

**Analysis of Multi-Layer Spatial Data to Increase the Accuracy of Thematic Soil Maps for
Agriculture**

Felippe Hoffmann Silva Karp

Department of Bioresource Engineering
Faculty of Agricultural and Environmental Sciences
Macdonald Campus of McGill University
21,111 Lakeshore Road, Ste. Anne de Bellevue, Québec H9X 3V9 Montreal, Canada

July 2024

A Thesis submitted to McGill University in partial fulfillment of the requirements of the degree
of Doctor of Philosophy

© Felippe Hoffmann Silva Karp, Canada, 2024

Dedication

To my beloved wife, Karen Florentino Correa. This thesis is as much yours as mine, your unlimited support, love, and encouragement kept me moving even through the most challenging times.

Table of Contents

Dedication	ii
Abstract	vi
Résumé.....	viii
Acknowledgments.....	x
Publications, Manuscripts, Conference Papers	xiii
Contribution of Authors	xvi
List of Tables.....	xvii
List of Figures	xix
List of Abbreviations.....	xxv
Chapter 1: Introduction	1
1.1. General Introduction and Problem Statement	1
1.2 Research objectives	4
1.3 Thesis Organization.....	4
Chapter 2: Review of Literature	8
2.1 Traditional soil characterization and its challenges for PA	8
2.2 Individual PSS for within-field soil characterization.....	10
2.3 PSS data fusion for within-field soil characterization: challenges and current developments	14
2.4 PSS Data Processing	21
Chapter 3: Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes.....	25
3.1 Introduction	25
3.2 Materials and Methods	26
3.3 Results and Discussion.....	35
3.4 Conclusion.....	43

3.5 Acknowledgments	43
3.6 References	43
Chapter 4: Comparative Study of Interpolation Methods for Low-Density Sampling.....	50
4.1 Introduction	50
4.2 Material and Methods.....	52
4.3 Results and Discussion.....	62
4.4 Conclusions	79
4.5 Acknowledgments	80
4.6 Reference.....	80
Chapter 5: Prediction of soil chemical properties using proximal soil sensing technologies and topography data: Methodology and a case study	86
5.1 Introduction	87
5.2 Material and Methods.....	88
5.3 Results and Discussion.....	95
5.4 Conclusion.....	108
5.5 Acknowledgments	108
5.6 References	109
Chapter 6: Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing.....	114
6.1 Introduction	114
6.2 Materials and Methods	116
6.3 Results and Discussion.....	123
6.4 Conclusion.....	141
6.5 Acknowledgments	141
6.6 References	141
Chapter 7: General Discussion.....	148

Chapter 8: Summary and General Conclusions	155
Chapter 9: Contribution to Knowledge.....	156
References	157

Abstract

Soil sampling and subsequent lab analyses have guided site-specific crop management. Precision agriculture (PA) developments have introduced a wide variety of proximal soil sensing (PSS) technologies to agriculture to improve soil mapping. However, valuable information can only be extracted from PSS after properly processing and analyzing its high-density data. Due to the complexity of the signal from soil sensing techniques, accurate and reliable PSS-based soil management decisions require specific skills and time from PA practitioners, which is a barrier to the adoption of PSS and PA. The automated analysis of PSS through decision support systems (DSS) and PSS data fusion (DF) have been suggested as solutions to improve the quality of soil mapping for PA purposes. Therefore, four studies were conducted to propose techniques to improve soil mapping and support the development of DSS for agricultural soil management.

The first study defined a framework for batch processing PSS data. It was determined that properly processing any spatial dataset requires the following steps: data projection, position offset correction, global and local filtering, and interpolation. The proposed batch-processing framework improved the value of PSS data by highlighting field spatial variability that erroneous data had previously masked.

A second study evaluated eleven interpolation approaches for low and extra-low sampling densities, including traditional methods, a newly proposed methodology, and a kriging-based approach. Field average never emerged as the basis for the best approach. Also, none of the evaluated interpolation procedures appeared to be best for all fields, soil properties, and sampling densities. The proposed kriging-based approach and inverse distance weighting (IDW) with the power parameter value of 1 emerged as the most robust approaches as they rarely yielded errors worse than those obtained using the field average.

A third study proposed and evaluated a method to predict within-field spatial variability of soil chemical properties through PSS and topography data fusion. Machine learning (ML) algorithms (support vector machine, random forest, and partial least squares) were trained using two sampling densities (0.4 and $3.5 \text{ ha} \cdot \text{sample}^{-1}$) and validated using an additional 20 independent soil samples. Differences between ML models or sampling densities were insignificant for a given soil property. However, the mean squared error (MSE) and the coefficient of determination (R^2) indicated that some models outperformed others. The definition of the evaluated ML algorithms does not consider the spatial location of the samples; thus, an inverse distance smoothing window (SW)

was applied to the predicted surfaces. The SW did not change predictions significantly but often led to decreased R^2 and increased MSE values.

Lastly, the potential of the fusion of PSS and topography data to validate and improve the delineation of soil boundaries was assessed. Using an existing high-resolution soil survey, an analysis of variance was performed to compare the distribution of data from each sensor within the delineated soil types. None of the sensors could differentiate all soils in the field. However, maps containing an overlay between sensors and soil models provided an important insight: overall, the soils were located correctly, but the boundaries needed to be adjusted. A final map containing well-delimited homogenous PSS-based zones was obtained using an unsupervised spatial clustering method.

Through the development of these four studies, potential frameworks and methods to be embedded in DSS were developed to maximize the value of the data and facilitate data analysis and the soil management decision-making process. This research indicated that using supervised and unsupervised ML to fuse PSS and topography data can improve spatial soil characterization, potentially providing more accurate PA soil management strategies.

Résumé

Le prélèvement d'échantillons de sol et les analyses de laboratoire ont guidé la gestion des sols en agriculture. Le développement de l'agriculture de précision (PA) a introduit des technologies de détection proximale des sols (PSS) pour améliorer la cartographie des sols. Cependant, les informations des PSS ne peuvent être obtenues qu'après un traitement et une analyse appropriés des données. Les signaux complexes des PSS nécessitent des compétences spécifiques et du temps, créant un obstacle à l'adoption des PSS et de la PA. L'analyse automatisée des PSS via des systèmes d'aide à la décision (DSS) et la fusion de données PSS (DF) pourraient améliorer la cartographie des sols en PA. Ainsi, quatre études menées visaient à développer des techniques soutenant les DSS pour la gestion des sols agricoles.

La première étude a porté sur la création d'un cadre pour le traitement par lots des données PSS. Ce cadre, basé sur la projection des données, la correction des décalages de position, le filtrage global et local, et l'interpolation, a amélioré la valeur des données PSS en mettant en évidence la variabilité spatiale masquée par des données erronées.

La deuxième étude a évalué onze approches d'interpolation pour des densités d'échantillonnage faibles et très faibles, y compris des techniques traditionnelles, une nouvelle méthodologie et une approche basée sur le krigeage. La moyenne des valeurs du champ n'a jamais émergé comme la meilleure approche. Aucune des méthodes évaluées n'était optimale pour tous les champs, propriétés du sol et densités d'échantillonnage. Le krigeage et la pondération inverse de la distance (IDW) avec un paramètre de puissance de 1,0 étaient les meilleurs, produisant rarement des erreurs pires que la moyenne du champ.

La troisième étude a testé une méthode pour prédire la variabilité spatiale des propriétés chimiques du sol à travers la fusion des données PSS et topographiques. Les algorithmes d'apprentissage automatique (ML) (support vector machine, random forest et partial least squares) ont été entraînés avec deux densités d'échantillonnage (0,4 et 3,5 ha· échantillon⁻¹) et validés avec 20 échantillons de sol indépendants. Les différences entre les modèles de ML ou les densités d'échantillonnage étaient insignifiantes pour une propriété donnée du sol. Cependant, l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R²) ont montré que certains modèles surpassaient les autres. Puisque les algorithmes de ML évalués omettent la localisation spatiale des échantillons, une fenêtre de lissage par distance inverse (SW) a été appliquée aux

surfaces prédites. La SW a conservé les prédictions de manière significative mais a souvent conduit à une diminution des valeurs de R^2 et à une augmentation des valeurs de MSE.

Enfin, la possible fusion des données PSS et topographiques pour valider et améliorer la délimitation des frontières des sols a été évaluée. En utilisant une enquête existante à haute résolution (1:5000), une analyse de variance a comparé la distribution des données de chaque capteur au sein des types de sol délimités. Aucun des capteurs n'a pu différencier tous les sols. Cependant, les cartes superposant les données des capteurs et les modèles de sol ont fourni des aperçus importants: les sols se situaient correctement, mais les frontières devaient s'ajuster. Une méthode de clustering spatial non supervisée a créé une carte finale avec des zones homogènes claires basées sur les PSS.

Ces quatre études ont développé des cadres et méthodes à intégrer dans les DSS pour maximiser la valeur des données, faciliter l'analyse et améliorer la prise de décision en gestion des sols. Cette recherche a montré que l'utilisation de ML supervisé et non supervisé pour fusionner les données PSS et topographiques peut améliorer la caractérisation spatiale des sols, offrant des stratégies de gestion des sols plus précises pour la PA.

Acknowledgments

First, I would like to thank God for giving me strength, energy, and wisdom during my academic training.

I want to acknowledge anyone directly or indirectly involved in this thesis's data collection, analysis, and development of the HyperLayer Data Project. The work presented in this thesis could not be accomplished without the support of a group of people. To the friends I made during my summers at Olds College, thank you for your support, our casual conversations during lunch (or during coffee breaks – for the coffee drinkers), and great research discussions. I especially thank my great friends Akshay Bhanot and Jyotish Prabhakaran Jayasree, who provided a great work environment while building the HyperLayer.ag platform. The HyperLayer project (consequently the data used for the development of this thesis) would not exist without the leadership of my co-advisor and friend, Dr. Alexei Melnitchouck, the great mind who began, raised funds, industry partners, and led this project at Olds College of Agriculture & Technology. Alex, I appreciate all the opportunities you gave me to interact with industry and government specialists and your ongoing advice, guidance, and support. I would also like to thank Dr. Joy Agnew, VP of Research at Olds College, for her support, ideas, and late afternoon meetings.

I wholeheartedly thank my advisor, Dr. Viacheslav Adamchuk, for his guidance and supervision. I was still an undergraduate student attending my first international conference when I met him (11th European Conference on Precision Agriculture in Scotland). It was the first time I got to “put the face” to the authors of the Precision Agriculture papers I was reading. At that time, I could not imagine that one day I would be working with such a famous person in the Precision Agriculture community. Dr. Adamchuk, thank you for the opportunity to work with you. Your experience, passion, and knowledge of Precision Agriculture and its tools are invaluable, and I appreciate your efforts in sharing your experiences with other students and me. I truly appreciate how you helped me practice and improve the delineation of achievable research questions. Your unique ideas, solutions, and guidance were essential for completing this thesis.

Dr. Pierre Dutilleul, I am incredibly grateful you agreed to be a supervisory committee member. The mathematical and geostatistical foundation you provided is invaluable. I was extremely fortunate to work with you and deepen my spatial and temporal analysis knowledge through our interactions these past three years, including your great AEMA 614 lectures. I appreciate the time

you invested in reviewing our research articles; it improved the quality of our work and my writing skills.

I also thank Mitacs and Telus Ag for their financial support, covering my stipend and project-related expenses through the Mitacs Accelerate project (IT24224) entitled “Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services.” This financial support was essential for pursuing my PhD at McGill University.

Special thanks to other industry partners who contributed to developing the sensing database for the HyperLayer sites. Trevor Thornton from Crop Care (Saint François Xavier, MB, Canada), Paul Raymer, and Zach Harmer from SoilOptix (Tavistock, ON, Canada) to facilitate the collection and access to the complete passive-gamma-ray datasets. Steve Larocque and his team from Beyond Agronomy (Three Hills, AB, Canada) carefully and consistently collected the soil samples following our strict research guidance. Doug Mackey, Senior Commercial Project Manager at xarvio (Münster, Germany), was the Canadian representative responsible for the participation and contribution of xarvio to the HyperLayer project.

I also acknowledge the direct or indirect contributions and friendship of past and present members of the Precision Agriculture and Sensor Systems (PASS) research team members: Dr. Md Saifuzzaman, Dr. Roberto Buelvas, Karim Abdalla, Andrés Rello, John Lan, and Guillaume Cloutier Boily.

I truly thank my friends and research mentors, Rodrigo Gonçalves Trevisan and André Freitas Colaço, who have always supported my academic endeavors and engaged in discussions about research objectives. I also thank Dr. Jeffrey Cardille and Dr. Vijaya Raghavan, who were my comprehensive exam committee members. I would also like to acknowledge “Escola Superior de Agricultura Luiz de Queiroz” (ESALQ – USP) and Louisiana State University for their contributions to my academic training, especially Drs. José Paulo Molin and Luciano Shozo Shiratsuchi, who advised me on this journey.

To my Brazilian, American, and Canadian friends, thank you for your support and friendship.

I thank all my family, especially my mom, Paula Emília (*in memoriam*), my dad, Willian, and my sister, Aline, who supported my eagerness to learn since I was young. Special thanks to my grandmother, Neuza, for believing in me and investing in my education throughout my life. Lastly, I thank my wife’s parents, Darcy and Ivone, for their unfailing love and direct support, or indirectly through my wife, Karen, who gave up on many things in life to support me during my academic

journey. No matter what happened in our lives, the constant moving to different places, or the many challenges I encountered as a graduate student, Karen was always by my side, encouraging and cheering me on. Even when I did not believe in myself, Karen did. Thank you for everything, sweetheart.

“If you want to go fast, go alone; if you want to go far, go together.” African Proverb

Publications, Manuscripts, Conference Papers

a. Conference Proceedings

1. Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2022). Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes. In: Proceedings of the 15th International Conference on Precision Agriculture (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture. <https://www.ispag.org/proceedings/?action=abstract&id=8792&title=Optimization+of+Batch+Processing+of+High-density+Anisotropic+Distributed+Proximal+Soil+Sensing+Data+for+Precision+Agriculture+Purposes>.
2. Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2023). Comparative study of interpolation methods for low-density sampling. In: Precision agriculture '23 (Vol. 34, pp. 563–569). The Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-947-3_71
3. Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2024). Predicting soil chemical properties using proximal soil sensing technologies and topography data: A case study. In: Proceedings of the 16th International Conference on Precision Agriculture (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture.
4. Lan, J, Adamchuk, V. I., Geddes, K., Kvezerelu, B., Karp, F. H. S., McGuire, S. (2024). Development of an Automated System for In situ Measurements of Soil Health Indicators. In: 2024 ASABE Annual International Meeting. St. Joseph, MI: American Society of Agricultural and Biological Engineers. doi:10.13031/aim.202401135.

b. Journal Publications

1. Karp, F. H. S., Adamchuk, V. I., Melnitchouck, A., Allred, B., Dutilleul, P., & Martinez, L. R. (2023). Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing. *Journal of Environmental and Engineering Geophysics*, 28(1), 45–61.
<https://doi.org/10.32389/JEEG22-018>
2. Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2024). Comparative study of interpolation methods for low-density sampling. *Precision Agriculture*.
<https://doi.org/10.1007/s11119-024-10141-0>
3. Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2025). Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes (in preparation).
4. Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2025). Prediction of soil chemical properties using proximal soil sensing technologies and topography data: Methodology and a case study (in preparation).

c. Books and Book Chapter

1. Adamchuk, V.I., Karp, F. H. S., Biswas, A. (2025). Developments in proximal soil sensing for precision agriculture. In: Stafford, J. (ed.) Precision agriculture for sustainability: Second edition. Burleigh Dodds Science (in press).

Contribution of Authors

This thesis comprises eight chapters, from which the manuscripts in chapters three to six were presented at scientific conferences or submitted to peer-reviewed journal publications. For these four manuscripts, the author of the thesis was the first author and was responsible for developing the research questions, new interpolation and analytical approaches and methodologies, programming custom scripts for data exploration and analysis, interpretation of the results, and writing and compiling this thesis. However, the work presented here would not be possible without the contributions of the co-authors. A detailed contribution for each one of the four manuscripts is given below.

In Chapter 3, all co-authors provided technical edits and comments that substantially improved the quality of the manuscript. Dr. Viacheslav Adamchuk suggested adding the projection of the data to a “custom localized Cartesian coordinate system” as an essential step, as some of the other procedures relied on accurate distance calculations. An expanded version of Chapter 3 will be submitted to *Computers and Electronics in Agriculture*. Chapter 4 has already been published in the *Precision Agriculture* journal; Dr. Pierre Dutilleul provided input from a geospatial statistics perspective and technical edits, Dr. Adamchuk suggested the testing of a new methodology and a modified kriging approach, and Dr. Alexei Melnitchouck guaranteed the data collection and access to it. In Chapter 5, all co-authors contributed with technical edits and comments that substantially improved the quality of the manuscript; an expanded version of this chapter will be submitted to the *Precision Agriculture* journal. Lastly, Chapter 6 was published in the *Journal of Environmental and Engineering Geophysics*; the contribution of the co-authors and journal reviewers in this manuscript was essential to improve its quality. Dr. Adamchuk contributed with discussions on the clustering algorithm and introductions to two other co-authors, Dr. Barry Allred and Luiz R. Martinez. Dr. Barry Allred is a specialist in ground penetrating radar and geophysics applied to agriculture and contributed with insights, comments, and technical edits. Luiz R. Martinez guided the proper data processing of the ground penetrating radar (GPR) data, Dr. Melnitchouck helped with data collection and availability, and Dr. Dutilleul provided technical edits, statistical and geostatistical comments, and support.

List of Tables

Table 2.1 Proximal soil sensing techniques functionality and measured soil properties.....	11
Table 2.2 Examples of studies on proximal soil sensing data fusion for soil characterization (a hyphen indicated the information was not available or not applicable for that study)	17
Table 3.1 Information for the different proximal soil sensors used for the evaluation of processing methodologies	34
Table 3.2 Geostatistical analysis results (Nugget, Sill: variogram model parameter estimates) and root mean squared error (RMSE) from ordinary kriging cross-validation for position offset and filtering procedures for each proximal soil sensor. Data is standardized to zero mean and unit variance. Bold and red-colored numbers represent the lowest values, and dashes indicate that the procedure was not applied or no point was removed.....	35
Table 3.3 Interpolation methods root mean squared error (RMSE) and coefficient of determination (R^2) resulted from 10-fold cross-validation for each proximal soil sensor.	42
Table 4.1 Description of the studied fields	52
Table 4.2 Summary table of all the different methods and approaches evaluated in this study, and their respective reference name used for figures and tables	60
Table 4.3 Descriptive statistics for plant available Potassium (K), plant available Phosphorus (P), and pH for all the fields, sampling densities, and validation samples (a hyphen indicates that values are not available).....	63
Table 4.4 Variogram model parameter estimates from a standard variogram fitting procedure for the four fields at the highest sampling density collection (Table 4.3), after data was standardized to a zero mean and a unit variance.....	65
Table 5.1 Description of the sensors, mapped variables, and collection settings adopted.....	91
Table 5.2 Descriptive statistics for soil testing results for plant-available potassium (K) and phosphorous (P), pH, and soil organic matter (OM) from two sampling densities and validation samples. (Modified from Karp et al., 2024).....	97

Table 5.3 Example of descriptive statistics for the elevation raster data and one variable from each PSS before (raw) and after applying the filtering procedure steps from Karp et al. (2022) (a hyphen indicates that the filtering procedure was not applied to that dataset).....	98
Table 5.4 Example of descriptive statistics from the 3.5 ha·sample ⁻¹ co-located training dataset and interpolated surface (15-meter raster) for elevation and one variable from each PSS	99
Table 6.1 Field AGRASID soil models classified during the soil survey performed in 2003. Information retrieved from Walker and Mcneil (2004).....	117
Table 6.2 Description of proximal soil sensors, most significant measured physical property, and mapped variables. EC _a – apparent electrical conductivity, RTK – Real-Time Kinematic.....	119
Table 6.3 Comparison of the data distribution before and after the global and local outlier removal process and percentage of data removed. GPR – Ground Penetrating Radar, Med. Inst. Ampl. – Median Instantaneous Amplitude, EM – Electromagnetic Induction, GC - Galvanic Contact, EC _a – apparent electrical conductivity, Hist. – Histogram.....	124
Table 6.4 Variogram model parameters for Ground Penetrating Radar (GPR) median instantaneous amplitude (Med. Inst. Ampl.) for every 0.1 m depth interval.	126
Table 6.5 Variogram model parameters for electromagnetic induction apparent electrical conductivity (EC _a) for shallow (0-0.75m) and deep (0-1.50m) layers.	128
Table 6.6 Variogram model parameters for galvanic contact apparent electrical conductivity (EC _a) for shallow (0-0.3 m) and deep (0-0.9 m) layers.	129
Table 6.7 Variogram model parameters for SoilOptix passive gamma ray for ¹³⁷ Cs, ²³² Th, ²³⁸ U, ⁴⁰ K, and Count Rate.	130
Table 6.8 One-Way ANOVA and Tukey-Kramer ad-hoc test results for each proximal soil sensor and terrain data. The rows of the table are ordered by the F-Test values, and different lowercase letters represent significant differences in the averages among the different soil models. GPR – ground penetrating radar, GC – galvanic contact, EM – electromagnetic induction, EC _a – apparent electrical conductivity	138

List of Figures

Figure 1.1 Use of precision agriculture technologies estimated by precision agriculture dealerships in the United States (Modified from Erickson & Lowenberg-DeBoer, 2023).....	2
Figure 1.2 Barriers to the growth and expansion of precision agriculture (Modified from Erickson & Lowenberg-DeBoer, 2023)	3
Figure 2.1 Data fusion levels. A – Data/signal level; B – Feature level; C – Decision Level.	15
Figure 2.2 Example of a precision agriculture dataset with different support, spatial location, and data type.	22
Figure 3.1 Illustration of the definitions of unidirectional (within-row) neighborhood (A) and omnidirectional (within- and between-rows) neighborhood (B)	31
Figure 3.2 Results of global and local filtering methodologies for galvanic contact (GC) sensor apparent electrical conductivity (EC_a). Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.....	38
Figure 3.3 Results of global and local filtering methodologies for electromagnetic induction (EMI) sensor apparent electrical conductivity (EC_a). Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.....	39
Figure 3.4 Results of global and local filtering methodologies for ground penetrating radar (GPR) Instantaneous Amplitude at 0-0.1m depth. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.....	40
Figure 3.5 Results of global and local filtering methodologies for γ -ray count rate. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.	41
Figure 4.1 Experimental sites: a – map of the Canadian province of Alberta with a zoomed-in window showing the distribution of the four experimental sites located within a region known as Central Alberta, and b – field boundaries and grid centroids representing the three sampling design	

densities – different shapes, and colors are used to represent the distribution of the sampling locations for 0.4 (solid black circles), 0.8 (hollow red circles), 3.5 (hollow blue squares) ha·sample⁻¹, and validation points (solid green diamonds); the latter are only available for Field 1 54

Figure 4.2 Flowchart explaining the proposed modification of the kriging-based approach to obtain interpolated surfaces from soil data collected at low and extra-low sampling densities; OK – ordinary kriging 57

Figure 4.3 Biplot of Mean Distance Gradient (DG) versus Mean Absolute Error (MAE) for 3,500 trials (blue dots), with the Pareto-optimal solutions (orange dots) and the optimal solution (purple triangle) 59

Figure 4.4 Box plots of interpolation errors for Field 1 at grid sampling densities of 0.4 ha·sample⁻¹ (panels a-c), 0.8 ha·sample⁻¹ (panels d-f), and 3.5 ha·sample⁻¹ (panels g-i). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: “Fitted variogram model” was removed from the analysis and does not appear in a few panels because the variogram model fitting algorithm failed to converge for that specific sampling density and soil property. Letters a and b indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test. 68

Figure 4.5 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the 0.4 ha·sample⁻¹ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps. 69

Figure 4.6 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the 0.8 ha·sample⁻¹ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps. 69

Figure 4.7 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the 3.5 ha·sample⁻¹ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps. 70

Figure 4.8 Box plots of interpolation errors for Field 2 at grid sampling densities of 0.8 ha·sample ⁻¹ (panels a-c) and 3.5 ha·sample ⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: “Fitted variogram model” was removed from the analysis and does not appear in most panels because the variogram model fitting algorithm frequently failed to converge. Letters a, b, c, d, and e indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.	72
Figure 4.9 Box plots of interpolation errors for Field 3 at grid sampling densities of 0.8 ha·sample ⁻¹ (panels a-c) and 3.5 ha·sample ⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Letters a, b, c, d, and e indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.	74
Figure 4.10 Box plots of interpolation errors for Field 4 at grid sampling densities of 0.8 ha·sample ⁻¹ (panels a-c) and 3.5 ha·sample ⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: Fitted variogram model” was removed from the analysis and does not appear in a few panels because the variogram model fitting algorithm failed to converge for that specific sampling density and soil property. Letters a and b indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.	76
Figure 4.11 Robustness and reliability analysis for the 11 interpolation procedures at the three different sampling densities. G values from Figures 4.4 and 4.8- 4.10 were grouped in the categories listed in the legend. “No Convergence” applies to the kriging approach “B – Fitted variogram model” when the model fitting algorithm failed to converge, and “Flat Variogram” applies to the modified kriging approach “J – Set Sill and Nugget” when estimated nugget-effect is the equal or higher than sample variance.	77
Figure 5.1 Centroid locations for the original sampling density (0.4 ha·sample ⁻¹ – solid black circles), selected samples to create the extra low-density sampling design (3.5 ha·sample ⁻¹ – hollow blue squares), and validation samples (solid green diamonds) (Modified from Karp et al., 2024)	89

Figure 5.2 Spearman's correlogram for all sensors and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; EC_a- apparent electrical conductivity; EM – Electromagnetic Induction sensor; CR – count rate..... 101

Figure 5.3 Spearman's correlogram matrix for soil properties, all sensors, and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; EC_a- apparent electrical conductivity; EM – Electromagnetic Induction sensor; CR – count rate. 102

Figure 5.4 Bar-dot plot for the coefficient of determination (R^2 ; bars) and mean squared error (MSE; points) for the 20 validation samples comparing partial least squares (PLSR; hatched bars and circle-shaped points), random forest (RF; crosshatched bars and triangle-shaped points), and support vector machines (SVM; bars with small dots and square-shaped points) as prediction algorithms for plant available potassium (K), phosphorus (P), pH, and soil organic matter (OM) for a given sampling density. A red border around a bar indicates the selected model for that sampling density..... 103

Figure 5.5 Box plots for the prediction errors for plant-available potassium (K), phosphorus (P), pH, and soil organic matter (OM). Results from the predictions based on the fusion of all data sources for two sampling densities, 0.4 and 3.5 ha· sample⁻¹, are indicated by a capital letter and a color. While partial least squares (PLSR), random forest (RF), and support vector machines (SVM) were tested for each scenario, only the best-performing algorithms are presented (refer to Figure 5.4 for all algorithms). MSE followed by different uppercase letters differ significantly at $\alpha = 0.05$ within the panel (between sampling densities), and different lowercase letters for the same sampling density but between "No Spatial Smoothing" and "IDW Moving Window" 105

Figure 5.6 Thematic maps for plant-available potassium (K; a-e), phosphorus (P; f-j), pH (k-o), and soil organic matter (OM; p-t) from the interpolation of the 0.4 ha·sample⁻¹ using ordinary kriging (first column of maps), and calibration model predictions for the 3.5 ha·sample⁻¹ (second and third columns) and 0.4 ha·sample⁻¹ (fourth and fifth columns) before and after the “IDW Moving Window” (IDW MW) was applied..... 107

Figure 6.1 Spatial distribution of the soil models within the study field – refer to Table 6.1 for details on each soil model.	118
Figure 6.2 Proximal soil sensors data acquisition characterization for (a) Galvanic Contact (GC) – Veris 3100, (b) Electromagnetic Induction (EM) – EM38-MK2, (c) Passive Gamma-Ray (γ -ray) – SoilOptix, and (d) Ground Penetrating Radar (GPR) – SIR-4000. Dist. Points – distance between consecutive points.	119
Figure 6.3 Alberta soil code Antler (ATL) sampled soil pedon and horizons depths observed during soil survey (Walker and Mcneil, 2004).	127
Figure 6.4 Ground penetrating radar (GPR) median instantaneous amplitude (Med. Inst. Amplitude) interpolated maps for every 0.1 m depths from 0 to 1.5 m. Each map has a different legend to highlight the spatial patterns represented by each depth layer.	127
Figure 6.5 Electromagnetic induction (EM) apparent electrical conductivity (EC_a) interpolated maps for (a) shallow and (b) deep layers. Each map has a different legend to highlight the spatial patterns represented by each depth layer.	129
Figure 6.6 Galvanic contact (GC) apparent electrical conductivity (EC_a) interpolated maps for (a) shallow and (b) deep layers. Each map has a different legend to highlight the spatial patterns represented by each depth layer.	130
Figure 6.7 SoilOptix passive gamma ray (γ -ray) maps for (a) ^{232}Th , (b) ^{40}K , and (c) Count Rate.	131
Figure 6.8 Maps for (a) digital elevation model and (b) terrain slope.	132
Figure 6.9 Correlation analysis for all proximal soil sensors, digital elevation model, and terrain slope. The upper triangular matrix presents the numerical values for the Person’s correlation coefficient (r), while the lower one shows a graphical representation of the correlation – a larger circle radius means a higher absolute r . Circles and numbers are filled with a color scale based on the r values. Empty cells represent the non-significant correlation at a significance level of 0.05. GC – galvanic contact apparent electro conductivity, GPR – ground penetrating radar, EM – electromagnetic induction apparent electro conductivity, CR – count rate.	133

Figure 6.10 Overlay of soil models and (a) Ground penetrating radar (GPR) median instantaneous amplitude (Med. Inst. Amplitude), (b) galvanic contact (GC) apparent electro conductivity (EC_a), (c) electromagnetic induction (EM) apparent electro conductivity, (d) terrain slope, (e) γ -ray for ^{40}K , and (f) elevation (m)..... 136

Figure 6.11 Proximal soil sensors and terrain data (a) spatial fuzzy c-means clustering result and (b) overlay with soil models. 139

List of Abbreviations

Abbreviation	Definition	Abbreviation	Definition
¹³⁷ Cs	Caesium-137 isotope	LiDAR	Light Detection and Ranging
²³² Th	Thorium-232 isotope	LOOCV	Leave One-Out Cross-Validation
²³⁸ U	Uranium-238 isotope	MAE	Mean Absolute Error
⁴⁰ K	Potassium-40 isotope	ML	Machine Learning
ANOVA	one-way analysis of variance	MSE	Mean Squared Error
AO	Adjusted Outlyingness Index	NIR	Near Infra-Red
CAAIN	Canadian Agri-Food Automation and Intelligence Network	NN	Nearest Neighbors
CI	Cone Index	OK	Ordinary Kriging
CR	Count Rate	OM	Soil Organic Matter
CRAD	Coregionalization Analysis with a Drift	P	plant-available Phosphorus
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	PLSR	Partial Least Squares
DEM	Digital Elevation Model	ppm	parts per million
DF	Data Fusion	PSS	Proximal Soil Sensing
DG	Distance Gradient	R ²	Coefficient of Determination
DSS	Decision Support System	RF	Random Forest
EC _a	Apparent Electrical Conductivity	RGB	Red-Green-Blue (image)
EMI	Electromagnetic Induction	RMSE	Root Mean Squared Error
FDR	Frequency Domain Reflectometry	RTK	Real-Time Kinematic
GC	Galvanic Contact	SD	Standard Deviation
GIS	Geographic Information System	SFCM	Spatial Weighted Fuzzy c-means
GNSS	Global Navigation Satellite System	SVM	Support Vector Machine
GPR	Ground Penetrating Radar	SW	Smoothing Window
Hist.	Histogram	UK	Universal Kriging
IDW	Inverse Distance Weighting	UTM	Universal Transverse Mercator
IDW MW	Inverse Distance Weighted Moving Window	vis-NIR	Visible and Near Infra-Red
IQR	Inter Quartile Range	WGS-84	The current version of the World Geodetic System
ISE	Ion Selective Electrode		
ISO	International Organization for Standardization		
K	plant-available Potassium		
L-BFGS-B	Limited Memory Algorithm for Bound-Constrained Optimization		

Chapter 1: Introduction

1.1. General Introduction and Problem Statement

Modern society demands more environmentally friendly agriculture practices, focusing on soil conservation and fertility, water and air quality, wildlife, and landscape protection. As a result, farmers are under pressure to comply (Dimitri et al., 2005). In addition to the need for environmental conservation, food production must increase to feed the growing world population (FAO, 2009). Consequently, Precision Agriculture (PA) can provide strategies to help farmers mitigate market pressures.

