of irrigation management zones and variable-rate irrigation scheduling customized to within-field variability in water-holding capacity. ECa survey data were used as input for the *QC* algorithm (Section 2.2) for soil sampling design, then – after collecting ground-truth samples – as features for ML model fitting.

The specific objectives of this study are to: (1) Develop and validate the quantile-cluster (*QC*) sampling algorithm for ECa-based soil texture mapping; (2) Evaluate optimal clay/silt cutoff thresholds (2-8 µm) for laser diffraction in ECa prediction contexts; (3) Compare classification versus regression approaches for spatial texture prediction; and (4) Assess the prediction accuracy and practical utility of ECa-based texture mapping for precision irrigation applications A field experiment was conducted to test the hypothesis that the integrated workflow – ECa survey, *QC* sampling design, high-resolution PSD analysis with entropy-based characterization, variance-based augmentation, and RF prediction – produces accurate and reliable soil texture maps suitable for precision irrigation applications, and that regression-based prediction of the continuous D index will achieve superior accuracy compared to direct texture classification. All data, source code, and supplementary materials for this study are publicly available via *GitHub* (see Appendix B for details).

The remainder of this thesis is organized as follows: Section 2 describes the study area, ECa survey, sampling designs, laboratory methods, and spatial modeling approaches; Section 3 presents results of sampling design approaches, ECa-texture correlations, and spatial predictions; Section 4 discusses findings in the context of existing literature, identifies their limitations, and provides conclusions and recommendations for future research.

Methodology

A field experiment was conducted in 2023 in a 6.5 ha crop field in the Hula Valley, Israel (33°08'23.4"N 35°35'15.1"E), a semi-arid agricultural region with mean annual precipitation of approximately 480 mm. The field, characterized by alluvial soil, is used for intercropping of winter wheat and summer crops. The plot is located on the edge of historic Lake Hula, which was drained in the 1950s. The site was historically a seasonal marshland (Es Salihiya map, Massad Cartography of Tel-Hai College 1942), and the current soil distribution reflects this fluvial history, with substantial textural heterogeneity across the field, resulting in significant spatial variability in soil texture and water-holding capacity, making the site well-suited for evaluating precision soil mapping approaches. The study followed a sequential workflow: (1) ECa survey for ancillary data collection, (2) *QC* algorithmic-based soil sampling design, (3) soil sample collection and laboratory analysis, (4) spatial prediction modeling, and (5) independent validation benchmark. This section describes each component in detail.

2.1 EC_a survey

The *EM38-MK2* (Geonics Ltd., Canada) is a dual-dipole electromagnetic induction sensor consisting of a nonconductive boom housing a 14.5 kHz electromagnetic signal generator with receiving coils located at 0.5 m and 1.0 m from the transmitter. The instrument operates on the principle of electromagnetic induction (EMI): a primary electromagnetic field generated by the transmitter coil induces eddy currents in conductive subsurface materials, which generate a secondary electromagnetic field detected by the receiver coils. The instrument can be operated in two orientations: horizontal dipole mode (*ECaH*), measuring apparent electrical conductivity (mS/m) at cumulative effective probing depths of approximately 0.375 m and 0.75 m (each representing depth-weighted average conductivity from surface to the specified depth), and vertical dipole mode (*ECaV*), measuring at approximately 0.75 m and 1.5 m depth. These depth ranges encompass the root zone of most field crops and provide information on both topsoil and subsoil textural properties.

We conducted the ECa survey in late April, 2023, approximately 4 days after the last significant precipitation event (~40 mm over a few days), when soil moisture was at near field-capacity based on antecedent rainfall. This timing was selected to maximize the influence of soil texture on ECa readings while minimizing confounding effects from spatial moisture variability (see Section 1.3). At the time of survey, wheat had already been harvested and left as windrows for field drying. Weather conditions were stable throughout the survey with air temperature of 22-24 °C, minimizing temporal

drift in ECa measurements. The survey was conducted using an *EM38-MK2* device mounted on a dedicated sled towed by an off-road vehicle (Figure 1), over soil covered by dry plant stubble, resulting in transect spacing of approximately 12 m. Measurements were operated by a tablet computer with designated software (*RTmap38MK2*; GEOMAR Software) connected to a GPS receiver (*S850*, Stonex, Milano, Italy). The measurement rate was set to one reading every 500 milliseconds at an average travel speed of approximately 10 km/h, resulting in measurement intervals of approximately 1.4 m along transects. The survey was pursued over 2 hours in both horizontal and vertical dipole orientations during separate passes. GPS positioning accuracy was ± 5 cm, adequate for the 1×1 m grid resolution of final maps. A total of 9,700 georeferenced ECa readings were obtained in the horizontal orientation and 11,100 readings in the vertical orientation in two separate passes, yielding approximately 20,800 total measurements across the 6.5 ha field. All spatial data were recorded and processed in UTM Zone 36N coordinate system using `R v4.3` (R Core Team 2025) and QGIS software (QGIS Development Team 2023). Spatial preprocessing and interpolation methods are described in Section 2.2.1.

Figure 1: ECa survey equipment and study site. (a) Survey in progress in the Hula Valley study site, April 2023, showing wheat residue windrows and transect pattern. (b) EM38-MK2 electromagnetic induction sensor mounted on a nonconductive sled, sensor shown in horizontal dipole orientation.

2.2 Soil sampling design

2.2.1 Preprocessing

Raw ECa measurement data ($n = 20,800$; Section 2.1) were preprocessed to prepare them for geostatistical analyses and sampling design. ECa measurement files (`.xyz`) were uploaded to a custom-developed Shiny web application (`R shiny` framework; Chang et al. 2024) for interactive preprocessing, sampling design, and visualization (implementation details are provided in Section 2.5). Data preprocessing involves compaction by the moving average technique (every n-th point) to reduce computational load while preserving spatial structure; normalization; outlier removal; and spatial

interpolation to create continuous layers for algorithmic input. To meet kriging assumptions of approximate normality, a log-transformation was applied to both *ECaH* layers (0.375 m and 0.75 m), resulting in near-zero skewness for both orientations. Empirical variograms were calculated for each ECa dataset using the `gstat` R package (Pebesma 2004). Theoretical variogram models (spherical) were fitted using weighted least squares with parameters: $nugget = 0.01,\ sill = 0.53,\ range = 1500\ m$ for *ECaH* and $nugget = 13,\ sill = 2643,\ range = 1858\ m$ for *ECaV*. The fitted variogram models provided the spatial correlation structure for subsequent ordinary kriging interpolation. The very low nugget suggests that the soil proximal survey successfully quantified the soil heterogeneity in terms of the ECa values and no nested structure exists. Ordinary kriging interpolation was performed on a 1×1 m grid covering a polygon defined by the field boundary, creating continuous ECa surfaces across the 6.5 ha study area. The *ECaH* at 0.75 m depth was selected as input for sampling design because it represents the main root zone. All four ECa maps (*ECaH* at 0.375 m and 0.75 m; *ECaV* at 0.75 m and 1.5 m) were retained as predictor variables for subsequent spatial modeling (Section 2.4).

2.2.2 QC: a novel algorithm for soil sampling design

We developed a novel computational method for designing efficient soil sampling schemes that simultaneously resolves two objectives: (1) maximizing geographic dispersion of samples across the field, and (2) ensuring representative coverage of the ancillary variable's (ECa) feature space. The quantile-cluster (*QC*) algorithm addresses the fundamental trade-off in sampling design between spatial coverage and feature space representation while maintaining computational efficiency suitable for iterative exploration of sampling strategies.

Algorithm inputs:

- X – Raster layer of ancillary variable (e.g., *ECaH0.75* interpolated layer)
- n_{min}, n_{max} – minimum and maximum sample sizes
- V – Kriging variance layer for QCvar mode [optional]
- u – Variance filtering parameter [optional]

Algorithm outputs:

- Set of sampling point coordinates for each sample size $n \in [n_{min}, n_{max}]$
- Performance metrics for each sample size (to be described in Section 2.2.3)

Step 1: Geographic Clustering. For a given sample size n, the study area is partitioned into n geographic clusters using the k-means algorithm (Hastie, Tibshirani, and Friedman 2009), which minimizes within-cluster variance in the coordinates space $\{x, y\}$. Each cluster represents a geographic region from which one sample point will be selected, ensuring spatial dispersion.

Step 2: Feature Space Quantiling. Independently, the ECa values across the study area are divided

into n quantiles, creating bins that partition the feature space into equal-frequency intervals. This ensures that the full range of ECa values – from lowest to highest – will be represented in the sample.
Step 3: Point Selection. Sample points are selected iteratively through a greedy procedure that aims to resolve multiple targets: For each geographic cluster (processed in order from clusters with smallest to largest number of ECa quantiles), a point is selected that (a) falls within that cluster, (b) belongs to a quantile not yet represented in the sample, and (c) is closest to the cluster centroid. This dual function balances feature space coverage (criterion b) with geographic dispersion (criteria a and c).
Step 4: Augmentation with Random Points. To capture short-range spatial variation of soil properties, 10% (i.e., $\lceil \frac{n}{10} \rceil$) additional random points are added within close range from selected points (within a maximum distance of $\frac{1}{3}$ of the minimal distance between centroids, that is, ~20 m in this study), following recommendations by Lark and Marchant (2018) for capturing nested variability structures. The contribution of adding just a few close-pair sample points was found to improve an objective function comprised of variogram uncertainty and prediction error variance metrics (Wadoux, Marchant, and Lark 2019).

The QC_{var} variant introduces a quality control constraint on the search space. Rather than selecting from all possible locations, QC_{var} restricts sampling to locations where the ECa kriging variance is below a threshold defined by the parameter u. For this study, $u = 0.15$ was set (i.e., 15$^{\text{th}}$ percentile), which corresponded to a search space of 9,320 pixels representing 15% of the field area. The reduced search space induced by QC_{var} may result in the impossibility to identify sampling points in one or several clusters, due to unmet requirement for one point per cluster and quantile; therefore, to enable convergence and adequacy to field data, the variance filter level is set by the adjustable parameter u.

Furthermore, both the QC and QC_{var} algorithms can be constrained to include only points within the interquartile range (0.25-0.75) of values per feature-space quantile, assuring a representative sampling, by accepting only points whose ECa values fall within the central portion of their assigned quantile (the middle 50% of values within that quantile bin). The example output presented in Figure 2 illustrates the QC algorithm's output for $n = 11$ samples, demonstrating both spatial dispersion of selected points, coverage of the ECa feature space, and ECa values at sampling points displayed within their representative quantile. Figure 3 presents the modified search-space for the same sample-size, reduced by ECa variance ($u = 0.15$).

The QC algorithm exhibits high computational efficiency, with runtime of approximately 1 second per sample size n on a PC with an `Intel i5-8250U` processor and 8GB of RAM. For a typical parameter sweep ($n_{min} = 11, n_{max} = 22$), total processing time is under 15 seconds, representing a $> 100\times$ speedup compared to the comprehensive optimization approach from which this algorithm was developed (Israeli et al. 2019). This computational efficiency enables interactive exploration of sampling strategies, allowing users to iteratively adjust parameters (quantile filtering, variance thresholds) and immediately assess resulting sampling configurations while still in the field following

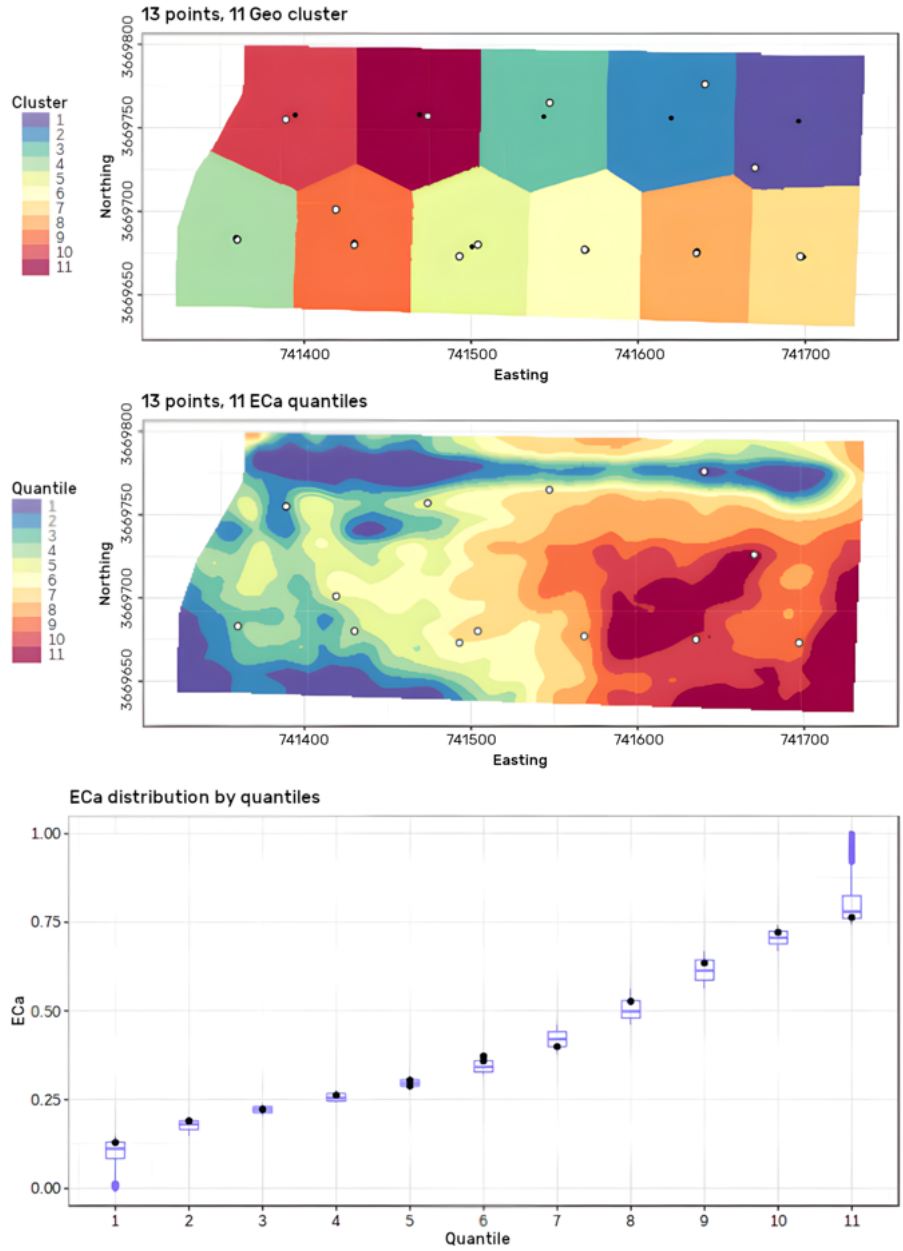

Figure 2: Example output from the QC algorithm for $n = 11$. Geographic space: Division into 11 clusters with cluster centers (centroids) shown as black dots (top). Feature space developed by kriging: Division into 11 ECa quantiles with 13 sampling points marked in white, comprised 11 primary + 2 random augmentation points (middle). Quantile representation: Distribution of normalized ECa values by quantile, with black dots representing ECa values at selected sampling points, demonstrating coverage across the full feature space range (bottom).

ECa survey. To evaluate the proposed sampling method against established approaches, three sampling designs were implemented with identical sample size ($n = 17$, as described in Section 2.2.3):

- ***QC***: Standard quantile-cluster algorithm selecting from the full study area.
- ***QC**$_{var}$*: quantile-cluster with variance constraint, restricting the search space to areas where ECa kriging variance $< 15^{\text{th}}$ percentile (0.007 mS/m^2), ensuring samples are placed only near actual ECa measurements.

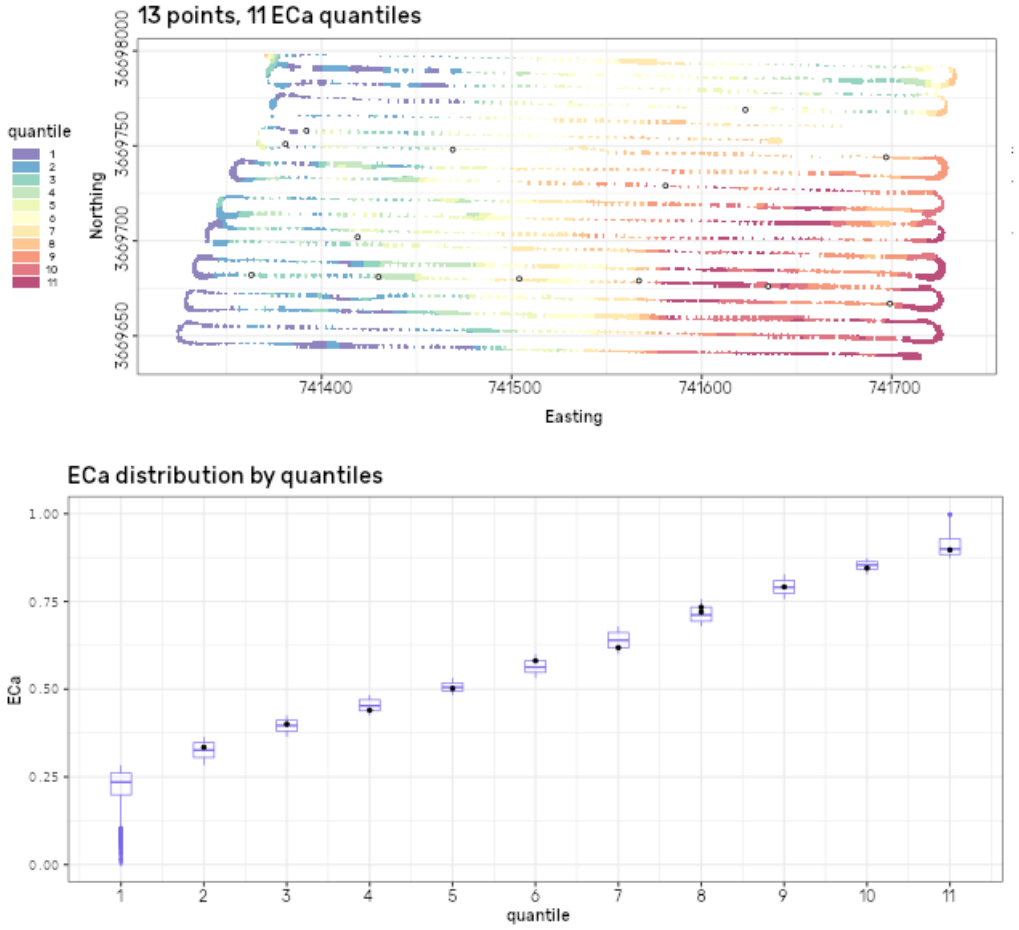

Figure 3: Example output from the QC_{var} algorithm for $n = 11$, displaying the search-space, reduced by variance: Division of the ECaH0.75 layer into 11 quantiles with 13 sampling points marked in dots, comprised 11 primary + 2 random augmentation points. It can be noticed that quantile #1 is not sampled.

- ***Grid***: Regular grid sampling with 67 m horizontal spacing and 49 m vertical spacing, serving as a design-based reference method.

Each sampling configuration was implemented in the field, with soil samples collected at designated locations for laboratory PSD analysis (Section 2.3). The three designs were evaluated based on ECa-PSD correlations and spatial prediction accuracy as described in Section 3.

2.2.3 Sample-size selection aided by information metrics

Determining optimal sample size requires balancing information gain against sampling costs. We employed multiple complementary metrics from Information Theory and geostatistics to objectively assess sampling design quality across sample sizes, enabling data-driven sample-size selection. These metrics evaluate different aspects of sampling design (spatial coverage, feature space representation, and predictive accuracy) and were selected based on their established use in digital soil mapping literature, collectively addressing the multiple objectives of sampling design:

$\bar{\chi}$ – A measure of spatial structure quality that assesses how well sampling points capture spatial autocorrelation in the ancillary variable (Glass 2003). The index $\bar{\chi}$ quantifies the ratio between spatial covariance and variance of ECa values at sampling locations, with values ranging from 0 to 1. Lower $\bar{\chi}$ values indicate stronger spatial structure in the sample, as the average covariance between point pairs approaches the total variance, suggesting effective capture of spatial dependence. The index is calculated as:

$$\bar{\chi} = \frac{1}{\sum_{l=1}^{T} m_l} \sum_{k=1}^{T} \left(\left| 1 - \frac{\left(\gamma_{ij} + \mathrm{cov}\{x_i, x_j\}\right)_k}{\mathrm{var}\{x\}} \right| m_k \right) \tag{2.1}$$

where x represents ECa values at sampling points, T is the number of point pairs, and γ_{ij} is the semivariance between points i and j. The term $cov(x_i, x_j)$ represents the covariance between paired observations, $var(x)$ is the sample variance, and m is the number of points pairs. As sample size increases, $\bar{\chi}$ typically decreases, indicating improved representation of spatial structure. However, $\bar{\chi}$ reaches a plateau when additional samples no longer substantially improve spatial structure characterization, suggesting an optimal sample size.