According to the definition of PA presented by the International Society for Precision Agriculture (ISPA, 2024): “Precision Agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual plant and animal data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production.” Therefore, by implementing PA, farmers would be able to meet the new demands.

As the definition shows, PA has its foundation in collecting and analyzing spatiotemporal data. From its inception, PA researchers and practitioners have developed and adapted tools and techniques that enable data collection and analysis to improve and/or propose new agricultural management strategies.

Over the years, with advances in technology and computational capacity, tools and techniques that ease and reduce the costs of agriculture data collection and mapping have been developed, increasing the volume and popularity of PA-focused georeferenced data and services. However, farmers tend to adopt technologies that are easy to implement, such as autosteering, while information that needs to be processed, e.g., data originating from sensors, has lower adoption rates (Lowenberg-DeBoer & Erickson, 2019).

A long-term survey in the United States evaluated the adoption of PA tools and techniques (Erickson & Lowenberg-DeBoer, 2023). During this survey, PA dealerships were asked to estimate the producers’ adoption of PA technologies. Figure 1.1, extracted from the 2023 edition of the above survey, supports the findings from Lowenberg-DeBoer & Erickson (2019) (i.e., farmers tend to adopt ready-to-use technologies). The adoption rates in Figure 1.1 present two distinct groups, one of which adoption rates were above 50% since 2021 and the other lower than 32%. For the

most adopted group of PA technologies, very little to no user input or data analysis is required, except for “Grid or zone soil sampling.” Also, it is worth noting that most of the newer harvesters already come equipped with yield monitors, and yet even with its high adoption rates as shown in Figure 1.1, this does not guarantee that the information provided by such a tool is processed and used by farmers.

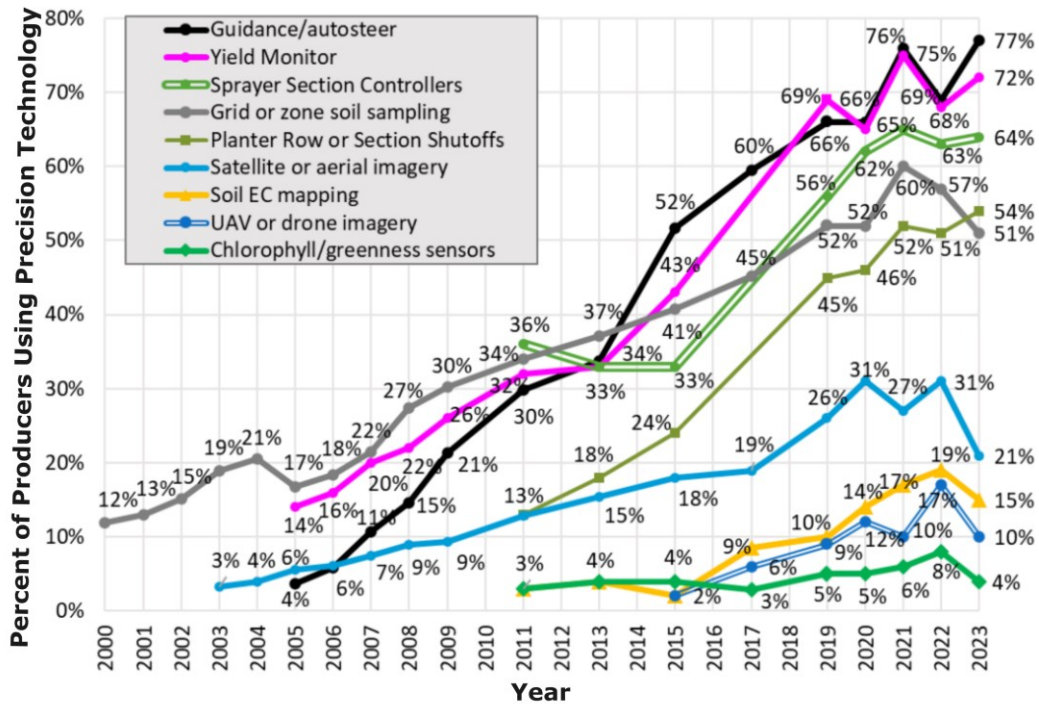


Figure 1.1 Use of precision agriculture technologies estimated by precision agriculture dealerships in the United States (Modified from Erickson & Lowenberg-DeBoer, 2023)

The same survey also evaluated the main barriers to PA's growth and expansion. In 2023, farm income and high costs with low benefits were the first and second most common barriers, respectively. The third was the time necessary to interpret and make decisions based on PA information (Figure 1.2). From this, it can be seen that the growth and expansion of PA depend on improving its techniques to yield higher benefits, lower costs, and simplify the decision-making process based on PA information. According to Zhai et al. (2020), developing decision support systems (DSS) can help overcome some of these barriers by automatically gathering, processing, and providing ready-to-use information supporting evidence-based decisions.

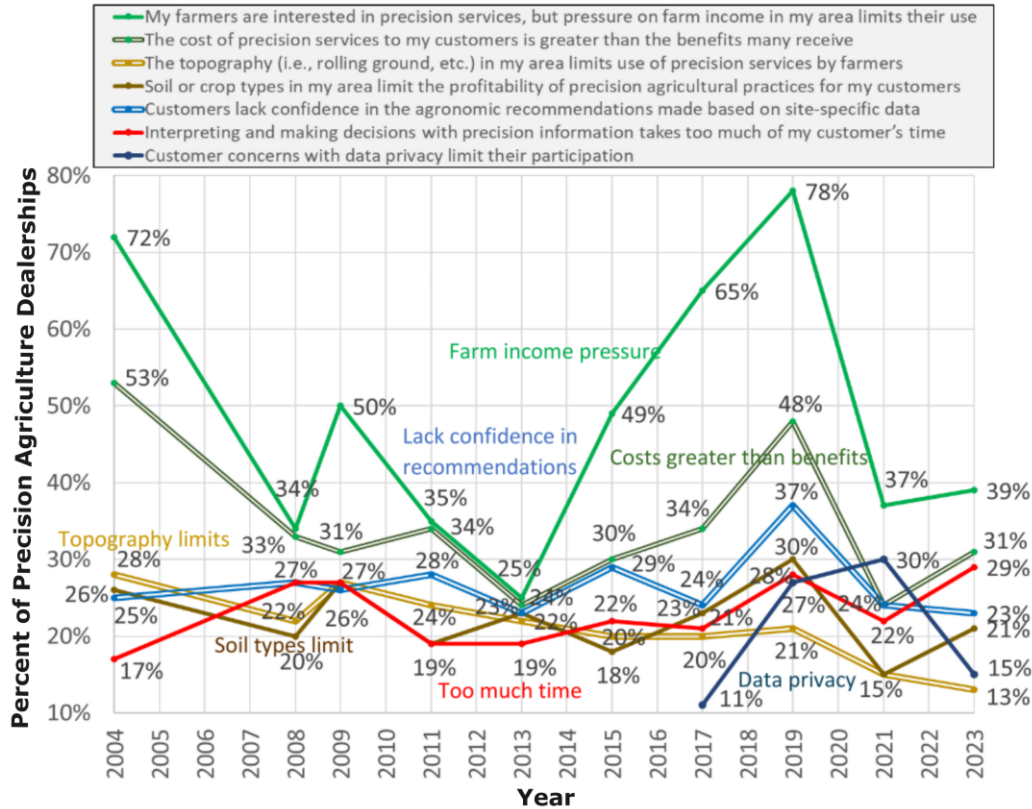


Figure 1.2 Barriers to the growth and expansion of precision agriculture (Modified from Erickson & Lowenberg-DeBoer, 2023)

Soil sampling provides farmers with important information on soil chemical and physical properties, which are commonly used to determine seed and fertilizer rates and make other management decisions. Thus, despite being expensive and labor-intensive, grid and zone soil sampling are among the most adopted PA practices (Figure 1.1). Also, it is necessary to process and analyze the data to obtain valuable information from soil sampling and its subsequent lab analysis. It is imperative to improve soil sampling and embed its analysis into DSS to benefit the PA community.

According to Adamchuk et al. (2011) and Gebbers (2018), proximal soil sensing (PSS) data fusion (DF) has the potential to improve the mapping of soil properties. Therefore, research has examined how DF can be applied to PA. Ji et al. (2019) evaluated the potential of using individual and fused data from multiple soil sensors to predict soil properties. These authors concluded that fused data predictions often outperformed individual sensors' predictions. Others have evaluated the DF of remote and proximal sensing and concluded that the proper selection of auxiliary

predictors and data fusion techniques have the capability of estimating the spatial distribution of soil chemical properties (Lachgar et al., 2024; Wang et al., 2022)

However, as Meng et al. (2020) reported, the DF process presents a few challenges: data imperfection, inconsistency, alignment, and heterogeneity. Research must be developed to overcome those challenges in an automated manner. Overall, a strategy and methodology that allow decision support systems to perform data fusion and provide valuable, easy, and fast-to-interpret insights to farmers concerning soil characteristics are needed. In addition, due to the complex response from PSS, further analysis of their interaction with soils should be conducted.

1.2 Research objectives

The overarching goal of this research was to propose and evaluate techniques to streamline the process and fusion of spatial data used for soil mapping in agriculture that could potentially be implemented in DSS. The following were the specific objectives:

- 1.1. To optimize the batch processing and rasterization of high-density proximal soil sensing data with anisotropic distribution to maximize the information value of sensor data;
- 1.2. To optimize universal interpolation methodology for low and extra-low-density soil sampling data;
- 1.3. To verify the ability to predict within-field soil chemical properties using proximal soil sensing and topographic data;
- 1.4. To validate and propose improvements to soil survey maps using proximal soil sensing.

1.3 Thesis Organization

This thesis is composed of eight chapters. **Chapter 1**, the current chapter, introduces the studied topic using relevant publications and concepts and presents the overall and specific objectives of the document. **Chapter 2** presents the state-of-the-art traditional and data fusion techniques for mapping soil characteristics for agricultural applications; the briefer literature review in **Chapters 1,3-6** is complemented by **Chapter 2** to satisfy the institution requirements for the thesis. **Chapter 3** referenced, analyzed, and evaluated most of the work developed concerning data pre-processing (e.g., filtering for removing outliers) when proposing a PSS batch-processing framework. **Chapter 4** proposes and evaluates different interpolation approaches to maximize the information value for low- and extra-low-density soil sampling, creating a baseline later to be used to compare with the results presented in the following chapter. **Chapter 5** proposes a method for data fusion that focuses on the calibration of PSS and topography data to predict soil

chemical properties. This method was evaluated using one of the fields from **Chapter 4**, while the relationship between soil chemical properties and PSS and topography was also assessed. **Chapter 5** focused on soil chemical properties; **Chapter 6** evaluated the ability of data fusion of PSS and topography to delineate soil zones within a field to validate and propose improvements to an existing high-resolution soil survey. **Chapter 7** provided a broader discussion covering the results from all manuscripts jointly, while an overall conclusion is covered in **Chapter 8**. Lastly, **Chapter 9** presents the contributions to knowledge provided by this thesis.

Connecting Text to Chapter 2

The “General Introduction and Problem Statement” section from **Chapter 1** initiates this thesis literature review by exposing the reader to the concept and barriers to the adoption of PA. The need for better approaches to soil mapping techniques that enable PA practices was highlighted. **Chapter 2** builds upon the topic introduced in **Chapter 1**. First, the most common soil mapping approaches and their challenges are discussed (which led to the development of Objective 1.2 – **Chapter 4**), followed by a comprehensive evaluation of literature using individual sensor systems for PA soil characterization. Lastly, the potential of DF of PSS for improving soil mapping was supported by previous research, and the existing knowledge gaps this thesis addresses were highlighted (which led to the development of Objectives 1.1, 1.3, and 1.4 – Chapters 3, 5, and 6, respectively).

A significant portion of **Chapter 2** was used as a baseline to develop a book chapter entitled “Developments in Proximal Soil Sensing for Precision Agriculture,” which has been accepted for publication as part of the second edition of the “Precision Agriculture for Sustainability” book. Publication:

Adamchuk, V.I., Karp, F. H. S., Biswas, A. (2025). Developments in proximal soil sensing for precision agriculture. In Stafford, J. (ed.). Precision agriculture for sustainability: Second edition. Burleigh Dodds Science (in press).

Abbreviations for Chapter 2

Abbreviation	Definition
Ca^{+2}	Extractable Calcium
Ca_{tot}	Total Calcium in Soil
CEC	Cation Exchange Capacity
CI	Cone-Index
CRAD	Coregionalization Analysis with a Drift
DF	Data Fusion
DSS	Decision Support System
EC_a	Apparent Electrical Conductivity
EMI	Electromagnetic Induction
FDR	Frequency Domain Reflectometry
GC	Galvanic Contact
GPR	Ground Penetrating Radar
ISE	Ion Selective Electrode
ISFET	Ion-Selective Field Effect Transistors
K^{+}	Extractable Potassium
K_{tot}	Total Potassium in Soil
LIBS	Laser-Induced Breakdown Spectroscopy
ML	Machine Learning
Na^{+}	Soluble Sodium
NIR	Near Infra-red
NO_3^{-}	Extractable Nitrogen
OM	Soil Organic Matter
PA	Precision Agriculture
PO_4	Extractable Phosphorous
PSS	Proximal Soil Sensing
RGB	Red-Green-Blue image
TDR	Time Domain Reflectometry
vis-NIR	Visible and Near-Infrared Spectroscopy
XRF	X-ray fluorescence

Chapter 2: Review of Literature

2.1 Traditional soil characterization and its challenges for PA

Accurate soil characterization is challenging and costly, as it usually involves the collection and subsequent analysis of soil samples, borings, and trenches. Viscarra Rossel & McBratney (1998) stated that mapping spatial variability at a fine scale is necessary for developing PA practices. However, they concluded that the high costs and uncertainty of soil sampling and analysis make this approach impractical. According to the same authors, the main factors influencing the uncertainty of the analysis of soil samples were the soil sampling procedure (e.g., choice of sampling location, inadequate tools, and sample handling), the inherent variability of soils, complex nature of the laboratory analysis requiring precise steps (quantitative inaccuracies from lab-to-lab), and the lack of quality assurance programs for laboratories.

The conclusion drawn by Viscarra Rossel & McBratney (1998) was supported by others, such as Jacobsen et al. (2002) who evaluated the repeatability of lab analysis by submitting identical samples to the same laboratory and found significant differences in reported results. Rains et al. (2001) sent the same sample to two different laboratories and found significantly different results. More recent studies indicate that the development of new techniques and improvements in analytical hardware for soil analysis combined with quality assurance programs might improve the soil analysis results. Low average inaccuracy was observed when Demattê et al. (2019) evaluated four laboratories that undergo constant quality assessment tests. On the other hand, the same authors reported that when lime prescription recommendation rates were calculated, they identified discrepancies in the recommendations; for example, for the same sample, the results from one lab indicated lime application was not required, while for another lab almost 2 t·ha⁻¹ would be recommended.

Even though the challenges related to soil characterization through soil sampling were identified in the early years of PA, many farmers and consultants still rely on this technique to plan and determine site-specific soil management. According to the 2023 Precision Agriculture Dealership Survey (Erickson & Lowenberg-DeBoer, 2023), site-specific soil management in the United States still relies heavily on soil sampling, with systematic soil sampling using grids as the most common approach. According to the same survey, sampling schemes with 1 sample·ha⁻¹ predominates. Surveys performed for Ontario and Western Canada also presented similar results with a predominant sampling grid cell of 1 ha (Mitchell et al., 2018; Steele, 2017). In other regions

of the world, such as Brazil, the most common sampling density among practitioners is usually <0.5 samples·ha⁻¹ (Cherubin et al., 2015, 2022). Molin (2017) reported that more than 10% of farmers who adopted grid soil sampling in 2013 collected 1 sample every 9 hectares, which may not be considered PA in some parts of the world (Lowenberg-DeBoer, 2022).

An interpolation procedure should be performed to obtain valuable information from these soil samples. Kriging is among the most common interpolation technique used for mapping soil characteristics in agriculture. This geostatistical method requires the estimation of parameters that are used to fit a model to an experimental variogram. Through a series of simulation scenarios, Webster & Oliver (1992) proved that the uncertainty of the estimation of variogram parameters greatly increased with the reduction in the number of samples and recommended that at least 100 samples should be used to obtain an adequate variogram parameter estimation. Since frequently 1 or no samples representing an area of 100×100 m are adopted by PA practitioners, a minimum of 100 might not be reached. Therefore, high-uncertainty interpolated surfaces would be obtained, which could affect site-specific recommendations, creating a challenging scenario for precisely managing the fields based on these results.

Stepień et al. (2013) compared prescription maps obtained from 1, 0.5, and 0.25 samples·ha⁻¹. These authors concluded that while the average amount of fertilizer applied did not change drastically, low sampling densities (0.5 and 0.25 samples·ha⁻¹) resulted in a large area of the field receiving fertilizer rates that did not match the soil requirements. Note that the correct fertilizer placement might not guarantee a significant reduction in fertilizer costs or higher crop yields. However, the soil as a non-renewable resource must be utilized wisely, and using soil nutrients without replenishing them may cause irreversible long-term effects. Similarly, the application of an excessive amount of nutrients will result in environmental contamination (Andraski et al., 2000).

The challenges described above also apply to high-resolution surveys used to classify and delineate soil types, as they require many samples, borings, and trenches to guarantee the proper classification and delineation of the soils in an area. Since the location of the boundaries is often delineated based on sparse observations (subsurface investigations) and subjective analysis performed by experts, there can be high uncertainty in the location of the soil boundary. For example, James et al. (2003) compared soil classification with apparent electroconductivity (EC_a) from an electromagnetic induction (EMI) sensor and observed some inconsistencies, so they

decided to resample the field. As a result, the authors determined that the previous soil boundaries needed to be adjusted.

Traditional methods for soil characterization have long been used to provide insightful information. However, these methods are time-consuming, cost- and labor-intensive. Consequently, one would often choose to reduce the density of samples, which, as mentioned above, would result in inaccurate maps and management decisions. The results presented by Wadoux et al. (2019) suggested that adding a sub-set of close-pair sampling can reduce the uncertainty of estimating the spatial variability for soil mapping. However, it is not a common approach in PA.

Viscarra Rossel & McBratney (1998) promoted a different approach that uses PSS to improve soil mapping characteristics in agriculture. Doolittle (1987), when discussing conventional methods for soil surveying, mentioned that “conventional methods of observing soils are often slow and tedious and produce incomplete data” and suggested using PSS to improve the results from soil surveys.

2.2 Individual PSS for within-field soil characterization

Viscarra Rossel et al. (2011) defined PSS as “the use of field-based sensors to obtain signals from the soil when the sensor’s detector is in contact with or close to (within 2 m) the soil”. This definition excludes any remote sensing or analysis developed in the laboratory. However, as mentioned by Viscarra Rossel et al. (2011), it is important to acknowledge that the development of PSS techniques begins in the laboratory (e.g., x-ray Fluorescence and vis-NIR).

PSS data is commonly used as ancillary information to create directed (guided) sampling schemes. Gonçalves et al. (2021) used soil apparent electrical conductivity data collected by a commercial PSS to create zones and determine sampling locations for a field. However, since the readings from PSS are influenced by the chemical and physical properties of the soil, they can be used to directly or indirectly measure or estimate soil properties.

A substantial body of literature has also investigated the potential of using individual PSS to measure or estimate soil properties and characteristics. A few examples of the most recent studies are the use of pulse acoustic wave sensors (Xu et al., 2021), GPR (Algeo et al., 2016), EMI (Badewa et al., 2018; Martini et al., 2017) for mapping soil moisture; GPR (Akinsunmade, 2021), EMI (Pentoś et al., 2022), load cells installed in planter discs (Badua and Sharda, 2023) for soil compaction assessment; GPR for delineation and identification of soil horizons (Ryazantsev et al.,

2022); ion selective electrodes (Silva and Molin, 2018) for soil pH mapping; step-frequency GPR for mapping subsurface morphology (Lombardi and Lualdi, 2019); mobile phone images (Fu et al., 2020), microscopy (Sudarsan et al., 2016), vis-NIR spectroscopy (Conway et al., 2022; Dhawale et al., 2021) for estimating or mapping soil organic matter (OM); vis-NIR spectroscopy (Eitelwein et al., 2022), portable x-ray florescence (Tavares et al., 2023), passive γ -ray spectrometry (Heggemann et al., 2017; Kassim et al., 2021), laser induced breakdown spectroscopy (Tavares et al., 2022), terahertz spectroscopy (Dworak et al., 2020; Sheikh et al., 2022) for measurement or prediction of soil chemical and physical properties.

A comprehensive analysis of the resources listed above, books, book chapters, and reviews dedicated to describing the different PSS techniques, their functionality, and their application in agriculture (Adamchuk et al., 2004, 2017; Adamchuk and Rossel, 2010; Allred et al., 2008; Behera et al., 2022; Gebbers, 2019; Molin and Tavares, 2019; Viscarra Rossel et al., 2011) guided the creation of Table 2.1 that describes the functionality of each PSS technique and the soil properties it can measure directly or indirectly.

Table 2.1 Proximal soil sensing techniques functionality and measured soil properties.

Sensing Technique	Physical Property Measured/Functionality	Soil Property (Direct - D; Indirect - I)
γ -ray spectrometry	Passive measurement of γ -rays emitted by isotopes in soil	D: K_{tot} I: Texture, Water Content, K^+
Neutron Scattering	Active sensing technique (radioactive source) that determines the density of slow neutrons (elastic neutron scattering) or measures γ -rays emitted from soil (inelastic) after soil is bombarded with neutrons	D: Elemental analysis (elastic) I: Soil Water Content (inelastic)
X-ray fluorescence (XRF)	Measures the fluorescence of atoms when irradiated with X-rays	D: Elemental Analysis (better detect elements with higher atomic numbers), Soil Minearology, Heavy Metals I: Texture, Soil Organic Matter (OM), pH, Cation Exchange Capacity (CEC), plant available nutrients

Table 2.1 cont.

Sensing Technique	Physical Property Measured/Functionality	Soil Property (Direct - D; Indirect - I)
Laser-Induced Breakdown Spectroscopy (LIBS)	Uses a laser to create a plasma on the soil surface, then measures the light emitted by the plasma.	D: Elemental Analysis I: CEC, plant-available nutrients, Soil Carbon, Texture
Imaging	Passive light detection and structure-from-motion	D: Texture, Color, Aggregates I: OM
Visible and Near-Infrared Spectroscopy (vis-NIR)	Operates on the principle of measuring the interaction between electromagnetic radiation and soil particles across a spectral range from about 350 nm to 2500 nm	D: Soil color I: OM, Mineral Composition, Soil Water Content, Texture, soil pH, CEC, nutrients, and heavy metals
Raman spectroscopy	Uses a laser beam to excite molecules, then when they return to their original state, with a small percentage of molecules returning to a vibrational excited state that generates photons, which are part of the Raman scattering	D: OM, Carbonates, PO ₄ I: Ca _{tot} , Ca ²⁺
Passive Microwave	Measures the natural microwave emissions from soil surfaces	I: Soil Water Content
Ground Penetrating Radar (GPR)	Analyze the reflected signals (time and strength) from the emitted high-frequency electromagnetic waves, which are related to the dielectric property of the soil	D: Subsurface features and objects I: Bulk Density, Water Content, Permeability, Compacted Layers, OM
Time Domain Reflectometry (TDR), Frequency Domain Reflectometry (FDR), and Capacitance	Travel time of an electromagnetic pulse (TDR), the frequency response of an electrical circuit (FDR), and the charge time of the capacitor (Capacitance). The three techniques rely on the dielectric properties of soil	I: Soil Water Content

Table 2.1 cont.

Sensing Technique	Physical Property Measured/Functionality	Soil Property (Direct - D; Indirect - I)
Galvanic Contact Resistivity (GC) and Electromagnetic Induction (EMI)	Measure soil apparent electrical conductivity (EC_a) using direct contact electrodes (GC) or electromagnetic fields (EMI)	I: Texture, Bulk Density, Water Content, Permeability, Salinity, pH, CEC
Cosmic-Ray Neutron	Passive counting of neutrons backscattered out of the soil	I: Soil Water Content
Thermocouples and Thermal Radiation	Measure soil temperature using the thermoelectric effect (thermocouples) or emitted thermal radiation	D: Temperature
Electrochemical [Ion-Selective Electrodes (ISE), Ion-Selective Field Effect Transistors (ISFET)]	Generates voltage or current upon a specific interaction between the sensor and a chemical solution	D: pH, K^+ , NO_3^- , Na^+ , and other ions
Mechanical (cone index, instrumented shafts, etc.)	Measures resistance of soil to penetration by a probe or tip or force required to pull a tillage/seeding implement	D: Soil Mechanical Resistance I: Soil Compaction, Bulk Density, Soil Strength
Acoustic	Uses sound waves to detect variations in soil texture and compaction	I: Soil texture, Compaction Layers, Soil Water Content
Gas Analysis	Detects and quantifies specific gases (e.g., CO_2), byproducts of microbial activity in soil	I: Biological activity, Soil Health

The information presented in Table 2.1 highlights the potential of PSS to map soil properties. However, the suitability of the sensing techniques for measuring chemical and physical properties differs. For instance, a soil scientist who wishes to determine the boundaries of different soil types

would be better equipped with GPR, GC/EMI, and mechanical or acoustic sensors rather than electrochemical sensors. On the other hand, an agronomist evaluating the soil pH levels of a field would obtain more accurate measurements using an electrochemical sensor (directly measures pH) than a GC/EMI (can be used to predict pH indirectly). In addition, many sensing techniques are limited in their use due to the need for sample preparation (e.g., XRF) or because their measurement is influenced by other soil properties (e.g., high clay and soil moisture attenuate the signal of GPR, which affects the readings and effective depth penetration; high EC_a regions identified by EMI and GC can be due to soil texture, moisture, or salinity).

Knowledge of the chemical and physical properties of soils is necessary to employ PA management strategies. While individual PSS can provide insights on the within-field variability that could be used to guide sampling or measure/predict some soil properties, “There appears to be no single sensor that can capture all relevant soil properties at the same time” (Gebbers, 2019). Therefore, the combination of data from multiple sensors (data fusion) can potentially overcome this challenge (Adamchuk et al., 2011; Gebbers, 2019; Viscarra Rossel et al., 2011).

2.3 PSS data fusion for within-field soil characterization: challenges and current developments

The remainder of this chapter will focus on describing the concept of data fusion and its use in agriculture, as well as the latest advances in fusing on-the-go and stop-and-go PSS for soil characterization.

The process of combining different sensing techniques is known as data fusion and is defined as “A process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance”(White, 1991). Data fusion can be characterized based on its levels of abstraction, which are often classified into three (Dalla Mura et al., 2015; Pantazi et al., 2020): raw data (low), feature (medium), and decision (high) level (Figure 2.1).

The use of DF for PA purposes is not a new concept. After searching the keywords “data fusion” and “agriculture” in Google Scholar and Scopus databases, Barbedo (2022) found almost 400 papers. The same author assessed the quality of the manuscripts and selected 119 documents to be analyzed. The following are some of the applications of DF in agriculture identified by Barbedo (2022): autonomous equipment guidance, fruit detection, plant phenotyping, crop monitoring,

disease detection, delineation of management zones, yield and biomass prediction, and soil properties/characteristics prediction.

According to Barbedo (2022), significant advances in DF for agriculture have been made. However, some challenges and limitations still need to be addressed, such as model overfitting due to spatial and temporal autocorrelation and the lack of research to develop appropriate methodologies to combine disparate spatial datasets.

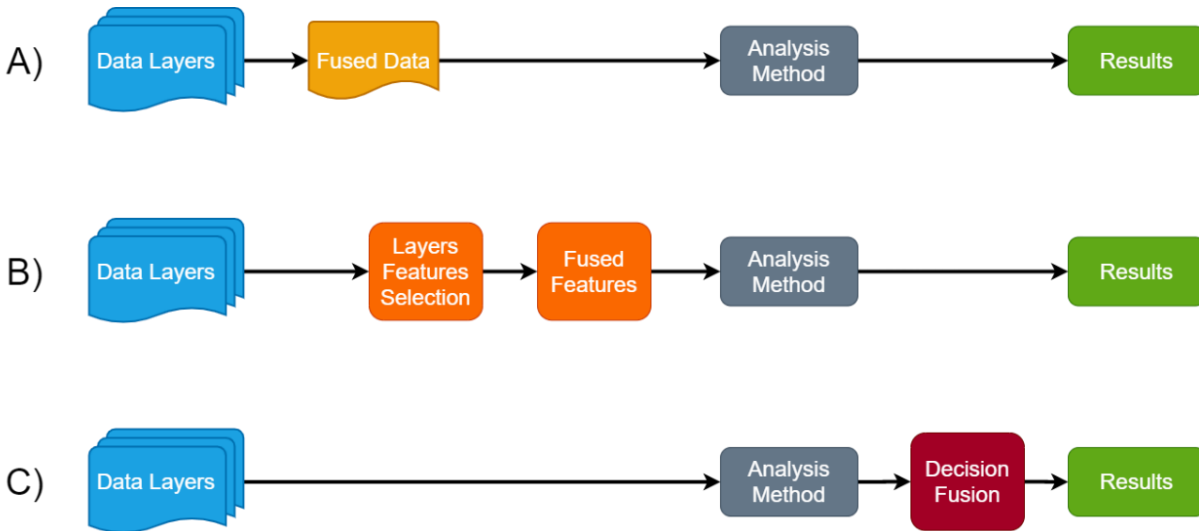


Figure 2.1 Data fusion levels. A – Data/signal level; B – Feature level; C – Decision Level.

Recent advances in on-the-go or stop-and-go PSS data fusion for soil characterization

Per the definition of PSS given by Viscarra Rossel et al. (2011), this sub-section focused on studies developed with PSS that are capable of collecting data in the field with an on-the-go or stop-and-go operation. Also, except for topography data, DF of PSS with other sensing technologies (e.g., remote sensing, plant proximal sensing, etc.) was excluded as it was considered out of the scope of the thesis.

White's (1991) general definition of DF can be contextualized to the field of PSS as follows: an individual PSS can respond to multiple environmental variables, but the combination of signals and responses from different PSS techniques may complement each other and improve soil mapping accuracy (Adamchuk et al., 2011; Gebbers, 2019).

The study developed by Conway et al. (2022) provides an example of how other environmental variables can affect the data originating from a single PSS technique. These authors evaluated the effect of volumetric water content on the predictions of OM when utilizing a vis-NIR spectrometry-based commercial sensor (SmartFirmer, Precision Planting, Tremont, IL, US).

Conway et al. (2022) concluded that the sensor could measure relative differences in OM in the soil, but the accuracy of the prediction of this soil property was affected by soil moisture. Based on these results, one could apply the concept of PSS DF and add a capacitance sensor, proven suitable for mapping soil water content (Adamchuk et al., 2009), to the tested commercial sensor, which could help reduce the moisture effect on the OM predictions.

While the DF of PSS in PA is not a new subject, research and development in this topic continue to grow with the development of computational power, analytical processes, and sensing technologies. Eleven studies that provide an overview of the latest advances and DF techniques for PSS were carefully selected and analyzed (Table 2.2).

These eleven PSS DF studies were categorized based on acquisition mode (multi-sensor platform or sequential acquisition), method of operation (on-the-go or stop-and-go), sensing techniques, fusion technique [machine learning (ML) or geostatistical-based], and contribution to knowledge (development of new sensors/platforms, evaluation of the potential to predict soil characteristics or delineation of management zones).

The main contribution of knowledge from three of these eleven studies was the development of new multi-sensor platforms. The data acquisition for four of the eleven studies was performed sequentially (i.e., different field operation activities); for each PSS data collected, there was a separate operational cost. Thus, the use and development of multi-sensor platforms can reduce costs by collecting all PSS data in one operation. Also, data processing can be simplified if the observations from different sources are co-located within the data-logging software.

Table 2.2 Examples of studies on proximal soil sensing data fusion for soil characterization (a hyphen indicated the information was not available or not applicable for that study)

Reference	Acquisition Mode	Operation	Sensing Techniques	Fusing technique	Contribution to Knowledge	Studied Soil Characteristics
Rezaei et al. (2022)	Multi-sensor	Stop-and-go	GC, Dielectric, Acoustic, and CI	ML	Development	Soil physical properties
Cao et al. (2024)	Multi-sensor	On-the-go	vis-NIR and RGB (camera)	ML	Development	Organic matter
Ahrends & Lajunen (2021)	Multi-sensor	On-the-go	GC, NIR, and topography	ML	Management zone	-
Vogel et al. (2022)	Multi-sensor	On-the-go	GC, Red and NIR reflectance, and ISE	ML + Geostatistics	Prediction	Base neutralizing capacity
Tavakoli et al. (2022)	Multi-sensor	On-the-go + Stop-and-go (pH)	GC, NIR, γ -ray, and ISE	Geostatistics	Prediction	Soil chemical and physical properties
Adamchuk et al. (2023)	Multi-sensor	Stop-and-go	FDR, soil-metal friction, air non-dispersive infrared, air permeameter, and CI	-	Development	Soil Health Indicators
Pei et al. (2019)	Multi-sensor	Stop-and-go	GC, vis-NIR, and CI	ML	Prediction	Soil moisture, chemical, and physical properties
Castrignanò et al. (2018)	Sequential	On-the-go	GPR, EMI, and topography	Geostatistics	Management zone	-

Table 2.2 cont.

Reference	Acquisition Mode	Operation	Sensing Techniques	Fusing technique	Contribution to Knowledge	Studied Soil Characteristics
De Benedetto et al. (2019)	Sequential	On-the-go	GPR, EMI, and topography	Geostatistics	Prediction	Soil water content
Koganti et al. (2023)	Sequential	On-the-go	γ -ray and topography	ML + Geostatistics	Prediction	Peat depth
Ji et al. (2019)	Sequential	On-the-go + Stop-and-go (vis-NIR)	EMI, vis-NIR, γ -ray, topography	ML	Prediction	Soil chemical properties

CI – Cone Index; EMI – Electromagnetic Induction; FDR – Frequency Domain Reflectometry; GC – Galvanic Contact; GPR – Ground Penetrating Radar; ISE – Ion Selective Electrode; ML – Machine Learning; RGB – Red-Green-Blue image; vis-NIR – Visible-Near Infrared spectrometry.

Among the different sensing techniques explored, EC_a is the most commonly measured physical property shown in nine studies but measured using different techniques: galvanic contact (GC), frequency domain reflectometry (FDR), and electromagnetic induction (EMI). While soil optical reflectance is another common technique (6 studies), sensors such as acoustic, RGB imaging cameras, soil-metal friction, air non-dispersive infrared, and air permeameter only show once across the different studies. These least frequent sensing techniques are not commonly used in PA soil characterization but highlight that new techniques have been constantly developed to improve and expand soil mapping. Sensing techniques with more than one occurrence in Table 2.2 are often used in PA.

Ground-penetrating radar (GPR), shown only twice in Table 2.2, is a well-established geophysics technique that has recently increased interest in PA. However, the complexity of the GPR signal and processing remain as barriers to further adoption of this technique. Ion-selective electrode (ISE), also only observed in two of the eleven studies, is not new to PA. Some commercially available sensor platforms utilize ISE (Adamchuk et al., 2005). However, difficulties during the on-the-go data collection and analysis (Schirrmann et al., 2011) could explain the lower popularity of this technique.

Although the eleven selected studies provide an overview of the latest advances in PSS DF, they do not cover several combinations of sensing techniques. Two sensing techniques that have recently gained popularity in PA are portable X-ray fluorescence spectroscopy and laser-induced breakdown spectroscopy. However, the limitation of most studies involving these techniques' DF is that they have been developed in a lab environment (Tavares et al., 2021a, 2021b).

Regarding the fusion technique, machine learning algorithms were found to be one of the most popular techniques used for PSS DF. Among the ML regression algorithms used by the studies in Table 2.2, multiple linear regression (e.g., Rezaei et al., 2022) is one of the simplest and convolution neural networks (e.g., Cao et al., 2024) one of the most complex. Also, representatives of ML classification algorithms are present in Table 2.2; k-means and fuzzy c-means unsupervised clustering algorithms were used by Ahrends & Lajunen (2021) to delineate management zones. Often, the performance of multiple ML algorithms for PSS DF was evaluated in these studies (e.g., Ji et al., 2019)

While ML is the most popular PSS DF technique, multivariate geostatistical methods based on a linear model of coregionalization (Journel & Huijbregts, 1976) and block cokriging (Chilès &

Delfiner, 2012; Goovaerts, 1997) have been suggested as one of the most appropriate and ‘statistically sound’ data fusion techniques for PSS DF due to its probabilistic framework’s ability to measure uncertainty (Castrignanò & Belmonte, 2023). This approach has been used in two of the eleven studies in Table 2.2; one focused on the definition of homogenous regions within a field (i.e., management zones - Castrignanò et al., 2018), and the other the estimation of soil water content (De Benedetto et al., 2019). However, this geostatistical approach has significant disadvantages that would limit its implementation in DSS for soil management practices: it is computationally intensive and requires the assumption of second-order stationarity. In most cases, the latter assumption is not satisfied for PSS data from agricultural fields (i.e., second-order stationarity: the means of a random variable is constant at all locations of the field, and the covariance depends only on the separation distances between observations - Chilès & Delfiner, 2012). To address this issue, a Coregionalization Analysis with a Drift (CRAD) method has been developed (Pelletier et al., 2009a, 2009b); in its “phase I,” a trend (drift) is estimated and removed from the data, which residuals are then conceptualized and modeled as second-order stationary processes in “phase II.” CRAD’s phase II coregionalization, and spatial and non-spatial correlation analysis are carried out. Although CRAD is a statistically sound method that overcomes one of the limitations of the approach proposed by Castrignanò & Belmonte (2023), it is still computationally expensive, and outputs must be analyzed by an expert, resulting in limited potential to be implemented in a DSS.

Some of the studies in Table 2.2 also used a combination of geostatistics and ML. Most utilized geostatistics to interpolate all the data to a common raster and later utilized the raster data to train ML algorithms. For example, Vogel et al. (2022) utilized ordinary block kriging to interpolate all point-based sensor data to a raster file, later utilized to train a stepwise multivariate linear regression.

Regarding the DF performance compared to individual sensors, studies such as Rezaei et al. (2022), which evaluated a few combinations of the four sensing techniques, have indicated that the fusion of all data sources significantly improved the prediction of soil physical properties. Similarly, Ji et al. (2019) concluded that the fusion of all sensors often performed best when predicting soil chemical properties. Such a conclusion is consistent in all of the other nine studies, while sometimes individual sensors emerge as the best estimator for a few soil properties and DF for others Pei et al. (2019).

2.4 PSS Data Processing

Raw proximal soil sensor data frequently requires substantial preprocessing to assure data quality and applicability to precision agriculture needs. Yield monitor data is an excellent example of data processing methodology established and studied by many researchers (Simbahan et al., 2004; Ping & Dobermann, 2005; Sudduth & Drummond, 2007). A similar research emphasis is needed for new and emerging proximal soil sensing solutions.

Generic strategies focused on spatial data processing have been presented in the past (Shekhar et al., 2003; Spekken et al., 2013; Singh and Lalitha, 2018). Although relatively few studies focused on PSS, Maldaner et al. (2022) developed an automated framework that took raw data as input and output an interpolated map. To date, many studies that require processing of PSS data often refer to yield monitor frameworks (e.g., Rodrigues et al., 2015).

When differences in the spatial location of the measurements taken by different instruments are insignificant or non-existent, the different DF levels can be directly applied and evaluated. For example, using a polyethylene cup of 31 mm in diameter, Tavares et al. (2021) assessed the performance of an array of sensors (diffuse reflectance, x-ray fluorescence, and laser-induced breakdown spectroscopy) and DF in predicting soil fertility attributes. Since the data from the three sensors were collected within a small area bounded by the cup's dimension, it was considered that all measurements were collocated and shared the same support. The term “support” here refers to the region (area, volume, shape, and direction) associated with each collected data value (Chilès & Delfiner, 2012).

However, PA data collected in the field using different instruments and techniques may present different support, spatial locations, and data types (vector and raster). Figure 2.2 exemplifies one of those cases. Unlike the case presented by Tavares et al. (2021), in which all data was collocated, applying DF to the Figure 2.2 dataset would be challenging and require the data to be brought to the same support first. Some researchers have used vector support to co-locate the data to overcome this challenge and perform DF.

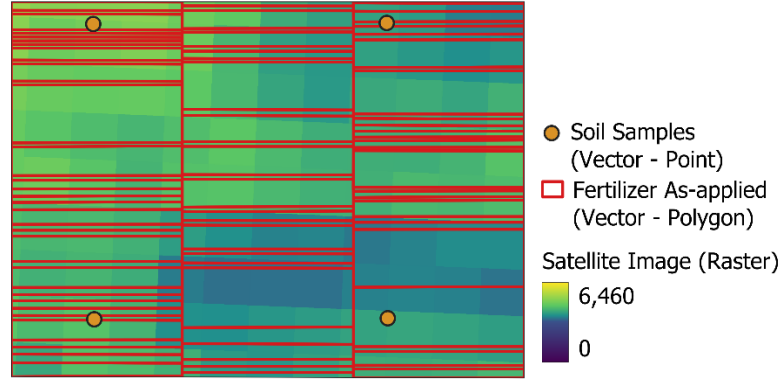


Figure 2.2 Example of a precision agriculture dataset with different support, spatial location, and data type.

Adamchuk & Wang (2007) developed DM_Comp, a software with the main function of “collocating multiple self-generated data layers in (1) points that belong to one of these layers (using user-defined fixed radius or nearest neighbor averaging) and (2) centers of rectangular grid cells (using grid cell averaging)”. To extract PSS data (point vector layer with continuous samples taken over a transect) to soil sampling locations (sparse point vector layer), Ji et al. (2019) used DM_comp, and the resulting collocated dataset was used to evaluate the potential of DF for measuring soil properties. Anastasiou et al. (2019), when collocating on-the-go data using a crop canopy radiometric sensor (point vector layer) and PSS (point vector layer), migrated the radiometric data to the soil sensor sample location by using only the nearest neighbor. Based on these references, collating the data into a chosen vector support has proved to be a possible solution to the DF for non-collocated datasets. Instead of using vector support, others utilized a raster by simply interpolating the data for different sensors (Vogel et al., 2022).

The research developed in the current thesis involved further evaluation of PSS combinations while methodologies and procedures to overcome some PSS DF challenges were proposed. A PSS data co-location procedure that includes general data pre-processing (e.g., removal of outliers) was proposed. Also, a methodology for DF of PSS that generates soil thematic maps was proposed and evaluated. Both proposed methodologies are automated and allow for easy implementation and use of the data and information provided by PSS.

Connecting Text to Chapter 3

The previous chapter presented the state-of-the-art use of PSS data fusion in soil characterization. Proper pre-processing of the data before performing the fusion is essential to guarantee the quality and accuracy of the prediction results. Therefore, through **Chapter 3**, a framework for processing PSS data was proposed, and based on a literature review, the most relevant methodologies were evaluated. A focus on procedures and methodologies that require minimal input from users was taken, facilitating the use and embedding of the framework in DSS and addressing some PA adoption barriers highlighted in **Chapters 1** and **2**.

The steps and methodologies presented in **Chapter 3** were necessary to develop **Chapters 5** and **6**. The content in **Chapter 3** was presented and published in the proceedings of the 15th International Conference on Precision Agriculture in Minneapolis, Minnesota, USA, in 2022.

Publication:

Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2022). Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes. In Proceedings of the 15th International Conference on Precision Agriculture (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture. <https://www.ispag.org/proceedings/?action=abstract&id=8792&title=Optimization+of+Batch+Processing+of+High-density+Anisotropic+Distributed+Proximal+Soil+Sensing+Data+for+Precision+Agriculture+Purposes>

Abbreviations for Chapter 3

Abbreviation	Definition
AO	Adjusted Outlyingness Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EC _a	Apparent Electrical Conductivity
EMI	Electromagnetic Induction
GC	Galvanic Contact
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPR	Ground Penetrating Radar
IDW	Inverse Distance Weighting
ISO	International Organization for Standardization
L-BFGS-B	Limited Memory Algorithm for Bound Constrained Optimization
NN	Nearest Neighbors
OK	Ordinary Kriging
PSS	Proximal Soil Sensing
R ²	Coefficient of Determination
RMSE	Root Mean Squared Error
RTK	Real-Time Kinematic
UK	Universal Kriging
UTM	Universal Transverse Mercator

Chapter 3: Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes

Felippe H. S. Karp, Viacheslav Adamchuk, Alexei Melnitchouck, Pierre Dutilleul

Abstract

The amount of spatial data collected in agricultural fields has been increasing over the last decade. Advances in computer processing capacity have resulted in data analytics and artificial intelligence becoming hot topics in agriculture. Nevertheless, the proper processing of spatial data is often neglected, and the evaluation of methods that efficiently process agricultural spatial data remains limited. Yield monitor data is a good example of a well-established methodology for data processing that could be used as a guide to determine data processing strategies. However, data processing methods for proximal soil sensors (PSS) are not as well-known as yield, even though sensors are widely used in precision agriculture and their data often applied in predictive models for soil spatial variability characterization. The main objective of this study was to identify suitable methodologies for processing PSS data and applying them to a particular dataset. It was determined that properly processing any spatial dataset required the following steps: (1) data projection, (2) position offset correction, (3) global and (4) local filtering, and (5) interpolation. Based on a literature review, the most suitable methods for each step are listed and discussed, and frameworks are proposed. These methods were applied to 4 different types of PSS data (gamma ray spectrometry, ground-penetrating radar, galvanic contact and electromagnetic induction soil apparent electric conductivity). To evaluate the accuracy of each processing framework, a cross-validation procedure was used. Overall, the proposed batch processing framework improved the value of PSS data by highlighting spatial variability in the field that was previously masked by the presence of erroneous data. Also, when performing an analysis in the measurement's maps, discrepancies were found between the raw versus processed data, thus, emphasizing the need for properly processed PSS datasets.

Keywords. Filtering, Interpolation, Proximal Soil Sensors, Processing Evaluation

3.1 Introduction

The amount of spatial data collected in agricultural fields has increased over the last decade, and with the increase in computational power, developments in machine learning, and data analytics, great advances have been made in extracting practical information from these data

(Colaço et al., 2021; Fulton and Port, 2018; Robertson et al., 2021; Willam et al., 2022). However, the importance of proper processing of spatial data is often neglected; this may be related to the lack of research focused on testing and/or developing different processing methodologies and frameworks to handle this type of data for precision agriculture practices.

Yield monitor data is a good example of a well-established processing methodology that has been discussed and studied by many researchers (Arslan and Colvin, 2002; Blackmore and Moore, 1999; Leroux et al., 2018; Lyle et al., 2014; Menegatti and Molin, 2004; Ping and Dobermann, 2005; Simbahan et al., 2004; Sudduth and Drummond, 2007; Sun et al., 2013; Vega et al., 2019). Alternatively, processing strategies for proximal soil sensors (PSS) are not as well-established as those for yield. To our knowledge, very little research exists on this topic, even though PSS data is widely used in research and often it is used as a guide to determine management zones within agricultural fields (Adamchuk et al., 2004; Ji et al., 2018; Saifuzzaman et al., 2019; Schepers et al., 2004).

Some researchers provided general strategies that focused on spatial data processing (Shekhar et al., 2003; Singh and Lalitha, 2018; Spekken et al., 2013) but they were not tested on PSS data. Others, like Maldaner et al. (2022), subjected PSS data to their methodology and obtained promising results, but still, the main focus of the processing is on spatial data filtering. Overall, studies involving the use of PSS often refer to yield monitor frameworks when processing soil sensing data (e.g., Rodrigues et al., 2015). Thus, there is a need to establish specific processing algorithms for PSS data, and considering the large amount of information already collected, special focus should be given to methodologies that allow the joint processing of multiple files (sensors data). Therefore, the main objective of this study was to identify suitable methodologies for batch processing PSS data and apply them to a particular dataset.

3.2 Materials and Methods

The literature review that we performed on processing agricultural spatial data led to the definition of important steps for processing on-the-go sensing data. These steps are detailed below. They are presented in the order in which they should be applied. It must be noted that in listing, testing, and evaluating the available methodologies for each of the processing steps, more attention was given to methodologies using the least possible variables, computational power, and user input. Because PSS data often presents an anisotropic distribution with a higher density of data

acquisition within a pass than between passes, more focus was also given to methods that consider the existence of within-row variability.

Local Projection

Most of the data acquired by PSS are exported from sensors systems with non-projected coordinates (geographic latitude and longitude). Converting this data to a projected coordinate system is often needed for the calculation of distances, which are necessary for other procedures involved in data processing, e.g., interpolation. Universal Transverse Mercator (UTM) is one of the most used projected coordinate systems. Sometimes, more localized systems, such as the Modified Transverse Mercator, are also used for specific locations.

However, a known issue with projection systems is distortion. In addition, a field might be located over two projection zones. In this case, one zone should be chosen to continue with the data processing, even though an error would be associated with this decision. To avoid and/or reduce these issues, we propose the use of a custom localized Cartesian coordinate system, described under the International Organization for Standardization 12188-1 Annex A (ISO, 2010). Equations 1 and 2 show the calculation for the conversion factors proposed in the standard above mentioned, while Equations 3 and 4 the conversation formulas

$$F_{lon} = \frac{\pi}{180^\circ} \left(\frac{a^2}{\sqrt{a^2 \cos^2 \varphi + b^2 \sin^2 \varphi}} + h \right) \cos \varphi \quad (2.1)$$

$$F_{lat} = \frac{\pi}{180^\circ} \left(\frac{a^2 b^2}{(a^2 \cos^2 \varphi + b^2 \sin^2 \varphi)^{\frac{3}{2}}} + h \right) \quad (2.2)$$

$$X = (Long - Long_{min}) * F_{lon} \quad (2.3)$$

$$Y = (Lat - Lat_{min}) * F_{lat} \quad (2.4)$$

where F_{lat} , F_{lon} are the location-specific conversion factors, in meters per degree, for latitude and longitude, respectively; φ is the latitude location for the center of the field, in degrees; h is the average elevation for the field above the ellipsoid, in meters; a is the semi-major axis of the ellipsoid, in meters; and b is the semi-minor axis of the ellipsoid, in meters; X and Y are the easting and northing projected coordinates, in meters; $Long_{min}$ and Lat_{min} are the geographic lowest values of longitude and latitude observed in the dataset, in degrees; $Long$ and Lat are the geographic longitude and latitude, in degrees.

Position Offset

Projection is not the only source of error for the geographic placement of PSS data. Often during data collection, the Global Navigation Satellite System (GNSS) receiver cannot be placed right above the sensor. A standard practice is to place it at a high location free of obstructions , (e.g., above the cabin of the vehicle), but still centered with the sensor. However, this procedure creates an offset between the data collected and the GNSS location recorded. Also, depending on the type of receiver used during data collection, a known pass-to-pass error can also affect the uncertainty of the geolocation of the data. Any such displacement in the data must be corrected. Even though some sensors' collection systems allow an adjustment of the offset distance between the GNSS receiver and sensor, this option is often neglected, or the adjustment is forgotten when making changes to the collection settings (e.g., using another vehicle). Thus, an automatic procedure was considered and tested to calculate and apply this correction to the data.

The methodology proposed by Lee et al. (2012) was employed for this purpose. Their methodology uses phase correlation analysis to automatically calculate the necessary offset. This analysis includes two main steps: rasterization of the points and application of the phase correlation analysis. According to the authors, the pixel size used in the first step can affect the results of the analysis. Thus, they proposed a methodology that automatically determines this value. After determining the best pixel size, a range of offset values was applied to the data, the phase correlation analysis performed, and the phase correlation coefficient obtained. The offset value representing the highest phase correlation coefficient was selected to be applied to the data. To our knowledge, this methodology has not been tested on the identification of offset values for PSS, but is employed in the yield processing software Yield Editor 2.0 (Sudduth et al., 2012), to estimate the delay for yield and moisture sensors.

Other methodologies (e.g., Chung et al., 2002) are available to estimate the offset mentioned above. However, based on the conclusions of Lee et al. (2012), phase correlation analysis produces results similar to those of the geostatistical method, while reducing the processing time, which is an important variable when batch processing data.

Operational, Global, and Local Filtering

Assuming that any data displacements or projection distortions are corrected or reduced, sensors' readings may still be subjected to random factors causing erroneous observed values, or outliers (e.g., a high reading by an electromagnetic induction sensor due to the presence of a metal

piece, temporary malfunctioning of the sensor because of field operations). Ignoring outliers in the data can lead to higher uncertainty on interpolation or predictive processes, and affect management decisions made based on the data. Many papers that focused on yield monitor data processing highlighted the issues in final thematic maps and decision-making procedures that may arise from the presence of erroneous measurements (Leroux et al., 2018; Lyle et al., 2014; Maldaner et al., 2022; Ping and Dobermann, 2005; Sudduth et al., 2012).

The literature on yield monitor data processing also emphasizes the need to perform not only a global outlier detection and removal step, but also a local analysis aimed at comparing an observation to a pre-determined neighborhood. In the conditions of application of Geostatistics, it is assumed that an observation is spatially correlated with those of its neighbors. Thus, if an observation is distinct from neighborhood values, that observation is considered a spatial outlier.

Accordingly, based on the developed frameworks for yield monitor data processing, we propose three major steps for global and local filtering of PSS data: operational (removal of operation related patterns, e.g., maneuvers, changes in travel speed), global-statistical (removal of erroneous and extreme values), and local-statistical (removal of spatial outliers).

Operational Filtering

Changes in travel speed and direction (heading), stops, etc. are inevitable during field data collection. However, these operational factors become part of the data and can affect its quality. Co-located points can occur due to a stop or error on the sensor's logger and affect the data interpolation. Maneuvers can cause the tilt or loss of contact between the sensor and soil, also generating erroneous measurements.