$\boldsymbol{D_{KL}}$ (Kullback-Leibler divergence) – A measure of distributional similarity (Kullback and Leibler 1951) that quantifies how much the sample distribution of ECa values differs from the population distribution across the entire field. D_{KL} originates from Information Theory and measures the statistical distance between two probability distributions. Lower D_{KL} values indicate the sample better represents the population's feature space, with $D_{KL} = 0$ representing a perfect distributional match. This measure is generally asymmetric, and is formally described as

$$D_{KL}(\mathbf{A} \| \mathbf{A}^{(p)}) = -\sum \mathbf{Pr}(\mathbf{A}) \log \left(\frac{\mathbf{Pr}(\mathbf{A}^{(p)})}{\mathbf{Pr}(\mathbf{A})} \right) \tag{2.2}$$

where $\mathcal{A}$ represents the set of georeferenced ECa values, $Pr(A^{(p)})$ is the cumulative distribution function (CDF) of ECa values in the sample P, and $Pr(\mathcal{A})$ is the CDF of ECa values across the entire field. D_{KL} is particularly sensitive to underrepresentation of rare ECa values (distribution tails). As sample size increases, D_{KL} generally decreases as the sample better captures the population's full distributional characteristics, including extreme values.

cLHS – a conditioned Latin Hypercube Sampling objective function developed by Minasny and Mcbratney (2006) – a composite metric evaluating both feature space coverage and correlation preservation between the sample and population. The *cLHS* objective function measures how uniformly the sample covers the feature space (ECa value range) and how well it preserves the correlation structure present in the population. Lower *cLHS* values indicate better sampling design, with the metric quantified as:

$$cLHS = (O_1 + O_2)/2n \tag{2.3}$$

where O_1 measures deviation from uniform feature space coverage using continuous histogram matching, and O_2 quantifies the difference in correlation structures between sample and entire population – for full mathematical formulation see Minasny and Mcbratney (2006). The division by 2 prescribes equal weight for each objective whereas division by n normalizes the metric by sample size. A value near 0 represents ideal uniform coverage across the ECa spectrum and perfect correlation preservation, while higher values indicate poorer representation.

Mean distance to centroid – The average Euclidean distance between sampling points and their respective cluster centroids in geographic space:

$$MeanDist = \sum_i dist(p_i, c_i) \tag{2.4}$$

Where p_i are coordinates of point in zone i, and c_i are the centroid's coordinates in zone i. This metric evaluates the spatial compactness of the sampling design, with lower values indicating samples are well-positioned near cluster centers, achieving good geographic dispersion while minimizing within-cluster variance.

Cross-validation metrics – a suite of geostatistical measures quantifying spatial prediction accuracy through leave-one-out cross-validation (LOOCV). For each sampling point, the value is predicted using all other points via kriging, then compared to the actual value. These metrics assess how well the sampling design enables accurate spatial interpolation:

- ***MPE* (Mean Prediction Error)** $= (1/n)\sum_i(\hat{z}_i - z_i)$: Measures systematic bias in predictions, with values near zero indicating unbiased estimation.
- ***MSPE* (Mean Squared Prediction Error)** $= (1/n)\sum_i((\hat{z}_i - z_i)^2$: Overall prediction accuracy; lower values indicate better performance.
- ***MSNE* (Mean Squared Normalized Error)** $= (1/n)\sum_i[(\hat{z}_i - z_i)/\sigma_i]^2$: Standardized prediction error accounting for kriging variance σ_i^2. Values near 1 indicate properly quantified uncertainty

where $\hat{z}_i$ is the kriged prediction at location i excluding that point from the model, z_i is the observed value, and n is the number of samples. Together, these metrics provide comprehensive assessment of both accuracy (*MSPE*) and uncertainty quantification (*MSNE*, *MPE*).

By calculating these metrics across a range of sample sizes ($n \in [n_{min}, n_{max}]$), one can identify trends and inflection points that suggest optimal sampling intensity. In theory, all metrics should improve as sample size grows, with diminishing returns beyond an optimal threshold. However, in practice, metrics may show conflicting patterns due to their different objectives: $\bar{\chi}$ emphasizes spatial structure, D_{KL} and *cLHS* focus on feature space coverage, mean distance prioritizes geographic dispersion, and cross-validation assesses predictive accuracy. An optimal sample configuration for one criterion may turn suboptimal for another. Therefore, sample-size selection requires multi-criteria evaluation rather than optimization of a single metric. We opt to plot all metrics against sample size to

visualize trends and identify inflection points, looking for an "elbow" or plateau where metrics cease to improve. Then, by considering practical constraints (field logistics, budget, laboratory capacity), we select a sample size that balances information gain with sampling effort, typically where most metrics show diminishing marginal returns.

For this study, we evaluated sample sizes from $n = 11$ to $n = 22$, with detailed results presented in Section 3.2. The selected sample size ($n = 17$) represents the extent where most metrics plateaued while remaining within budget constraints.

The *QC* algorithm design directly addresses multiple metrics: geographic clustering optimizes mean distance to centroid, quantile stratification minimizes D_{KL} and improves *cLHS*, and the combined approach aims to balance these often-competing objectives. The QC_{var} variant additionally aims to minimize error by restricting samples to low-variance regions where spatial predictions are most reliable.

2.3 Soil-sampling and particle-size analysis

Following the ECa survey and *QC* algorithmic-based sampling design (Sections 2.1-2.2), soil samples were collected in the field to provide ground-truth data for PSD analysis and spatial model calibration. Soil sampling occurred in April 2023, approximately one week after the ECa survey, to maintain similar soil moisture conditions and ensure temporal consistency between ECa measurements and soil characterization.

Soil samples were collected in the field at the 17 sampling points determined by each of the three designs (*QC*, QC_{var}, *Grid*), for a total of 51 soil samples at a depth of 0-20 cm (representing the surface horizon and plow layer), and 34 samples at the subsoil horizon in a depth of 40-60 cm (*QC* and *Grid* only, QC_{var} was excluded from subsoil sampling due to operational constraints) – summing up to 85 in total. At each GPS located point, soil pits were excavated using a backhoe (Figure 4), and samples were collected from pit walls following field evaluation of diagnostic horizons using standard protocols. At each depth interval, approximately 500 grams of soil were collected. Samples were placed in labeled paper bags and transported to the laboratory within hours for processing.

Sample preparation: Soil samples were air-dried at room temperature for ~14 days until reaching constant mass. Dried samples were disaggregated using a motorized soil grinder (Gilson Co. Inc., OH, USA) and sieved through a 2 mm mesh. Approximately 1.0 ± 0.1 g of the <2 mm fraction was weighed into a 50 mL Erlenmeyer flask for subsequent chemical pretreatment according to standard protocols for the laser diffraction analysis. Organic matter removal: Samples were treated with 10% v/v hydrogen peroxide (H_2O_2) solution (approximately 30 mL) and allowed to react for 24 hours in a fume hood. Reaction completion was indicated by cessation of effervescence, then excess liquid was carefully decanted. Following OM removal, samples were treated with approximately 10 mL

Figure 4: Soil sampling collection at the study site using a backhoe excavator (left); soil samples air-drying before PSD measurement (center); and soil sample pretreatment: a 1 g sample stirred with hydrogen peroxide solution for organic matter removal.

3.7% v/v hydrochloric acid (HCl) to dissolve carbonates. Samples were allowed to settle for 3 days to ensure complete reaction and settling of soil particles. After decantation of excess acid, samples were washed once with distilled water with a settling period of 3 days, to remove residual chemicals. Prior to measurement, 1-2 drops of 1% sodium hexametaphosphate $(NaPO_3)_6$ solution were added as a dispersing agent to prevent particle aggregation and ensure measurement of primary particles rather than aggregates (Crouvi et al., 2008).

Laser diffraction analysis: PSD measurements were performed using the *Mastersizer 3000* (Malvern Panalytical Ltd., UK) with the *Hydro EV* wet dispersion unit. Each soil sample was thoroughly homogenized by vortexing for 30 seconds, then 4-6 drops of the soil suspension were added to the dispersion unit until reaching optimal obscuration (10-20%). Ultrasonic dispersion was applied for 180 seconds immediately before measurement to ensure complete particle dispersion. For each soil sample, three analytical replicates (independent sub-samples from the same prepared soil) were measured, with six technical replicates (repeated measurements of the same subsample) per analytical replicate, totalling 18 measurements per sample. This replication design enabled assessment of both within-sample heterogeneity (analytical variance) and instrument measurement error (technical variance). Results were recorded at full resolution (101 size classes spanning 0.01-3,500 µm) for subsequent entropy index calculation (Section 1.4). Complete analysis from field collection to final PSD results required approximately 10 weeks. *Mastersizer* measurements took approximately 15 minutes per sample including technical replicates.

Data processing: Raw particle size distributions (101 classes) from the *Mastersizer 3000* software were exported to textual CSV file formats for analysis. The 6 technical replicates and the 3 analytical replicates were averaged for each soil sample. Then, particle size classes were aggregated into traditional USDA texture classes (clay <2 µm, silt 2-50 µm, sand 50-2000 µm) using adjustable thresholds (2-8 µm for clay/silt boundary) as described in Section 2.4.4. Finally, for each sample,

entropy indices ($H_{Shannon}$ and balanced entropy D) were calculated from the 101-class distribution (excluding >2,000 µm fractions).

The final dataset comprised 85 samples with associated particle size distributions, texture class assignments, and entropy indices, ready for spatial modeling (Section 2.4) and validation analysis (Section 3).

2.4 Spatial Prediction Methodology

Spatial prediction aimed to generate continuous field-scale maps of soil texture from the discrete soil sample measurements ($n = 17$), using ECa as the predictor variable. This required addressing the key challenge of limited sample size for geostatistical modeling and ML training.

2.4.1 Data Augmentation via Interpolation

A sample size of $n = 17$ is insufficient for both reliable variogram estimation (which typically requires at least 100 samples; Webster and Oliver 1992) and robust ML model training (which benefits from larger training datasets). To address this limitation while maintaining spatial representativeness, a data augmentation approach was employed:

- **Interpolate** soil properties from measured points onto a regular grid across the field
- **Filter** the interpolated grid to retain only locations with low prediction uncertainty (high confidence)
- **Use the filtered dataset** for ML model training and testing

This approach, presented in Figure 5, balances the need for larger training datasets against the risk of incorporating unreliable interpolated values.

2.4.2 Variance-Based Filtering

Rather than using all interpolated grid cells (which would include highly uncertain predictions far from sample points), we retained only locations where interpolation uncertainty was relatively low. Specifically, we selected grid cells where the interpolation variance was below the (1/q)-th quantile of the variance distribution (analogous to the QC_{var} sampling constraint described in Section 2.2.2). For example, $q = 10$ (1/10 = 10th percentile) retains cells in the lowest 10% of variance, while $q = 500$ (1/500 = 0.2nd percentile) retains only cells in the lowest 0.2% of variance (highest confidence, smallest dataset). We systematically evaluated q values of $\{10, 100, 200, 500\}$ to assess the trade-off between training dataset size and data quality.

Interpolation methods differed by target variable:

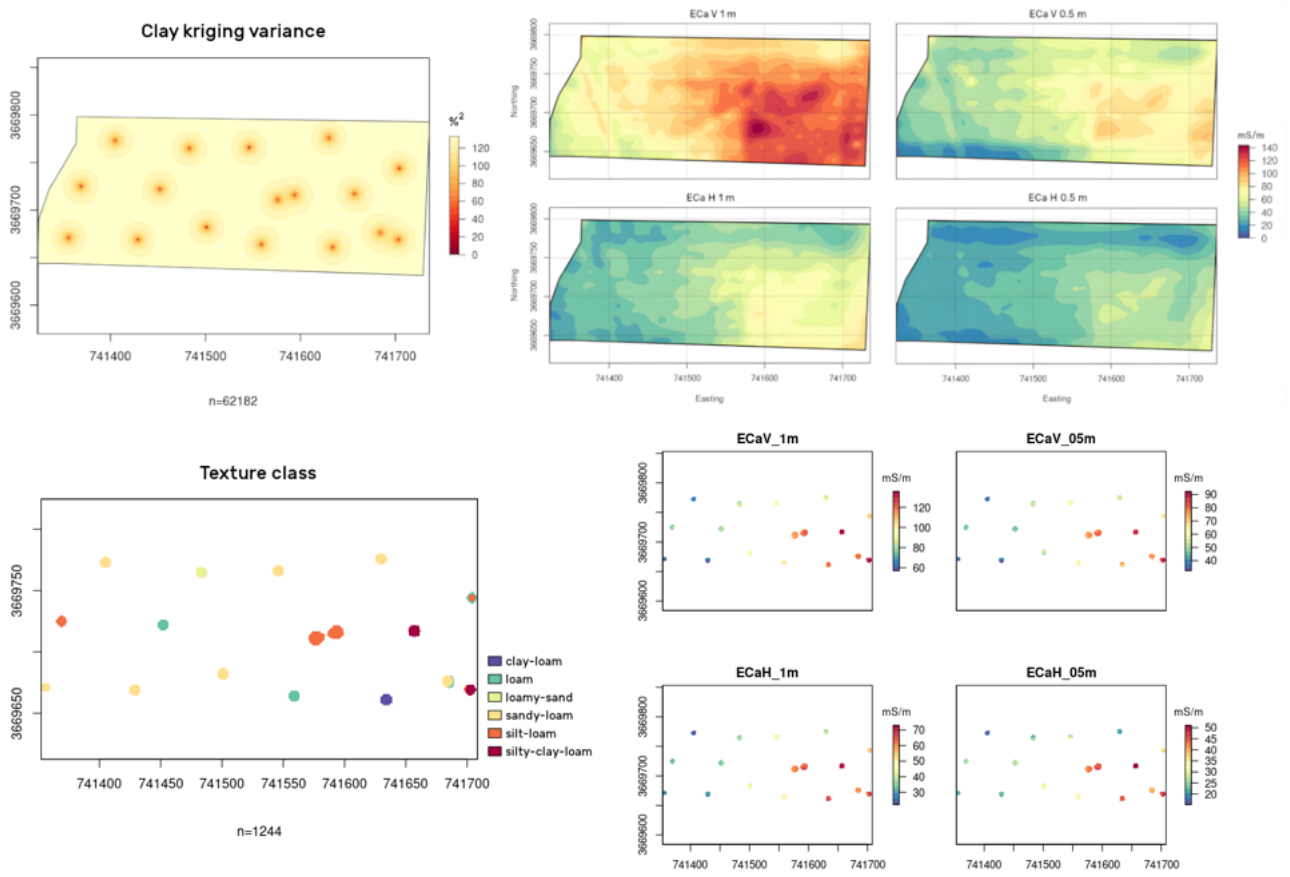

Figure 5: Spatial prediction pipeline for QC sample in horizon B: Kriging variance of clay content ($N = 17$) interpolation onto a 1x1 m grid ($N = 62,218$), used for filtering by variance (top-left); Interpolated ECa layers ($N = 62,218$, top-right); ECa layers filtered by kriging variance with factor $q = 50$ ($N = 1,244$, bottom-right); and soil texture class, calculated from interpolated clay and sand, then reduced by kriging variance ($q = 50$, $N = 1,244$, bottom-left).

- for **categorical texture classes**: Inverse Distance Weighting (IDW) was applied separately to clay, silt, and sand fractions. The interpolated fractions were normalized to sum to 100%, then classified into USDA texture classes based on the triangular distribution (USDA, (2017)).
- for **continuous *D* index**: Ordinary kriging was used with automatic variogram parameters (`gstat` R package; Pebesma 2004), as kriging provides both predictions and associated variance estimates needed for variance-based filtering.

IDW provides more stable fraction predictions but doesn't produce variance estimates; therefore, clay, as the primary texture determinant, was interpolated separately via kriging to enable variance-based filtering for the classification approach.

The unified training dataset comprised:

- **Spatial coordinates** {x, y}
- **Predictor variables**: ECa layers (*ECaH* at 0.375 m and 0.75 m; *ECaV* at 0.75 m and 1.5 m)
- **Particle-size measurements:**: Interpolated and normalized *clay*, *silt*, and *sand* fractions (%)
- **Response variables**:

  - Assigned USDA texture class (categorical: 12 classes)
  - Interpolated D index (continuous)

- **Quality metric**: Interpolation variance (for filtering)

After variance-based filtering, the training dataset contained 124 to 6,200 grid cells (depending on q value), representing a 7-fold to 365-fold increase over the original 17 samples.

2.4.3 Random Forest Model Training

The Random Forest algorithm (Breiman 2001) was implemented via the `randomForest` and `caret` R packages (Liaw and Wiener 2002; Kuhn 2008). The response variable was either soil texture class (classification) or D index (regression), and four scaled ECa layers served as predictor variables. The hyperparameters were set to defaults, with a relatively small number of trees $\{20, 50\}$, since this range proved sufficient and to keep the computational requirements modest. The filtered dataset was randomly split into a *training set* containing 75% of data points for model fitting and a *test set* which holds 25% of data points for hyperparameter tuning and initial model assessment. 10-fold cross-validation was applied to the training set to assess within-training-set performance and model stability. Once a model was fitted, soil texture classes were predicted across the entire field ($n = 62,180$) at 1 m resolution. For the regression approach, predicted D values required conversion to texture classes through a two-stage process: Calibration curves relating D to clay ($R^2 = 0.72$) and sand ($R^2 = 0.85$) content were fitted for each parameter configuration using the 17 measured soil samples via linear regression, whereas silt content was calculated as complementary to 100, then texture class was assigned based on the triangular classification diagram (USDA, (2017)). This approach enables regression on a continuous variable (D) that may better capture distributional information, then translates back to categorical texture classes for practical interpretation.

2.4.4 Model Comparison Framework

To identify optimal model configuration, a comprehensive parameter sweep was conducted across multiple factors. The parameter space included:

- **Model type**: *Classification* of texture class vs. *Regression* of D index
- **Clay/silt threshold**: $\{2, 3, 4, 5, 6, 7, 8\}$ µm
- **Variance filter** q: $\{10, 100, 200, 500\}$
- **Number of trees**: $\{20, 50\}$
- **Initialization seed**[1]: $\{123, 234, ..., 012\}$ (10 values)

[1]Seed values were chosen for compactness and ease of reproduction. For modern pseudo-random number generators such as the Mersenne Twister (Matsumoto and Nishimura 1998), every admissible seed initializes the generator into a distinct point of its state space, and the resulting sequences are statistically equivalent regardless of seed magnitude.

These settings induced 1,120 model runs per sampling design/depth combination, and 5,600 runs in total across 3 sampling designs (*QC, QC_{var}, Grid*) and two horizons (0-20 cm, 40-60 cm, except QC_{var}). Models were run sequentially on a Linux workstation with `Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30 GHz` featuring 48 processing units, with total processing time of approximately 70 hours, resulting in an average runtime of ~45 seconds per model configuration, or ~0.75 milliseconds per grid cell prediction. Performance was evaluated using cross-design validation, where models trained on one sampling design were validated using samples from a different design:

- *QC*-trained models validated on *Grid* samples ($n = 17$)
- *Grid*-trained models validated on *QC* samples ($n = 17$)
- *QCvar*-trained models validated on *Grid* samples ($n = 17$)

This approach ensures complete spatial and methodological independence between training and validation sets, while assessing whether sampling design influences prediction accuracy.