The removal of such operational factors is important to increase overall data quality. Four filters are proposed to remove such data points. For the operational filters involving the setting of thresholds for the identification of extreme values (maneuvers and travel speed), a non-parametric methodology was adopted to reduce the need for user inputs. This methodology uses the identification of outliers proposed by Hubert and Van Der Veen (2008), which identifies extreme values by calculating an adjusted outlyingness index (AO) and cutoff, both based on a skewness-adjusted boxplot. The four filters can be detailed as follows:

1. Co-located points: any points with the same coordinates are removed.

2. Headland: a negative 5-m buffer is applied to an auto-generated field's boundary and any observation outside the buffered area is removed. A pre-generated boundary could be used in this step but was avoided because it would be another user input.
3. Maneuvers: differences in calculated heading direction between consecutive points are computed. In sequence, the adjusted boxplot, AO, and threshold are obtained, and any observation exceeding the threshold is considered a maneuver and is removed.
4. Travel speed: not every data collection system exports information on travel speed or frequency of data collection, but the latter is usually constant, allowing the use of distances between consecutive points to estimate changes in travel speed. Thus, a procedure similar to the maneuver filtering is adopted, while the Euclidean distance between consecutive points is calculated and used in the analysis for outliers. Eventually, points that are too close (lower travel speed) or too far (higher travel speed) are removed.

Global and Local Statistical Filtering

Two recently published methodologies for spatial data filtering include global and local steps and they were selected for testing. Other methodologies are described in the literature, but the two evaluated here consider spatial outliers in one direction (the data collection direction) and in all directions horizontally (Figure 3.1).

One methodology was proposed by Leroux et al. (2018), and can be summarized in 11 steps:

1. Global statistical filtering: based on Hubert and Van Der Veecken's (2008) AO, cutoff, and skewness-adjusted boxplot (described under Operational Filtering).
2. Median calculation for unidirectional and omnidirectional neighborhoods from each observation within two-pass widths.
3. *Outlyingness* metric calculation: $h_A = f_A - g_A$, where f_A is the observation value, and g_A is the median of the neighborhood values for the two types of neighborhoods.
4. Repeat Steps 2 and 3 for two more neighborhood settings: three and four times the pass width.
5. Average the results of unidirectional and omnidirectional *outlyingness* metrics over the three neighborhood settings and do a biplot (unidirectional vs. omnidirectional *outlyingness* metrics).

The next steps no longer use spatial information (coordinates), but the distribution observed in the biplot obtained from the unidirectional vs. omnidirectional *outlyingness* metrics.

6. Construct the matrix of distances between points in the biplot, and estimate the most frequent distance (ϵ) using Kernel Density Estimation.
7. Determine the number of neighbors within ϵ for each observation, estimate their frequency by Kernel Density Estimation, and identify the first local minimum.
8. Apply a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996) to the biplot, using ϵ as neighborhood radius and the number of local minimum found in Step 7 as the minimum number of points.
9. Extract the cluster containing normal observations (excluding outliers).
10. Refine the detection of outliers by repeating Steps 2–9 and comparing the observations flagged as outliers with their neighborhood after removing the previously flagged outliers.
11. Outliers flagged twice are considered spatial outliers and removed from the data.

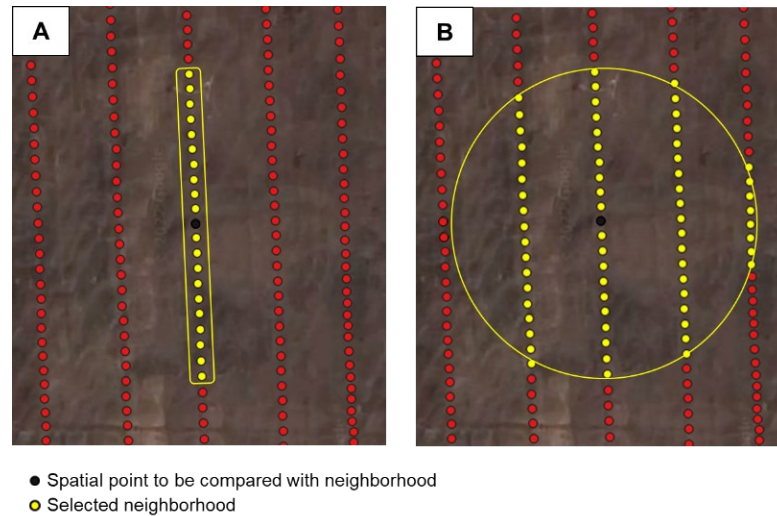


Figure 3.1 Illustration of the definitions of unidirectional (within-row) neighborhood (A) and omnidirectional (within- and between-rows) neighborhood (B)

The second methodology that was reviewed in greater detail was proposed by Maldaner et al. (2022) and used a simpler approach than the first while requiring more user inputs. For the global and local filters, upper and lower bounds are defined as the median plus and minus a coefficient times the median. The coefficient can be different for the global and local filters. At the local

filtering step, the median is calculated for the observations in the neighborhood, in a way similar to Step 2 for the first methodology but by using a radius equal to 2.5 times the pass width. An observation value found to be below the lower bound or above the upper bound is flagged as an outlier. This methodology removes outliers following the sequence: global, local unidirectional, and local omnidirectional.

Both methodologies use the distance between two adjacent passes (pass width) to determine the neighborhood for local filtering. This distance can be obtained easily using a Geographic Information System (GIS). However, this requires another user input. Considering that the position offset methodology of Lee et al. (2012) calculates a pixel size that corresponds to 50 to 80% of the pass width, we propose 2 times the pixel size estimated at the position offset step to be used as an estimate of the pass width.

Interpolation

Finally, in the process of using PSS data as a decision-making tool or for any other application in precision agriculture, point data must be converted to a map format. Interpolation is used for this. Many interpolation methods are available and can be applied. The most common for agricultural data are: ordinary kriging (OK), universal kriging (UK), inverse distance weighting (IDW), and nearest neighbors (NN). After selection of the most appropriate filtering, the filtered PSS data were submitted to these four interpolation procedures.

Even though normality is not an assumption about the data distribution for geostatistical analysis, a common practice is to test for normality and if rejected, to apply a data transformation. This can help improve the results of the OK and UK procedures, as they are based, in different ways, on mean values (Isaaks and Srivastava, 1989; Schossler et al., 2019). We, therefore, recommend that for interpolation based on kriging, the data distribution be tested for normality, and when rejected, a Box-Cox transformation (Box and Cox, 1964) be used.

The IDW interpolation requires a maximum number of neighbors and a weight applied to distances (Shepard, 1968). Either parameter can be user-defined or automatically optimized. A common approach is to divide the dataset into 'test and train' subsets and evaluate different options and combinations of parameter values in relation to prediction errors. The combination of parameter values providing the smallest prediction error is selected (Barbulescu et al., 2020). This procedure can be computationally costly and time-consuming. Thus, for datasets with >1000 points (size of most PSS datasets), a threshold distance is recommended to reduce the

computational cost for IDW interpolation (Hengl, 2009). We also propose the use of a Limited Memory Algorithm for Bound Constrained Optimization through an L-BFGS-B algorithm (Byrd et al., 1995), to obtain optimized parameters for IDW interpolation.

Evaluation

Data from four PSS instruments were used to evaluate the steps and different methodologies described above. Table 3.1 presents the different sensors, characteristics of the data collection (pass swath and distance between consecutive points), data density, and variables evaluated. The data from galvanic contact apparent electrical conductivity (GC) and ground penetrating radar (GPR) are from a 43-ha field, while the γ -ray and electromagnetic induction apparent electrical conductivity (EMI) come from an 82-ha field.

Raw data was standardized to zero mean and unit variance to facilitate interpretation of results. The evaluation of position offset and filtering steps (global and local) took place by comparing the changes in the estimated parameters of fitted variogram models and the average root mean squared error (RMSE) from a 10-fold cross-validation using OK interpolation.

To compare interpolation methods, data distribution normality was tested first, and a Box-Cox transformation was applied as required. Transformed data was used for each interpolation method, including IDW and NN in addition to OK and UK. No Box-Cox transformation was applied before interpolation because, by definition, a Box-Cox transformation modifies the data distribution, which would affect the comparison among methodologies in their efficiency to remove global and local outliers.

For each interpolation method, a 10-fold cross-validation was performed, and the RMSE and the coefficient of determination (R^2) between predicted and observed values were reported. For any kriging procedure in this study, three models (exponential, spherical, and Gaussian) were fitted by weighted least squares to the empirical variogram. The best variogram model was selected based on the error sum of squares. Evaluation of frameworks and methodologies was performed with customized scripts and functions written in the R language (R Core Team, 2021).

Table 3.1 Information for the different proximal soil sensors used for the evaluation of processing methodologies

Sensor Model	Manufacturer Provider	Technique	Swath (m)	Point Distance (m)	Samples·ha⁻¹	Nb. Obs.	Variables
EM38-MK2	Geonics Limited (Mississauga, Ontario-CA)	EC _a - Electromagnetic Induction (EMI)	25	4.1	108	8852	EC _a Deep (0-1.5 m)
SIR-4000 (400 MHz Antenna)	GSSI (Nashua, New Hampshire- USA)	Ground Penetrating Radar (GPR)	34	5	63	2714	Instantaneous Amplitude (0-0.1 m)
SoilOptix	SoilOptix (Tavistock, Ontario-CA)	Passive Gamma- Ray (γ -ray)	12	12	61.5	5048	Count Rate
Veris 3100	Veris Technologies (Salinas, Kansas- USA)	EC _a - Galvanic Contact (GC)	18	3.3	147	6331	EC _a Shallow (0-0.3 m)

3.3 Results and Discussion

Table 3.2 Geostatistical analysis results (Nugget, Sill: variogram model parameter estimates) and root mean squared error (RMSE) from ordinary kriging cross-validation for position offset and filtering procedures for each proximal soil sensor. Data is standardized to zero mean and unit variance. Bold and red-colored numbers represent the lowest values, and dashes indicate that the procedure was not applied or no point was removed.

Process	EMI ^a – 0-1.5 m			GPR ^b – 0-0.1 m			γ -Ray – Count Rate			GC ^c – 0-0.3 m		
	Nugget	Sill	RMSE	Nugget	Sill	RMSE	Nugget	Sill	RMSE	Nugget	Sill	RMSE
Raw	0.050	0.802	0.190	0.658	0.949	0.817	0.896	0.970	0.975	0.141	1.167	0.246
Offset Applied	0.047	0.802	0.192	0.633	0.942	0.814	-	-	-	-	-	-
Operational	0.034	0.517	0.203	0.347	0.680	0.666	0.855	0.989	0.982	0.102	1.150	0.233
Global Leroux et al. 2018	-	-	-	-	-	-	0.857	0.983	0.980	-	-	-
Local Leroux et al. 2018	0.004	0.138	0.092	0.122	0.314	0.424	0.575	0.688	0.797	0.015	0.534	0.134
Global Maldaner et al. 2022	0.012	0.444	0.179	0.262	0.444	0.565	0.748	0.831	0.904	0.056	0.665	0.203
Local Maldaner et al. 2022	0.000	0.282	0.129	0.061	0.226	0.295	0.225	0.382	0.538	0.000	0.600	0.102

^a EMI – electromagnetic induction apparent electrical conductivity, ^bGPR – Ground Penetrating Radar, ^cGC – galvanic contact apparent electrical conductivity

Position Offset

Offset values ranging from -25 to 25 m in 0.5 m steps were used for the phase correlation analysis for each PSS dataset. It was found that a negative shift of -3 m and -5 m needed to be applied to the EMI and GPR data, respectively; for GC, the highest phase correlation coefficient was obtained at 0 m, so no offset should be applied. For these three sensors, a well-determined distribution for the phase correlation coefficient was obtained with a peak at the estimated offset value. For γ -ray, high variability in the phase correlation coefficient was observed, which resulted in a lack of convergence to a specific offset. A visual analysis of the raw data from the γ -ray sensor showed a high variability at short distances, that is, a weaker spatial correlation. This can also be seen in the high values of the Nugget relative to the Sill (Nugget representing 92.3% of Sill) in Table 3.2. The position offset methodology under evaluation assumes the presence of spatial correlation in the data, so the results obtained for the γ -ray sensor data could be related to its weaker spatial correlation. Dashes for GC and γ -ray in Table 3.2 indicate that no offset was applied to the data, and the geostatistical analysis results and RMSE values are the same as for the Raw dataset, that is, when the data is standardized to zero mean and unit variance (without Box-Cox transformation).

Due to the low computational cost of the position offset methodology, multiple interactions can be assessed, together with an estimation of the variability. When a high variability in the interactions is observed, a warning can be issued to the user, or an automatic decision based on a threshold can be taken on whether or not to use the estimated offset (Lee et al., 2012; Sudduth et al., 2012).

One may also be intrigued by the need to apply a negative offset to the data from EMI and GPR. The data from these two sensors were collected by contractors, who used proprietary software to combine the GNSS data with the sensor readings. Small inconsistencies in this combination could have caused a shift in the data, resulting in the need for a negative offset. For GPR, the negative offset could also be associated with a pass-to-pass error because the data was collected with a push-cart (>15 min between the beginning and end of a pass), and data was not collected using a Real-Time Kinematic (RTK) level receiver.

Moreover, the offsets estimated for the EMI and GPR sensors are close to their average distance between consecutive points (Table 3.1), so the results could be related to the data collection density. Comparing geostatistical analysis results before and after the position offset (Table 3.2), a slight

decrease in the Nugget effect for both sensors can be observed, indicating a small increase in the capability of the model to capture variability at shorter distances, even though there is a small increase in the interpolation RMSE.

To further assess the uncertainty about the effectiveness of this methodology in determining the need for a position offset in the data, a simple complementary analysis was performed by simulating a negative shift of -7 m in the GC data (between the GNSS antenna placement and the sensor). As a result, the phase correlation analysis indicated the need for a positive offset of 7 m. Accordingly, this methodology shows potential for the identification of a need for a position offset in PSS data, while being very suitable for batch processing because of its computational efficiency and autonomy. Future work will focus on simulating more datasets with a variety of offsets and evaluating the performance of the methodology to identify and correct them.

Operational, Global, and Local Filtering

Operational

The operational filtering methodology removed about 30%, 17%, 22%, and 18% of the observations from EMI, GPR, γ -ray, and GC, respectively. In view of geostatistical analysis results in Table 3.2, RMSE decreased for GPR and GC, and slightly increased for EMI and γ -ray. The Nugget effect presented to be smaller for all the PSS, which shows that the fitted variogram model captures more spatial dependence from the data. The blue dots presented in the Removal Step maps in

Figure 3.2 represent the points removed by the Operational Filter, mostly observations made at the borders of the field, in maneuvers or turns, and points too close or too distant from its consecutives were removed, which represents a good performance of the filter. Similar results can be observed in the maps for the other PSS (Figures 3.2 – 3.5).

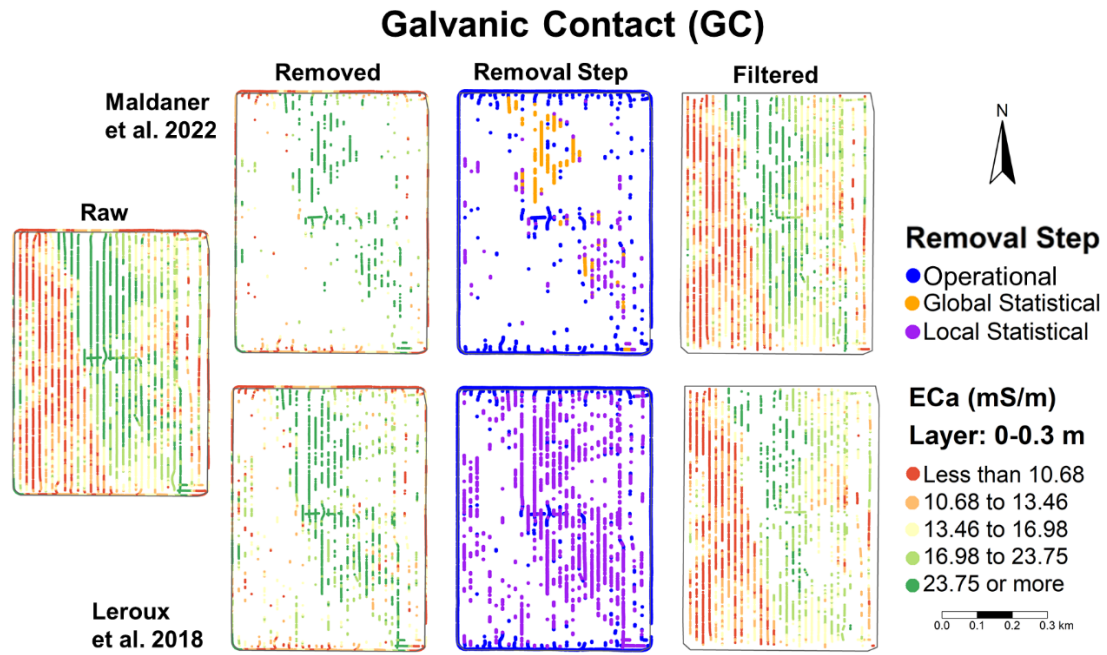


Figure 3.2 Results of global and local filtering methodologies for galvanic contact (GC) sensor apparent electrical conductivity (EC_a). Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

Global Statistical

The methodology applied by Leroux et al. (2018) for the removal of global statistical outliers did not identify any extreme values for three of the four PSS datasets. This same methodology only removed 3 points from the γ -ray data and dashes in Table 3.2 indicate that no analysis was performed for this step because the results would be the same as for the Operational step. However, the methodology proposed by Maldaner et al. (2022) removed about 0.4%, 2%, 2.3%, and 4% of the observations from EMI, GPR, γ -ray, and GC, respectively, in addition to the percentages for Operational filter. The removed points in this step for the GC sensor can be observed in

Figure 3.2 under Global Statistical (orange dots) in the Removal Step maps. Since no observations were removed by this step when using Leroux et al.'s (2018) methodology, no points for this class can be observed on the map in this case.

Maldaner et al.'s (2022) methodology requires a user-defined coefficient for this filtering step. Considering that our study focuses on the batch processing of PSS data, we tested multiple values

for this coefficient on each of the sensors. We found that the filtering strategy used by this method is sensitive to the value of the coefficient, which tends to be proportional to the variability in the data. To obtain the results above, the coefficients used for EMI, GPR, γ -ray, and GC were 0.2, 1.6, 0.2, and 1.6, respectively. Spekken et al. (2013), who had formerly proposed a methodology very similar to Maldaner et al. (2022), also reported such behavior when selecting the threshold to determine the presence of spatial outliers. Despite this disadvantage, there was an overall reduction of the data variance (Sill), Nugget effect, and interpolation RMSE for all four PSS, indicating an increase in the data quality.

Local Statistical

The Local Statistical filter is the last step towards the improvement of data quality. At this step, it is possible to fully assess the differences between the filtering methodologies under evaluation. Table 3.2 shows that the lowest values for Nugget, Sill, and RMSE are for Maldaner et al.'s (2022) methodology for GPR and γ -ray, while Leroux et al.'s (2018) methodology provides the lowest Sill for GC and the lowest Sill and RMSE for EMI. Through the local filtering, Leroux et al. (2018) and Maldaner et al. (2022) methodologies respectively removed an additional 15%, 7.7%, 6.7%, and 25% of the observations, and 3%, 21%, 28%, and 3% of the observations, for EMI, GPR, γ -ray, and GC.

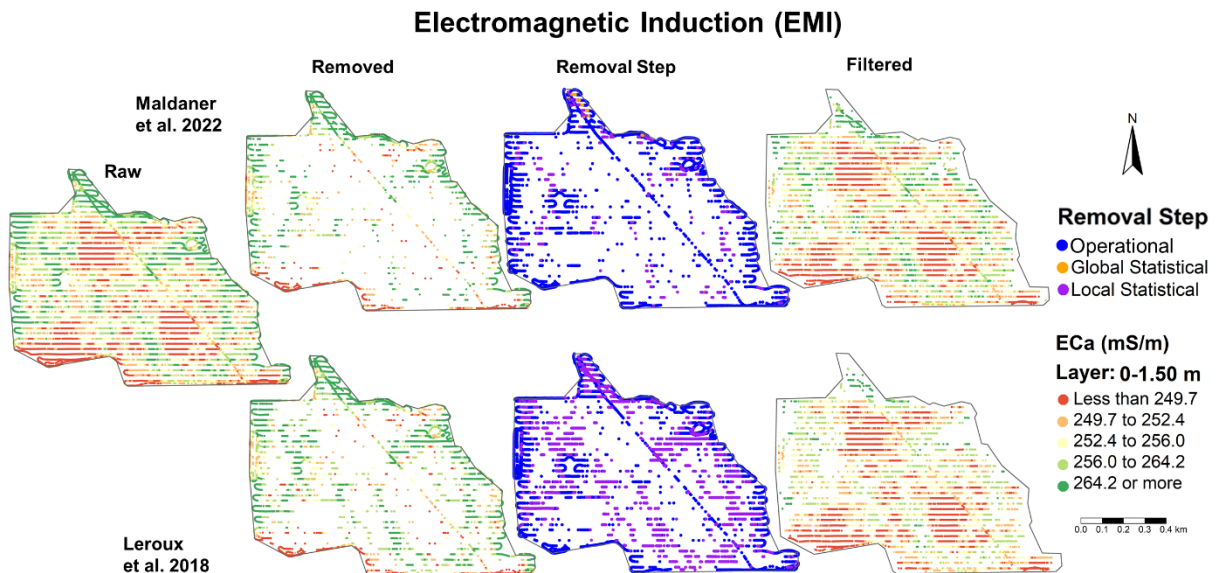


Figure 3.3 Results of global and local filtering methodologies for electromagnetic induction (EMI) sensor apparent electrical conductivity (EC_a). Results are presented for the methods of Maldaner

et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

Although the geostatistical analysis results support that the methodology of Leroux et al. (2018) performs better filtering for two of the four sensors (GC and EMI), a visual inspection of the removed and filtered maps (Figures 3.2 and 3.3– maps for Leroux et al., 2018) suggests possible excessive removal of observations during the Local Statistical step. By comparing the Raw and Removed data points, the removal of some true patterns in the field can be observed when the Leroux et al. (2018) methodology was used (e.g., excessive removal of high EC_a values for the EMI sensor in Figure 3.3).

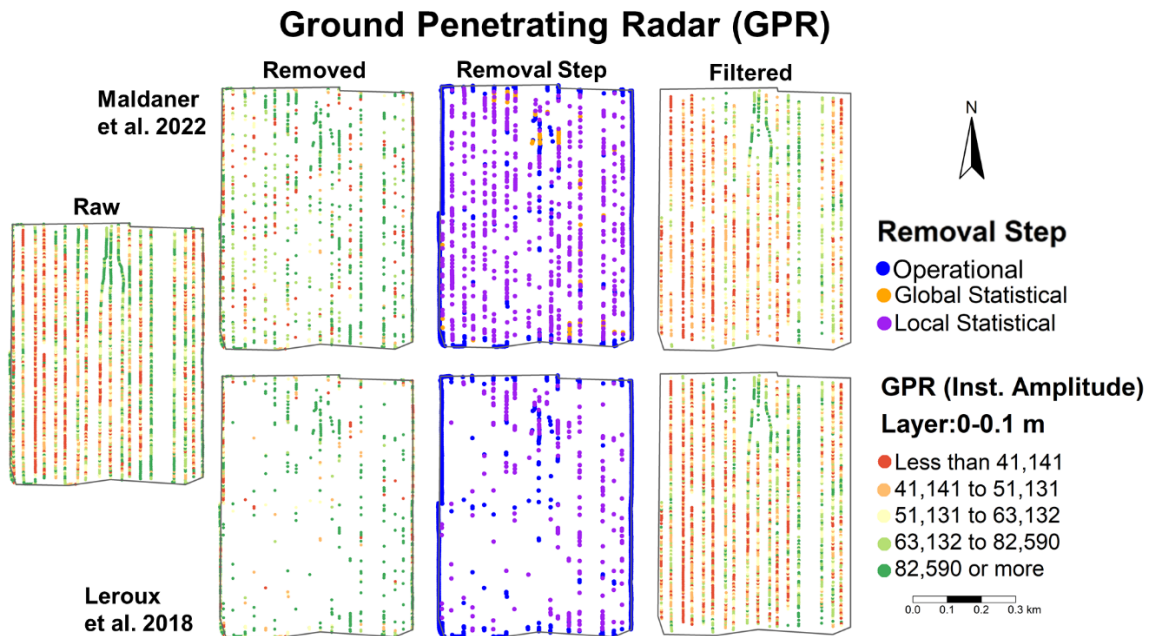


Figure 3.4 Results of global and local filtering methodologies for ground penetrating radar (GPR) Instantaneous Amplitude at 0-0.1m depth. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps

The issue raised in the previous paragraph might cause the removal of important patches in the field which could affect further analysis using the filtered data. In addition, the results presented for GPR and γ -ray for Leroux et al. (2018) (Figures 3.4 and 3.5) suggest insufficient removal of spatial outliers. The maps obtained for these sensors when using this methodology still present some noise and do not reveal a well-delimited spatial distribution (Figures 3.4 and 3.5). On a

separate note, the identification of spatial outliers based on the non-parametric methodology of Leroux et al. (2018) requires a few more extra steps than Maldaner et al. (2022) methodology, and therefore, is more computationally costly.

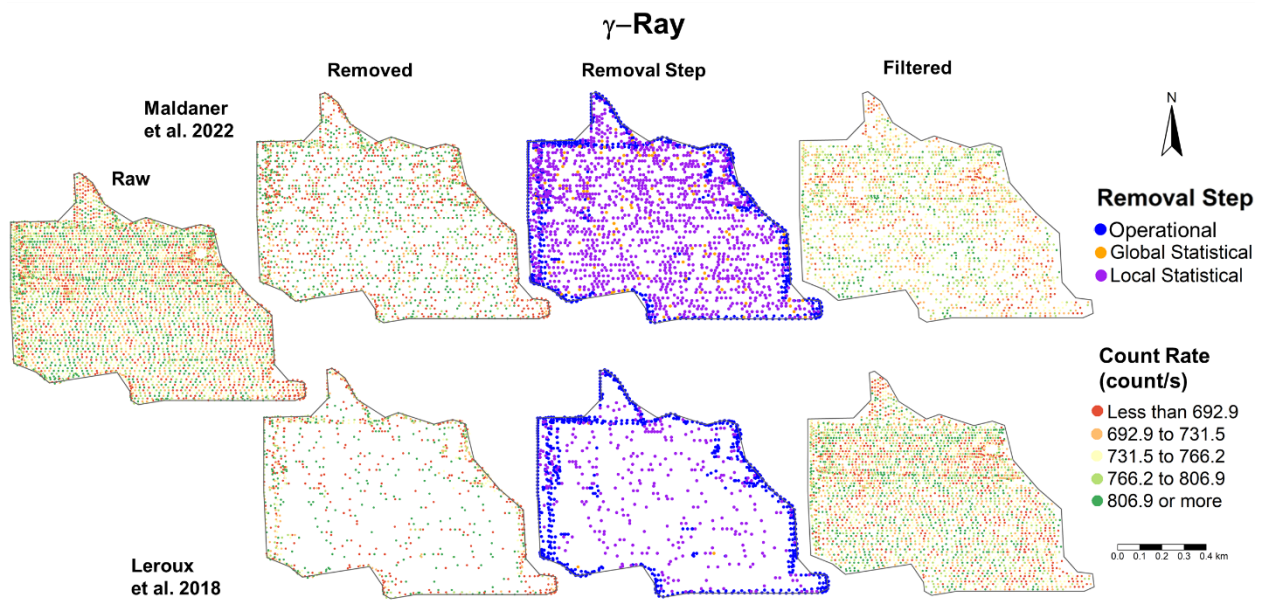


Figure 3.5 Results of global and local filtering methodologies for γ -ray count rate. Results are presented for the methods of Maldaner et al. (2022), top, and Leroux et al. (2018), bottom. Points removed at each filtering step are classified at the Removal Step maps.

Based on the above analyses and results, Maldaner et al. (2022) methodology appears to be the most suitable to process the particular dataset used in the present study. Accordingly, its results were used for interpolation. However, as briefly mentioned during the presentation of the Global Statistical step, a disadvantage of this methodology is its sensitivity to a user-defined coefficient. When considering data batch processing, the need for a user-defined coefficient affects the flow and efficiency of the process. For the Local Statistical step, the same sensitivity as for the Global Statistical step was observed, and after trying a number of values, 0.05, 0.4, 0.1, and 0.5 were selected as the best coefficients for EMI, GPR, γ -ray, and GC, respectively.

Overall, the two methodologies presented advantages and disadvantages. The proposed filtering steps (operational, global, and local statistical) improved the data quality; nevertheless, there is still a need for the development and evaluation of more efficient, non-parametric methodologies for global and local statistical filtering, in order to improve the batch processing of these steps.

Interpolation

The four PSS datasets, after applying all of the above steps, had the normality of their distributions tested ($\alpha=0.05$). It was rejected for all variables, so each set of data was submitted to a Box-Cox transformation. The four interpolation methods were applied to the transformed data. Results are reported in Table 3.3, where the values for OK may differ from those in Table 3.2 since no data transformation was performed at the evaluation step.

Table 3.3 Interpolation methods root mean squared error (RMSE) and coefficient of determination (R^2) resulted from 10-fold cross-validation for each proximal soil sensor.