Overall classification accuracy was defined as:

$$Accuracy\ (\%) = \frac{Number\ of\ correctly\ predicted\ texture\ classes}{Number\ of\ validation\ points} \times 100 \tag{2.5}$$

For regression models, accuracy was assessed by first converting predicted D values to texture classes (via calibration curves), then calculating classification accuracy on the resulting classes. This facilitated direct comparison between classification and regression approaches by utilizing a common metric.

While the data augmentation step uses interpolation, the subsequent ML model serves a different purpose: learning the ECa-texture relationship from the augmented dataset, then applying this relationship to predict texture at all locations based on ECa values alone. The interpolation step expands the training dataset but introduces uncertainty; variance filtering mitigates this by retaining only reliable interpolated values. The final predictions are based on ECa measurements (which are available everywhere at 1 m resolution), not on proximity to soil samples, distinguishing this approach from pure spatial interpolation. Detailed results of the model comparison, including optimal parameter selection and prediction accuracy across sampling designs, are presented in Section 2.4.

2.5 Web-Based Application for Workflow Implementation

To facilitate the integrated workflow described in Sections 2.1-2.4 and enable broader application of the *QC* sampling methodology, a custom web-based application was developed using the `shiny` framework (Chang et al. 2024), an open-source R package for creating interactive web applications. The application provides an accessible interface for the computationally intensive procedures involved in ECa-based sampling design and soil texture mapping, enabling users without programming expertise to deploy the methodology.

Technical implementation: The application framework comprises R `shiny` (version 1.10) with supporting packages including `shinydashboard`, `ggplot2`, `caret`, `gstat`, and `sf`. Applications are hosted on shinyapps.io and on local institutional servers and accessible via web browser with no local software installation required. Source code is available on GitHub (see Appendix B), enabling local deployment via R/RStudio for offline use.

Current status and limitations: This *proof-of-concept* system is configured for the UTM Zone 36N coordinate system (Israel/Middle East region). Adaptation to other regions requires minor configuration changes to accommodate different UTM zones or coordinate reference systems. The application is optimized for ECa survey data; adaptation for other ancillary variables (satellite imagery, terrain data) requires code modification. Maximum tested field size is approximately 10 ha (6.5 ha in this study).

The workflow comprises two integrated web applications:

- `App-1` – **Sampling Design Tool**: Workflow from raw ECa survey data to optimized sampling plans
- `App-2` – **Texture Analysis and Prediction Tool**: Integrates PSD measurements with ECa data for spatial texture mapping, including prediction models and benchmarking

Both applications were integral to the workflow described in Sections 2.1-2.4. Detailed technical specifications, user interfaces, and workflows are provided in Appendix B.

Implementation in this study: ECa survey data ($n = 20,800$ points) and PSD measurements ($n = 85$ samples) were processed using these applications. The complete workflow – from ECa preprocessing and QC sampling design (*App-1*) through PSD analysis and spatial prediction (*App-2*) – was executed as described in Sections 2.1-2.4. Most of the figures in this thesis were generated using the built-in visualization tools.

2.6 Methodological Assumptions and Additional Considerations

This section describes the primary assumptions and "design choices" that were made throughout this work, and justify them in light of our working hypotheses.

2.6.1 Key Assumptions

The core assumption underlying the proposed approach is that ECa measurements, when collected under near-field-capacity conditions, primarily reflect soil textural variability rather than just transient moisture differences. Spatial stationarity of soil texture patterns is assumed within field boundaries.

2.6.2 Organic Matter and Salinity Considerations

Organic matter (OM) was measured using Loss on Ignition (Heiri, Lotter, and Lemcke 2001): samples were oven-dried at 105°C, weighed, heated to 550°C overnight, and reweighed. OM measurements ($n = 85$) showed $mean = 5.29\%$, $SD = 1.06\%$, $CV = 20\%$, and moderate correlation with ECa ($R = 0.54$). Despite this correlation, OM was excluded from prediction models to achieve model parsimony focusing on ECa-texture relationships. Soil organic matter may also somewhat hinder the correlation. These highly recalcitrant materials mostly contain humic substances. Their physical chemistry is difficult to ascertain and varies greatly in the Hula Valley because of the drainage and human activities. For soils with highly variable OM (e.g., peatlands), OM should likely be incorporated.

Soil salinity was not formally measured (e.g. by EC_e, SAR). The Hula Valley site is irrigated by freshwater with no visual salinity indicators, and successful ECa-texture correlations (Section 3.3) suggest salinity is not dominant. However, this represents a study limitation. Our methodology is most appropriate for non-saline or mildly saline soils; saline conditions require additional characterization to separate the contribution of soil water content and salinity to ECa signal (Autovino et al. 2025).

2.6.3 Model Transferability

ECa relationships with soil properties are inherently site-specific (Heil and Schmidhalter 2017). This study does not propose a universal prediction model, but rather a transferable five-step procedure: (1) ECa survey under standardized conditions, (2) QC-based sampling design, (3) high-resolution PSD analysis with entropy characterization, (4) site-specific model calibration, and (5) independent validation. Naturally, in practice one should expect different optimal parameters such as clay/silt thresholds, ECa-texture correlations, and q values across sites. The QC web framework (Section 2.5) facilitates rapid site-specific model development.

2.6.4 Pilot Studies and Limitations

The QC methodology was tested at several additional crop fields and orchard sites. At Conqueiros (Alentejo, Portugal), as well as at Beeri (Israel), the QC algorithm performed as expected but lacked independent validation samples for prediction. At Neve Yaar (Israel), all samples were classified as a single texture class (silt-loam), demonstrating that homogeneous fields require simplified approaches.

Results

3.1 ECa Survey Results and Spatial Patterns

The ECa survey conducted in late April 2023 (Section 2.1) successfully characterized electromagnetic conductivity patterns across the 6.5 ha study field. A total of 9,700 ECa readings were obtained in horizontal orientation and 11,100 readings in vertical orientation during two separate passes, providing high-density spatial coverage (approximately 3,200 measurements per hectare). Following quality control procedures (Section 2.2.1), *ECaH* values were log transformed to achieve a distribution closer to normality. After data compaction by moving average at every 5^{th} point, about 2,000 measurements per layer served as kriging inputs, which were then used for sampling design (Section 2.2) and soil texture prediction (Section 2.4). ECa measurements showed substantial spatial variability across the field (Table 1).

The coefficient of variation ranged from 24% to 35%, indicating moderate to high spatial heterogeneity, suitable for the *QC* sampling approach. ECa values showed a clear depth trend, increasing nearly 3-fold from shallow ($mean = 31\ mS/m$ at $0.375\ m$) to deep ($mean = 92\ mS/m$ at $1.5\ m$) measurements, suggesting either increasing clay content with depth, higher moisture retention in subsoil, or both. This vertical stratification is consistent with the alluvial depositional history of the site (Henares, Donselaar, and Caracciolo 2020; Section 2.1). While ECa magnitude increased with depth, spatial patterns showed strong similarity across depths, with Pearson's correlation between layers of $R(ECaH0.375 \sim ECaH0.75) = 0.97$ and $R(ECaH0.75 \sim ECaV1.5) = 0.93$, indicating that shallow and deep texture patterns are spatially coherent. This supports the use of shallow ECa (0.75 m horizontal) as the primary input for sampling design (Section 2.2).

The interpolated ECa maps reveal distinct spatial zones within the field (Figure 6): Lower ECa zone – located in the western portion of the field, covering approximately 30% of the area, likely

Table 1: Summary statistics for raw ECa measurements across four depth settings: mean, standard deviation, coefficient of variation, minimum and maximum values.

Layer	Orientation	Depth	Mean	SD	CV	Min	Max
		m	mS/m	mS/m	%	mS/m	mS/m
ECaH0.375	Horizontal	0.375	31.96	11.23	35.1	8.76	62.55
ECaH0.75	Horizontal	0.75	45.48	16.05	35.3	13.77	84.86
ECaV0.75	Vertical	0.75	56.4	17.6	31.3	6.88	137.59
ECaV1.5	Vertical	1.5	92.34	22.7	24.6	29.82	145.64

corresponding to sandier soil textures; Higher ECa zone – concentrated in the south-eastern portion, representing approximately 30% of the area, suggesting finer-textured (clay-rich) soils; and transitional zones covering approximately 40% of the area show gradual transitions between extremes rather than sharp boundaries, consistent with alluvial deposition patterns and the paleochannel features observed at the site (Section 2.1). The spatial continuity and structure visible in these maps justified the geostatistical interpolation and sampling design approaches employed (Section 2.2).

Variogram analysis (Section 2.2.1) revealed structured spatial dependence with a range exceeding 1,500 m. This substantial spatial structure – exceeding the field dimensions (~400 m × 160 m) – indicates that spatial correlation extends beyond the sampled area, which is advantageous for kriging as it suggests minimal edge effects. These ECa spatial patterns and their relationship to soil texture are evaluated through correlation analysis (Section 3.3) and serve as the foundation for spatial prediction models (Section 3.4).

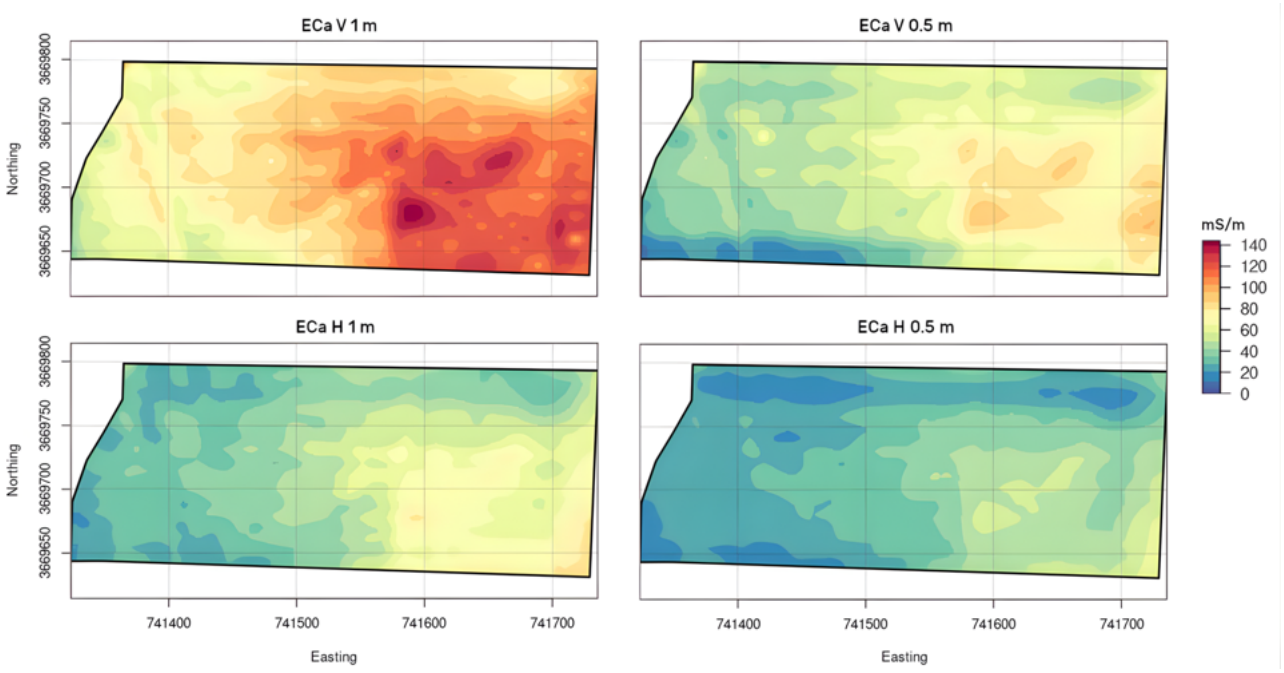

Figure 6: Interpolated ECa maps of the 6.5 ha study field. Maps show apparent electrical conductivity (mS/m) for four depth configurations: (a) ECaV at 1.5 m, (b) ECaV at 0.75 m, (c) ECaH at 0.75 m, and (d) ECaH at 0.375 m depth. All maps are based on ordinary kriging with 1 m grid resolution ($n = 62,180$ cells) from approximately 2,000 quality-controlled measurement points per layer (after compaction from original 9,700 horizontal and 11,100 vertical measurements). Note the systematic increase in ECa values with depth and the spatially coherent patterns across all depths. Warmer colors indicate higher ECa values; cooler colors indicate lower ECa values.

3.2 Sampling Design Optimization and Particle Size Distribution Analysis

3.2.1 Sampling Design Optimization

Following ECa interpolation (Section 3.1), the *QC* algorithm was applied to generate optimized sampling plans using *ECaH0.75* as the input layer. Sample sizes ranging from 11 to 22 points were evaluated to identify the optimal balance between information gain and sampling effort (Section 2.2.3). Figure 7 presents seven information metrics calculated for each candidate sample size. The metrics did not simultaneously converge on a single optimal sample size, reflecting the multi-criterion nature of sampling design optimization:

- $\bar{\chi}$ showed minimum at $n = 17$, suggesting optimal feature-space coverage at this sample size.
- **Kullback-Leibler divergence (D_{KL})**: While the observed minimum occurs at $n = 11$ (D_{KL} = -2, likely due to limited sample representation of distribution tails), the next local minimum is at $n = 17$.
- **Conditioned Latin Hypercube Sampling metric (*cLHS*)**: Optimal value observed at $n = 17$, indicating efficient stratification.
- **Mean distance to centroid**: Generally decreased with sample size, as expected, with a minimum at 18.
- **Cross-validation metrics (*MPE, MSNE, MSPE*)**: Showed optimal values at different sample sizes ($n = 15$, 13, and 22 respectively), indicating no clear consensus from validation statistics.

Given the differing optimal points across metrics (as to be expected in multi-criterion problems), $n = 17$ was selected based on the following considerations: (1) Earliest majority, as $n = 17$ showed the first instance where multiple metrics ($\bar{\chi}$ and *cLHS*) simultaneously achieved optimal values; (2) Best *cLHS* score among all sample sizes, indicating superior spatial and feature-space stratification; (3) Reasonable scores for cross-validation metrics; (4) Beyond $n = 17$, marginal improvement in most metrics was minimal (diminishing returns); and (5) a reasonable balance between information gain and field sampling effort.

Three sampling approaches were implemented for the selected sample size ($n = 17$) to enable comparative evaluation (Figure 8):

- ***QC*** – quantile-cluster design (blue points):
  - **Algorithm**: *QC* with quantile stratification (Section 2.2.2)
  - **Input layer**: *ECaH0.75*
  - **Initialization**: deterministic (seed is a benchmarking variable)
  - **Spatial distribution**: dispersed

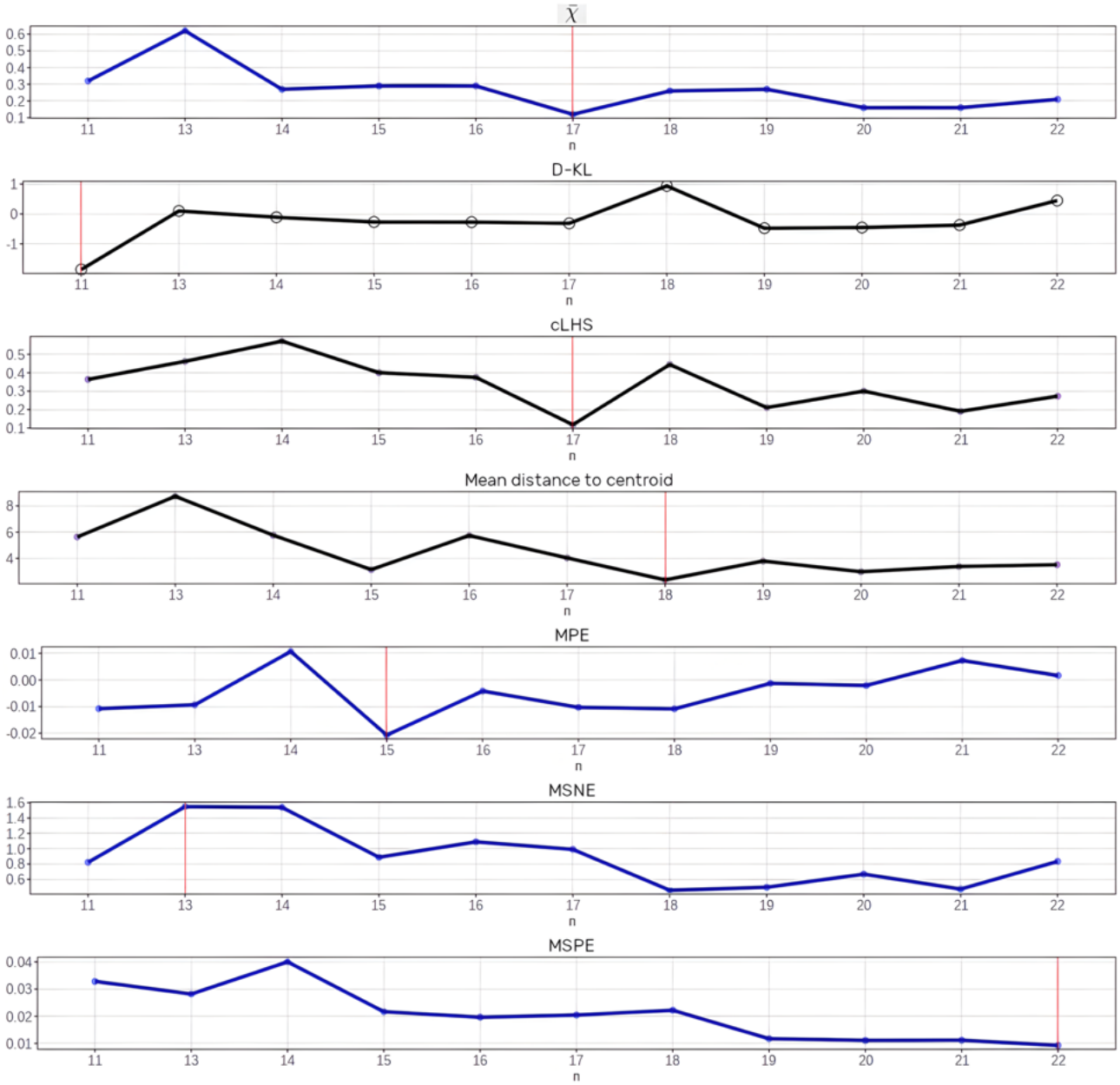

Figure 7: Information metrics for evaluating QC sampling plan quality across sample sizes ($n \in \{11, \ldots, 22\}$). (a) $\bar{\chi}$ index assessing spatial model quality, (b) Kullback-Leibler divergence (D_{KL}) measuring distributional similarity, (c) Conditioned Latin Hypercube Sampling metric (cLHS) evaluating feature stratification, (d) Mean distance to centroids quantifying dispersion, (e) Mean Prediction Error (MPE), (f) Mean Squared Normalized Error (MSNE), and (g) Mean Squared Prediction Error (MSPE). Red vertical lines indicate the sample size that achieves optimal value for each metric. Note the lack of consensus across metrics, with different optimal sample sizes per the different metrics. $n = 17$ was selected based on simultaneous optimality of $\bar{\chi}$ and cLHS. All metrics calculated from the ECaH0.75 input layer.

  - **Feature-space coverage**: designed to capture full range of ECa values with equal representation in each quantile

- ***QC**$_{var}$* – Variance-filtered *QC* design (red points):
  - **Algorithm**: *QC* with variance-based filtering ($u = 0.15$)
  - **Input layer**: *ECaH0.75* with kriging variance

  - **Constraint**: samples restricted to locations with kriging variance $< 15^{th}$ percentile
  - **Spatial distribution**: dispersed, more concentrated in high-confidence areas

- ***Grid*** – Regular systematic design (yellow points):
  - **Algorithm**: Regular points (QGIS Development Team 2023)
  - **Grid spacing**: approximately 60 m between points
  - **Adjustment**: manually adjusted to fit field boundary
  - **Spatial distribution**: uniform coverage regardless of ECa patterns
  - **Purpose**: serve as traditional sampling reference for comparison

Visual inspection of Figure 8 reveals that the 3 designs are distributed across the full ECa gradient (from low-ECa western zone to high-ECa southeastern zone). Soil sampling included 51 sampling locations (17 points $\times$ 3 designs) in two depths: Horizon A (0-20 cm), representing the topsoil layer, $n = 51$ samples and Horizon B (40-60 cm), representing the subsoil, $n = 34$ samples (*QC* and *Grid* only). In total, 85 samples were collected (51 surface + 34 subsoil).