Interpolation Method	EMI ^a (0-1.5 m)		GPR ^b (0-0.1 m)		γ -ray (count rate)		GC ^c (0-0.3 m)	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
NN ^d	0.320	0.899	0.855	0.406	1.149	0.120	0.200	0.960
IDW ^e	0.258	0.934	0.699	0.516	0.875	0.235	0.159	0.975
OK ^f	0.228	0.949	0.665	0.558	0.858	0.263	0.203	0.960
UK ^g	0.231	0.947	0.662	0.563	0.859	0.262	0.143	0.980

^aEMI – electromagnetic induction apparent electrical conductivity, ^bGPR – Ground Penetrating Radar, ^cGC – galvanic contact apparent electrical conductivity, ^dNN – Nearest Neighbor, ^eIDW – Inverse Distance Weighting, ^fOK – Ordinary Kriging, ^gUK – Universal Kriging

Overall, kriging presented the best performance for all four different PSS data, while NN was the worst. Even though OK and UK performed better than the other two, automatic calculation and model fitting to a variogram is still challenging (Oliver and Webster, 2014). Thus, to properly apply geostatistics to specific data, there would be a need for some user analysis of the variogram and model. For example, analyzing the data from GC and its distribution on the X and Y coordinates, a trend was observed. In this case, UK is the most suitable kriging method instead of OK. However, an analysis of the data distribution and the variogram had to be performed to identify the trend and model it.

On the other hand, by automatically optimizing the IDW parameters similar results to kriging were obtained. For GC data, where a trend was observed, IDW even outperforms OK. In this

scenario, when focusing on batch processing, IDW might be a good option. However, there are still some other interpolation options that could be evaluated, such as spatial random forest (Sekulić et al., 2020), support vector machines, and their combination with IDW (Willam et al., 2022). Although, these also need the optimization of their hyper-parameters, a time-consuming and computationally costly procedure.

For future work, this framework should be tested with simulated datasets, while implementing and/or developing more robust methodologies for the spatial outlier detection step. Newly developed interpolation methodologies, using machine learning and their combination with traditional interpolation methods, should also be tested.

3.4 Conclusion

A framework for batch processing of high-density anisotropic data from proximal soil sensors (PSS) was proposed and tested. It was found that data quality improved after applying the framework which used different methodologies at each step. However, the development of more computationally efficient and autonomous methods is required, as the methods available in the literature require expert thresholds or they are not robust enough for use with data with such variability as the PSS data collected in agricultural fields.

3.5 Acknowledgments

We would like to thank all Olds College crew involved in the collection of the proximal soil sensor. SoilOptix Inc. (Tavistock, ON, Canada) is also thanked for providing the complete gamma ray analysis. This research is part of the project "Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services" supported by Mitacs through the Mitacs Accelerate program.

3.6 References

- Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44(1), 71–91. <https://doi.org/10.1016/j.compag.2004.03.002>
- Arslan, S., & Colvin, T. S. (2002). Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precision Agriculture*, 3(2), 135–154. <https://doi.org/10.1023/A:1013819502827>
- Barbulescu, A., Bautu, A., & Bautu, E. (2020). Optimizing inverse distanceweighting with particle swarm optimization. *Applied Sciences* (Switzerland), 10(6). <https://doi.org/10.3390/app10062054>

- Blackmore, S., & Moore, M. (1999). Remedial Correction of Yield Map Data. *Precision Agriculture*, 1(1), 53–66. <https://doi.org/10.1023/A:1009969601387>
- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5), 1190–1208. <https://doi.org/10.1137/0916069>
- Chung, S. O., Sudduth, K. A., & Drummond, S. T. (2002). Determining Yield Monitoring System Delay Time with Geostatistical and Data Segmentation Approaches. *Transactions of the ASAE*, 45(4), 915–926. <https://doi.org/10.13031/2013.9938>
- Colaço, A. F., Richetti, J., Bramley, R. G. V., & Lawes, R. A. (2021). How will the next-generation of sensor-based decision systems look in the context of intelligent agriculture? A case-study. *Field Crops Research*, 270(February), 108205. <https://doi.org/10.1016/j.fcr.2021.108205>
- Ester, M., Kriegel, H.-P., Sander, J., & Xiaowei, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Menlo Park, California: American Association for Artificial Intelligence (AAAI).
- Fulton, J. P., & Port, K. (2018). Precision Agriculture Data Management. In *Precision Agriculture Basics* (pp. 169–187). <https://doi.org/10.2134/precisionagbasics.2016.0095>
- Hengl, T. (2009). A Practical Guide to Geostatistical Mapping. <http://spatial-analyst.net/book/>
- Hubert, M., & Van Der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, 22(3–4), 235–246. <https://doi.org/10.1002/cem.1123>
- Isaaks, E. H., & Srivastava, R. M. (1989). *Applied Geostatistics*. NY: Oxford University Press, Inc.
- ISO. (2010). ISO 12188-1:2010 Tractors and machinery for agriculture and forestry — Test procedures for positioning and guidance systems in agriculture — Part 1: Dynamic testing of satellite-based positioning devices.

- Ji, W., Biswas, A., Adamchuk, V., Perron, I., Cambouris, A., & Zebarth, B. (2018). Proximal soil sensing-led management zone delineation for potato fields. 14th International Conference on Precision Agriculture: 24 - 27 June, 1–14.
- Lee, D. H., Sudduth, K. A., Drummond, S. T., Chung, S. O., & Myers, D. B. (2012). Automated Yield Map Delay Identification Using Phase Correlation Methodology. *Transactions of the ASABE*, 55(3), 743–752. <https://doi.org/10.13031/2013.41506>
- Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., & Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, 19(5), 789–808. <https://doi.org/10.1007/s11119-017-9555-0>
- Lyle, G., Bryan, B. A., & Ostendorf, B. (2014). Post-processing methods to eliminate erroneous grain yield measurements: Review and directions for future development. *Precision Agriculture*, 15(4), 377–402. <https://doi.org/10.1007/s11119-013-9336-3>
- Maldaner, L. F., Molin, J. P., & Spekken, M. (2022). Methodology to filter out outliers in high spatial density data to improve maps reliability. *Scientia Agricola*, 79(1), 1–7. <https://doi.org/10.1590/1678-992x-2020-0178>
- Menegatti, L. A. A., & Molin, J. P. (2004). Removal of errors in yield maps through raw data filtering. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 8(1), 126–134. <https://doi.org/10.1590/s1415-43662004000100019>
- Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *CATENA*, 113, 56–69. <https://doi.org/10.1016/j.catena.2013.09.006>
- Ping, J. L., & Dobermann, A. (2005). Processing of yield map data. *Precision Agriculture*, 6(2), 193–212. <https://doi.org/10.1007/s11119-005-1035-2>
- R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria. <https://www.r-project.org/>
- Roberton, S. D., Lobsey, C. R., & Bennett, J. M. L. (2021). A Bayesian approach toward the use of qualitative information to inform on-farm decision making: The example of soil compaction. *Geoderma*, 382(August 2020), 114705. <https://doi.org/10.1016/j.geoderma.2020.114705>
- Rodrigues, F. A., Bramley, R. G. V., & Gobbett, D. L. (2015). Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in

- contrasting soils. *Geoderma*, 243–244, 183–195.
<https://doi.org/10.1016/j.geoderma.2015.01.004>
- Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., et al. (2019). Clustering tools for integration of satellite remote sensing imagery and proximal soil sensing data. *Remote Sensing*, 11(9). <https://doi.org/10.3390/rs11091036>
- Schepers, A. R., Shanahan, J. F., Liebig, M. A., Schepers, J. S., Johnson, S. H., & Luchiari, A. (2004). Appropriateness of Management Zones for Characterizing Spatial Variability of Soil Properties and Irrigated Corn Yields across Years, 195–203.
- Schossler, T. R., Mantovanelli, B. C., de Almeida, B. G., Freire, F. J., da Silva, M. M., de Almeida, C. D. G. C., & Freire, M. B. G. dos S. (2019). Geospatial variation of physical attributes and sugarcane productivity in cohesive soils. *Precision Agriculture*, 20(6), 1274–1291. <https://doi.org/10.1007/s11119-019-09652-y>
- Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, 12(10), 1–29. <https://doi.org/10.3390/rs12101687>
- Shekhar, S., Lu, C. T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, 7(2), 139–166. <https://doi.org/10.1023/A:1023455925009>
- Shepard, D. (1968). A two- dimensional interpolation function for irregularly-spaced data. In R. B. S. Blue & A. M. (Eds. . Rosenberg (Eds.), *Proceedings of the 1968 ACM National Conference* (pp. 517–524). New York: ACM Press.
- Simbahan, G. C., Dobermann, A., & Ping, J. L. (2004). Screening Yield Monitor Data Improves Grain Yield Maps. *Agronomy Journal*, 96(4), 1091–1102. <https://doi.org/10.2134/agronj2004.1091>
- Singh, A. K., & Lalitha, S. (2018). A novel spatial outlier detection technique. *Communications in Statistics - Theory and Methods*, 47(1), 247–257. <https://doi.org/10.1080/03610926.2017.1301477>
- Spekken, M., Anselmi, A. A., & Molin, J. P. (2013). A simple method for filtering spatial data. In J. V. Stafford (Ed.), *Precision agriculture'13: Proceedings of the 9th European conference on precision agriculture* (pp. 259–266). Wageningen: The Netherlands: Wageningen Academic Publishers.

- Sudduth, K. A., & Drummond, S. T. (2007). Yield Editor: Software for Removing Errors from Crop Yield Maps. *Agronomy Journal*, 99(6), 1471–1482. <https://doi.org/10.2134/agronj2006.0326>
- Sudduth, K. A., Drummond, S. T., & Myers, D. B. (2012). Yield editor 2.0: Software for automated removal of yield map errors. *American Society of Agricultural and Biological Engineers Annual International Meeting 2012, ASABE 2012*, 4(12), 3378–3391. <https://doi.org/10.13031/2013.41893>
- Sun, W., Whelan, B., McBratney, A. B., & Minasny, B. (2013). An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. *Precision Agriculture*, 14(4), 376–391. <https://doi.org/10.1007/s11119-012-9300-7>
- Vega, A., Córdoba, M., Castro-Franco, M., & Balzarini, M. (2019). Protocol for automating error removal from yield maps. *Precision Agriculture*, 20(5), 1030–1044. <https://doi.org/10.1007/s11119-018-09632-8>
- Willam, G., Domingos, P., Magalhães, S., Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., et al. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*. <https://doi.org/10.1007/s11119-022-09880-9>

Connecting Text to Chapter 4

In **Chapter 3**, a framework to process high-density anisotropic PSS data was proposed. Chapter 4 shifted the focus to the low- and extra-low-density data traditionally collected for soil characterization. The collection, lab analysis, and interpolation of georeferenced soil sampling to map soil chemical properties for agricultural management decisions is a common PA practice, as highlighted in **Chapter 2**. Therefore, **Chapter 4** aimed to propose and evaluate robust interpolation approaches that maximize the value of low- and extra-low-density sampling designs.

All eleven methods in **Chapter 4** could be added to a DSS, allowing the user to choose the best one based on a validation procedure. However, it would still be time-consuming and require a skilled user. A single or combination of interpolation approaches should then be selected to automate this process; thus, a robustness measure was used to evaluate methods that rarely yielded results worse than average.

Preliminary results from **Chapter 4** were presented and published at the proceedings of the 14th European Conference on Precision Agriculture in Bologna, Italy, in 2023 and Precision Agriculture '23. An expanded version was then published in the *Precision Agriculture* journal and reproduced with Springer Nature's permission as **Chapter 4** of this thesis.

In the published manuscript, the word “hollow” was misspelled as “hallow” in the caption for Figure 1 (Figure 4.1 of this thesis). The analytical methods used by the lab to measure pH, plant-available P, and K were soil-water suspension (1:1 ratio), weak Bray, and ammonia acetate extraction, respectively.

Publications:

Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2023). Comparative study of interpolation methods for low-density sampling. In *Precision Agriculture '23* (Vol. 34, pp. 563–569). The Netherlands: Wageningen Academic Publishers.

https://doi.org/10.3920/978-90-8686-947-3_71

Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2024). Comparative study of interpolation methods for low-density sampling. *Precision Agriculture*.

<https://doi.org/10.1007/s11119-024-10141-0>

Abbreviations for Chapter 4

Abbreviation	Definition
CAAIN	Canadian Agri-Food Automation and Intelligence Network
DG	Distance Gradient
DSS	Decision Support System
IDW	Inverse Distance Weighting
K	plant-available Potassium
LOOCV	Leave One-Out Cross-Validation
MAE	Mean Absolute Error
MSE	Mean Squared Error
N:S	Nugget-Effect-to-Sill Ratio
OK	Ordinary Kriging
P	plant-available Phosphorus
PA	Precision Agriculture
ppm	parts per million
PSS	Proximal Soil Sensing
SD	Standard Deviation

Chapter 4: Comparative Study of Interpolation Methods for Low-Density Sampling

F. H. S. Karp, V. Adamchuk, P. Dutilleul, A. Melnitchouck

Abstract

Given the high costs of soil sampling, low and extra-low sampling densities are still being used. Low-density soil sampling usually does not allow the computation of experimental variograms reliable enough to fit models and perform interpolation. In the absence of geostatistical tools, deterministic methods such as inverse distance weighting (IDW) are recommended but they are susceptible to the “bull’s eye” effect, which creates non-smooth surfaces. This study aims to develop and assess interpolation methods or approaches to produce soil test maps that are robust and maximize the information value contained in sparse soil sampling data. Eleven interpolation procedures, including traditional methods, a newly proposed methodology, and a kriging-based approach, were evaluated using grid soil samples from four fields located in Central Alberta, Canada. In addition to the original 0.4 ha·sample⁻¹ sampling scheme, two sampling design densities of 0.8 and 3.5 ha·sample⁻¹ were considered. Among the many outcomes of this study, it was found that the field average never emerged as the basis for the best approach. Also, none of the evaluated interpolation procedures appeared to be the best across all fields, soil properties, and sampling densities. In terms of robustness, the proposed kriging-based approach, in which the nugget effect estimate is set to the value of the semi-variance at the smallest sampling distance, and the sill estimate to the sample variance, and the IDW with the power parameter value of 1.0 provided the best approaches as they rarely yielded errors worse than those obtained with the field average.

4.1 Introduction

In the 2021 Precision Agriculture Dealership Survey (Erickson and Lowenberg-Deboer, 2021), grid and zone sampling are among the most often offered and adopted precision services in the United States. According to the same survey, the use of grids with the common cell size of 1 ha predominates over zone sampling, while many continue to sample using larger grids.

Thus, a scenario in which a field of ≥ 100 ha is soil sampled using grid cells of >1 ha would not be uncommon. In such a case, the total number of soil samples would be smaller than 100, whereas 100 is recommended as the minimum sample size to obtain adequate experimental variograms for soil properties (Webster and Oliver, 1992). It follows that to produce an interpolated surface, at least some of the available deterministic interpolation methods should be considered. Among them,

inverse distance weighting (IDW) is one of the most popular as it has no limitations on the number of samples, is computationally efficient, and easy-to-apply (Kravchenko, 2003).

As IDW is an exact interpolator (*i.e.*, it predicts values identical to observed values at sampling locations), a known issue with maps obtained with this method is the influence of isolated extreme values on their surroundings, which creates an effect called “bull’s eye.” Although many modifications of the original IDW method (Shepard, 1968) have been tested and compared to other interpolation methods (Franke and Nielson, 1980; Robinson and Metternicht, 2006), further evaluation and improvement of interpolation techniques for low-density soil sampling is needed.

One could argue that when measurements are sparse or weakly correlated in space, an interpolation method using co-variables observed at a higher resolution (*e.g.*, satellite imagery, soil sensing techniques) may improve interpolation accuracy (Goovaerts, 1999). This is a valid point, but it assumes that the co-variables are spatially cross-correlated with the target variable (Isaaks and Srivastava, 1989; Webster and Oliver, 2007). If this assumption is proven to be wrong for co-variables for which historical or free-access datasets are available, new information must be collected, which may not be feasible due to additional costs or lack of time.

Clearly, sampling soils with low-density grids imposes difficulties in extracting useful information from the data. Thus, more efficient and accurate sampling strategies for precision agriculture practitioners and the industry are still required. Since the 1990s (Larocque et al., 2007; Laslett and McBratney, 1990; Wadoux et al., 2019), research results have supported that slight modifications made to a grid sampling design (*e.g.*, by collecting extra samples close to grid sampling nodes) can improve the estimation of variogram model parameters and decrease kriging interpolation errors. Nevertheless, the 2021 survey mentioned above suggests that, in large part, soil sampling data may continue to be collected in agricultural fields using the traditional grid sampling design at low or extra-low resolution.

Therefore, methodologies or approaches that maximize the value of datasets with too small a size to use classical geostatistical methods must be developed and explored further. Sobjak et al. (2023) used samples collected at a density of approximately 3 samples·ha⁻¹ to build and test an automated process to improve the selection of the parameters that maximize the accuracy for ordinary kriging (OK) and IDW. These authors reported an improvement in the interpolation accuracy when the best interpolation parameters were identified using a newly proposed assessment index (*i.e.*, effective spatial dependence index). Others have focused on evaluating the

potential of machine learning algorithms for interpolating soil properties. Hengl et al. (2018) proposed an approach that uses the distance between samples as co-variables in a random forest model to perform spatial interpolation. Pereira et al. (2022) reported a potential improvement in the interpolation accuracy of soil properties when using a combination of support vector machines and IDW to interpolate samples collected from one field at densities ranging from 1.4 to 5.7 samples·ha⁻¹ when compared to OK and IDW. However, there is still a lack of comprehensive comparison of different interpolation methods and the definition of universal approaches for interpolating low-density soil sampling datasets.

The objective of this study was to develop alternative interpolation procedures and to assess them in comparison with other methods to produce soil test maps that are robust and maximize the information value contained in datasets collected with low soil sampling density.

4.2 Material and Methods

Study Area and Data

Soil samples were collected from four fields in Central Alberta, Canada (Figure 4.1a). The four fields were mainly grown under wheat, barley, and canola crop rotation. A summary of the field characteristics is presented in Table 4.1.

Table 4.1 Description of the studied fields

Field	Area (ha)	Central Coordinates	Predominant Soil Subgroup ASIC (2001)
1	43	51°46'11.1"N 114°05'19.2"W	Orthic Black Chernozem
2	92	51°46'22.8"N 114°00'57.4"W	Orthic Black Chernozem
3	126	51°39'23.8"N 114°15'20.0"W	Orthic Black Chernozem
4	66	51°40'27.2"N 112°47'54.0"W	Southwest - Orthic Humic Vertisol Northeast - Vertic Dark Brown Chernozem and Orthic Dark Brown Chernozem

All soil samples from 0 to 0.15 m deep were collected during Fall 2022, using grids of cells of about 0.4 ha. A total of 128 (including 20 independent validation samples), 216, 274, and 144 samples were collected from Fields 1, 2, 3, and 4, respectively (Figure 4.1b). The locations for the validation samples from Field 1 were chosen based on the spatial variability of the field data observed in previous years; see the diamond shapes in Figure 4.1b for Field 1. These validation samples were collected on the same day and under the same conditions as the grid samples from this field. All samples were sent to the same laboratory for the analysis of their chemical properties. The reported values for pH, plant-available phosphorus (P), and potassium (K) were used to compare the interpolation approaches and methods. P and K values were reported in parts per million (ppm), while the pH followed the scale from 0 (most acidic) to 14 (most basic). These three soil chemical properties were chosen to evaluate the interpolation methods due to their relevance in the soil management practices performed in precision agriculture. Thematic maps from pH are often used to determine areas with lower pH values, which can be amended through site-specific lime application. Similarly, prescription maps can be generated by mapping the spatial distribution of P and K to address and improve the levels of these two macronutrients in regions within the field where their availability is below the ideal values for crop development.

From the original $0.4 \text{ ha} \cdot \text{sample}^{-1}$ sampling scheme, samples were gradually removed to create sampling design densities of 0.8 and $3.5 \text{ ha} \cdot \text{sample}^{-1}$, thus, broadening the scope of the comparison. Only regular grid designs were considered because they still predominate over other sampling designs (Erickson and Lowenberg-Deboer, 2021). The three sampling designs for the four fields are presented in Figure 4.1b.

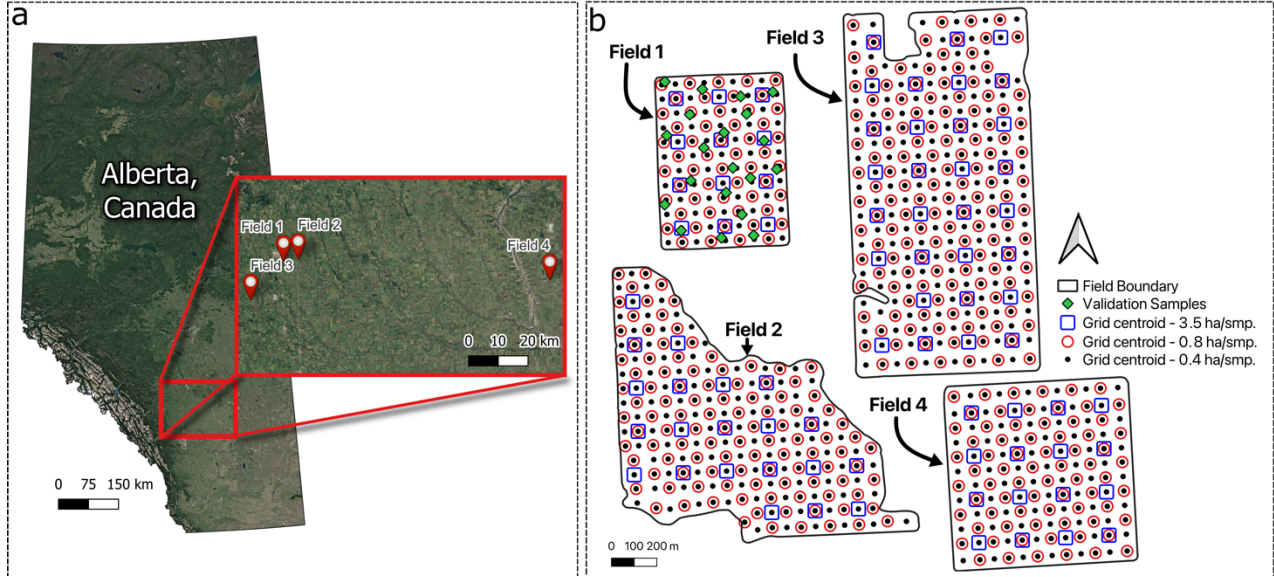


Figure 4.1 Experimental sites: a – map of the Canadian province of Alberta with a zoomed-in window showing the distribution of the four experimental sites located within a region known as Central Alberta, and b – field boundaries and grid centroids representing the three sampling design densities – different shapes, and colors are used to represent the distribution of the sampling locations for 0.4 (solid black circles), 0.8 (hollow red circles), 3.5 (hollow blue squares) ha-sample⁻¹, and validation points (solid green diamonds); the latter are only available for Field 1

Conventional Interpolation Methods

The three datasets from each of the four fields were submitted to Shepard’s original IDW algorithm (Shepard, 1968). This algorithm requires setting two parameters: the power and the search neighborhood. The power value controls the influence of the closest samples to an interpolated location – the higher the power, the greater the influence of the samples closer to the interpolated point. The search neighborhood is defined by the number of points used for estimating the value at the interpolation location; for example, with a search neighborhood of 4, the interpolated value is determined from the 4 closest sample data. This study used power values of 1 and 2 and a search neighborhood equal to all available neighbors (108, 216, 274, and 144 for Fields 1, 2, 3, and 4, respectively) to evaluate Shepard’s original IDW.

An approach, called “Optimal IDW” hereafter, was also evaluated. This approach uses brute-force search and Leave One-Out Cross-Validation (LOOCV) to assess a wide range of combinations of values for the power and search neighborhood parameters. Power values between 1 and 5, in increments of 0.2, and a search neighborhood from 4 to all available neighbors, in

increments of 1, were tested. Based on the Mean Absolute Error (MAE) from this procedure, a combination of both parameters that minimized the MAE was selected. In addition, a separate approach in which the power value is set to 0 and the search neighborhood to 1 was evaluated as “Nearest neighbor.” Both approaches described above are variants of Shepard’s original IDW algorithm (Shepard, 1968).

The local modified Shepard’s IDW interpolator (Franke and Nielson, 1980), another IDW variant, uses estimates from a locally fitted polynomial and a limited neighborhood for the inverse distance weights calculation to address some of the caveats imposed by the original Shepard’s IDW (*e.g.*, the influence of isolated extreme values). This interpolator requires setting two independent parameters that control the neighborhood size for a local quadratic polynomial fitting and inverse distance weights. According to Franke & Nielson (1980), properly setting these two parameters can significantly influence the performance of this method. In implementing the local modified Shepard’s IDW, a brute-force search was employed to select the combination of parameters that minimized the MAE from a LOOCV. Neighborhood sizes varying from 4 to all available neighbors, in increments of 1, were tested for both parameters.

The original OK, called “Fitted variogram model” hereafter, was evaluated as an interpolation option in the geostatistical approach. Thus, using the R library *gstat* (Pebesma, 2004), experimental variograms were computed for all datasets from all fields, soil properties, and sampling designs. Spherical, exponential, and Gaussian models were fitted to each experimental variogram, and the best model was selected based on the residual sum of squares for a weighted least-squares fitting procedure. Even though Webster & Oliver (1992) recommend 100 as the minimum sample size to obtain adequate experimental variograms for soil properties, “Fitted variogram model” was used to broaden the discussion of results when this minimum number was not reached in this study.

In the following two subsections, two new methods are proposed for interpolating low-density soil sample data, one model-based and the other model-free

A modification of the kriging-based approach

Due to limited information about variability below the shortest sampling distance and the small number of pairs of observations available to estimate semi-variances at a number of distance lags, fitting a model to an experimental variogram computed from a sparse dataset is challenging and highly uncertain. Accordingly, the proposed approach sets the sill and the nugget effect to values

directly derived from the sample data without passing by the nugget-effect and sill estimates obtained from a fitted variogram model.

When data are not correlated, the sample variance is the classical estimator of the population variance. Barnes (1991) already raised the awareness that if the data are evenly distributed in an area with dimensions greater than the range of spatial correlation, the sample variance could be a reasonable first estimate of the sill. Based on a different reason leading to the same outcome, instead of relying on an experimental variogram and variogram model parameter estimates that are uncertain due to a small sample size (Larocque et al., 2007), it is proposed that a sample variance estimated under the assumption of independence should be used as an alternative to a sill estimate obtained from an experimental variogram that does not completely represent the correlation structure of the data.

The challenges mentioned above for the sill concern the estimation of the nugget effect in a similar way, more particularly the absence (lack) of direct (indirect) information in the data about the behavior of the semi-variance function at distances smaller than the shortest sampling distance (between grid nodes if regular). Thus, estimating the nugget effect accurately by fitting a variogram model is at the least very difficult or practically impossible, so it is proposed that the semi-variance estimate at the shortest sampling distance be used as the nugget-effect estimate in that case. This nugget-effect estimator is likely to be biased upwards, and the nugget-effect estimates might even approach the sample variance, used as the sill estimate. This would then be interpreted as there is no spatial correlation in the data, and a flat, pure nugget-effect variogram model would be adopted. The expectation, however, is that the generated interpolated surface would be better than the field average and as good as the field average otherwise. For comparison purposes, interpolated surfaces were also produced with the nugget effect set to 0 (“Set Sill, and Nugget=0”), the strict minimum value in theory for the parameter in a variogram model.

Finally, the alternative estimates of the sill and the nugget effect described above are inserted in the equation of a spherical variogram model, and an estimate of the range of spatial correlation is obtained by fitting the equation to the experimental variogram by least squares. A spherical variogram model is preferred to a member of the Matérn family, such as the exponential and Gaussian variogram models, because the semi-variance reaches the process variance only asymptotically in them. The steps for this modification of the kriging-based approach (“Set Sill and Nugget”) are summarized in Figure 4.2.