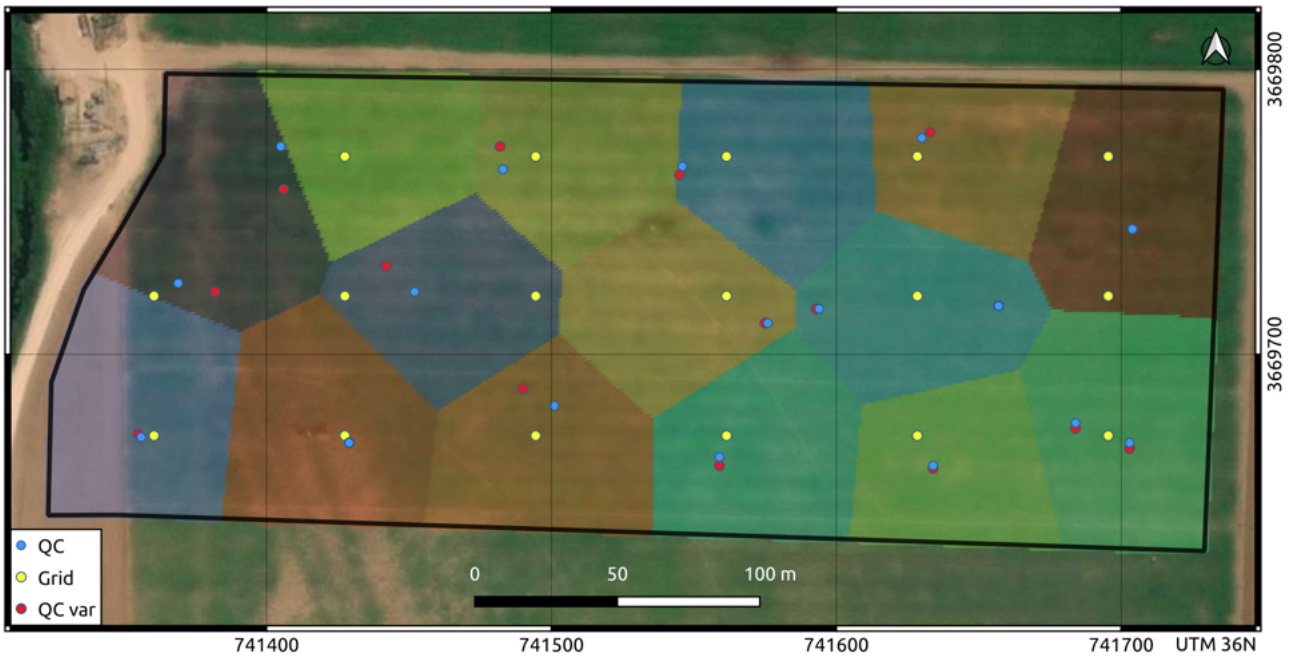

Figure 8: Comparison of three sampling designs ($n = 17$ points each) in the study area. (a) QC design (blue dots) with quantile-based stratification, (b) Grid design (yellow dots) with regular spacing, and (c) QC_{var} design (red dots, behind) with variance-based filtering. Surface color marks the division into 15 clusters implemented by QC and QC_{var}.

3.2.2 Particle Size Distribution Analysis

Following soil samples drying and pretreatment (Section 2.3), high-resolution PSD was measured using laser diffraction (*Malvern Mastersizer 3000*), providing 101 discrete size classes spanning 0.01 to 3,500 µm (Figure 9. The complete distributions enabled: (1) calculation of entropy index D (Eq. 1.3) from the full 101-class distribution and (2) derivation of traditional texture fractions (clay, silt,

Table 2: Texture class distribution by sampling design, $n = 17$ for each design, Horizon A: 0-20 cm depth; Horizon B: 40-60 cm depth)

Texture Class	QC A	QC B	Grid A	Grid B	QCvar A	Total	% of Samples
Clay Loam	0	1	0	0	0	1	1%
Loam	9	3	9	9	9	39	46%
Sandy Loam	8	7	5	6	2	28	33%
Silt Loam	0	3	3	2	6	14	17%
Loamy Sand	0	1	0	0	0	1	1%
Silty Clay Loam	0	2	0	0	0	2	2%
Total classes	17	17	17	17	17	85	100%

sand) using variable clay/silt cutoff parameters (2-8 µm). Analysis with the 2 µm clay/silt cutoff

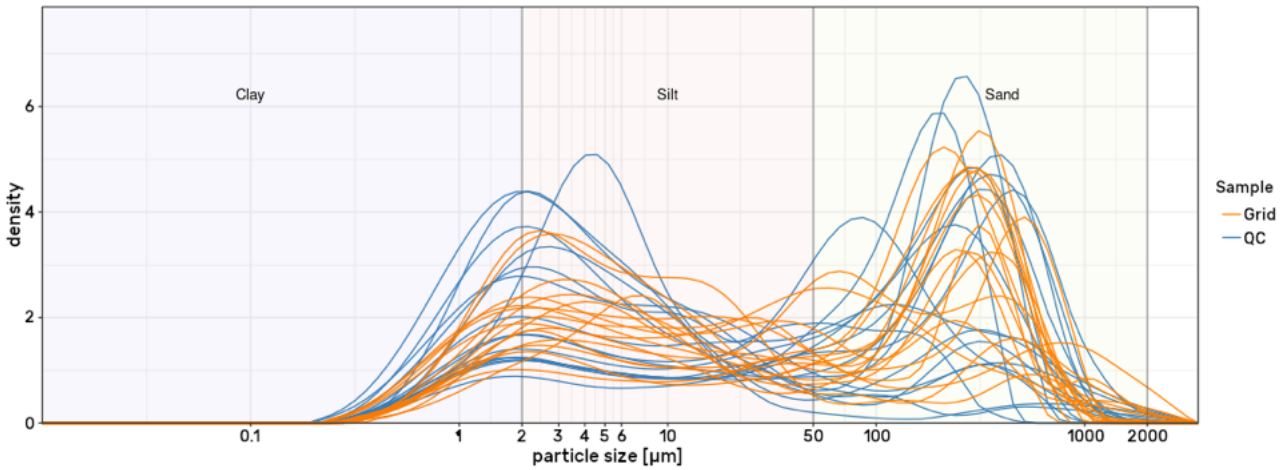

Figure 9: Density plot of raw particle size distributions for Horizon B by sampling design, portraying the 101-classes distribution for Grid sample (orange lines) and QC (blue lines), in the range $[0.01 - 3,500\ \mu m]$, vertical lines dividing PSD fractions by USDA (2017) classification. It is noticeable that QC sample has more variability compared to Grid sample.

(conventional USDA standard, though alternative cutoffs are evaluated in Section 3.4.4) revealed substantial textural heterogeneity across the field (Figure 10, Table 2).

In total, 6 out of 12 possible USDA texture classes were observed, ranging from loamy-sand to silty clay loam, whereas the dominant class is loam, representing 46% of samples. Across all samples ($n = 85$), particle size fractions showed:

- **Clay** content (<2 µm): $mean = 15.3\% \mp 5.6\%$, range $[2 - 38\%]$, $CV = 37\%$
- **Silt** content (2-50 µm): $mean = 39.8\% \mp 11.2\%$, range $[21 - 68\%]$, $CV = 28\%$
- **Sand** content (>50 µm): $mean = 44.6\% \mp 15\%$, range $[6 - 70\%]$, $CV = 34\%$
- **Entropy index *D***: $mean = 0.57 \mp 0.05$, range $[0.42 - 0.67]$, $CV = 9\%$

The moderate to high coefficients of variation confirm substantial textural heterogeneity across the field, justifying the precision sampling approach.

Comparison of Horizon A (0-20 cm) and Horizon B (40-60 cm) for QC and Grid designs ($n = 34$ paired samples) revealed:

- **Horizon A** ($n = 34$): Dominant classes were loam (18 samples, 53%) and sandy loam (13 samples, 38%)
- **Horizon B** ($n = 34$): More diverse texture distribution with increased finer textures: loam (12 samples, 35%), sandy loam (13 samples, 38%), silt loam (5 samples, 15%), plus clay loam (1), loamy sand (1), and silty clay loam (2)
- **Vertical stratification**: Subsoil showed greater representation of silt loam and clay-rich classes (9/34 = 27% vs. 3/34 = 9% in topsoil), with several samples shifting toward finer textures with depth

This vertical stratification is consistent with the ECa depth trend (Section 3.1, where deeper ECa measurements showed ~3-fold higher values) and reflects typical alluvial profile development with finer particle accumulation in subsoil horizons (Henares, Donselaar, and Caracciolo 2020).

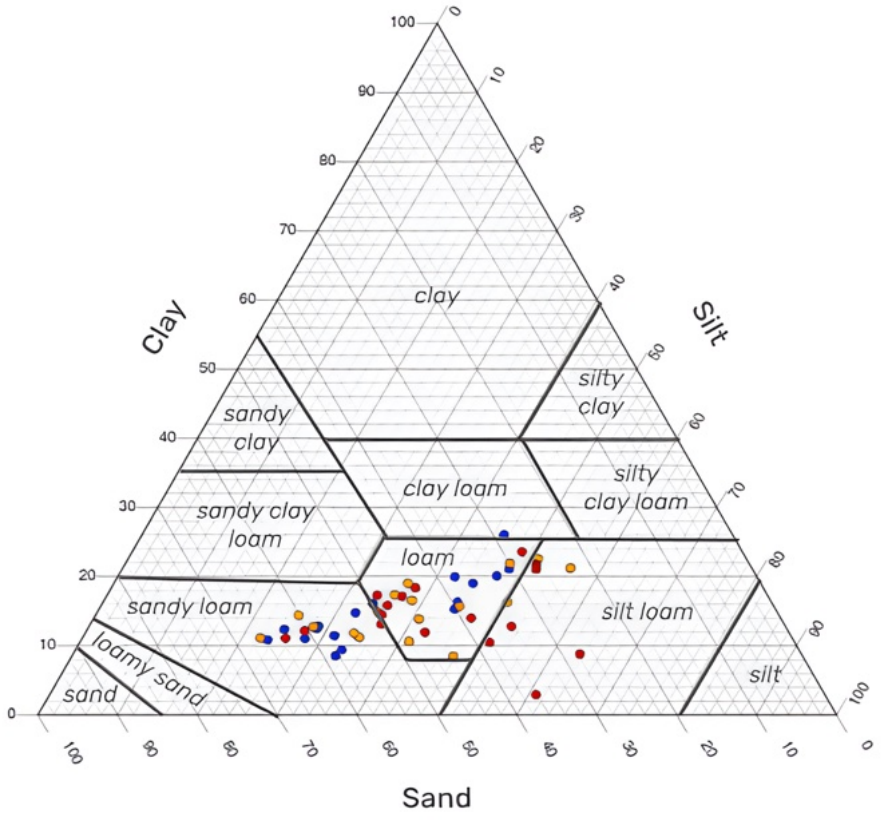

Figure 10: Particle size distribution results plotted over USDA soil texture triangle ($n = 51$, depth 0-20 cm). Points represent soil samples from three designs: QC (blue), Grid (orange), and QC_{var} (red). Clay/silt boundary set at 2 µm (conventional USDA standard). Six texture classes observed (Table 2): loamy sand, sandy loam, loam, silt loam, clay loam, and silty clay loam. These particle size distributions and texture classifications provide the foundation for ECa-texture correlation analysis (Section 3.3) and spatial prediction model development and validation (Section 3.4).

Table 3: Pearson's R correlation coefficient: ECa layers vs. PSD classes (clay/silt cutoff = 2 µm). Values in bold text indicate statistically significant correlations ($|R| > 0.482$, $p < 0.05$ for $n = 17$).

Sample	Horizon	Fraction	ECa H 0.37 m	ECa H 0.75 m	ECa V 0.75 m	ECa V 1.5 m
QC	A	Clay	0.38	**0.51**	0.48	0.48
		Silt	0.46	**0.62**	0.48	**0.51**
		Sand	-0.46	**-0.62**	**-0.51**	**-0.53**
	B	Clay	**0.67**	**0.74**	**0.69**	**0.71**
		Silt	**0.67**	**0.77**	**0.7**	**0.7**
		Sand	**-0.71**	**-0.81**	**-0.74**	**-0.75**
Grid	A	Clay	0.22	0.26	0.11	0.11
		Silt	0.03	0.21	-0.03	0.04
		Sand	-0.1	-0.25	-0.06	0
	B	Clay	**0.52**	**0.63**	**0.57**	**0.52**
		Silt	0.38	**0.58**	0.39	0.31
		Sand	-0.45	**-0.63**	-0.48	-0.4
QCvar	A	Clay	0.46	**0.58**	0.42	0.44
		Silt	0.35	0.26	0.28	0.3
		Sand	**-0.55**	**-0.52**	-0.46	**-0.49**

3.3 ECa-Texture Correlations and Sampling Design Comparison

3.3.1 Correlation Analysis Overview

Prior to developing spatial prediction models, the strength of relationships between ECa measurements and soil texture properties were evaluated across the three sampling designs. This correlation analysis serves two purposes: (1) validate the fundamental assumption that ECa reflects textural variability under controlled moisture conditions (Section 2.6), and (2) assess whether sampling design influences the observed ECa-texture relationships, which could affect prediction model accuracy.

Pearson correlation coefficients (R) were calculated between four ECa layers (*ECaH0.375, ECaH0.75, ECaV0.75, ECaV1.5*) and three texture fractions (clay, silt, sand) derived using 2 µm clay/silt cutoff, for each sampling design (QC, $Grid$, QC_{var}) separately, at each depth (horizons A, B), based on $n = 17$ sampling points per design per depth. Correlation strength interpretation follows conventional guidelines (Cohen, 1988): $|R| < 0.3$ (weak), $0.3 - 0.7$ (moderate), > 0.7 (strong). Statistical significance was assessed at $\alpha = 0.05$.

3.3.2 Correlations with Traditional Texture Fractions

Table 3 presents Pearson correlation coefficients between ECa layers and texture fractions (clay/silt cutoff = 2 µm) for all design-depth combinations.

It is apparent that Horizon B (40-60 cm) consistently showed stronger correlations than Horizon A (0-20 cm) across all designs, with mean R for clay-ECa: Horizon A $= 0.39 \pm 0.15$, Horizon B

$= 0.64 \pm 0.09$ (64% stronger), and mean R for sand-ECa: Horizon A $= -0.38 \pm 0.21$, Horizon B $= -0.63 \pm 0.18$ (66% stronger). This depth-dependent correlation pattern likely reflects reduced management interference in subsoil (no tillage, less compaction variability), greater textural stability in B horizon, stronger moisture gradient in deeper measurements where texture more strongly controls retention, and/or better soil-sensor contact during ECa survey when shallow soil is drier.

The *ECaH0.75* layer generally showed strongest correlations across designs and depths, with mean $R(ECaH0.75) = 0.56$ vs. $R(ECaH0.375) = 0.45$, (24% stronger). This justifies the selection of *ECaH0.75* as the primary input for sampling design (Section 3.3.3). Clay and sand showed strong negative correlation, as expected (R for clay-sand $= -0.6$ in Horizon A, -0.86 in Horizon B, not shown), reflecting the inverse relationship between fine and coarse fractions. Sand and clay showed similar correlations with ECa in most cases, with mean R for clay-ECa $= 0.47$ vs. sand-ECa $= -0.47$, while silt-ECa correlations were generally slightly weaker and more variable.

3.3.3 Sampling Design Comparison

Sampling design significantly influenced correlation strength:

In **Horizon A** (topsoil, $n = 17$), *QC* design achieved moderate yet statistically significant correlations between *ECah0.75* and clay ($R = 0.51,\ p = 0.04$), akin to the sand fraction ($R = -0.62,\ p = 0.01$). Correlations with ECaH0.75 by the QC_{var} design were of similar strength as for *QC* in Horizon A, with clay ($R = 0.58,\ p = 0.01$) and sand ($R = -0.52,\ p = 0.03$). Grid design in Horizon A showed insignificant correlations between *ECaH0.75* with clay ($R = 0.26,\ p = 0.31$) and sand ($R = -0.25,\ p = 0.32$), displaying correlations that are 51-60% weaker than QC/QC_{var}. Statistical comparison by Fisher's Z-transformation revealed that for clay-*ECaH0.75* relations, *QC* vs *Grid* scored $Z = 1.15$ ($p = 0.13$), indicating a non-significant trend toward stronger correlations. QC_{var} vs *Grid* had a similar difference, whereas *QC* vs QC_{var} scored $Z = 0.31$ ($p = 0.76$) indicating a non-significant difference.

At the subsoil **Horizon B** ($n = 17$), *QC* design achieved strong correlations between *ECaH0.75* and the soil fractions, where clay ($R = 0.74,\ p = 0.001$), sand ($R = -0.81,\ p < 0.001$) and silt ($R = 0.77,\ p < 0.001$) correlations are highly significant and strong. *Grid* design in Horizon B showed moderate-strong correlations between *ECaH0.75*, clay ($R = 0.63,\ p = 0.01$) and sand ($R = -0.63,\ p = 0.01$), which are significant but 15-22% weaker than *QC*. Statistical comparison by Fisher's Z-transformation found a consistent but insignificant trend between *QC* vs. *Grid* for clay, with $Z = 0.62$ ($p = 0.27$), while for sand, the trend was marginally significant with $Z = 1.07$ ($p = 0.14$).

While *Grid* sampling achieved statistically significant correlations in Horizon B, *QC* and QCvar designs consistently showed stronger ECa-texture relationships, suggesting better feature-space coverage despite identical sample size ($n = 17$). These results demonstrate that *QC* and QC_{var} designs captured ECa-texture gradients more efficiently than uniform grid. Stratified sampling (*QC*) outperformed uniform spatial sampling (*Grid*) for correlation analysis, whereas variance filtering (QC_{var}) maintained

strong correlations while restricting to high-confidence areas. These correlation differences likely translate to prediction accuracy differences (evaluated in Section 3.4).

3.3.4 High-Resolution Particle Size Correlation Patterns

To examine ECa-texture relationships in greater detail, correlations were calculated between ECa layers and all 101 discrete particle size fractions as measured by the laser diffraction analysis (Figure 11).

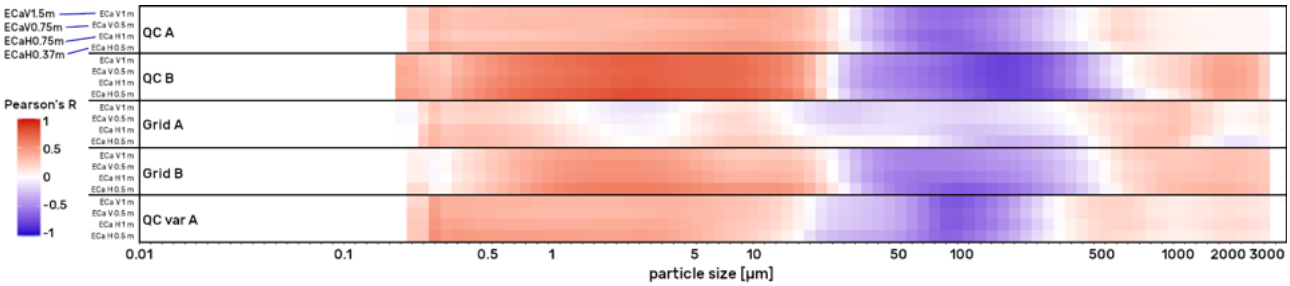

Figure 11: Correlation between 4 ECa layers and 101 particle size classes, 17 sampling points in each sample design (QC, Grid, QC_{var}), at two depths (A: 0-20 cm, B: 40-60 cm). Orange color corresponds to positive correlation (1) and purple corresponds to negative correlation (-1), white means no correlation. A bimodal pattern with positive peak at 2-5 µm and a negative peak at 60-200 µm is apparent across all designs, as the QC Horizon B sample (2nd from top) shows the strongest correlations.

Figure 11 demonstrates that **maximum positive correlation** occurred in the 1-5 µm size range (overlapping clay/silt boundary) as *QC* Horizon B showed the strongest peak: R = 0.81 with *ECaH0.75* at **2.4 µm** (above the conventional 2 µm clay/silt boundary). This fine particle range consistently showed higher correlations than either pure clay (<2 µm) or pure silt (2-50 µm). This is likely due to ECa responding most strongly to the fine-silt + clay fraction that controls both water retention and electrical conductivity. **Peak negative correlations** occurred in the 60-200 µm size range (fine to medium sand), with *QC* Horizon B showing strongest negative peak with *ECaV0.75* at **143 µm** (R = -0.88). This fine sand range (125-250 µm) showed stronger correlations than total sand fraction (>50 µm).

Inspection by design and depth patterns (Figure 11) reveals that QC Horizon B (second panel from top) achieved the strongest correlations overall (R range: -0.88 to +0.81), with a clear bimodal pattern displaying a positive peak at 2-3 µm and a negative peak at 100-200 µm, with sharp transitions between positive and negative zones, plateauing to 0 correlation at ~20 µm. This indicates a high signal-to-noise ratio. Similar peak correlations, with a dampened bimodal pattern, are noticeable at *QC* Horizon A (top panel, *R* range: -0.86 to $+0.75$). *Grid* Horizon B (3rd panel from top) showed similar correlations (*R* range: -0.84 to $+0.73$ for *ECaH0.75*), with a bimodal pattern present and more variability across ECa layers. However, *Grid* Horizon A displayed medium-weak correlations (*R* range: -0.38 to $+0.5$), with pattern barely discernible, indicating high noise relative to signal. QC_{var}

Table 4: Pearson's correlation coefficient R between ECaH0.75 layer and clay content (mean, median and maximal) at sampling points, for Horizons A and B, by clay/silt cutoff size.

Clay/silt cutoff	**Horizon A**			**Horizon B**		
µm	**mean**	**median**	**max**	**mean**	**median**	**max**
2	0.45	0.51	0.58	0.69	0.69	0.74
3	0.46	0.56	0.61	0.74	0.74	0.78
4	0.47	0.6	0.64	0.76	0.76	0.8
5	0.49	0.63	0.67	0.77	0.77	0.82
6	0.51	0.65	0.69	0.78	0.78	0.82
7	0.52	0.67	0.7	0.78	0.78	0.82
8	0.53	0.67	0.7	0.78	0.78	0.82

Horizon A (bottom panel) achieved medium-strong correlations (R range: -0.84 to $+0.62$), similar to *QC* A, stronger than *Grid* A.

All four ECa layers showed similar correlation patterns within each design-depth combination. *ECaH0.75* and *ECaV1.5* generally showed slightly stronger correlations, whereas pattern shapes were consistent even when magnitudes differed. These results suggest that: (1) traditional clay/silt boundary (2 µm) may not be optimal for ECa prediction – peak occurs slightly above this boundary; (2) fine silt (2-5 µm) contributes strongly to ECa signal; (3) fine sand (~60-200 µm) is key discriminator, more than total sand; (4) *QC* sampling better captured these relationships than Grid sampling; and (5) depth matters more than design, as even Grid B outperformed *QC* A.

3.3.5 Effect of Clay/Silt Cutoff on Correlation Strength

Given that peak correlations occurred in the 1-5 µm range rather than below the conventional 2 µm boundary, we systematically evaluated how clay/silt cutoff selection affects ECa-clay correlations. Clay content was recalculated using cutoff values of $\{2,3,4,5,6,7,8\}$ µm, and correlations with *ECaH0.75* were computed for each design-depth combination (Table 4).

In horizon A ($n = 51$, all designs pooled), correlation grew by +18% from 2 µm (mean $R = 0.45 \pm 0.16$) to 8 µm cutoff (mean $R = 0.53 \pm 0.13$), while for horizon B the trend was similar, from 2 µm (mean $R = 0.69 \pm 0.08$) to 8 µm cutoff (mean $R = 0.78 \pm 0.06$), representing an increase of +13%. **All designs showed increasing correlations with larger cutoffs**, as improvements plateaued around 6-8 µm, whereas the *QC* design showed the strongest improvement ($R = 0.51$ß0.67 and $R = 0.74$ß0.82 for horizons A and B, respectively). It appears that fine silt particles (2-5 µm) contribute to ECa signal similarly to clay, as laser diffraction may classify some particles differently than traditional sedimentation methods. Therefore, for ECa-based prediction, a 4-6 µm cutoff may be more appropriate than conventional 2 µm. This cutoff sensitivity is explored further in prediction models (Section 3.4). Based on these findings, prediction models were tested with cutoffs of 2-8 µm (Section 3.4.4) to identify optimal classification for this field and measurement technique.

3.3.6 Summary and Implications for Spatial Prediction

The correlation analysis revealed that strong ECa-texture relationships exist within the study data ($|R|$ up to 0.88), validating the fundamental premise for ECa-based texture mapping. A dominant depth effect is evident, as Horizon B correlations were 64-66% stronger than Horizon A across all designs. The correlations demonstrate that sampling design matters, as QC and QC_{var} achieved stronger correlations ($R = 0.51 - 0.74$) than Grid ($R = 0.26 - 0.63$) with identical sample size.

High-resolution PSD provides an interesting perspective (Figure 11), with peak correlations occurring at 2-5 µm (fine silt/clay boundary) and 60-200 µm (fine sand), not at traditional fraction boundaries. Clay/silt cutoff sensitivity analysis showed that increasing the threshold from 2 to 6-8 µm improves correlations by 12-18%. Finally, the *ECaH0.75* layer (horizontal orientation, depth of up to 0.75 m) emerged as an optimal predictor, consistently demonstrating strongest correlations across designs and depths.

The sufficient correlation strength ($|R| > 0.5$ for most cases) supports the development of ML prediction models (Section 3.4), with QC and QC_{var} expected to yield higher prediction accuracy than *Grid* due to stronger correlations, and Horizon B predictions likely more accurate than Horizon A. Testing multiple clay/silt cutoffs (2-8 µm) is warranted given the sensitivity exemplified.

With $n = 17$ samples per design, correlations of $R > 0.48$ are statistically significant ($p < 0.05$), meaning most of QC and QC_{var} correlations are significant, while Grid correlations at Horizon A are not. This has important implications for prediction model reliability (Section 3.4).

3.4 Spatial Prediction Model Performance

3.4.1 Model Configuration and Evaluation Overview

Following the correlation analysis (Section 3.3), which confirmed strong ECa-texture relationships, Random Forest models were established, by training and testing, to predict soil texture across the entire field from ECa measurements. The objectives were to: (1) evaluate prediction accuracy across different model configurations, (2) compare *classification* vs. *regression* approaches, (3) assess the impact of sampling design on prediction performance, and (4) identify optimal model parameters for this field, substantiating a general methodology for prediction of soil texture or other spatial properties. Experimental design: A systematic benchmarking procedure evaluated 1,120 model configurations per sampling design per depth (Section 2.4.4), varying:

- **Response variable**: Classification (texture class) vs. Regression (D index)
- **Clay/silt cutoff**: $\{2,3,4,5,6,7,8\}$ µm (7 values)
- **Variance filter** (q): $\{10,100,200,500\}$ (4 values)
- **Number of trees**: $\{20,50\}$ (2 values)

- **Initialization seed**: $\{123, 234, ..., 102\}$ (10 values)

Total models evaluated: 1,120 configurations $\times$ 5 design-depth combinations (*QC-A, QC-B, Grid-A, Grid-B, QC$_{var}$-A*) $\approx$ **5,600 model runs**.

Each model generated field-scale predictions at 1×1 m resolution ($n \approx 62,000$ grid cells) covering the entire 6.5 ha study area, and validated using independent samples from a different design (cross-design validation, Section 2.4.4):

- *QC* models validated on *Grid* samples ($n = 17$)
- *Grid* models validated on *QC* samples ($n = 17$)
- *QCvar* models validated on *Grid* samples ($n = 17$)

The evaluation accounts for the following accuracy metric:

$$Classification\ accuracy\ (\%) = \frac{correctly\ predicted\ classes}{number\ of\ validation\ points} \times 100 \tag{3.1}$$

3.4.2 Overall Model Performance

Validation accuracy showed substantial variability across the 5,600 model runs (Figure 12), with range spanning 0% (complete failure) to 76% (best performance), median of 36%, and mean accuracy of $37\% \pm 14\%$. The wide distribution ($0 - 76\%$) reflects the strong influence of parameter choices on model performance, highlighting the importance of systematic evaluation. Figure 12 presents accuracy distributions for each design-depth combination.

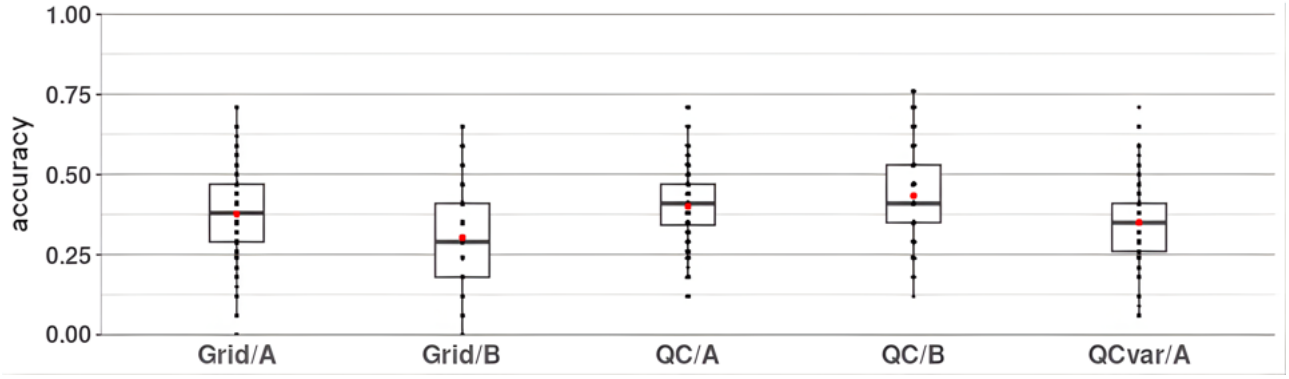

Figure 12: Distribution of validation accuracy across all model configurations by sampling design and horizon. Box plots show median (line), interquartile range (box), mean (red dot), and outliers. $N = 1,600$ configurations per design-depth combination (varying response variable, clay/silt cutoff, variance filter, trees, seed). Notably, QC/Horizon B achieved the highest maximum accuracy (76%).

QC Horizon B achieved the best overall performance, reflecting strong ECa-texture correlations ($R = 0.74 - 0.81$, Table 3), with median 41%, mean 43% and maximum accuracy of 76%. ***QC Horizon A***, the second-best overall, with median accuracy 41%, mean 40%, reaching a maximum of 71%. ***Grid Horizon B*** displayed a moderate performance, reflecting weaker correlations than *QC* ($R = 0.63$ vs 0.74), with median accuracy 29%, mean 30% and maximum of 65%, yielding the poorest performance,

with 40% less accuracy on average compared to *QC Horizon B* (30% and 43%, respectively), and lowest maximal accuracy overall. ***Grid Horizon A*** performed surprisingly quite well, with a median and mean accuracies of 38%, and 71% at maximum (despite $R = 0.26$ correlations). ***QC$_{var}$ Horizon A*** reached a median and mean accuracies of 35%, with maximal accuracy of 71%, similar to *QC-A*.

Key findings suggest that depth and sampling design strongly influenced accuracy, with *QC-B* outperforming all others, consistent with correlation patterns (Section 3.3). Since most configurations yielded low-moderate quality predictions, we opted to focus on the maximal attained accuracy.

3.4.3 Classification vs. Regression Approach Comparison

Two modeling approaches were systematically compared (Figure 13): (1) ***classification*** – direct prediction of USDA texture class from ECa, and (2) ***regression*** – prediction of D index from ECa, then conversion to texture class via calibration curves (Section 2.3).

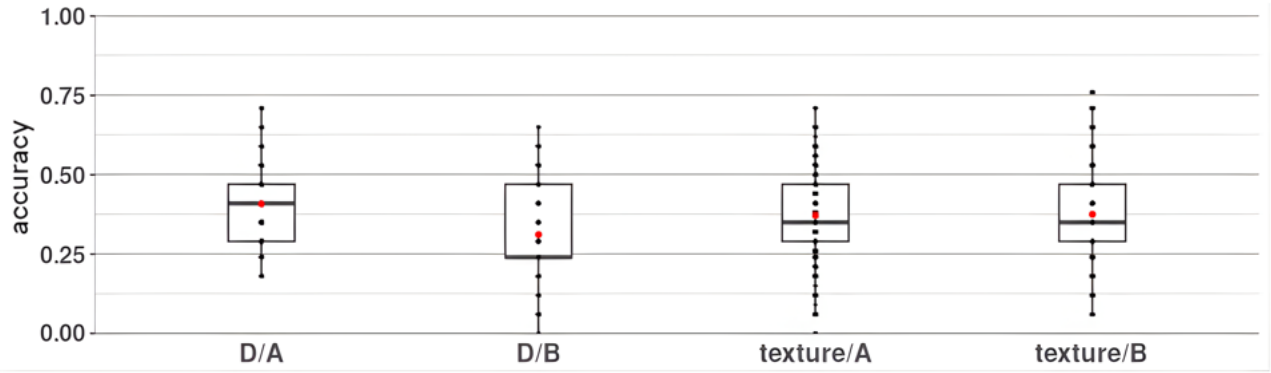

Figure 13: Comparison of classification vs. regression approaches across all sampling designs and horizons. Box plots show accuracy distribution for models using (left) regression on D index with conversion to texture classes vs. (right) direct classification of texture classes. $N \approx 1,400$ models per category. Notably, classification outperforms regression in Horizon B (76% vs 65% maximum) but both achieve similar performance in Horizon A (71% maximum for both).

Accuracy by horizon:

- **Horizon A** ($N = 3,360$ model runs):
  - Classification maximum: 71% (12/17 correct)
  - Regression maximum: 71% (12/17 correct)
  - Median accuracy: Classification = 35%, Regression = 41%

  As the overlapping interquartile ranges show, the two methods' performance does not differ significantly.

- **Horizon B** ($N = 2,240$ model runs):
  - Classification maximum: 76% (13/17 correct)
  - Regression maximum: 65% (11/17 correct)
  - Median accuracy: Classification = 35%, Regression = 24%

Table 5: Maximum validation accuracy by clay/silt cutoff, horizon, and response variable.

clay/silt cutoff	Horizon		Response variable	
µm	**A**	**B**	***D***	***Texture class***
2	71%	59%	71%	65%
3	71%	53%	71%	59%
4	47%	47%	47%	47%
5	71%	47%	59%	71%
6	71%	65%	65%	71%
7	65%	71%	65%	71%
8	53%	76%	65%	76%

In this horizon, classification achieved 17% higher accuracy at its peak compared to regression (76% vs 65%), but without significant difference as a whole.

Classification outperformed regression in Horizon B. This could stem from calibration uncertainty because *D*-to-*texture* conversion introduces error ($R^2 = 0.72 - 0.85$, Section 2.4.3), or due to class boundaries, as direct classification may better capture discrete texture class transitions. In Horizon A, approaches tied, likely because of moderate correlations ($R < 0.62$) limit both approaches – in this respect, *Grid-A* with low ECa-texture correlations (< 0.26), performed on the same scale as *QC-A*. Surface variability in this horizon might affect both approaches equally.

According to this perspective, the classification approach is recommended for high-correlation scenarios (Horizon B), while both approaches are viable when correlations are moderate (Horizon A).

3.4.4 Effect of Clay/Silt Cutoff on Prediction Accuracy

The clay/silt cutoff parameter (2-8 µm) showed horizon-dependent effects on accuracy (Table 4, Figure 14).

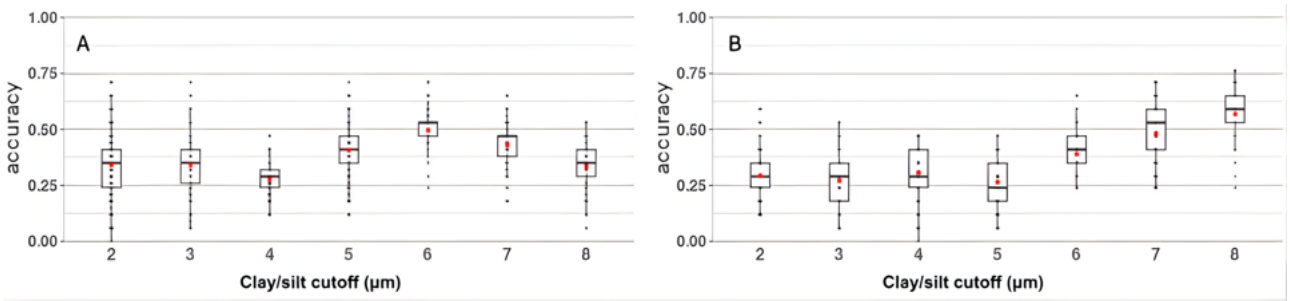

Figure 14: Validation accuracy for all samples by the clay/silt cutoff parameter in the A (left) and B (right) horizons.

Horizon A patterns show high variability with optimal cutoffs at 2-6 µm (multiple optima at 71% accuracy), while the poorest performance observed for 4 µm (47%), Horizon A displayed multiple local optima without a clear monotonic trend.

Horizon B presented optimal cutoff at 8 µm (76% accuracy for classification), exhibiting a general trend of increasing accuracy with larger cutoffs (2 µm: 59% $\rightarrow$ 8 µm: 76%), where both classification and regression show similar trends.

The 8 µm optimal cutoff for Horizon B aligns with correlation analysis (Section 3.3.5, Table 4), where correlations improved 13% from 2 to 8 µm. This supports the hypothesis that fine silt particles (2-5 µm) contribute to ECa signal similarly to clay, likely due to water holding capacity, which is reflected in ECa sensitivity to these fractions of particles with high specific surface area.

The lack of clear trend in Horizon A reflects weaker overall correlations – modest ECa-texture relationships ($R = 0.45 - 0.53$, Table 4) driven by greater noise – management effects and spatial variability. It is also possible that parameter interactions are at play, as cutoff size effects may be confounded with other parameters.

These findings implies that for similar alluvial soils measured with laser diffraction, 6-8 µm cutoff may be more appropriate than conventional 2 µm for ECa-based prediction, particularly in subsoil horizons.

3.4.5 Effect of Variance Filter on Training Dataset Size and Accuracy

The variance filter parameter q (Section 2.4.2) controls training dataset size by restricting samples to locations with low kriging variance. Four q values were tested: $\{10, 100, 200, 500\}$, corresponding to decreasing dataset sizes (6,200, 620, 310 and 124 cells). Figure 15 presents accuracy vs. the q parameter for both horizons.

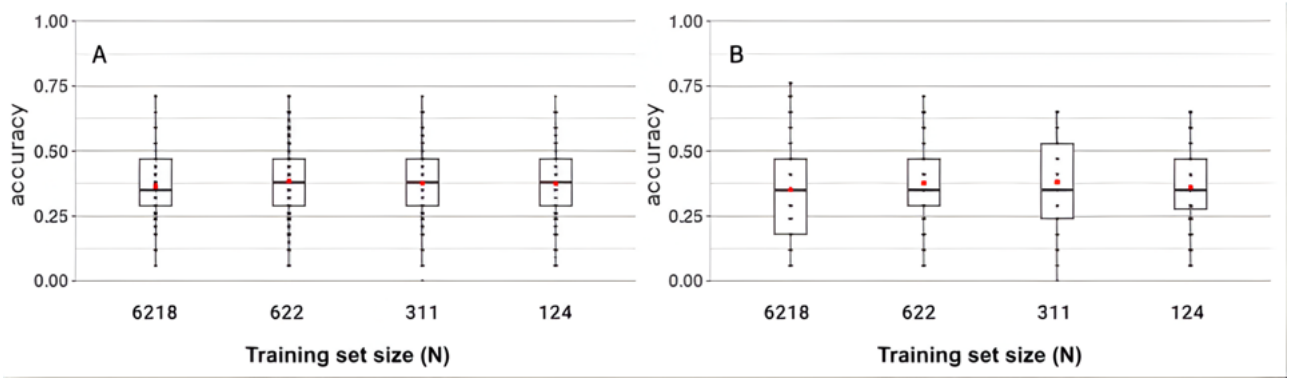

Figure 15: Validation accuracy for all samples by the q parameter (training-set size) in the A (left) and B (right) horizons. Horizon A shows no variation, while in Horizon B a trend of decreasing maximum values is apparent as training sets get smaller

Results demonstrate a clear positive trend in Horizon B (Figure 15, right panel) where accuracy increases with larger training datasets (smaller q, more data). While mean accuracies show minimal variation across q values (range: $35 - 38\%$), the maximum accuracy improved by 18% (from 65% for $q = 500$ to 76% for $q = 10$). It shows that larger training datasets are beneficial when correlations are strong ($R > 0.7$).