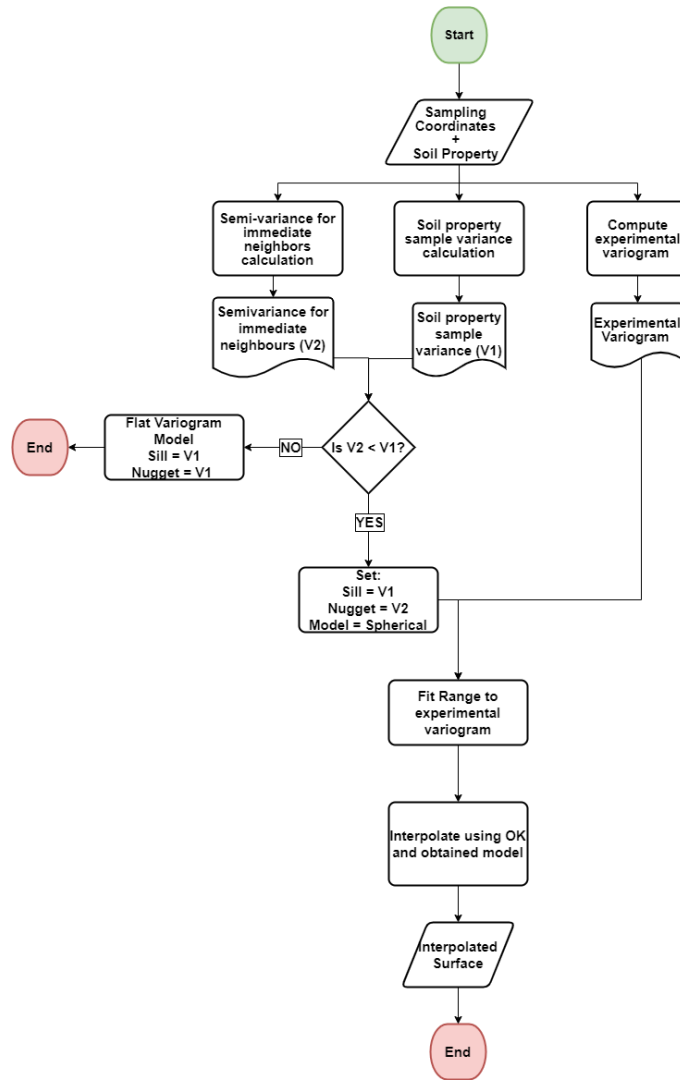


Figure 4.2 Flowchart explaining the proposed modification of the kriging-based approach to obtain interpolated surfaces from soil data collected at low and extra-low sampling densities; OK – ordinary kriging

Proposed modification of the original IDW

In this paper, a new model-free method is also proposed: IDW Smoothed. It postulates that a better, more representative interpolated surface can be produced by jointly minimizing the difference between an optimized value and the observed value at each sampling location and a quantity called “distance gradient” (DG), which indicates the rate of change of measurements with distance. Therefore, the following Pareto multi-objective optimization procedure is proposed:

1. Define a search domain. The minimum and maximum values from the observed values are calculated for a specific soil test. The precision for the values reported in the lab results is

determined and used as an incremental value to define the search domain (*e.g.*, minimum, maximum, and precision of laboratory analysis of 1, 3, and 0.1, respectively, were determined; thus, a search domain is defined by all numbers between 1 and 3, in increments of 0.1)

2. Randomly select an array of values from inside the search domain with the same sample size (number of observations) as the input data. The newly selected values are referred to as “trial values.”
3. Calculate the Mean Absolute Error (MAE) between the observed and trial values. Using the trial values and the matrix of geographic distances between each sampling location and the $n - 1$ other sampling locations, calculate the distance gradients:

$$DG_k = \frac{\sum_{i=1, i \neq k}^n \frac{|x_k - x_i|}{d(x_k, x_i)}}{n-1} \text{ for } k = 1, \dots, n \quad (4.1)$$

where n is the number of sampling locations, x_k is the trial value for the sampling location (k) for which the distance gradient is calculated in Eq. (4.1), x_i denotes the trial value for sampling location i , and $d(x_k, x_i)$ is the geographic distance between sampling locations k and i .

Obtain the mean DG by averaging the n DGs.

4. Repeat Steps 2 and 3 for a user-set number of times (trials), with the objective of minimizing the MAE and mean DG. The two objectives are plotted in Figure 4.3.
5. Obtain the optimal solution (shown with the purple triangle in Figure 4.3) by selecting the solution closest to the origin of the biplot from the Pareto optimal solutions (*i.e.*, the orange dots in Figure 4.3).
6. Produce a continuous surface by using the optimized values and their respective spatial coordinates in an IDW with $n - 1$ neighbors for each sampling location and a small distance power value (*e.g.*, 1).

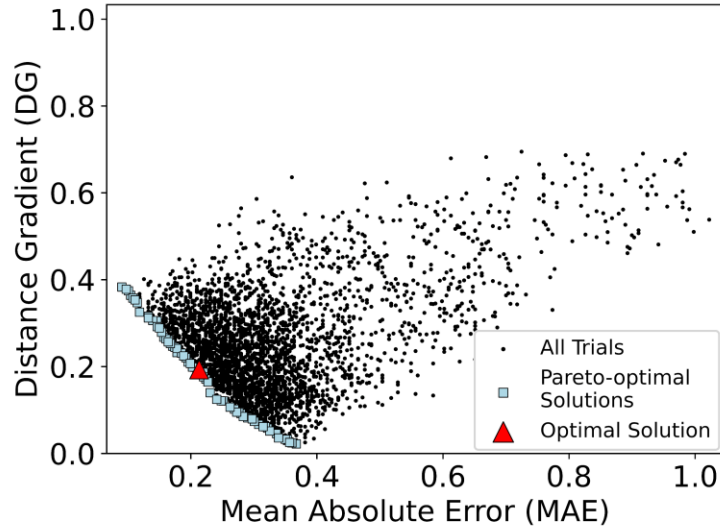


Figure 4.3 Biplot of Mean Distance Gradient (DG) versus Mean Absolute Error (MAE) for 3,500 trials (blue dots), with the Pareto-optimal solutions (orange dots) and the optimal solution (purple triangle)

A customized script written in Python 3 programming language was developed to implement the multi-objective optimization process for the proposed methodology. The Optuna library (Akiba et al., 2019) and Nondominated Sorting Genetic Algorithm II (Deb et al., 2002) were used to obtain the Pareto front with 3,500 trials and select the optimal solution for each soil property. In the definition of the search domain, increments were 0.1 for pH and 1 ppm for the K and P tests. To compare this method with the original IDW (Shepard, 1968), values of 1 and 2 were used for the power parameter (Step 6).

A summary of the 10 interpolation approaches and methods described above and the interpolation procedures used in each are provided in Table 4.2. In addition, for each sampling density, the field average was used as a benchmark.

Table 4.2 Summary table of all the different methods and approaches evaluated in this study, and their respective reference name used for figures and tables

Reference Name	Base Method	Approach
Average	Average	Field average calculated using the available sample data
Fitted variogram model	Ordinary Kriging	Fit a model to an experimental variogram
IDW ^a P:1 ^b	Shepard's IDW	Power value = 1, and search neighborhood = all neighbors
IDW P:2	Shepard's IDW	Power value = 2, and search neighborhood = all neighbors
Optimal IDW	Shepard's IDW	Optimize with respect to power and the size of the search neighborhood
Nearest Neighbor	Nearest Neighbor	Attribute values to the closest neighbor
IDW Modified Shepard	Local modified Shepard's IDW	Optimize with respect to the size of the neighborhood for local polynomial fitting and inverse distance weight calculation
Set Sill and Nugget	Ordinary Kriging	Use a spherical variogram model with the sill = the sample variance and the nugget effect = the semi-variance estimate at the shortest sampling distance, and estimate the range
Set Sill, and Nugget=0	Ordinary Kriging	Use a spherical variogram model with the sill = the sample variance and the nugget effect = 0, and estimate the range
IDW Smoothed P:1	Shepard's IDW and optimization of sample values	Power value = 1, and search neighborhood = all neighbors
IDW Smoothed P:2	Shepard's IDW and optimization of sample values	Power value = 2, and search neighborhood = all neighbors

^aIDW: Inverse Distance Weighting; ^bP:” – IDW power parameter

Data Analysis

For a given sampling density and soil property (pH, K, P), the data were standardized to a zero mean and a unit variance before performing the interpolation, except for IDW Smoothed where the data span is part of the optimization. The standardized data were back-transformed for the report of results.

The Mean Squared Error (MSE) and the G criterion (Eq. 4.2) proposed by Agterberg (1984) were used to assess the performance of the interpolation procedures. The G criterion, which involves the MSE, compares the residuals for a specific approach with the residuals obtained if the field average was used:

$$G = \left(1 - \frac{MSE}{MSE_{Average}}\right) \times 100 \quad (4.2)$$

where MSE is the mean of squared errors for the evaluated method, while $MSE_{Average}$ denotes the mean of squared error if the field average was used instead for the interpolation. Positive G values indicate an advantage of using the evaluated interpolation method over the field average. Negative G values imply that the field average and the associated flat surface provide a more accurate interpolated surface. A zero G value means equivalency.

Independent validation samples were only available for Field 1, which allowed for validation at all three sampling densities. For Fields 2, 3, and 4, the samples discarded to produce the 0.8 and 3.5 ha·sample⁻¹ sampling grids were used to validate the results for these two densities only (Figure 4.1). Cross-validation could have been used but was considered to be out of the scope of the study, and biased results could be obtained by cross-validation in the case of low-density sampling designs (Wadoux et al., 2021).

A pairwise Levene's test was employed to evaluate the standard errors calculated through the validation samples. When the above-mentioned statistical test rejected the null hypothesis (homogeneity of variances) at a significance level of 0.05 for a pair of interpolation procedures, their interpolation accuracy was considered significantly different since there was heterogeneity in the variance of their standard errors. Data processing, interpolation, and statistical analysis were performed using customized scripts written in the R language (R Core Team, 2022).

4.3 Results and Discussion

The descriptive statistics for the different grid sampling designs and validation samples are reported in Table 4.3. For Field 1, the mean and median values of the K and P variables are slightly smaller for the validation set than for the grid samples, whereas the contrary is observed for pH. For Fields 2, 3, and 4, where the samples removed from the 0.4 ha·sample⁻¹ grid sampling design were used for validation, the descriptive statistics for grid and validation samples are very similar. From the highest sampling density down to 0.8 and 3.5 ha·sample⁻¹, the number of samples is reduced by about 50% and 90%, respectively. However, the means and medians calculated from the grid samples in a given field for a given soil property present only slight changes across the different sampling densities. This indicates that representative samples from the underlying surface were collected regardless of the sampling density.































The experimental variogram from the highest sampling density available (0.4 ha·sample⁻¹) and a weighted least-squares fitting procedure were used to obtain variogram model parameters estimates for each field and soil property (Table 4.4). The obtained values were considered the best available estimates of the spatial structure of the underlying surface from which the samples were collected. Note that for this sampling density, the number of samples available was higher than 100 (Table 4.3), as Webster & Oliver (1992) recommended.

An analysis of Table 4 presents differences and similarities in the spatial variability of soil properties within and across fields. For example, for Field 1, an exponential model was selected for K. In contrast, a spherical model was selected for P and pH, indicating a difference in the behavior of the spatial correlation in the data. For this same field, all soil properties presented a strong spatial structure (classification modified from Cambardella et al., 1994), whereas, for Field 2, a strong spatial structure is only observed for pH and weak for K and P. In general, the different spatial structure classes in Table 4.4 highlight the differences in the spatial variability across the fields and soil properties, an important data characteristic when evaluating different interpolation methods.

Table 4.3 Descriptive statistics for plant available Potassium (K), plant available Phosphorus (P), and pH for all the fields, sampling densities, and validation samples (a hyphen indicates that values are not available)

Field	Density	Grid Samples						Validation Samples				
		SP ^a	n ^b	Mean	SD ^c	Med ^d	Histo-gram	n	Mean	SD	Med.	Histo-gram
1	0.4 ha-sample ⁻¹	K	108	133.9	37.6	130.5		20	128.0	31.5	122.5	
		P		34.0	14.8	33.5			29.3	18.4	24.0	
		pH		7.37	0.49	7.30			7.58	0.79	7.75	
	0.8 ha-sample ⁻¹	K	54	155.3	47.8	147.0		20	128.0	31.5	122.5	
		P		39.0	21.6	35.0			29.3	18.4	24.0	
		pH		7.45	0.51	7.40			7.58	0.79	7.75	
	3.5 ha-sample ⁻¹	K	12	127.8	40.7	136.5		20	128.0	31.5	122.5	
		P		31.3	15.4	33.0			29.3	18.4	24.0	
		pH		7.39	0.51	7.20			7.58	0.79	7.75	
2	0.4 ha-sample ⁻¹	K	216	123.4	35.7	119.0		-	-	-	-	-
		P		16.1	6.1	15.0		-	-	-	-	-
		pH		7.42	0.45	7.50		-	-	-	-	-
	0.8 ha-sample ⁻¹	K	110	122.3	28.7	119.0		106	124.5	41.9	121.0	
		P		16.1	6.5	15.0			16.0	5.8	15.0	
		pH		7.42	0.45	7.50			7.42	0.46	7.50	
	3.5 ha-sample ⁻¹	K	21	120.8	23.3	118.0		195	123.7	36.8	119.0	
		P		14.5	4.1	14.0			16.2	6.3	15.0	
		pH		7.39	0.47	7.40			7.43	0.45	7.50	

Table 4.3 cont.

Field	Density	Grid Samples						Validation Samples				
		SP ^a	<i>n</i> ^b	Mean	SD ^c	Med. ^d	Histo-gram	<i>n</i>	Mean	SD	Med.	Histo-gram
3	0.4 ha·sample ⁻¹	K	274	125.9	48.7	117.5		-	-	-	-	-
		P		15.4	9.4	12.0		-	-	-	-	-
		pH		7.32	0.76	7.60		-	-	-	-	-
	0.8 ha·sample ⁻¹	K	136	127.6	52.2	118.0		138	124.2	45.1	116.5	
		P		15.7	9.3	12.5			15.0	9.5	12.0	
		pH		7.31	0.77	7.60			7.32	0.75	7.60	
	3.5 ha·sample ⁻¹	K	30	127.6	51.8	117.5		244	125.7	48.4	117.5	
		P		15.9	11.2	12.0			15.3	9.2	12.0	
		pH		7.33	0.78	7.70			7.32	0.75	7.60	
4	0.4 ha·sample ⁻¹	K	144	363.6	82.2	346.0		-	-	-	-	-
		P		20.9	9.7	18.5		-	-	-	-	-
		pH		6.58	0.59	6.60		-	-	-	-	-
	0.8 ha·sample ⁻¹	K	72	354.2	80.4	342.5		72	373.1	83.5	350.5	
		P		19.7	9.2	18.0			22.0	10.1	20.0	
		pH		6.57	0.57	6.50			6.60	0.61	6.70	
	3.5 ha·sample ⁻¹	K	16	379.1	102.5	358.0		128	361.7	79.6	345.0	
		P		19.6	8.5	16.5			21.0	9.8	19.5	
		pH		6.56	0.51	6.65			6.59	0.60	6.60	

^a Soil Property, ^b*n* – number of samples, ^cSD – Standard Deviation, ^dMed. - Median

Table 4.4 Variogram model parameter estimates from a standard variogram fitting procedure for the four fields at the highest sampling density collection (Table 4.3), after data was standardized to a zero mean and a unit variance

Field	Density	Soil Property	Model Type	Nugget Effect	Sill	Range (m)	N:S ^a	Spatial Structure ^b
1	0.4 ha·sample ⁻¹	K	Exponential	0.21	1.07	93	0.20	Strong
		P	Spherical	0.18	1.12	400	0.16	Strong
		pH	Spherical	0.00	1.16	237	0.00	Strong
2	0.4 ha·sample ⁻¹	K	Exponential	0.70	1.10	289	0.64	Weak
		P	Gaussian	0.94	1.50	1693	0.63	Weak
		pH	Spherical	0.21	1.04	244	0.20	Strong
3	0.4 ha·sample ⁻¹	K	Exponential	0.33	1.17	288	0.28	Medium
		P	Spherical	0.43	1.16	635	0.37	Medium
		pH	Spherical	0.06	1.20	557	0.05	Strong
4	0.4 ha·sample ⁻¹	K	Spherical	0.48	1.02	294	0.47	Medium
		P	Exponential	0.33	1.27	316	0.26	Medium
		pH	Exponential	0.45	1.12	256	0.40	Medium

^aN:S – Nugget-Effect-to-Sill Ratio, ^bSpatial structure classification based on the Nugget-Effect-to-Sill ratio, $N:S < 0.25$ = strong spatial structure, $0.25 \leq N:S \leq 0.6$ = medium spatial structure, and $N:S > 0.6$ = weak spatial structure (classification modified from Cambardella et al., 1994 , originally with 0.75 instead of 0.6 for the weak spatial structure)

Box plots showing the distribution of interpolation errors for the 11 procedures evaluated for Field 1 (3 variables x 3 sampling densities) are presented in Figure 4.4. In each panel, box plots are sorted by increasing values of MSE from top to bottom. The letters a and b beside the reported MSE values correspond to the results of Levene’s test; two MSE values that are not followed by the same letter differ significantly at $\alpha = 0.05$. The largest number of significant differences among MSE values are for variable P at the sampling densities of 0.4 and 0.8 ha·sample⁻¹ (Figure 4.4, panels b and e). The G values (on the left in box plots) reveal advantages and disadvantages relative to the field average. For example, in Figure 4.4g, G values indicate that “IDW Smoothed P:1 and P:2,” “Set Sill and Nugget,” and “IDW P:1” generated surfaces with MSEs similar to or better than

“Average.” In contrast, interpolating data using “IDW P:2,” “Set Sill, and Nugget=0,” “Optimal IDW,” “Nearest Neighbor,” and “IDW Modified Shepard” results in less accurate surfaces than when the field average is used. No box plot is presented for “Fitted variogram model” in Figure 4.4, panels g and i, because the model fitting algorithm failed to converge.

Only “IDW P:1” and “Set Sill and Nugget” show positive G values in all 9 panels of Figure 4.4, meaning absolute superiority over “Average” in Field 1. With the field average set aside, some interpolation procedures proved to be more or less reliable than others. For example, based on MSE values, “Optimal IDW” and “Nearest Neighbor” surfaces were found to be the most accurate (Figure 4.4b) or the least accurate (Figure 4.4a), although Levene’s test detected no statistically significant differences in Figure 4.4a. Overall, with a lower sampling density, the advantage of using the evaluated interpolators over the field average is reduced. Also, as expected (Larocque et al., 2007; Webster & Oliver, 1992), fitting a variogram model becomes more difficult or even practically impossible at the lowest sampling density of $3.5 \text{ ha} \cdot \text{sample}^{-1}$; the variogram model fitting algorithm then failed to converge for two of the three soil properties sampled from Field 1 (number of samples: 12; Table 4.3).

Concerning the alternative interpolation methods proposed in this paper, “Set Sill and Nugget” shows signs of lesser performance than “Set Sill, and Nugget = 0” based on the G values from Figure 4.4, panels h and i. That might indicate some advantage to setting the nugget-effect value to 0. However, the G value of “Set Sill, and Nugget = 0” is negative in Figure 4.4g (K, lowest sampling density), whereas the corresponding G value for “Set Sill and Nugget” is positive. Leading to the assumption that for Field 1, “Set Sill and Nugget” performed as expected, producing either better results than the field average (i.e., positive G values) or as good as the field average (i.e., $G = 0$; Figure 4.4i), while the same does not apply for “Set Sill, and Nugget = 0”. Thus, even though “Sill and Nugget” might not be the highest performing interpolator in all situations, it might be a reliable interpolation procedure for dataset from low and extra-low sampling densities since it did not produce results worse than average. As for “IDW Smoothed P:1” and “IDW Smoothed P:2”, slightly negative G values are observed only for pH at the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ sampling density (Figure 4.4i).

The interpolated surfaces for P – the only variable that presented some statistical significance among the different interpolation procedures and for which the variogram fitting converged for all sampling densities – are presented in Figures 4.5-4.7. These maps exemplify the effect of the 10

different interpolation procedures in the spatial variability of P for all three sampling densities. Clearly, some procedures generate rougher surfaces, with abrupt changes between neighboring values (*e.g.*, Nearest Neighbor – Figures 4.5g, 4.6g, and 4.7g), while others generate smoother surfaces, presenting little to no spatial variability (*e.g.*, IDW Smoothed P:1 – Figure 4.5d).

From a soil management perspective, challenges can be encountered in both extreme scenarios mentioned above. For example, the “Nearest Neighbor” interpolated surface would generate variable rate maps with neighboring regions with abrupt changes in the amount of P fertilizer. These sudden changes in the prescribed fertilizer rate, combined with the rate adjustment lag existing in most variable rate applicators (Fulton et al., 2005), would reduce the quality and accuracy of the application. On the other hand, extremely smooth surfaces (*e.g.*, “IDW Smoothed P:1”; Figure 4.5d) could lead to the recommendation of a uniform rate application of P, even though there is a clear spatial variability of this nutrient in Field 1 (Figure 4.5a). Applying the same rate of P throughout Field 1 would result in a surplus and deficiency of this macronutrient in the field, negatively affecting the crop yield and potentially contaminating surface and groundwater (Liu et al., 2021). Although the levels of plant-available P in Field 1 are classified as medium to very-high for western Canadian soils (Alberta Ministry of Agriculture and Irrigation, n.d), the above-described scenarios highlight the impact of interpolation procedures on soil management practices.

A visual comparison of Figures 4.5- 4.7 shows similarities among “Set Sill, and Nugget = 0”, “IDW Modified Shepard,” and “Optimal IDW” maps, and that smoother maps are obtained from “IDW Smoothed P:1 and P:2” in comparison to the original IDW. Also, compared to “Set Sill, and Nugget = 0” and “Fitted variogram model,” “Set Sill and Nugget” tends to generate smoother surfaces, an effect of the higher nugget-effect estimates obtained from the latter procedure, a well-known behavior in geostatistics (Chilès and Delfiner, 2012).

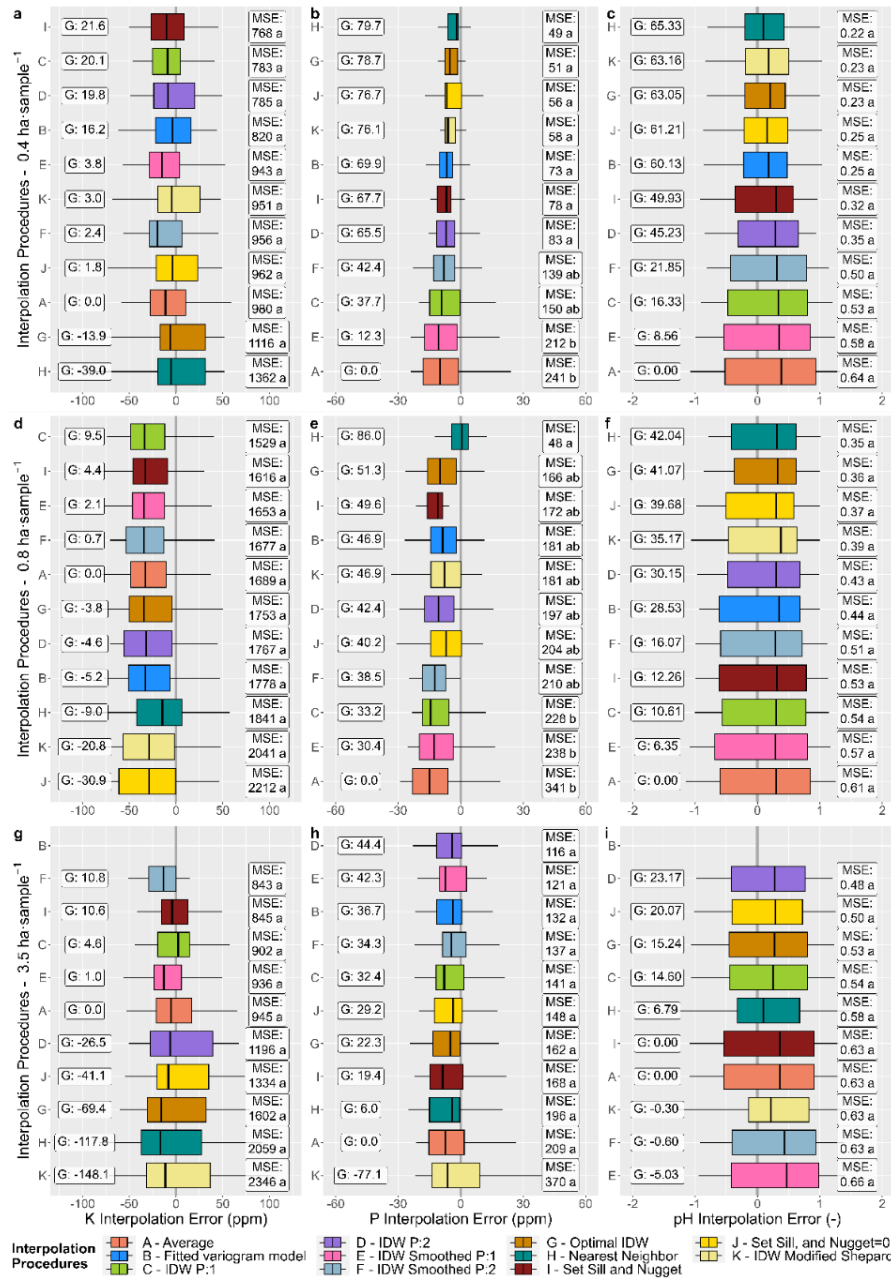


Figure 4.4 Box plots of interpolation errors for Field 1 at grid sampling densities of 0.4 ha-sample⁻¹ (panels a-c), 0.8 ha-sample⁻¹ (panels d-f), and 3.5 ha-sample⁻¹ (panels g-i). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: “Fitted variogram model” was removed from the analysis and does not appear in a few panels because the variogram model fitting algorithm failed to converge for that specific sampling density and soil property. Letters a and b indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.

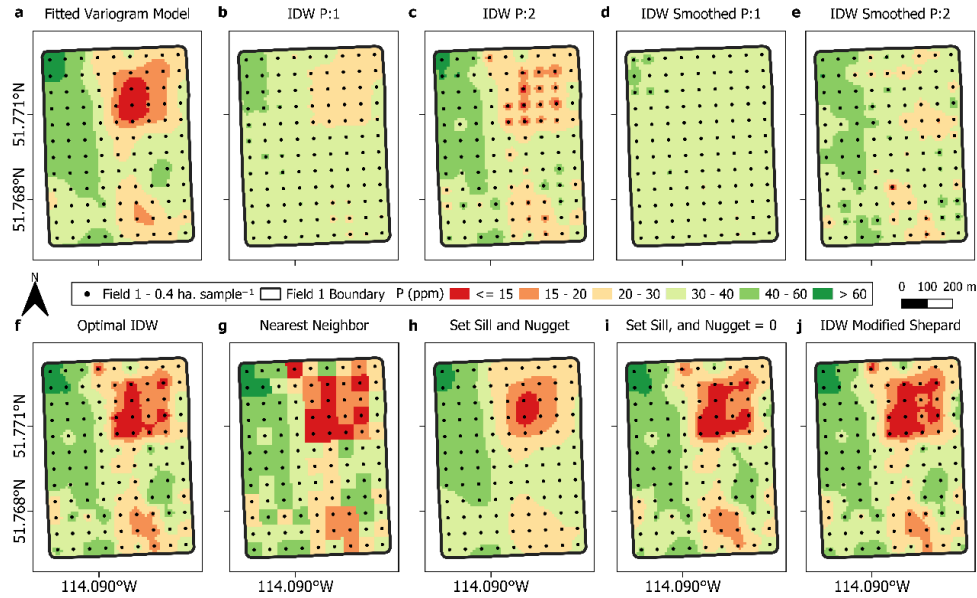


Figure 4.5 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the $0.4 \text{ ha} \cdot \text{sample}^{-1}$ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps.

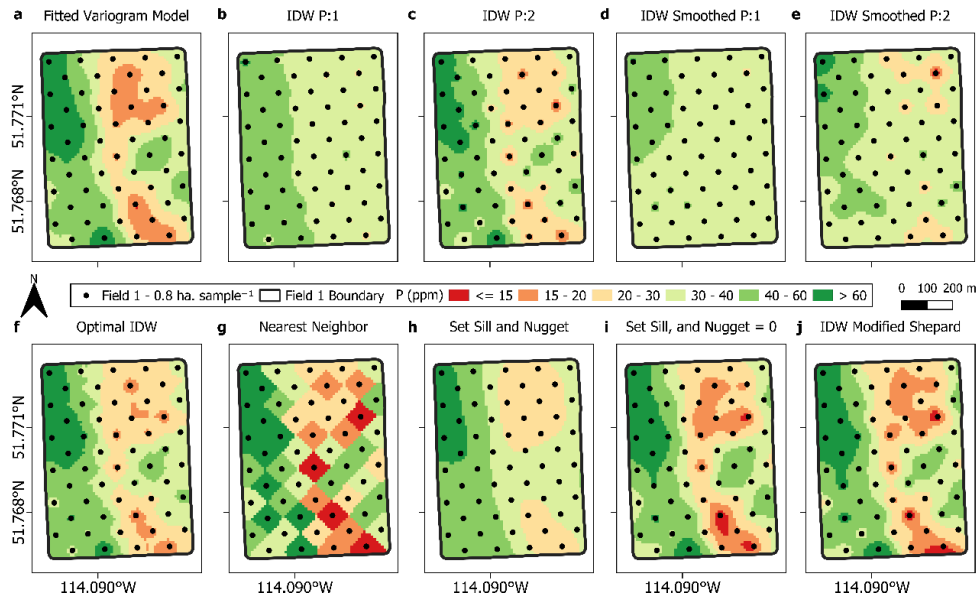


Figure 4.6 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the $0.8 \text{ ha} \cdot \text{sample}^{-1}$ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps.