For Horizon A (Figure 15, left panel), accuracy remained relatively stable across q values, displaying high variability with large error bars at all q values. Apparently, when correlations are moderate ($R \approx 0.5$), data quality (variance filtering) doesn't clearly improve predictions; other factors dominate.

Additional Random Forest configuration parameters showed minimal systematic effect on performance. Varying the number of trees (20, 50) showed similar mean, median and maximal accuracy (35%, 37%, 76%, respectively). The 10 values of random initialization seed yielded mean accuracies with mean 37% and median 35%, while maximal accuracies of $> 71\%$ attained by all seeds, except one seed, which resulted in 65% accuracy.

These findings indicate that 20 trees are sufficient for this application, likely due to moderate feature count (4 ECa layers) and relatively simple ECa-texture relationships. The random seed initialization accuracies show that predictions are stable across random initializations, indicating robust feature-response relationships, sufficient training data and consistent model convergence. These null findings validate the experimental design by showing that the chosen primary factors (depth, design, cutoff) constitute real systematic effects, not artifacts of RF stochasticity. Moreover, modest computational resources (20 trees) are adequate for model fitting, and single model runs might be representative, although variability among seeds contributes to the diversity of results.

3.4.6 Optimal Model Configuration and Predictions

The best overall model (76% accuracy) was attained by the following classification model configuration: prediction of texture class with *QC* Sampling design at Horizon B (40-60 cm), clay/silt cutoff set to 8 µm, variance filter (q) set to 10 (minimal filtering, largest training dataset), with either 20 or 50 trees. Training data comprised ~6,200 filtered cells from *QC-B* design, with validation score of 13/17 Grid samples correctly predicted (76%). The predicted map (Figure 16) delineates the field into 5 soil texture classes, with *sandy clay loam* in the north and western area (~55%), *clay* in the southeast (~30%), while the rest of the study was tagged as *clay loam*, *sandy clay* and *sandy loam*. Finer clayey textures are concentrated in the southeastern high-ECa zone (Figure 6). Four soil samples were incorrectly classified (23% error rate): three *sandy clay loam* points were predicted as *clay loam* and one *clay loam* was predicted as *clay* in the transitional zone.

The best regression model (71% accuracy) was obtained by several configurations with the following common parameters: prediction of D by the *QC* sampling design data at Horizon A (0-20 cm), with clay/silt cutoff at 2-3 µm, variance filter (q) set to 100, which implies a training data size of 622 cells. The predicted regression maps (Figure 17) show the D index map (top panel), with a range of 0.51 to 0.67 and mean of 0.57 ± 0.04. The values of D are lowest in the southeast, moderate in the western part, and highest in the transitional zones. The converted texture map (bottom panel) consists of only two texture classes: *loam* (~65%) and *sandy loam* (~35%). This configuration classified 12/17 *Grid* samples correctly predicted (71%) and is one of the best performers in Horizon A.

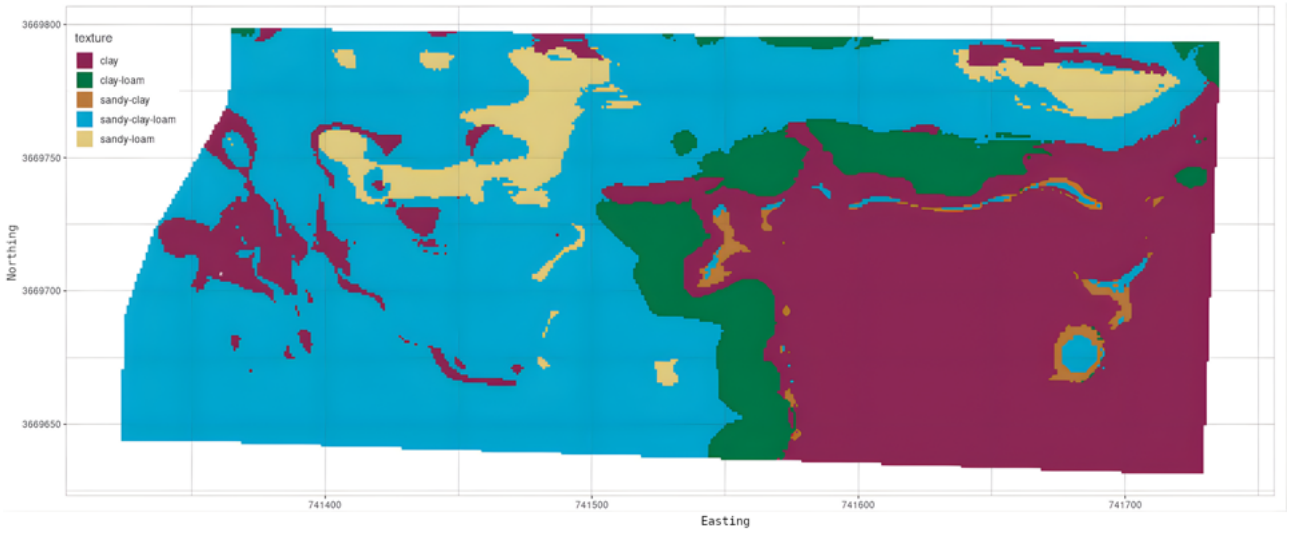

Figure 16: Predicted soil texture map with the highest overall accuracy (0.76): classification model by the QC sample at horizon B, clay/silt limit set to 8 µm. Hula Valley, 2023.

Table 6: Details of the best performing models for categorical and continuous approaches.

Metric	**Classification (QC-B)**	**Regression (QC-A)**
Accuracy	76%	71%
Horizon	B	A
Cutoff	8 µm	2 µm
ECa-texture R	0.74-0.81	0.51-0.62
Number of classes predicted	5	2
Dominant class	Sandy clay loam	Loam

Table 6 describes the configurations of the best classification and regression models. The top performing models consists of different horizons and approaches, suggesting that optimal strategy varies by depth – classification for strong-correlation subsoil, either approach for moderate-correlation soils.

3.4.7 Sampling Design Impact on Prediction Accuracy

The three tested sampling designs achieved moderate performance in general, with mean accuracies across all configurations of $41\% \pm 11\%$ (*QC*), $35\% \pm 14\%$ (*Grid*) and $35\% \pm 12\%$ (QC_{var}). However, the maximal attained accuracies with optimal configuration were higher:

- **QC**: 76% (Horizon B), 71% (Horizon A)
- **Grid**: 65% (Horizon B), 71% (Horizon A)
- **QCvar**: 71% (Horizon A only)

QC design advantages include stronger correlations captured (Section 3.3), with $R = 0.51 - 0.81$ vs. *Grid* $R = 0.26 - 0.63$; better feature-space coverage of ECa gradient; and higher maximum accuracy achieved (76% vs 71%).

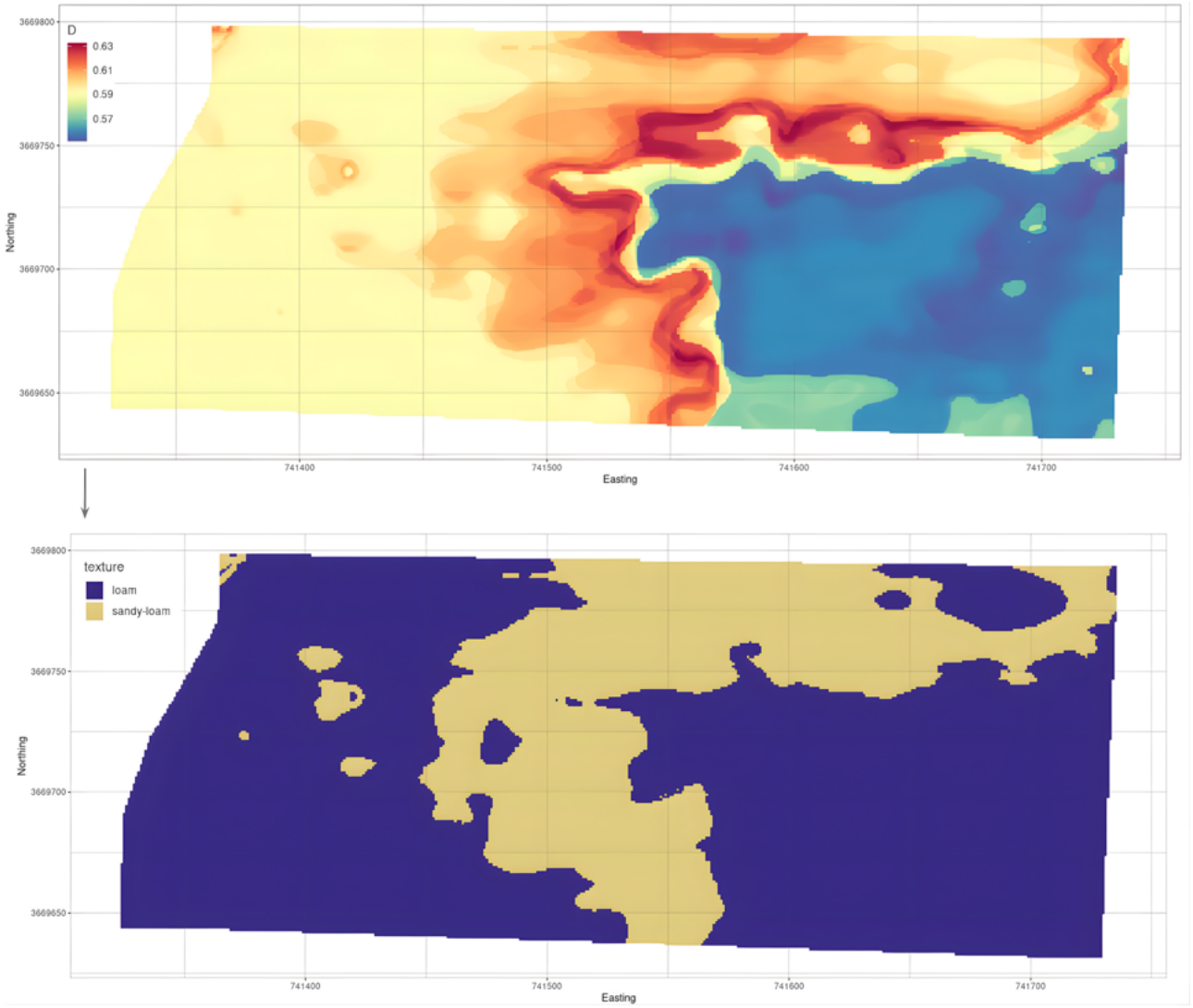

Figure 17: Predicted soil texture map with the highest attained accuracy for the regression approach (71%), by the QC sample at horizon A, clay/silt limit of 2 µm: The predicted variable D (top), and the corresponding texture map (bottom). Hula Valley, 2023.

The *Grid* design performance was characterized by moderate accuracy despite uniform spatial sampling. Correlation with *ECaH0.75* was significant in Horizon B ($R = 0.63$) but not in Horizon A ($R = 0.26$).

QC_{var} design performed similarly to *QC* in Horizon A (where both sampled), whereas variance filtering did not compromise accuracy, although the limited data (Horizon A only) prevented full comparison.

In this field experiment, sampling design influenced the accuracy of prediction and *QC* achieved the highest accuracy (76%), validating the optimization approach (Section 3.2). However, all designs achieved >60% accuracy with optimal configurations, suggesting that sample size ($n = 17$) and horizon matter more than design for this moderately heterogeneous field.

Discussion and Conclusions

This study developed and validated a novel quantile-cluster (QC) sampling design algorithm for ECa-based soil texture mapping. The procedure was tested in a field experiment, compared to a reference method, and prediction maps were evaluated through extensive parameters benchmarking.

The QC algorithm achieved 76% texture classification accuracy for Horizon B, outperforming traditional *Grid* sampling (65%) with identical sample size ($n = 17$), showing stronger ECa-texture relationships, with Pearson correlations up to $R = 0.74$ (clay~ECa) and $R = -0.81$ (sand~ECa). This advantage likely stems from QC's dual operation, which combines feature-space coverage and stratification of samples across ECa quantiles. This computational approach ensures equal representation of the full soil variability gradient, from sandy (low ECa) to clayey (high ECa) extremes, while *Grid* sampling, constrained by geometric regularity, may under-sample portions of the feature space (Brus and Heuvelink 2007). Spatial balance is achieved by the geographical clustering constraints with preference to sample centroids that prevents congestion of sampling points in certain areas. The addition of 10% close-pair points, as suggested by Lark and Marchant (2018), further improves short-range variability characterization for variogram modeling.

The observed ECa-texture correlations represent the upper range of reported relationships in the literature. For example, McCutcheon et al. (2006) reported weak Pearson correlation between clay and ECa ($R = 0.34$), but provided evidence for its temporal changes with fluctuations in soil moisture content. Perry et al. (2007) showed substantial variation in correlations, with Pearson coefficients ranging from 0.4 to over 0.8 in a semi-arid climate crop field. Siddique (2020) reported Pearson correlation of $0.66 - 0.78$ between clay and ECa, while higher values are common (e.g., $R = 0.84$; Hedley et al. 2004). A strong correlation was demonstrated by Rathnayaka et al. (2018) with an $R = 0.9$ for similar properties in a rice paddy. In this study, correlation strength was up to $R = 0.74$ (ECa-clay) and $R = -0.81$ (ECa-sand), both in Horizon B by the QC design. These strong correlations were likely influenced by the timing of near-field-capacity that minimizes effects of moisture variation, as well as the alluvial soil characteristics, in which texture is a dominant ECa driver.

Cousin et al. (2022) reviewed multiple studies in which the prediction of AWC components from ECa varied substantially with R^2 in the range of $0.20 - 0.80$. In this context, the accuracy obtained for texture class prediction in this study (76%) corresponds to the highest results and is comparable to an expert manual classification (70-85%, Vos et al. 2016), which are difficult to scale. Moreover, an accuracy of 76% can be considered as the maximum attainable given ECa-texture correlations of $R \approx 0.7 - 0.8$, and further emphasis should be put on improving signal-to-noise ratio, e.g., by timing

the survey to near field-capacity conditions (Corwin and Lesch 2013).

Wadoux et al. (2019) reported that the cLHS sampling method, which stratifies the feature-space into quantiles for maximum spectrum coverage (Minasny and Mcbratney 2006), was inefficient for generating digital soil maps with Random Forest, whereas spreading the sample uniformly in the space spanned by important covariates improves spatial prediction. Ma et al. (2020) found that feature-space coverage sampling (*FSCS*) method is more effective than *cLHS* for predicting soil classes based on multiple environmental features. However, these methods do not account for geographical coverage, hence points could be spatially clustered. In other studies (Lark and Marchant 2018; Wadoux, Marchant, and Lark 2019) it was found that using a scheme in which 10% of the sampling units are taken at short distances is a robust sampling strategy, which addresses short-distance variation – also known as the "nugget" effect – at little cost; Consequently, the sample balances between mapping the general area (coverage) with understanding fine-scale spatial correlation, leading to better overall geostatistical models (Lark and Marchant 2018). Our study extends these approaches by integrating ECa-based stratification with spatial dispersion constraints, the addition of variance filtering (QC_{var}) to enhance the quality of ECa values at sampling locations, coupled with a comprehensive comparison of parameters.

The variance-filtered *QC* (QC_{var}), which was conducted only in Horizon A, achieved similar accuracy to *QC* (71% both), despite restricting samples to low-variance areas (15^{th} percentile). This suggests that high-confidence zones with low kriging variance still capture sufficient variability for model training, and that variance filtering can reduce field sampling effort without accuracy loss (Figure 3). However, the lack of QC_{var} data in Horizon B limits full evaluation. Future studies should implement QC_{var} at both depths to assess whether variance filtering compromises precision in strong-correlation scenarios.

The cross-design validation approach (QC validated on Grid samples and vice versa) provides stringent assessment, as models must generalize across different spatial configurations. The 76% accuracy under this rigorous validation suggests that QC captures fundamental relationships (not spatial artifacts), as well as that ECa-texture patterns are spatially stationary within the field. Given the high variability and site-specific nature of soil properties, these models are not designed to be directly transferable; rather, they should be fitted and adjusted per-site, as outlined in this study, although model robustness can be aided by transfer learning between local sites (Viscarra Rossel et al. 2024).

The systematic clay/silt cutoff evaluation (2-8 µm range) identified 6-8 µm as the best fit for ECa-based prediction in Horizon B, with a monotonic trend of improvement across cutoff values, however, no preference was observed for Horizon A. For historical and operational reasons, the traditional USDA standard, which is based on Stokes' Law sedimentation, prescribes a clay/silt cutoff at 2 µm. Studies have shown (Beuselinck et al. 1998; Eshel et al. 2004) that LD-based methods yield different clay content values compared to sedimentation-based methods, depending on particle morphology, due

to different physical principles of interpretations. Fisher et al. (2017) reviewed studies reporting 4-9 µm as correspondent to 2 µm sedimentation clay, due to particle shape effects, since laser diffraction measures spherical equivalent diameter, while sedimentation measures hydraulic diameter. Since flatter clay particles settle slower, they may be classified as larger by sedimentation techniques. Rather than forcing LD to match sedimentation, we tuned the cutoff for the application (prediction by ECa) through systematic benchmarking. The obtained 6-8 µm threshold, in line with Crouvi et al. (2018) and Fisher et al. (2017) may represent the effective particle fractions controlling electrical conductivity. The high correlation between ECa and 2-8 µm size particles can be explained by their specific surface area, which is still large enough for CEC effects (Parry et al. 2011).

The high-resolution correlation analysis using 101 classes (Figure 11) reveals particle-scale relationships that are hidden in traditional 3 classes fraction analyses. The highest positive correlation was observed at 2.4 m ($R = 0.81$), not at *clay* ($< 2m$) but at *fine silt* boundary, suggesting that 2-3 µm particles disproportionately influence ECa. The strongest negative correlation was observed at 143 m ($R = -0.88$), at the *fine sand* fraction (125-250 µm), which appears to be more influential than *coarse sand*.

These analyses imply that the traditional clay/silt boundary (2 µm) may not be optimal for ECa-based classification using laser diffraction methods for PSD analysis, suggesting that *fine silt* (2-8 µm) should be grouped with *clay* for prediction with ECa. Moreover, focusing on specific sand fractions (e.g., 100-200 µm) could improve AWC models, as the high-resolution analysis identified specific particle sizes driving ECa response. Therefore, it is advised herein that when using laser diffraction for ECa-based texture mapping, the cutoff should be calibrated to maximize ECa-texture correlations rather than attempting to match sedimentation-based standards.

A central hypothesis of this study was that regression on the continuous D index (Martín, Rey, and Taguas 2004) might outperform direct classification of texture classes. The rationale was that D captures continuous distribution information that discrete classes may omit, as well as that regression can capture subtle gradients which classification forces into discrete bins. However, the benchmarking results showed that classification achieved higher maximum accuracy (76% vs. 71%) suggesting that multiple transformations in the regression approach introduce cumulative error: D calculation from PSD; prediction across the field; then D to clay ($R^2 = 0.72$) and sand ($R^2 = 0.85$) for texture classification. Furthermore, direct classification may better capture discrete boundaries inherent in the USDA system, and fundamentally, this is a classification problem by nature. The modest difference indicates that both approaches are viable, with choice depending on specific application, for example, continuous D for AWC modeling (Cousin et al. 2022) vs. discrete classes for management zones (Heil and Schmidhalter 2017).

To expand the scarce soil sampling, the ground-truth points were augmented by a common practice of interpolation (Dos Santos et al. 2025), then filtered by variance, a process that proved to yield decent

accuracy. However, with the fitted model, the training set points are also used for prediction to the entire field, including the training points themselves, a practice which violates standard train-test separation principles in Learning practices (Hastie, Tibshirani, and Friedman 2009). To address this concern, we conducted cross-design validation, using an independent dataset of the same size as the sampling design experiment ($n = 17$; Section 2.4). This small sample-size intrinsically leads to prediction errors, which our proposed probabilistic heuristic aims to minimize. Despite the validation limitations inherent to small sample sizes, the consistency of results in multiple parameter configurations and the agreement between correlation strength and prediction accuracy (Sections 3.3-3.4) support the robustness of our findings.

Limitations of QC_{var}'s reliance on interpolated points. The QC_{var} procedure restricts candidate locations to those interpolated points whose ECa kriging variance falls below a chosen threshold. While this restriction yields a candidate set on which the kriging surrogate is locally reliable, it carries three inherent limitations that are worth making explicit. First, candidates with low kriging variance are by construction spatially close to existing measurements, and are therefore highly correlated with the observed samples; they contribute less truly novel information than candidates drawn from under-explored regions would. Second, because QC_{var}'s candidate set is confined to the well-instrumented portion of the domain, the procedure is intrinsically *exploitative*: it refines the sampling design within the already-sampled neighborhood, but does not expand spatial coverage into regions that were sparsely covered by the ECa survey. Third, since the RF / surrogate model (Section 2.4.1) is trained partly on values produced by ordinary kriging, it inherits some of the variogram-based smoothness assumptions of the kriging model, rather than learning purely from independent measurements.

This study is single-site, 6.5 ha irrigated crop field in Hula Valley, Israel, with alluvial soils on a former seasonal marshland in a semi-arid climate. The ECa survey captured specific conditions (late April, near field capacity), yet ECa patterns are mostly temporally stable (Gonçalves et al. 2025), however, ECa-texture correlations depend on moisture conditions, salinity, and OM. It should be noted that the optimal clay/silt cutoff (6-8 μm) may be site- or sensor-specific. For the same reasons, *QC* advantage over *Grid* may vary with field heterogeneity.

Since water consumption is a major farming consideration in water-scarce regions (Nikolaou et al. 2020), the proposed procedure can be cost-effective as a basis for precision irrigation practice. Soil texture maps can be converted through pedotransfer functions into AWC maps, which can then drive management zones delineation. The *QC* algorithm and prediction steps are scalable, using open-source software (Section 2.5; Appendix B) implemented using commodity hardware. The web-based application enables practitioners without programming experience to apply the methodology, facilitating wider adoption in precision agriculture.

Although variability would arise under different settings, the proposed approach of per-site model fitting and parameter tuning (such as clay/silt cutoff) could be easily adapted. Regarding laser

diffraction specificity, the results are specifically tied to *Malvern Mastersizer 3000*, therefore, cutoff optimization may differ for other LD instruments. Similarly, it should be mentioned that ECa measurements are *EM38-MK2* specific results. Under the assumption that salinity is not an issue in the study plot, we have not measured salinity parameters which is highly recommended for future applications. Furthermore, if mapping for AWC, it is advised to analyze bulk density and OM in addition to texture.

Despite these limitations, this study provides rigorous cross-validation by independent designs, with comprehensive parameter evaluation (i.e., benchmarking over 5,600 models), transparent reporting of all constraints, and generalizable methodology even if specific parameters may vary. Combining ECa-based water content mapping with salinity assessment (Autovino et al. 2025) might be a valuable extension of our methodology for salt-affected agricultural systems.

Future work should address these limitations through: (1) multi-site validation across diverse soils and climates to assess methodology transferability; (2) larger validation sets ($n > 30$) for adequate statistical power to detect design differences; (3) direct AWC measurements by pressure plate or time domain reflectometry (TDR) to validate which cutoff yields best water retention prediction; (4) sensor comparisons with multiple ECa instruments and remote sensing satellite imagery, to explore data fusion approaches; and (5) integration with crop models to demonstrate irrigation scheduling improvements. These extensions would further validate the transferability of the methodology and allow optimization for specific crops, soil types, and management objectives.

This study demonstrated that strategic integration of algorithmically-guided sampling design (*QC* algorithm), high-resolution PSD analysis (101 size classes), systematic parameter optimization (5,600 model configurations), and rigorous cross-validation can achieve state-of-the-art accuracy (76%) for ECa-based texture mapping at practical cost. The generalizable framework (i.e., tunable parametric settings) represents the key contribution, enabling site-specific model development for precision agriculture applications worldwide. As water scarcity intensifies globally, such tools for efficient soil characterization become increasingly critical for sustainable agricultural intensification.

Bibliography

Al-agele, H. A. et al. (2021). "A Variable Rate Drip Irrigation Prototype for Precision Irrigation". In: *Agronomy* 11.12, p. 2493. DOI: 10.3390/agronomy11122493.

Allen, J. R. L. and D. M. Thornley (2004). "Laser granulometry of Holocene estuarine silts: effects of hydrogen peroxide treatment". In: *The Holocene* 14.2, pp. 290–295. DOI: 10.1191/0959683604hl681rr.

Amemiya, M. (1965). "The Influence of Aggregate Size on Soil Moisture Content-Capillary Conductivity Relations". In: *Soil Science Society of America Journal* 29, pp. 744–748. DOI: 10.2136/sssaj1965.03615995002900060039x.

Arya, L. M. and J. F. Paris (1981). "A Physicoempirical Model to Predict the Soil Moisture Characteristic from Particle-Size Distribution and Bulk Density Data". In: *Soil Science Society of America Journal* 45, pp. 1023–1030. DOI: 10.2136/sssaj1981.03615995004500060004x.

Autovino, D. et al. (2025). "An in-situ methodology to separate the contribution of soil water content and salinity to EMI-based soil electrical conductivity". In: *EGUsphere* 2025, pp. 1–30. DOI: 10.5194/egusphere-2025-2696.

Beuselinck, Laurent et al. (1998). "Grain-size analysis by laser diffractometry: comparison with the sieve-pipette method". In: *Catena* 32.3-4, pp. 193–208.

Bezdek, J. C. (1973). "Cluster Validity with Fuzzy Sets". In: *Journal of Cybernetics* 3.3, pp. 58–73. DOI: 10.1080/01969727308546047.

— (1975). "Mathematical models for systematics and taxonomy". In: *Proceedings of the 8th International Conference on Numerical Taxonomy*. Freeman.

Bhadha, J. et al. (2017). "Raising Soil Organic Matter Content to Improve Water Holding Capacity". In: *EDIS*. DOI: 10.32473/edis-ss661-2017.

Biswas, A. and Y. Zhang (2018). "Sampling Designs for Validating Digital Soil Maps: A Review". In: *Pedosphere* 28.1, pp. 1–15. DOI: 10.1016/S1002-0160(18)60001-3.

Blaschek, M. et al. (2019). "Prediction of soil available water-holding capacity from visible near-infrared reflectance spectra". In: *Scientific Reports* 9, p. 12833. DOI: 10.1038/s41598-019-49226-6.

Bouyoucos, G. J. (1962). "Hydrometer Method Improved for Making Particle Size Analyses of Soils". In: *Agronomy Journal* 54.5, pp. 464–465. DOI: 10.2134/agronj1962.00021962005400050028x.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324.

Bruand, A. and D. Tessier (2000). “Water retention properties of the clay in soils developed on clayey sediments: significance of parent material and soil history”. In: *European Journal of Soil Science* 51, pp. 679–688. DOI: 10.1111/j.1365-2389.2000.00338.x.

Brus, D. J. (2019). “Sampling for digital soil mapping: A tutorial supported by R scripts”. In: *Geoderma* 338, pp. 464–480. DOI: 10.1016/j.geoderma.2018.07.036.

Brus, D. J. and G. B. M. Heuvelink (2007). “Optimization of sample patterns for universal kriging of environmental variables”. In: *Geoderma* 138.1-2, pp. 86–95. DOI: 10.1016/j.geoderma.2006.10.016.

Brus, D. J., B. Kempen, and G. B. M. Heuvelink (2011). “Sampling for validation of digital soil maps”. In: *European Journal of Soil Science* 62, pp. 394–407. DOI: 10.1111/j.1365-2389.2011.01364.x.

Caliński, T. and J. Harabasz (1974). “A dendrite method for cluster analysis”. In: *Communications in Statistics* 3.1, pp. 1–27. DOI: 10.1080/03610927408827101.

Cámara, J., A. Lázaro-López, and V. Gómez-Miguel (2016). “Introducing heterogeneity measurements in terroir studies. Application in the Região Demarcada do Douro (N Portugal)”. In: *The 11th International Terroir Congress*. Linfield college. McMinnville, Oregon.

Chagas, C. et al. (2016). “Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions”. In: *Catena* 139, pp. 232–240. DOI: 10.1016/j.catena.2016.01.001.

Chang, W. et al. (2024). *shiny: Web Application Framework for R*. URL: https://shiny.posit.co.

Chen, S. et al. (2022). “Digital mapping of GlobalSoilMap soil properties at a broad scale: a review”. In: *Geoderma* 409, p. 115567. DOI: 10.1016/j.geoderma.2021.115567.

Clay, S. (2001). “Factors influencing spatial variability of soil apparent electrical conductivity”. In: *Communications in Soil Science and Plant Analysis* 32, pp. 2993–3007. DOI: 10.1081/CSS-120001102.

Corwin, D. L. and S. M. Lesch (2013). “Protocols and guidelines for field-scale measurement of soil salinity distribution with ECa-directed soil sampling”. In: *Journal of Environmental and Engineering Geophysics* 18.1, pp. 1–25. DOI: 10.2113/jeeg18.1.1.

Corwin, D. L. and E. Scudiero (2019). “Mapping soil spatial variability with apparent soil electrical conductivity (ECa) directed soil sampling”. In: *Soil Science Society of America Journal* 83.1, pp. 3–4. DOI: 10.2136/sssaj2018.06.0228.

Cousin, I. et al. (2022). “Available water capacity from a multidisciplinary and multiscale viewpoint. A review”. In: *Agronomy for Sustainable Development* 42, p. 46. DOI: 10.1007/s13593-022-00774-8.

Crouvi, O. et al. (2018). “Middle to late Pleistocene shift in eolian silts contribution into Mediterranean soils at the fringe of the Negev loess, Israel”. In: *Quaternary Science Reviews* 191, pp. 101–117. DOI: 10.1016/j.quascirev.2018.04.030.

De Feudis, C., C. Ferré, and R. Comolli (2025). “Practical insights for ECa-based soil mapping: Case studies in croplands and vineyards”. In: *Smart Agricultural Technology* 10. DOI: 10.1016/j.atech.2024.100697.

Dharumarajan, S. and R. Hegde (2022). “Digital mapping of soil texture classes using Random Forest classification algorithm”. In: *Soil Use and Management* 38, pp. 135–149. DOI: 10.1111/sum.12668.

Ding, J. et al. (2020). “Using Apparent Electrical Conductivity as Indicator for Investigating Potential Spatial Variation of Soil Salinity across Seven Oases along Tarim River in Southern Xinjiang, China”. In: *Remote Sensing* 12.16, p. 2601. DOI: 10.3390/rs12162601.

Domsch, H. and A. Giebel (2004). “Estimation of Soil Textural Features from Soil Electrical Conductivity Recorded Using the EM38”. In: *Precision Agriculture* 5, pp. 389–409. DOI: 10.1023/B:PRAG.0000040807.18932.80.

Dos Santos, A. P. et al. (2025). “Random forest algorithm applied to model soil textural classification in a river basin”. In: *Environmental Monitoring and Assessment* 197.3, p. 330. DOI: 10.1007/s10661-025-13786-0.

Eshel, G. et al. (2004). “Critical evaluation of the use of laser diffraction for particle-size distribution analysis”. In: *Soil Science Society of America Journal* 68.3, pp. 736–743. DOI: 10.2136/sssaj2004.7360.

Evett, S. and G. Parkin (2005). “Advances in Soil Water Content Sensing: The Continuing Maturation of Technology and Theory”. In: *Vadose Zone Journal* 4. DOI: 10.2136/vzj2005.0099.

Fisher, P. et al. (2017). “Adequacy of laser diffraction for soil particle size analysis”. In: *PLOS ONE* 12.5, e0176510. DOI: 10.1371/journal.pone.0176510.

Fortes, R., S. Millán, M. H. Prieto, et al. (2015). “A methodology based on apparent electrical conductivity and guided soil samples to improve irrigation zoning”. In: *Precision Agriculture* 16, pp. 441–454. DOI: 10.1007/s11119-015-9388-7.

Friedman, S. P. (2005). “Soil properties influencing apparent electrical conductivity: a review”. In: *Computers and Electronics in Agriculture* 46.1-3, pp. 45–70. DOI: 10.1016/j.compag.2004.11.001.

Fukuyama, Y. (1989). “A new method of choosing the number of clusters for the fuzzy c-mean method”. In: *Proceedings of the 5th Fuzzy Systems Symposium*, pp. 247–250.

García-Gaines, R. A. and S. Frankenstein (2015). *USCS and the USDA soil classification system: Development of a mapping scheme*. Tech. rep. Vicksburg, MS, USA: U.S. Army Engineer Research and Development Center.

García-Gutiérrez, C., Y. Pachepsky, and M. Á. Martín (2018). "Technical note: Saturated hydraulic conductivity and textural heterogeneity of soils". In: *Hydrology and Earth System Sciences* 22, pp. 3923–3932. DOI: 10.5194/hess-22-3923-2018.

Gavlak, Ray et al. (2003). "Soil, plant and water reference methods for the western region". In: *WCC-103 Publication, Fort Collins, CO*, pp. 1–207.

Ghodgaonkar, A. et al. (2025). "Realizing low-energy drip irrigation via a 1-dimensional model of low-pressure drip emitters". In: *Scientific Reports* 15.1, p. 41063. DOI: 10.1038/s41598-025-24988-4.

Glass, H. J. (2003). "Method for assessing quality of the variogram". In: *Journal of the Southern African Institute of Mining and Metallurgy* 103.1, pp. 43–51.

Gonçalves, L. A. et al. (2025). "Spatial and temporal variability of soil apparent electrical conductivity". In: *Precision Agriculture* 26, p. 10. DOI: 10.1007/s11119-024-10209-x.

Gorączko, A. and S. Topoliński (2020). "Particle Size Distribution of Natural Clayey Soils: A Discussion on the Use of Laser Diffraction Analysis (LDA)". In: *Geosciences* 10, p. 55. DOI: 10.3390/geosciences10020055.

Gozdowski, D. et al. (2015). "Prediction accuracy of selected spatial interpolation methods for soil texture at farm field scale". In: *Journal of Soil Science and Plant Nutrition* 15, pp. 639–650. DOI: 10.4067/S0718-95162015005000033.

Gundim, A. da S. et al. (2023). "Precision irrigation trends and perspectives: a review". In: *Ciência Rural* 53.8, e20220155. DOI: 10.1590/0103-8478cr20220155.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer Series in Statistics. Springer, pp. 509–510. DOI: 10.1007/978-0-387-84858-7.

He, W. et al. (2024). "Digital Mapping of Soil Particle Size Fractions in the Loess Plateau, China, Using Environmental Variables and Multivariate Random Forest". In: *Remote Sensing* 16.5, p. 785. DOI: 10.3390/rs16050785.

Hedley, C. B. et al. (2004). "Rapid identification of soil textural and management zones using electromagnetic induction sensing of soils". In: *Australian Journal of Soil Research* 42, pp. 389–400. DOI: 10.1071/SR03149.

Heil, K. and U. Schmidhalter (2017). "The Application of EM38: Determination of Soil Parameters, Selection of Soil Sampling Points and Use in Agriculture and Archaeology". In: *Sensors* 17.11, p. 2540. DOI: 10.3390/s17112540.

Heiniger, R. W., R. G. McBride, and D. E. Clay (2003). "Using soil electrical conductivity to improve nutrient management". In: *Agronomy Journal* 95.3, pp. 508–519.

Heiri, O., A. F. Lotter, and G. Lemcke (2001). “Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results”. In: *Journal of Paleolimnology* 25, pp. 101–110. DOI: 10.1023/A:1008119611481.

Henares, S., M. E. Donselaar, and L. Caracciolo (2020). “Depositional controls on sediment properties in dryland rivers: Influence on near-surface diagenesis”. In: *Earth-Science Reviews* 208. DOI: 10.1016/j.earscirev.2020.103297.

Hengl, T. et al. (2018). “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables”. In: *PeerJ* 6:e5518. DOI: https://doi.org/10.7717/peerj.5518.

Hirotatsu, M., T. Yusuke, and I. Toshiyuki (2015). “Origin of the soil texture classification system used in Japan”. In: *Soil Science and Plant Nutrition* 61.4, pp. 688–697. DOI: 10.1080/00380768.2014.998594.

Israeli, A. et al. (2019). “Statistical learning in soil sampling design aided by pareto optimization”. In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO ’19)*. New York, NY, USA: ACM, pp. 1198–1205. DOI: 10.1145/3321707.3321809.

Jiang, Q. et al. (2019). “Characterising dryland salinity in three dimensions”. In: *Science of the Total Environment* 682, pp. 190–199.

Kalumba, M. et al. (2022). “Machine Learning Techniques for Estimating Hydraulic Properties of the Topsoil across the Zambezi River Basin”. In: *Land* 11.4, p. 591. DOI: 10.3390/land11040591.

Kelley, J. et al. (2017). “Mapping Soil Texture by Electromagnetic Induction: A Case for Regional Data Coordination”. In: *Soil Science Society of America Journal* 81, pp. 923–931. DOI: 10.2136/sssaj2016.12.0432.

Kerry, R. and M. A. Oliver (2007). “Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood”. In: *Geoderma* 140.4, pp. 383–396. DOI: 10.1016/j.geoderma.2007.04.019.

Koncagül, E., R. Connor, and V. Abete (2024). *The United Nations World Water Development Report 2024: water for prosperity and peace; facts, figures and action examples*. Tech. rep. UNESCO World Water Assessment Programme. URL: https://unesdoc.unesco.org/ark:/48223/pf0000388952.

Krause, Andreas, Ajit Singh, and Carlos Guestrin (2008). “Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies”. In: *Journal of Machine Learning Research* 9, pp. 235–284.

Kuhn, Max (2008). “Building Predictive Models in R Using the caret Package”. In: *Journal of Statistical Software* 28.5, pp. 1–26. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Kullback, S. and R. A. Leibler (1951). “On information and sufficiency”. In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86.

Lakhiar, I. A. et al. (2024). “A Review of Precision Irrigation Water-Saving Technology under Changing Climate for Enhancing Water Use Efficiency, Crop Yield, and Environmental Footprints”. In: *Agriculture* 14.7, p. 1141. DOI: 10.3390/agriculture14071141.

Lark, R. and B. Marchant (2018). “How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters?” In: *Geoderma* 319, pp. 89–99.

Liaw, A. and M. Wiener (2002). “Classification and Regression by randomForest”. In: *R News* 2.3, pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

Loiseau, T. et al. (2021). “Density of soil observations in digital soil mapping: A study in the Mayenne region, France”. In: *Geoderma Regional* 24, e00358. DOI: 10.1016/j.geodrs.2021.e00358.

Lück, E. et al. (2009). “Electrical conductivity mapping for precision farming”. In: *Near Surface Geophysics* 7, pp. 15–25. DOI: 10.3997/1873-0604.2008031.

Ma, T. et al. (2020). “Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps”. In: *Geoderma* 370, p. 114366. DOI: 10.1016/j.geoderma.2020.114366.

Makó, A. et al. (2017). “Pedotransfer functions for converting laser diffraction particle-size data to conventional values”. In: *European Journal of Soil Science* 68. DOI: 10.1111/ejss.12456.

Martín, M. A., J. M. Rey, and F. J. Taguas (2001). “An entropy-based parametrization of soil texture via fractal modelling of particle-size distribution”. In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 457, pp. 937–947.