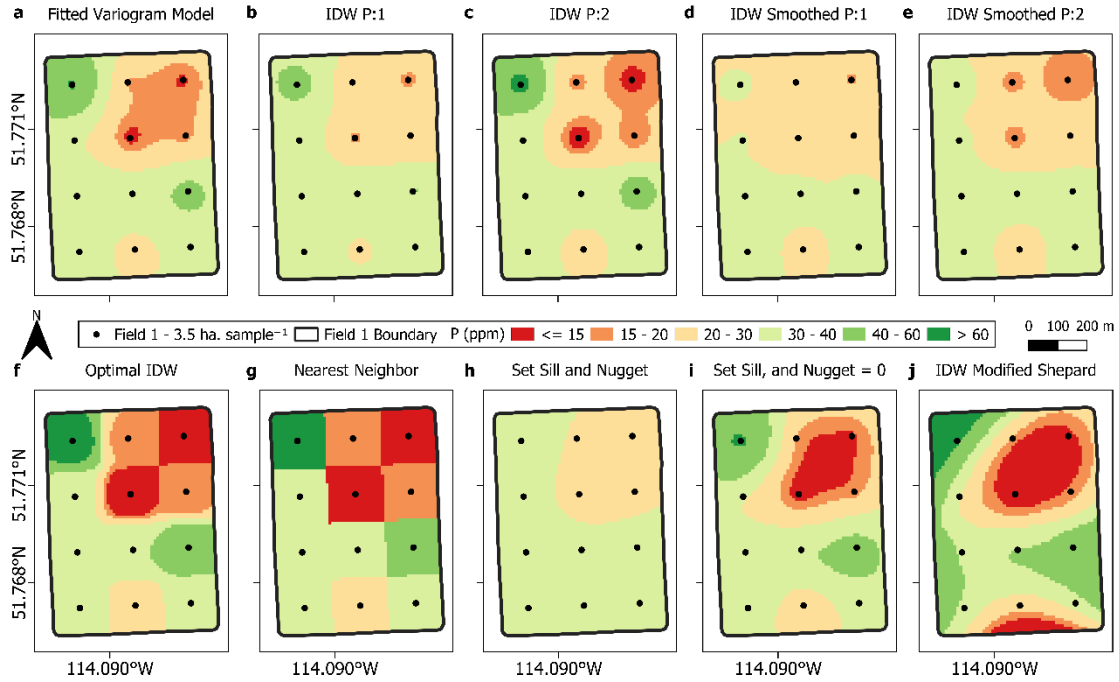


Figure 4.7 Interpolated maps from Field 1 representing the spatial variability for phosphorous (P) using the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ sampling density and 10 different interpolation procedures (maps for the “Average” procedure were not included as it does not show spatial variability). All the maps share the same legend – depicted between the two rows of maps.

Box plots for Field 2 are presented in Figure 4.8. Due to the similarity/dissimilarity in spatial structure of the three soil properties in this field (weak for K and P, and strong for pH; Table 4.4), the behavior of the G criterion is not only influenced by sampling density, but also by the spatial correlation present in the data. At the $0.8 \text{ ha} \cdot \text{sample}^{-1}$ sampling density (i.e., the higher evaluated density in Field 2), the G values are higher for pH (Figure 4.8c) than for K and P (Figures 4.8a and 4.8b). At the lower density in Field 2 ($3.5 \text{ ha} \cdot \text{sample}^{-1}$), the G values of pH show a clear worsening in performance of the interpolation procedures relative to the field average for this soil property (see Figure 4.8c vs. Figure 4.8f), while the G values of K and P rather tend to show a statu quo (see Figures 4.8a and 4.8b vs. Figures 4.8d and 4.8e).

The above result is consistent with findings reported in the literature. For example, Kravchenko (2003) compared kriging and IDW by simulating surfaces with various spatial structures and sampling densities, and found that for surfaces with stronger spatial structures, a reduction in sampling density produces a drop of the G values, whereas for weaker spatial structures, only slight changes are observed. This does not imply that the selection of a reliable interpolator can be

neglected for soil properties with weak spatial structure. Indeed, the results for P in Figure 4.8e indicate that depending on the choice of the interpolation procedure, the G value can be as low as -264%, the exception to the statu quo rule mentioned above. Also, the data are not 'known' before sampling, nor is their spatial correlation.

Besides "IDW Smoothed P:1", all the other interpolation procedures are less accurate than the field average in Figure 4.8e. In particular, the G value for the modified kriging-based procedure "Set Sill and Nugget" is slightly negative (-13.6%), while "Set Sill, and Nugget=0" is one the worst when compared to "Average" ($G = -106.0\%$). It is noteworthy that for a soil property with a stronger spatial structure, or equivalently a smaller nugget effect (*e.g.*, pH in Field 2; see Table 4.4), setting the nugget effect value to zero yields a more accurate surface than "Set Sill and Nugget." In contrast, for properties with a larger nugget effect (*e.g.*, P and K in Field 2; see Table 4.4), setting this model parameter (or its estimate) to zero produced less accurate interpolated surfaces than when the semi-variance estimate at the shortest sampling distance was used. For example, for P sampled at the density of $3.5 \text{ ha} \cdot \text{sample}^{-1}$ in Field 2, the interpolated surface of "Set Sill and Nugget" is significantly more accurate than that of "Set Sill and Nugget = 0" based on Levene's test (Figure 4.8e), even though both their G values are negative. The interpolation procedure "IDW Smoothed P:1" is the only one that produced surfaces more accurate than those of "Average" for all three soil properties in Field 2 at the extra-low sampling density of $3.5 \text{ ha} \cdot \text{sample}^{-1}$, but not so much at the $0.8 \text{ ha} \cdot \text{sample}^{-1}$ density (Figure 4.8).

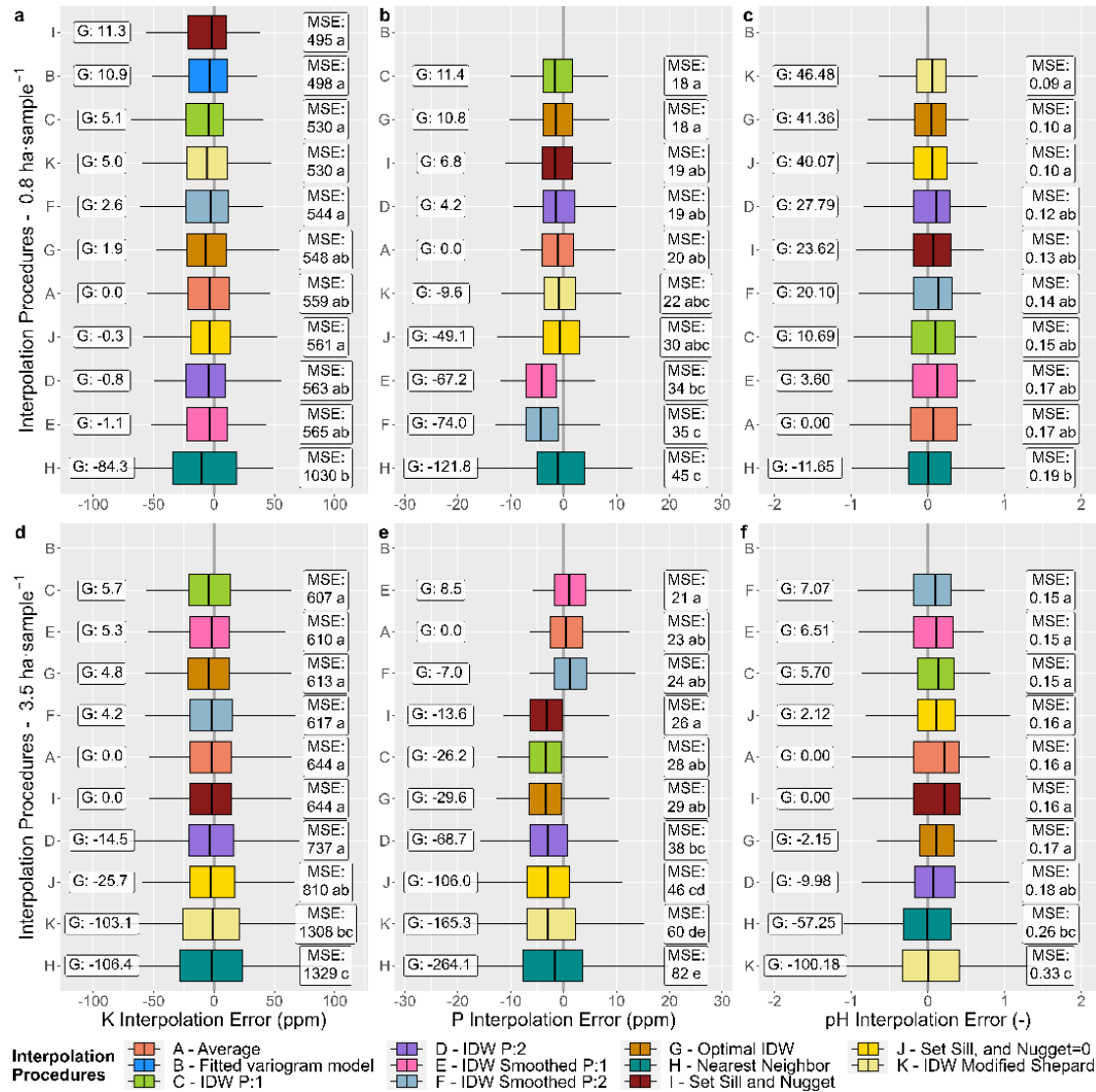


Figure 4.8 Box plots of interpolation errors for Field 2 at grid sampling densities of 0.8 ha·sample⁻¹ (panels a-c) and 3.5 ha·sample⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: “Fitted variogram model” was removed from the analysis and does not appear in most panels because the variogram model fitting algorithm frequently failed to converge. Letters a, b, c, d, and e indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.

Soil properties K and P show medium spatial structure in Field 3, while the spatial structure for pH appears strong (Table 4.4). Field 3 is the largest in size of the four study sites, thus yielding the largest number of observations at the sampling densities of 0.8 and 3.5 ha·sample⁻¹. Since a

medium to strong spatial structure and a larger number of sample data reduce the uncertainty in variogram analysis, including the estimation of model parameters (Larocque et al., 2007), the variogram model fitting algorithm converged without exception for Field 3 (Figure 4.9).

Fields 1 and 2 results indicated limited advantages reflected by positive G values for some interpolation procedures over “Average,” with little evidence for statistical significance. For Field 3, many of the MSEs computed from the interpolation errors show a significant improvement in interpolation accuracy over the field average. Except for P sampled at the extra-low density in Field 3 (Figure 4.9e), all the panels in Figure 4.9 identify a group of procedures yielding interpolated surfaces declared significantly more accurate than “Average” by Levene’s test. “Fitted variogram model” and “Set Sill and Nugget” are the only two interpolation procedures that consistently appear in these groups.

Following the interpretation rules developed and applied for Fields 1 and 2, the results obtained for Field 3 can be presented and discussed as follows. The G value of “Set Sill, and Nugget=0” is greater than that of “Set Sill and Nugget” for pH at both sampling densities in Field 3, which was expected as the estimated nugget effect was close to zero (Table 4.4). Since the estimated nugget effect for K and P were higher than for pH in Field 3 but did not reach 0.5 for a maximum value of 1.0 (Table 4.4), it is ‘as expected’ that “Set Sill and Nugget” performed better than “Set Sill, and Nugget=0” for one of these two soil properties (K), but not the other (P) at the low sampling density (Figure 4.9b). Levene’s test for P detected no significant difference between “Set Sill and Nugget” and “Set Sill, and Nugget=0” at both sampling densities in Field 3. Concerning the interpolation procedures “IDW Smoothed P:1” and “IDW Smoothed P:2”, their G values were mostly positive, except for K at $0.8 \text{ ha} \cdot \text{sample}^{-1}$ (Figure 4.9a); in the comparisons with “Average,” “IDW Smoothed P:2” consistently presents an advantage over “IDW Smoothed P:1” in Field 3, with no significant difference between the two.

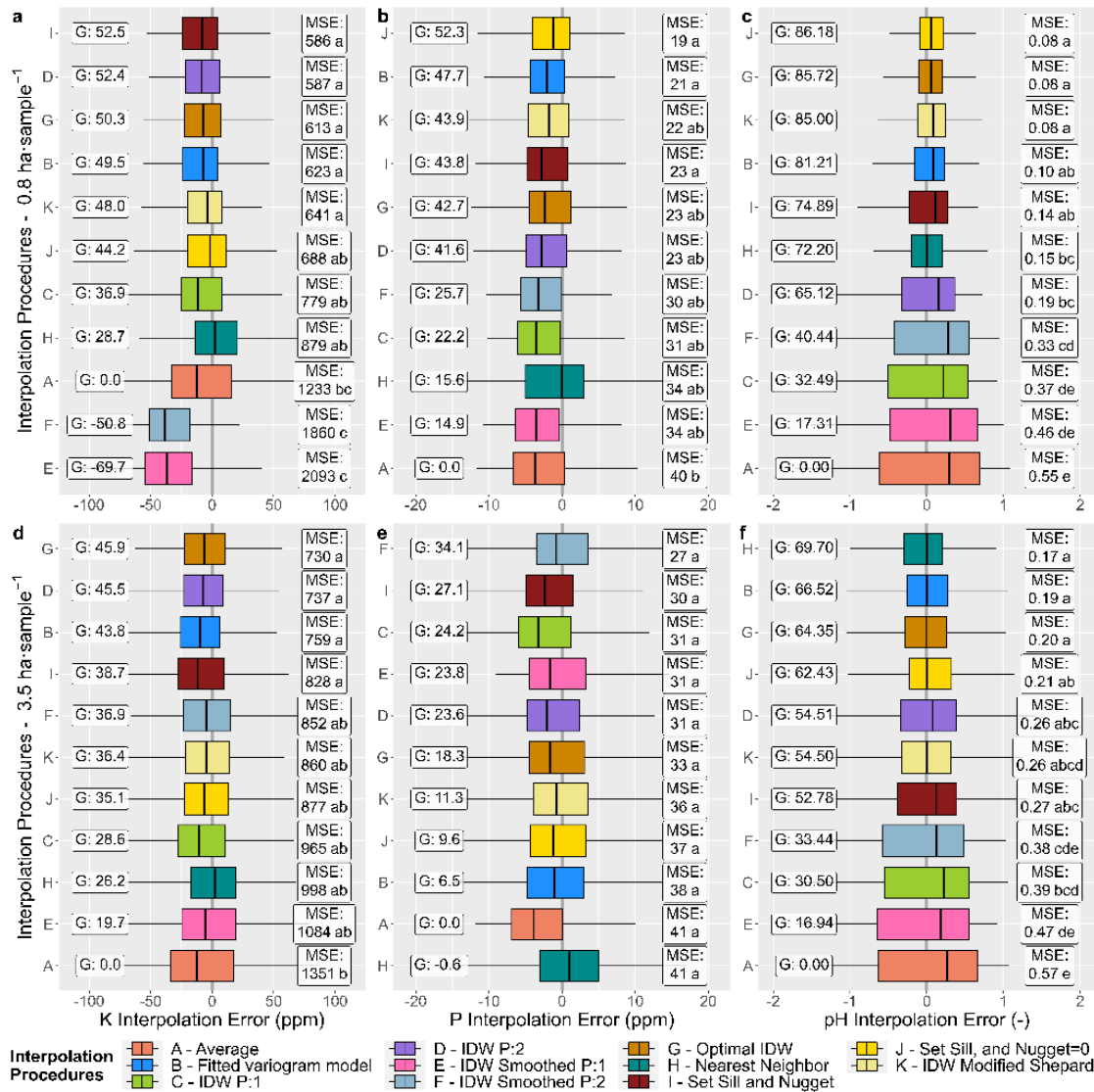


Figure 4.9 Box plots of interpolation errors for Field 3 at grid sampling densities of 0.8 ha·sample⁻¹ (panels a-c) and 3.5 ha·sample⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Letters a, b, c, d, and e indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene's test.

Several of the results obtained for Field 4 (Figure 4.10), located approximately 100 km east of the other three fields (Figure 4.1a), resemble some of those obtained for Field 1 (Figure 4.4). For example, there is no significant difference among most of the interpolation procedures that are compared at their respective sampling densities. Notably, “IDW Smoothed P:1 and “P:2” consistently yielded positive G values for all the soil properties and densities. The variogram model

fitting algorithm failed to converge for K and pH at the extra-low sampling density (Figures 4.10d and 4.10f). For the same soil properties and sampling density, a flat variogram was obtained with “Set Sill and Nugget.” For obvious reasons, “Set Sill and Nugget” presents a clear advantage over “Set Sill, and Nugget = 0” in these two cases.

Based on the results presented, none of the interpolation procedures consistently emerged as the best for all fields, soil properties, and sampling densities, suggesting that the optimal interpolator will change depending on the spatial structure present in the data and sampling density. This conclusion is consistent with the findings reported in the literature (Kravchenko, 2003; Kravchenko and Bullock, 1999; Robinson and Metternicht, 2006). It is noteworthy that for all fields, the field average never emerged as being the basis for the best approach.

Concluding the investigation at this point would leave the research question, “What is a robust and reliable interpolation procedure that maximizes the value of low-density soil sampling data?” unanswered. The keywords here are “reliable” and “robust,” and not “best,” as the presented results suggest that without having prior knowledge about the spatial structure of the data and the collection of validation samples, it would be difficult, if not impossible, to know which interpolation procedure would result in the most accurate surface.

Considering that a reliable and robust interpolator for low and extra-low sampling density is based on a procedure that will hardly yield surfaces less accurate than the field average and possibly represent the most accurate surface. The G values reported in the box plots (Figures 4.4, 4.8-4.10) were classified in three groups - $G < 0$ (surface less accurate than field average), $0 \leq G < \text{Highest } G$ (surface is as accurate as the field average but not the highest G value), and Highest G (interpolation approach with the most accurate surface), and the frequency of each category was assessed for the procedures. The results are presented in Figure 4.11.

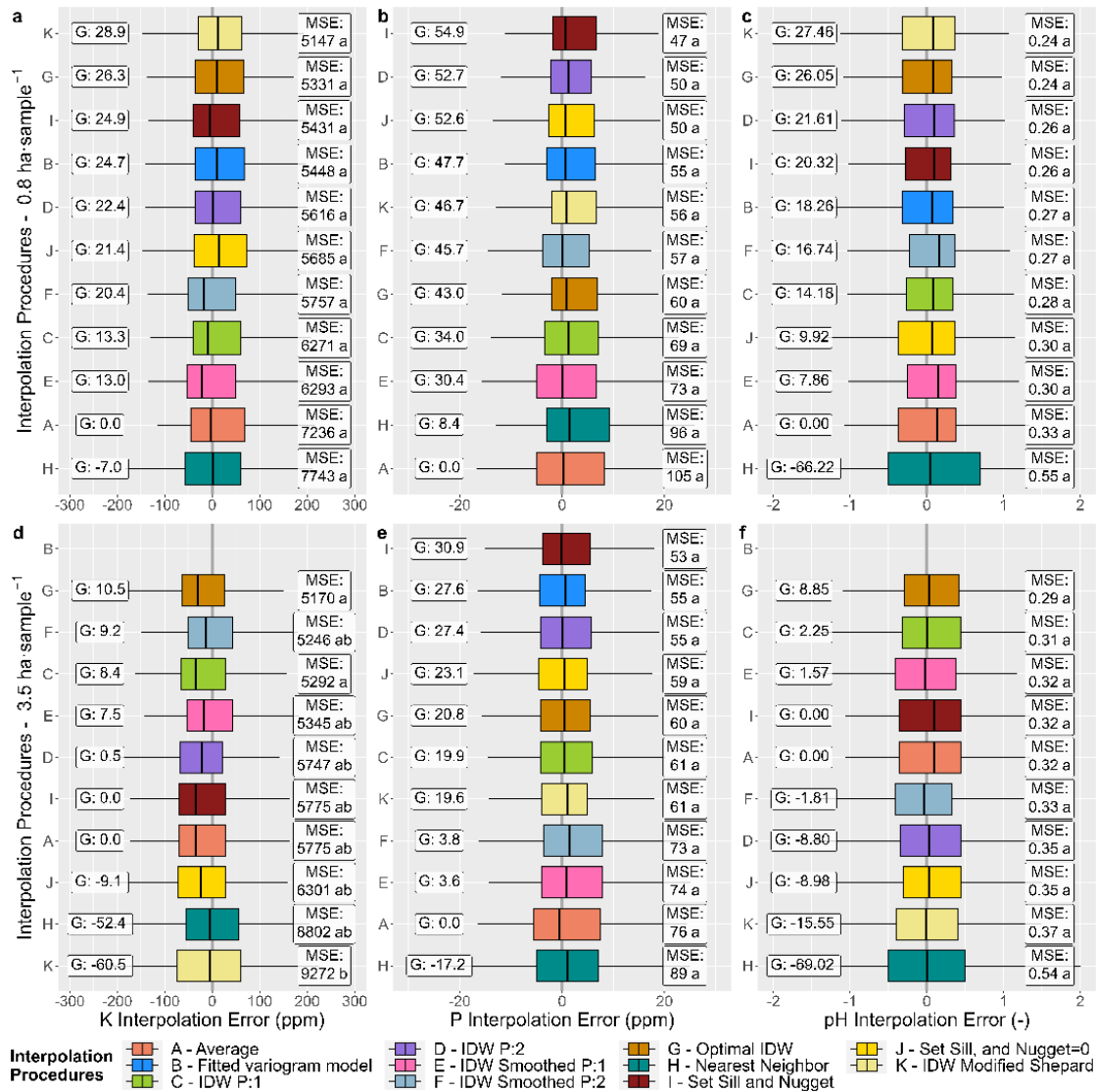


Figure 4.10 Box plots of interpolation errors for Field 4 at grid sampling densities of 0.8 ha-sample⁻¹ (panels a-c) and 3.5 ha-sample⁻¹ (panels d-f). Results are presented for a total of 11 interpolation procedures, each identified by a capital letter and a color; for details, see Table 4.2. Note: Fitted variogram model” was removed from the analysis and does not appear in a few panels because the variogram model fitting algorithm failed to converge for that specific sampling density and soil property. Letters a and b indicate the differences in MSE among interpolation procedures that are declared statistically significant ($\alpha = 0.05$) with Levene’s test.

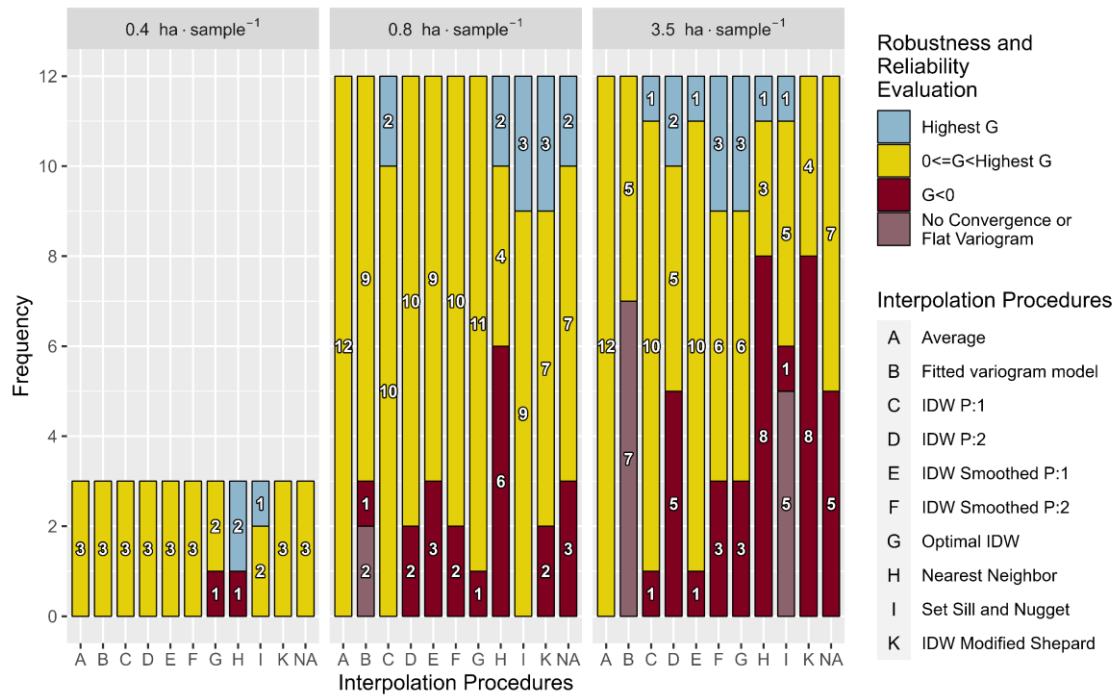


Figure 4.11 Robustness and reliability analysis for the 11 interpolation procedures at the three different sampling densities. G values from Figures 4.4 and 4.8- 4.10 were grouped in the categories listed in the legend. “No Convergence” applies to the kriging approach “B – Fitted variogram model” when the model fitting algorithm failed to converge, and “Flat Variogram” applies to the modified kriging approach “J – Set Sill and Nugget” when estimated nugget-effect is the equal or higher than sample variance.

Focusing the analysis on the two lower sampling densities in Figure 4.11 and evaluating the different procedures by their robustness and reliability, following the definition provided earlier, the most robust procedure at 0.8 ha · sample⁻¹ would be “Set Sill and Nugget.” For the extra-low sampling design, determining which procedure is the most robust proves to be challenging, as negative G values or “No convergence or Flat Variogram” are observed in all procedures with no exceptions.

One might argue that the “Fitted variogram model” could be considered the most robust procedure for the 3.5 ha · sample⁻¹ sampling density, as it did not yield any negative G values. If, in cases where the fitting algorithm failed to converge, the field average would be used instead, this would be a plausible affirmation. However, 7 out of 12 maps would lack spatial information. Moreover, over or underestimation of the variogram model parameters during the fitting procedure

could produce highly inaccurate surfaces – a scenario not covered by the dataset used in this paper but supported by the findings from Webster & Oliver (1992) and Larocque et al. (2007) which suggest that a variogram originating from a low number of samples can present a wide confidence interval, leading to high uncertainty in the estimation of the model parameters.

The “Set Sill and Nugget” results are not too different from the “Fitted variogram model,” for 5 out of 12 maps, a flat variogram would be used to obtain the kriging estimates, producing the same result as for the field average (assuming global neighborhood – all available neighbors – was used). In contrast to the “Fitted variogram model,” “Set Sill and Nugget” once generated the most accurate surface for the 3.5 ha·sample⁻¹ but also presented a negative G value – a specific case where most of the procedures also presented $G < 0$ (Figure 4.8e). Notably, both model-based approaches presented difficulties in consistently estimating model parameters that would generate an interpolated surface with a representation of the field spatial variability. Two other potential candidates for a robust interpolation procedure for this extra-low-density design were “IDW P:1” and “IDW Smoothed P:1”. Since “IDW P:1” was among the most robust methods for the 0.8 ha·sample⁻¹ sampling density, this method could be selected as the most robust. However, by using IDW, the issues with non-smooth surfaces due to the “bull’s eye” effect would remain, affecting the quality of prescription maps created based on the IDW surface.

Further analysis of the results presents that when the “Set Sill and Nugget” results in “pure nugget effect” (Figures 4.4g, 4.4i, 4.8d, 4.8f, and 4.10d), “IDW P:1” and “IDW Smoothed P:1” presented positive G values, with an exception for the latter which presented slightly negative G value for pH from Field 1 sampled at 3.5 ha·sample⁻¹ (Figure 4.4g). Thus, “Set Sill and Nugget” could be used when some spatial structure is still available in the data, but when a flat variogram is estimated, either “IDW P:1” or “IDW Smoothed P:1” could be employed to generate an interpolated surface that would at least provide a general trend of that soil property in the field. To make the newly proposed interpolation procedures readily available to precision agriculture practitioners, they could be implemented in existing open-source tools such as the Smart-Map (Pereira et al., 2022a) plugin for QGIS and AgDataBox-Map (Michelon et al., 2019) web platform.

Even though the results indicate a potential for the use of the combination of “Set Sill and Nugget” with “IDW P:1” or “IDW Smoothed P:1,” the analysis was still limited to soils available in Central Alberta. Also, since the samples used for validation were not consistent for all sampling

densities (except for Field 1), a formal statistical analysis of the effect of sampling density in each interpolation procedure was not performed, as it could generate biased results.

Therefore, future research should further explore the above-suggested combination of procedures and compare these to some of the already available machine learning-based spatial interpolation methodologies (Hengl et al., 2018; Pereira et al., 2022b). In addition, the robustness of the different interpolation methods could be evaluated from a perspective of the risks (economic and environmental) related to the soil management decisions that the practitioners would adopt based on the resulting interpolated surfaces. Based on the results from the comprehensive analysis presented in this paper, a focus should not be on identifying the best overall method but on the one that would be the most robust (*e.g.*, lower economic and environmental risks) in a wide range of scenarios. Also, for more representative results and identification of a “universal” solution for interpolation, soil samples from more fields, different sampling designs, and parts of the world should be included in the analysis.

4.4 Conclusions

A total of eleven interpolation procedures were undertaken – including a newly proposed methodology, two proposed approaches, and the use of field average – but none emerged as the best interpolator across all different fields, soil properties, and sampling densities. In terms of robustness, the proposed modification to the kriging approach, IDW, and IDW Smoothed with a power parameter of 1 appear among the most robust approaches, as they rarely yielded errors worse than when using the field average. In addition, when the kriging-based approach estimated a flat, “pure nugget effect,” variogram interpolated surfaces using IDW and IDW Smoothed often presented an advantage over using the field average, which indicates that a combination of these procedures could lead to interpolated surfaces that would maximize the value of low-density sampling designs.