— (2004). “An entropy-based heterogeneity index for mass-size distributions in Earth science”. In: *Ecological Modelling* 182, pp. 221–228. DOI: 10.1016/j.ecolmodel.2004.04.002.

Martinez, G. et al. (2009). “Can Apparent Electrical Conductivity Improve the Spatial Characterization of Soil Organic Carbon?” In: *Vadose Zone Journal* 8, pp. 586–593. DOI: 10.2136/vzj2008.0123.

Massad Cartography of Tel-Hai College (1942). *Es-Salihiya, 1:20,000*. Survey of Palestine, Jaffa. URL: https://maps.telhai.ac.il.

Matsumoto, Makoto and Takuji Nishimura (1998). “Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator”. In: *ACM Transactions on Modeling and Computer Simulation* 8.1, pp. 3–30.

McCutcheon, M. C. et al. (2006). “Effect of Soil Water on Apparent Soil Electrical Conductivity and Texture Relationships in a Dryland Field”. In: *Biosystems Engineering* 94.1, pp. 19–32. DOI: 10.1016/j.biosystemseng.2006.01.002.

Mgohele, R. N. et al. (2024). “Prediction of soil texture using remote sensing data. A systematic review”. In: *Frontiers in Remote Sensing* 5. DOI: 10.3389/frsen.2024.1461537.

Michael-Mertens, F., S. Pätzold, and G. Welp (2008). "Spatial heterogeneity of soil properties and its mapping with apparent electrical conductivity". In: *Journal of Plant Nutrition and Soil Science* 171, pp. 146–154. DOI: 10.1002/jpln.200625130.

Minasny, B. and A. Mcbratney (2006). "A Conditioned Latin Hypercube Method for Sampling in the Presence of Ancillary Information". In: *Computers & Geosciences* 32, pp. 1378–1388. DOI: 10.1016/j.cageo.2005.12.009.

Moreno-Maroto, J. M. and J. Alonso-Azcárate (2022). "Evaluation of the USDA soil texture triangle through Atterberg limits and an alternative classification system". In: *Applied Clay Science* 229, p. 106689. DOI: 10.1016/j.clay.2022.106689.

Nemhauser, George L., Laurence A. Wolsey, and Marshall L. Fisher (1978). "An analysis of approximations for maximizing submodular set functions—I". In: *Mathematical Programming* 14.1, pp. 265–294.

Nikolaou, G. et al. (2020). "Implementing Sustainable Irrigation in Water-Scarce Regions under the Impact of Climate Change". In: *Agronomy* 10.8, p. 1120. DOI: 10.3390/agronomy10081120.

Olmstead, L. B., L. T. Alexander, and H. E. Middleton (1930). *A pipette method of mechanical analysis of soils based on improved dispersion procedure*. Technical bulletin. Washington D.C.: U.S. Dept. of Agriculture.

Pace, L. et al. (2024). "Soil Mapping of Small Fields with Limited Number of Samples by Coupling EMI and NIR Spectroscopy". In: *Soil Systems* 8.4, p. 128. DOI: 10.3390/soilsystems8040128.

Parry, S. A. et al. (2011). "Is silt the most influential soil grain size fraction?" In: *Applied Geochemistry* 26, S119–S122. DOI: 10.1016/j.apgeochem.2011.03.045.

Pebesma, E. J. (2004). "Multivariable geostatistics in S: the gstat package". In: *Computers & Geosciences* 30, pp. 683–691. DOI: 10.1016/j.cageo.2004.03.012.

Perry, C. et al. (2007). "Use of VERIS Soil EC Sensor for Mapping Soil Texture in Georgia Cotton Fields". In: *Beltwide Cotton Conferences*. New Orleans, Louisiana. URL: https://www.cotton.org/beltwide/proceedings/2005-2022/data/conferences/2007/papers/6498.pdf.

QGIS Development Team (2023). *QGIS Geographic Information System*. QGIS Association. URL: https://www.qgis.org.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rai, R. K., V. P. Singh, and A. Upadhyay (2017). "Soil Analysis". In: *Planning and Evaluation of Irrigation Projects*. Academic Press. Chap. 17, pp. 505–523. DOI: 10.1016/B978-0-12-811748-4.00017-0.

Rathnayaka, R. A., U. Vitharana, and W. Balasooriya (2018). "Detailed mapping of soil texture of a paddy growing soil using multivariate geostatistical approaches". In: *Tropical Agricultural Research* 29, p. 300. DOI: 10.4038/tar.v29i4.8257.

Ray, S. and S. Majumder (2024). “Water management in agriculture: Innovations for efficient irrigation”. In: *Modern Agronomy*, pp. 169–185.

Rodrigues, H. et al. (2024). “Remote and Proximal Sensors Data Fusion: Digital Twins in Irrigation Management Zoning”. In: *Sensors* 24.17, p. 5742. DOI: 10.3390/s24175742.

Saurette, D. D. et al. (2024). “Sample Size Optimization for Digital Soil Mapping: An Empirical Example”. In: *Land* 13.3, p. 365. DOI: 10.3390/land13030365.

Saxton, K. E. and W. J. Rawls (2006). “Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions”. In: *Soil Science Society of America Journal* 70, pp. 1569–1578. DOI: 10.2136/sssaj2005.0117.

Schmidinger, J., V. Barkov, et al. (2024). “Which and how many soil sensors are ideal to predict key soil properties: A case study with seven sensors”. In: *Geoderma* 450, p. 117017. DOI: 10.1016/j.geoderma.2024.117017.

Schmidinger, J., I. Schröter, et al. (2024). “Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming”. In: *Precision Agriculture* 25, pp. 1529–1555. DOI: 10.1007/s11119-024-10122-3.

Siddique, M. N. A. (2020). “Potential of soil sensor (EM38) measurements for soil fertility mapping”. In.

Staff, Soil Science Division (2017). “Soil survey manual”. In: *USDA Handbook* 18. Ed. by K. Scheffe C. Ditzler and H.C. Monger (eds.) URL: https://www.nrcs.usda.gov/resources/guides-and-instructions/soil-survey-manual.

Stępień, M. et al. (2015). “Assessment of soil texture class on agricultural fields using ECa, Amber NDVI, and topographic properties”. In: *Journal of Plant Nutrition and Soil Science* 178. DOI: 10.1002/jpln.201400570.

Svensson, D. N., I. Messing, and J. Barron (2022). “An investigation in laser diffraction soil particle size distribution analysis to obtain compatible results with sieve and pipette method”. In: *Soil and Tillage Research* 223. DOI: 10.1016/j.still.2022.105450.

Veihmeyer, F. J. and A. H. Hendrickson (1928). “Soil Moisture At Permanent Wilting Of Plants”. In: *Plant Physiology* 3.3, pp. 355–357. DOI: 10.1104/pp.3.3.355.

Violino, S. et al. (2023). “A data-driven bibliometric review on precision irrigation”. In: *Smart Agricultural Technology* 5, p. 100320. DOI: 10.1016/j.atech.2023.100320.

Viscarra Rossel, R. A. et al. (2024). “An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning”. In: *Earth-Science Reviews* 254, p. 104797. DOI: 10.1016/j.earscirev.2024.104797.

Vos, C. et al. (2016). “Field-based soil-texture estimates could replace laboratory analysis”. In: *Geoderma* 267, pp. 215–219. DOI: 10.1016/j.geoderma.2015.12.022.

Wadoux, A. M., D. J. Brus, and G. B. M. Heuvelink (2019). “Sampling design optimization for soil mapping with random forest”. In: *Geoderma* 355, p. 113913. DOI: 10.1016/j.geoderma.2019.113913.

Wadoux, A. M., B. P. Marchant, and R. M. Lark (2019). “Efficient sampling for geostatistical surveys”. In: *European Journal of Soil Science* 70, pp. 975–989. DOI: 10.1111/ejss.12797.

Webster, R. and M. A. Oliver (1992). “Sample adequately to estimate variograms of soil properties”. In: *Journal of Soil Science* 43, pp. 177–192. DOI: 10.1111/j.1365-2389.1992.tb00128.x.

Webster, Richard and M. Oliver (Jan. 2007). “Geostatistics for Environmental Scientists: Second Edition”. In: *Geostatistics for Environmental Scientists: Second Edition*. DOI: 10.1002/9780470517277.

Wösten, J. H. M., Y. A. Pachepsky, and W. J. Rawls (2001). “Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics”. In: *Journal of Hydrology* 251.3-4, pp. 123–150. DOI: 10.1016/S0022-1694(01)00464-4.

Xing, Y. and X. Wang (2024). “Precise application of water and fertilizer to crops: challenges and opportunities”. In: *Frontiers in Plant Science* 15. DOI: 10.3389/fpls.2024.1444560.

Zhang, Y. et al. (2022). “Comparison of sampling designs for calibrating digital soil maps at multiple depths”. In: *Pedosphere* 32, pp. 588–601. DOI: 10.1016/S1002-0160(21)60055-3.

Appendix A

Formulation of the Quantile-Cluster (QC) algorithm

Input:

Raster layers or a georeferenced table of ancillary data such as ECa in the area of interest.

n_{min}, n_{max} - minimal and maximal number of points to consider.

Computational steps:

1. A greedy search of optimal design for $n = \{n_{min}, n_{min}+1, ...n_{max}\}$ sampling points.
   - `k-means` - divide the geometric space into n clusters g.
   - Division of the feature space into n quantiles of one of the layers f.
   - [optional] - remove values outside the interquartile range (0.25-0.75).
   - Loop ($i = 1 \dots n$)
     - For each geometric cluster g_i - find an unsampled feature-space quantile f
     - If the cluster centroid is an unsampled quantile f - select the centroid
     - If centroid is already a sampled f - select the closest point to centroid in an unsampled f
   - Add 10% random points at close range to existing points (nested sampling)
2. Calculate statistical indices for the results of (1) - quantify the quality by sample-size.
   - For each sample of size n - calculate information metrics: D_{KL}, *cLHS*, *Mean distance to centroid*, *MPE, MSPE, MSNE*.

Output:

A sampling plan for each sample-size n, ranked by sample quality:

- A set of points of size n - characterized by coordinates $\{x, y\}$.
- Quantitative evaluation of sample's quality.

Appendix B

Web Applications Technical Specifications

This appendix provides detailed technical specifications for the two web applications developed to implement the methodology described in Sections 2.1–2.4. The applications enable users without programming expertise to apply the QC sampling design algorithm and perform ECa-based soil texture mapping.

Software Information:

- Framework: R Shiny (version 1.10)
- License: MIT
- Source code:

System Requirements:

For hosted applications (via web browser): a modern web browser, Internet connection, no local software installation required.
For local deployment: R $\geq$ 4.0.0; RStudio; Main R packages: `shiny`, `shinydashboard`, `ggplot2`, `plotly`, `caret`, `gstat`, `sf`, `tidyverse`; Recommended 8+ GB RAM for typical workflows.

B.1. *App-1* - Soil Sampling Design Tool

Demo: https://aist.shinyapps.io/qc-sampling
Source code: https://github.com/assafis/qc-sampling
Purpose: Streamline the workflow from raw ECa survey data to ready-to-go sampling plans.

Input requirements:

- **ECa survey data**: CSV or text files with columns: X, Y, ECa values (*Q0.5, Q1.0*). UTM coordinate system. E.g., the `.xyz` files produced by *RTmap38MK2* ECa survey software.
- **Field boundary**: Polygon shapefile (`.shp`) defining study area

Processing workflow:
Step 1: Data sources

- Project creation and metadata entry
- ECa file upload with automatic column detection

- Field boundary upload and coordinate system verification

Step 2: Data exploration and preprocessing

- Data compaction: Moving average smoothing with user-specified window size to reduce noise and improve performance (default: by a factor of 30)
- Interactive visualization of raw ECa distributions (histograms, spatial maps)
- Distribution assessment: normality tests, outlier detection
- Transformation options (log, left/right tail truncation) with real-time distribution preview

Step 3: Spatial interpolation

- Variogram modeling interface with manual or automatic fitting
- Kriging interpolation with grid resolution 1 m or 10 m
- Interpolation and kriging variance maps preview
- Export as raster files (`.tif`)

Step 4: Management zone delineation (optional)

- Clustering algorithms: fuzzy c-means
- Number of zones: User-specified – guided by cluster validity indices – an ensemble of metrics to guide selection of optimal zone count:
- **Partition Coefficient** (Bezdek 1973)
- **Partition Entropy** (Bezdek 1975)
- **Fukuyama-Sugeno** index (Fukuyama 1989)
- **Calinski-Harabasz** criterion (Caliński and Harabasz 1974)
- **Zones visualization** and **export**

Step 5: QC sampling design Algorithm parameters

- **Input ECa layer**: selection from ECa maps within the project
- **Sample size range**: minimum and maximum values (n_{min}, n_{max})
- **Stratification method**:
  - *Quantiles* - equal-frequency bins in feature space
  - *Clusters* - feature-space clustering approach
- **Initialization**: Deterministic (fixed seed, e.g., 123) or random (seed = current timestamp)
- **Variance filtering (QC_{var} mode)**:
  - Enable/disable variance-based filtering
  - q parameter (q^{th} percentile threshold)
  - Displays a map of training set points for each q value
- **Inter-quartile filtering**: exclude extreme values (optional)
- **Random point augmentation**: add ~10% points within a close range (up to $1/3$ of the minimal distance between centroids, see Section 2.2)

Outputs:

- **Sampling plans**: Point coordinates for each sample size in range $[n_{min}, n_{max}]$
- **Performance metrics** for each sample size (see Section 2.2.3)
- **Interactive visualization**:
  - Map view showing sampling points overlaid on ECa layer
  - Information metrics vs. sample size plots, enabling identification of inflection points
  - ECa coverage by quantile box plots
- **Export formats**:
  - **CSV** file with point coordinates (`.csv`)
  - **Raster file** (`.tif`) for GIS processing
  - **Summary statistics** table (`.csv`)

User workflow time: Processing from data upload to final sampling plan typically takes less than an hour for a 6 Ha field, depending mostly on the compaction factor that precedes the kriging operation.

Figure 18: Screenshot of the shiny web application app-1, a tool for soil sampling design by ECa data, displaying a section of QC sample design results for $n \in \{11-22\}$ sampling points.

B.2. *App-2*: Soil Texture Analysis and Prediction Tool

Demo: https://aist.shinyapps.io/st-map Source code: https://github.com/assafis/st-map

Purpose: Integrate PSD measurements with ECa data for spatial texture mapping

Input requirements:

This app uses projects created with *App-1* (ECa maps, field polygons; Appendix B.1) or externally generated raster layers (`GeoTIFF` files) as input for spatial modeling.

- **PSD data:** Laboratory results from Mastersizer – `CSV` format with sample ID, 101 size classes distribution and *D* index (Section 1.4)
- **Sample locations:** Coordinates matching PSD samples
- **ECa layers:** Interpolated surfaces
- **Field boundary:** Study area polygon

Processing workflow: In a common menu, users can select a dataset, sample, horizon and clay/silt limit (2-8 µm).

Step 1: PSD data analysis

- **Texture triangle visualization:** Interactive USDA triangle with sample points plotted
- **Distribution display:**
  - Histograms of clay, silt, sand fractions, and the *D* index
  - Full 101-class distributions for individual samples
  - Summary statistics by texture class
- **Entropy calculation:** Automatic computation of *Shannon H* and *balanced entropy D* indices

Step 2: ECa-PSD correlation analysis

- **Correlation matrices:**
  - ECa vs. traditional fractions (clay, silt, sand) and entropy indices (H, D)
  - ECa vs. all 101 size classes

Step 3: Spatial interpolation

- **Variable selection:** Individual fractions (clay, silt, sand), texture classes (categorical) or entropy indices (H, D)
- **Interpolation method:**
  - IDW
  - Ordinary kriging
- **Variance mapping:** kriging variance maps for uncertainty assessment
- **Preview:** Interactive map comparison

Step 4: Random Forest prediction

- **Model configuration:**
  - **Input layers:** selection from available ECa layers (*H 0.375m*, *H 0.75m*, *V 0.75m*, *V 1.5m*)
  - **Response variable:**
    - **Classification:** USDA texture class
    - **Regression:** *D* index (with automatic conversion to classes)
  - **Variance filter** (q)**:** optional filtering of training data by interpolation uncertainty
  - **Smoothing:**
    - Input layer smoothing (moving window filter, user-specified size)

        * Prediction smoothing (spatial post-processing for practical results)
    - **Validation set**:
        * **Preferred:** independent dataset from alternative sampling design (*Grid / QC / QC$_{var}$*) for cross-design validation
        * **Alternative:** random split of the dataset's own samples when independent data are unavailable
    - **Seed:** Random / deterministic

Model execution processing time is typically 30 seconds for 6.5 ha field.

Step 5: Results visualization and analysis Performance metrics:

- **Accuracy metrics:** overall accuracy, class-specific accuracies, kappa
- **Regression metrics:** $RMSE$, R^2 for D predictions

Spatial visualization: predicted texture classes at 1 m resolution (for regression, a map of D is also displayed)

Benchmark comparison:

- **Parameter sweep results:**
    - Accuracy vs. clay/silt threshold
    - Accuracy vs. variance filter (q)
    - Accuracy vs. number of trees
    - Classification vs. regression comparison
- **Multiple perspectives:**
    - Grouped by sampling design
    - Grouped by depth
- Statistical visualizations in bar charts

Export capabilities:

- **Prediction maps:** GeoTIFF Raster file (`.tif`)
- **Performance reports:** tables (`.csv`), figures (`.png`)

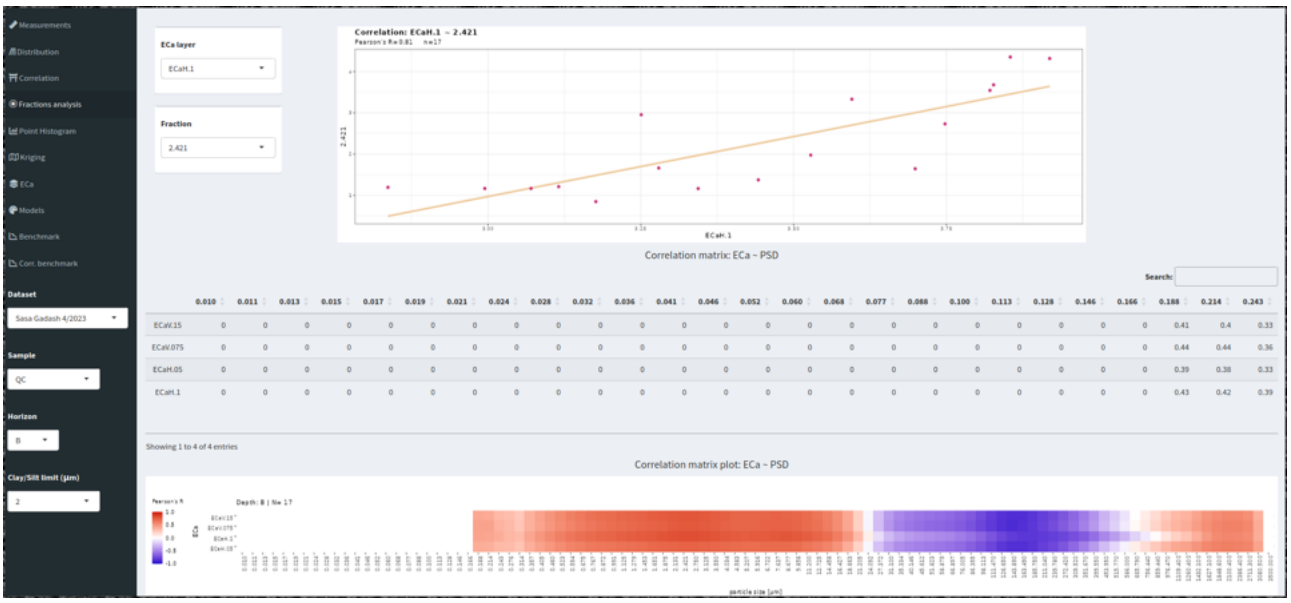

Figure 19: Screenshot of the shiny web application app-2, a tool for soil texture analysis and prediction, showing correlation analysis at resolution of 101-fraction size classes.