A few other important outcomes were identified while performing this extensive comparison of interpolation methods and approaches. Among all fields, soil properties, and sampling density, at least one interpolation method always yielded a surface more accurate than the field average (not necessarily producing a statistically significant difference). Moreover, the best interpolation procedure is tied to the sampling density and spatial structure present in the data. Finally, forcing the nugget to be zero when there is poor information about the behavior of the variogram at short distances is a high-risk decision, as it can lead to a low accuracy result if the spatial correlation of

the underlying surface is weak – information that cannot be determined based on low and extra low-density data.

Therefore, PA practitioners should avoid using interpolation tools that tend to force the nugget to be zero unless the variogram estimates calculated from the data provide enough direct (indirect) information about the absence of a nugget effect. Also, to identify a “universal” interpolation approach for low-density sampling designs, a focus should be given to methods that hardly produce results worse than average and not necessarily the best results. Also, before making any soil management decisions based on thematic maps, practitioners should carefully evaluate different interpolation procedures through validation samples (when available) or cross-validation to avoid causing economic or environmental impacts.

4.5 Acknowledgments

We thank Olds College of Agriculture & Technology and Olds College Center for Innovation for providing the infrastructure, support, and data for the development of this project. We also thank the Canadian Agri-Food Automation and Intelligence Network (CAAIN), Mitacs, and Telus for the funds provided for the project and data collection. This research is part of the “Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services” project funded through the Mitacs Accelerate program.

4.6 Reference

- Agterberg, F. P. (1984). Trend Surface Analysis. In *Spatial Statistics and Models* (pp. 147–171). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-3048-8_8
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). New York, USA: ACM. <https://doi.org/10.1145/3292500.3330701>
- Alberta Ministry of Agriculture and Irrigation. (n.d.). Phosphorus management in crops. <https://www.alberta.ca/phosphorus-management-in-crops>. Accessed 23 February 2024
- Barnes, R. J. (1991). The variogram sill and the sample variance. *Mathematical Geology*, 23(4), 673–678. <https://doi.org/10.1007/BF02065813>
- Cambardella, C. A., Moorman, T. B., Novak, J. M., Parkin, T. B., Karlen, D. L., Turco, R. F., & Konopka, A. E. (1994). Field-Scale Variability of Soil Properties in Central Iowa Soils. *Soil Science Society of America Journal*, 58(5), 1501–1511. <https://doi.org/10.2136/sssaj1994.03615995005800050033x>

- Chilès, J.-P., & Delfiner, P. (2012). *Geostatistics*. Hoboken, NJ, USA: Wiley.
<https://doi.org/10.1002/9781118136188>
- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.
<https://doi.org/10.1109/4235.996017>
- Erickson, B., & Lowenberg-Deboer, J. (2021). 2021 Precision Agriculture Dealership Survey.
https://ag.purdue.edu/digitalag/_media/croplife-report-2021.pdf. last accessed Jan. 10 2022
- Franke, R., & Nielson, G. (1980). Smooth interpolation of large sets of scattered data. *International Journal for Numerical Methods in Engineering*, 15(11), 1691–1704.
<https://doi.org/10.1002/nme.1620151110>
- Fulton, J. P., Shearer, S. A., Higgins, S. F., Darr, M. J., & Stombaugh, T. S. (2005). Rate Response Assessment from Various Granular VRT Applicators. *Transactions of the ASAE*, 48(6), 2095–2103. <https://doi.org/10.13031/2013.20086>
- Goovaerts, P. (1999). Geostatistics in soil science: State-of-the-art and perspectives. *Geoderma*, 89(1–2), 1–45. [https://doi.org/10.1016/S0016-7061\(98\)00078-0](https://doi.org/10.1016/S0016-7061(98)00078-0)
- Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 2018(8). <https://doi.org/10.7717/peerj.5518>
- Isaaks, E. H., & Srivastava, R. M. (1989). *Applied Geostatistics*. NY: Oxford Univeristy Press, Inc.
- Kravchenko, A., & Bullock, D. G. (1999). (1999) Comparative Study of Interpolation Methods for Mapping Soil Properties, A (AJ), 400, 393–400.
- Kravchenko, A. N. (2003). Influence of Spatial Structure on Accuracy of Interpolation Methods. *Soil Science Society of America Journal*, 67(5), 1564–1571.
<https://doi.org/10.2136/sssaj2003.1564>
- Larocque, G., Dutilleul, P., Pelletier, B., & Fyles, J. W. (2007). Characterization and quantification of uncertainty in coregionalization analysis. *Mathematical Geology*, 39(3), 263–288.
<https://doi.org/10.1007/s11004-007-9086-8>
- Laslett, G. M., & McBratney, A. B. (1990). Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes—a case study for soil pH. *Journal of Soil Science*, 41(3), 451–471. <https://doi.org/10.1111/j.1365-2389.1990.tb00079.x>

- Liu, L., Zheng, X., Wei, X., Kai, Z., & Xu, Y. (2021). Excessive application of chemical fertilizer and organophosphorus pesticides induced total phosphorus loss from planting causing surface water eutrophication. *Scientific Reports*, 11(1), 1–8. <https://doi.org/10.1038/s41598-021-02521-7>
- Michelon, G. K., Bazzi, C. L., Upadhyaya, S., de Souza, E. G., Magalhães, P. S. G., Borges, L. F., et al. (2019). Software AgDataBox-Map to precision agriculture management. *SoftwareX*, 10, 100320. <https://doi.org/10.1016/j.softx.2019.100320>
- Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, 30(7), 683–691. <https://doi.org/10.1016/j.cageo.2004.03.012>
- Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Coelho, A. L. de F., Costa, M. M., & Grift, T. (2022a). Smart-Map: An Open-Source QGIS Plugin for Digital Mapping Using Machine Learning Techniques and Ordinary Kriging. *Agronomy*, 12(6), 1350. <https://doi.org/10.3390/agronomy12061350>
- Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Santos, N. T., & Fernandes-Filho, E. I. (2022b). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*, (0123456789). <https://doi.org/10.1007/s11119-022-09880-9>
- R Core Team. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria. <https://www.r-project.org/>
- Robinson, T. P., & Metternicht, G. (2006). Testing the performance of spatial interpolation techniques for mapping soil properties. *Computers and Electronics in Agriculture*, 50(2), 97–108. <https://doi.org/10.1016/j.compag.2005.07.003>
- Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In R. B. S. Blue & A. M. Rosenberg (Eds.), *Proceedings of the 1968 23rd ACM national conference on* - (pp. 517–524). New York, USA: ACM Press. <https://doi.org/10.1145/800186.810616>
- Sobjak, R., de Souza, E. G., Bazzi, C. L., Opazo, M. A. U., Mercante, E., & Aikes Junior, J. (2023). Process improvement of selecting the best interpolator and its parameters to create thematic maps. *Precision Agriculture*, 24(4), 1461–1496. <https://doi.org/10.1007/s11119-023-09998-4>
- Wadoux, A. M. J. C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457(August). <https://doi.org/10.1016/j.ecolmodel.2021.109692>

- Wadoux, A. M. J. C., Marchant, B. P., & Lark, R. M. (2019). Efficient sampling for geostatistical surveys. *European Journal of Soil Science*, 70(5), 975–989. <https://doi.org/10.1111/ejss.12797>
- Webster, R., & Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43(1), 177–192. <https://doi.org/10.1111/j.1365-2389.1992.tb00128.x>
- Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. *Vadose Zone Journal* (Vol. 1). Wiley. <https://doi.org/10.1002/9780470517277>

Connecting Text to Chapter 5

Chapter 4 determined a baseline for errors from a surface obtained from interpolating low and extra-low sampling designs without ancillary data. In **Chapter 5**, a methodology was proposed and evaluated to perform the fusion of PSS and topography data to predict the spatial variability of within-field soil chemical properties. The framework from **Chapter 3** to process PSS data and the range of spatial correlation from the modified kriging-based approach described in **Chapter 4** were utilized during the development of **Chapter 5**. Note that OM was not included in **Chapter 4** as it allowed for a more succinct discussion focusing on the interpolation methods. However, OM is an important soil chemical property often used when determining fertilizer prescription rates (e.g., OM is used to calculate the mineralization rate of organic nitrogen, which is used to adjust the rates for nitrogen fertilizer). Therefore, since only one field was evaluated in **Chapter 5**, OM was added to the analysis.

The results presented in **Chapter 5** were presented and published in the proceedings of the 16th International Conference on Precision Agriculture, which was held in Manhattan, Kansas, USA, in 2024.

Publication:

Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2024). Predicting soil chemical properties using proximal soil sensing technologies and topography data: A case study. In Proceedings of the 16th International Conference on Precision Agriculture (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Abbreviations for Chapter 5

Abbreviation	Definition
¹³⁷ Cs	Caesium-137 isotope
²³² Th	Thorium-232 isotope
²³⁸ U	Uranium-238 isotope
⁴⁰ K	Potassium-40 isotope
CAAIN	Canadian Agri-Food Automation and Intelligence Network
CR	Count Rate
DEM	Digital Elevation Model
EC _a	Apparent Electrical Conductivity
EMI	Electromagnetic Induction
GNSS	Global Navigation Satellite System
GPR	Ground Penetrating Radar
IDW	Inverse Distance Weighting
IDW MW	Inverse Distance Weighted Moving Window
K	plant-available Potassium
LiDAR	Light Detection and Ranging
ML	Machine Learning
MSE	Mean Squared Error
OK	Ordinary Kriging
OM	Soil Organic Matter
P	plant-available Phosphorus
PA	Precision Agriculture
PLSR	Partial Least Squares
ppm	parts per million
PSS	Proximal Soil Sensing
R ²	coefficient of determination
RF	Random Forest
RTK	Real-Time Kinematic
SD	Standard Deviation
SVM	Support Vector Machine
SW	Smoothing Window

Chapter 5: Prediction of soil chemical properties using proximal soil sensing technologies and topography data: Methodology and a case study

Felippe H. S. Karp, Viacheslav Adamchuk, Pierre Dutilleul, Alexei Melnitchouck, Asim Biswas

Abstract

Using proximal soil sensors (PSS) is widely recognized as a strategy to improve the quality of agricultural soil maps. Nevertheless, the signals captured by PSS are complex and usually relate to a combination of processes in the soil. Consequently, there is a need to explore further the interactions at the source of the information provided by PSS. The objectives of this study were to examine the relationship between proximal sensing techniques and soil properties and evaluate the feasibility of using data fusion to improve the mapping of soil chemical properties with extra-low sampling densities. Field data from ground penetrating radar, passive gamma-ray spectrometry, apparent electrical conductivity, resistance to penetration, and elevation were collected from a 43-ha site in Central Alberta, Canada. Soil sampling (originally with 0.4 ha·sample⁻¹ density) and subsequent lab analysis provided information on soil organic matter, pH, and plant-available phosphorous (P) and potassium (K). After pre-processing and co-locating the sampling and sensor data, soil properties and some PSS data were correlated. Samples were then removed until a density of 3.5 ha·sample⁻¹ was reached, thus creating an extra-low sampling density. Using PSS and topography data as predictors of the soil properties, machine learning (ML) algorithms (support vector machine, random forest, and partial least squares) were trained for each sampling density and validated using an additional 20 independent soil samples. Differences between ML models or sampling densities were insignificant for a given soil property. However, the mean squared error (MSE) and the coefficient of determination (R^2) indicated that some models outperformed others. Models with an R^2 value above 0.5 were for P and pH with the 0.4 ha·sample⁻¹ density and for P when the extra-low sampling design was applied. The definition of the evaluated ML algorithms does not consider the spatial location of the samples, which, from a mapping perspective, can create spatial inconsistencies; thus, to minimize this effect, an inverse distance smoothing window (SW) was applied to the predicted surfaces. The SW did not change predictions significantly, but often led to decreased R^2 and increased MSE values.

Keywords Data fusion, machine learning, sensor calibration, soil fertility mapping

5.1 Introduction

Within-field soil mapping is crucial to understanding, planning, and applying efficient, responsible, and accurate soil management practices. Soil sampling and its subsequent lab analysis is the most traditional approach for soil mapping and is often used for soil fertility assessment in agriculture (Gebbers, 2018). A single composite sample from a field might provide insights into its average fertility levels but not its internal variability. Thus, grid and zone samplings are standard precision agriculture (PA) practices to evaluate the within-field variability. In the 2023 PA dealership survey, a long-term study conducted in the United States (Erickson and Lowenberg-DeBoer, 2023), PA retailers estimated that 51% of their local market area uses georeferenced soil sampling (i.e., grid or zone sampling).

Based on these numbers, precision agriculture practitioners use grid and zone sampling as a within-field fertility mapping approach. On the other hand, the U.S. dealerships involved in the above-mentioned survey estimated that 49% of the area did not use such a service. Such a high percentage of non-adoption could be attributed to the labor-intensive, time-consuming, and expensive nature of soil sampling, which induces farmers to opt for only a composite sample for the whole field or, in extreme cases, not to collect samples.

Therefore, there exists a challenge to develop and evaluate time- and cost-efficient methods for assessing the within-field soil variability to increase the adoption of PA strategies for best soil management practices (e.g., variable rate fertilizer application). Adamchuk et al. (2011) and Gebbers (2018) suggested the fusion of proximal soil sensors (PSS) as a potential solution to this challenge, leading others to evaluate this approach. Ji et al. (2019) assessed the potential of machine learning algorithms (ML) using different combinations of apparent electrical conductivity (EC_a), passive- γ -ray spectrometry, visible and near-infrared spectroscopy, and topography as predictors for soil chemical properties. These authors reported an improvement in the prediction when sensors were fused compared to when used individually. Saifuzzaman et al. (2021) fitted multivariate linear models to predict soil chemical properties using EC_a and topographic derivatives, obtaining similar findings as Ji et al. (2019) (i.e., there is a potential for data fusion to predict and map soil properties).

Despite the potential of PSS to improve agricultural soil maps, the signals captured by such sensors are complex and usually relate to a combination of processes occurring in the soil (Gebbers, 2018). Also, previous research results, such as by Ji et al. (2019) and Saifuzzaman et al.

(2021), were obtained using sampling densities of approximately $0.25 \text{ ha} \cdot \text{sample}^{-1}$, higher than commonly used by PA practitioners ($1 \text{ ha} \cdot \text{sample}^{-1}$ - Erickson and Lowenberg-DeBoer, 2023). In addition, over the years, other geophysical techniques, such as ground penetrating radar (GPR), gained interest from the agricultural community, so their interactions with soil and its chemical properties must be investigated thoroughly.

Consequently, there is a need to explore the interactions at the source of the information provided by PSS and understand how the fusion of these data could provide better insight into soil spatial variability. The objectives of this study were to examine the relationship between proximal sensing data and soil properties, evaluate the feasibility of using the fusion of PSS and topography to improve the mapping of soil chemical properties and assess the effect of higher and lower sampling densities on the calibration model performance.

5.2 Material and Methods

Dataset description

A total of 128 samples [108 from a $0.4 \text{ ha} \cdot \text{sample}^{-1}$ sampling design (Figure 5.1– solid black circles) and an additional 20 independent validation samples (Figure 5.1 – solid green diamonds)] were collected in 2022 from a 43-ha field in Central Alberta, Canada. Samples were removed from the original grid until a density of $3.5 \text{ ha} \cdot \text{sample}^{-1}$ was obtained (Figure 5.1 – hollow blue squares), creating an extra-low density sampling design. All samples were collected under the same conditions and sent to the same laboratory for analysis. Samples collected from the topsoil layer (0-0.15 m) had their analysis results for plant-available potassium (K) and phosphorous (P), pH, and soil organic matter (OM) used to evaluate the prediction potential of PSS and topography. Amonia acetate extraction, weak Bray, soil-water solution (1:1 ratio), and loss of ignition (at 360°C) were the analytical methods utilized to obtain, K, P, pH, and OM, respectively. The above-described dataset is a subset of the one described by Karp et al. (2023a, 2024).

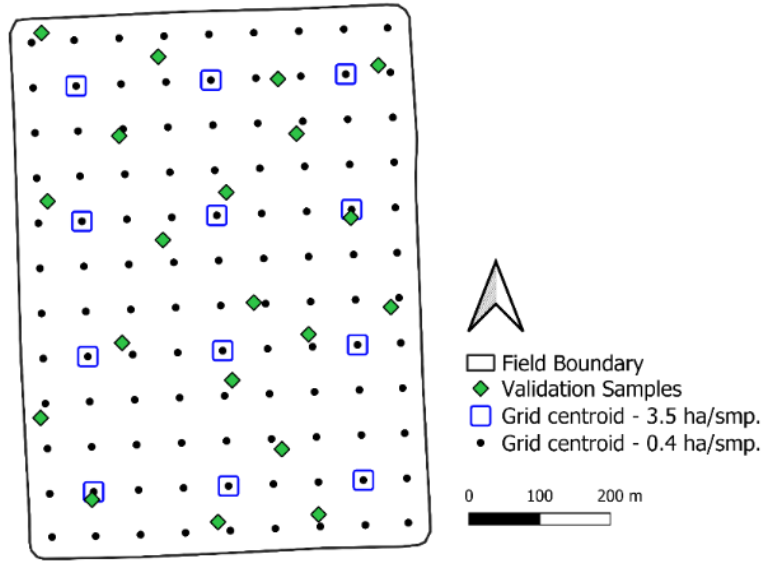


Figure 5.1 Centroid locations for the original sampling density ($0.4 \text{ ha} \cdot \text{sample}^{-1}$ – solid black circles), selected samples to create the extra low-density sampling design ($3.5 \text{ ha} \cdot \text{sample}^{-1}$ – hollow blue squares), and validation samples (solid green diamonds) (Modified from Karp et al., 2024)

PSS data for soil resistance to penetration, EC_a , dielectric permittivity, and passive γ -ray for ^{137}Cs , ^{232}Th , ^{238}U , and ^{40}K were collected using the tools described in Table 5.1. None of these sensors were used to directly estimate the lab data. The field's topography was assessed by creating its digital elevation model (DEM) using an unmanned aerial vehicle-mounted light detection and ranging (LiDAR) paired with a real-time kinematic correction (RTK) enabled global navigation satellite systems (GNSS) receiver. Hereafter, the five tools and variables described in Table 5.1 will be collectively referred to as "sources" and "predictor variables," respectively.

Due to time constraints, such as weather conditions and the limited time window between harvest and first snowfall or last snowfall and crop development, the data from the different sources was not collected during the same season, but as follows: γ -ray in Spring 2019, ground penetrating radar (GPR) in Summer 2020, penetrometer in Summer 2021, electromagnetic induction (EMI) in Fall 2021, and DEM in Spring 2022. From 2019-2022, the field was cultivated with annual crops under a rotation of wheat, barley, and canola. Within this timeframe, no significant soil disturbing operations (e.g., field leveling, subsoiling) were performed, and the field was seeded and fertilized using uniform rates.

Processing for sensing data

EM38-MK2 and SoilOptix (Table 5.1) data did not require specific pre-processing strategies before further procedures or analysis, while the other three data sources did. All data manipulation and filtering are described in the following sub-sections.

Penetrometer Pre-Processing

The original data from the S600 penetrometer provides pressure (kPa) measurements from 0 to 0.6 m deep at 0.01 m intervals. Therefore, a total of 61 measurements are obtained for every sampled location. To reduce the number of variables and improve the data quality (i.e., reduce the effect of outliers), a mean boxcar with a window of 0.1 m (10 vertical measurements) was applied to the data. This process resulted in a dataset of 6 depth intervals of 0.1 m.

LiDAR Pre-Processing

LiDARMill (Phoenix LiDAR Systems, Austin, Texas, USA) was used to process the LiDAR data. This software automatically performed necessary data corrections, generated 788 points·m⁻² georeferenced point cloud, and exported a 0.1-meter DEM raster. Finally, a custom R script (R Core Team, 2022) opened the DEM raster and calculated the fields' topography derivatives: slope, aspect, and curvature using the library *spatialEco* (Evans and Murphy, 2021).

Table 5.1 Description of the sensors, mapped variables, and collection settings adopted

Sensor Model	Manufacturer/ Provider	Technique	Swath width (m)	Density (points·ha ⁻¹)	Variables
S600 Penetrometer	Skok Agro (Vinnytsia, Vinnytsia Oblast-UA)	Cone Index	64	8	Resistance to penetration measured in pressure from 0 to 0.6 m (intervals of 0.01 m)
EM38-MK2	Geonic (Mississauga, Ontario-CA)	Electromagnetic Induction (EMI)	23	118	EC _a Shallow (0-0.75 m), EC _a Deep (0-1.5 m)
SoilOptix	SoilOptix (Tavistock, Ontario-CA)	Passive Gamma-Ray (γ-ray)	12	69	¹³⁷ Cs Count, ²³² Th Count, ²³⁸ U Count, ⁴⁰ K Count, Count Rate (CR)
SIR-4000 (400 MHz Antenna)	GSSI (Nashua, New Hampshire-USA)	Ground Penetrating Radar (GPR)	34	14,700	Soil Profile Amplitude through changes in dielectric permittivity
RECON-A	Phoenix LiDAR Systems (Austin, Texas-USA)	Light Detection and Ranging (LiDAR) + GNSS-RTK ^a	-	~7.8·10 ⁶	Elevation ^b

^aGNSS-RTK – Real Time Kinematic enabled Global Navigation Satellite System receiver; ^bElevation – a product of processing the GNSS-RTK location with the measured LiDAR distances

Ground Penetrating Radar Pre-Processing

The SIR-400 GPR unit provided separate files for GNSS and GPR readings. To open and process the data, a custom Python script and the library *readgssi* (Nesbitt et al., 2022) were used. Below is a simplified description of the GPR processing; a detailed description can be found in Karp et al. (2023b).

Due to a higher collection frequency for the GPR than for the GNSS, linear interpolation was applied to the original coordinates, guaranteeing the georeferencing of all the sensor readings. In sequence, the GPR signal was processed by automatically identifying (first-break method; Sensors & Software Inc, 2003) and setting time-zero, using a "dewow" filter, removing background noise, applying a Hilbert transformation (calculates the signal envelope – instantaneous amplitude), and converting the signal travel time to relative depth (field estimated dielectric constant of 12.79).

The GPR unit provided 512 vertical readings from the soil at every sampling location. Thus, the GPR instantaneous amplitude was subjected to a boxcar median with a 0.1 m window size. The maximum processing depth was set to 2 m, resulting in 20 depth layers. Since the density of the GPR data was very high within the collection transect (1 sample every 0.02 m – hence the high collection density in Table 5.1), a boxcar median was also applied in the direction of travel for the data collection. A 5 m distance between consecutive sampling points was achieved using a 250 samples window.

General PSS data filtering procedure

After completing the specific processing for the individual data sources, all PSS data were filtered to reduce the effect of outliers and maximize the data quality. The filtering procedure followed the steps suggested by Karp et al. (2022): (1) project the dataset to a custom localized Cartesian coordinate system, (2) identify and apply a position offset (e.g., the distance between the GNSS receiver and sensor), (3) operational filtering (i.e., removal of maneuvers, abrupt speed changes), (4) global and local statistical filtering.

Dataset co-location

To investigate the predicting capabilities of soil chemical properties using PSS and topography data, all the data must be co-located. Two approaches were adopted for the data co-location: one focused on building a dataset for training the predictive algorithms, and the other on predicting the spatial distribution of the soil property.

During the soil sampling activity, a GNSS receiver was used to record the location where the core samples were collected. At each core location, three subsamples were taken within a 5-meter radius. This intrinsic characteristic of the sampling method defined the first co-location method. A 5-meter buffer was applied to the recorded core locations, and the median of PSS or topography observations within the buffered area was calculated and attributed to the corresponding sampling location. The final dataset comprised 41 columns: 6 depth intervals of soil resistance to penetration from the penetrometer, 20 depth intervals of instantaneous amplitude from GPR, two depth ranges of EC_a from EMI, five count rates (total count rate plus the four separate nuclides) from γ -ray, four from topography (elevation, curvature, aspect, and slope), and the soil analysis for P, K, pH, and OM.

The above-described approach potentially minimizes issues with the change of support and co-location inaccuracies; however, it does not provide a continuous surface. Thus, in the second approach, inverse distance weighting (IDW) was used to interpolate the PSS data to a 15-meter resolution raster. For the topography data, the 0.1-m raster was downsampled to the same 15-meter raster using the median resampling method from *gdalwarp* (GDAL/OGR contributors, 2024). This approach resulted in a 37-band raster containing only the predictor variables.

Data preliminary analysis and predictive modeling

The descriptive statistics for PSS, the two soil sampling densities, and validation samples were calculated in a preliminary data analysis. The original 0.4 ha·sample⁻¹ co-located dataset was used to study the relationships between predictors and the soil properties by correlation analysis.

Thereafter, all the data were standardized to a zero mean and a unit variance for homogeneity purposes. Partial least squares (PLSR; Wold et al., 1983), support vector machine (SVM; Platt, 2000), and random forest (RF; Breiman, 2001) algorithms were evaluated with training from the co-located data of the 0.4 ha·sample⁻¹ and 3.5 ha·sample⁻¹ sampling designs. The three algorithms were implemented with a customized Python script using the library *scikit-learn* (Pedregosa et al., 2012). The Python library *Optuna* was used to tune the model parameters individually for a given soil variable, ML, and sampling density.

Most ML models benefit from larger datasets, and when exposed to a small number of observations and many predictors, they can overfit the training dataset (i.e., reduce the capability of generalizing the model). While a sampling density of 0.4 ha·sample⁻¹ was available for this study site, coarser sampling designs are more common among PA practitioners. According to

Erickson and Lowenberg-DeBoer (2023), two common barriers to the adoption of PA are related to the farm income and the PA service costs. Even though the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ sampling density provided only 12 samples, which might affect model performance, it broadened the discussion and produced realistic and practical results. From an economic perspective, using the original sampling density could defeat the option of collecting PSS and elevation data.

None of the three ML algorithms mentioned above includes a spatial structure for the data, which is likely to be an issue when using the trained models to predict a continuous surface. Therefore, the application of an inverse distance weighted (IDW) moving window was evaluated.

A window size and a matrix of weights are required to define the moving window. The window size limits the neighborhood of cells used to estimate the value at the center of the window. An estimate of the range of spatial autocorrelation was used as a basis to define the window size, assuming a circular shape and isotropy. Low-density sampling designs often will not provide enough information to obtain variogram estimates with traditional methods, so the approach proposed by Karp et al. (2024) was adopted. When a flat, pure-nugget effect variogram model was calculated, the minimum distance between sampling locations determined the window size. The formula $1/(\text{distance to window center})^2$ determined the weights. Map cells with a distance from the center of the window, greater than half the range of spatial autocorrelation, were removed, yielding a circular neighborhood. Finally, the weights were normalized so that the sum of their values was each time equal to 1.

Predictive modeling comparison

Mean Squared Error (MSE) and the coefficient of determination (R^2) were used to assess the predictive results' performance. For each soil property, three ML algorithms, two sampling densities, and two predicted surface treatments ("No Spatial Smoothing" and "IDW Moving Window") were evaluated, resulting in 12 different predictive results per soil property. These results were assessed by extracting the predicted results from the locations where the 20 independent validation samples were collected (a spatial operation) and then comparing the predictions to the laboratory analysis for a given soil property. The squared errors from the results were compared through a pairwise Levene's test; when the null hypothesis was rejected (homogeneity of variances), meaning there was heterogeneity in the variance of the squared errors, the two evaluated prediction results were considered different.

The effect of ML algorithms on prediction accuracy was evaluated independently for each sampling design and given soil property. A multi-objective decision-making logic was adopted to guarantee the selection of the most robust ML algorithm:

1. Models that presented statistically significantly lower errors were selected. If no statistically significant difference was observed, all models were selected.
2. If only one model was selected in Step 1, go to Step 5. Otherwise, MSE was standardized to a scale between 0 and 1.
3. A score was calculated using the formula $R^2 + (1 - \text{standardized MSE})$ and ranked in a descending order.
4. The model with the highest score was selected as the best predictor. The selected model was considered the most robust ML algorithm for the given dataset.

Using R^2 and MSE simultaneously avoids selecting models with low MSE and low R^2 and high R^2 and high MSE while selecting accurate models that best explain the variance of the soil property in the validation samples.

The most robust models for the two sampling densities were then compared, analyzing the effect of sampling density on the prediction of soil chemical properties. The IDW Moving Window effect was evaluated in sequence for a given density. Finally, thematic maps were generated for predicted surfaces and compared to surfaces obtained through ordinary kriging interpolation of the 0.4 ha·sample⁻¹ sampling design. All data, interpolation, and statistical analysis were performed using custom scripts written in the R language, and all standardized data and predictions were back transformed to report the results.

5.3 Results and Discussion

Descriptive Statistics

The descriptive statistics for the validation samples and the original and extra-low soil sampling designs are presented in Table 5.2. Similar standard deviations (SD), means, and medians are observed for the original (0.4 ha·sample⁻¹) and extra-low (3.5 ha·sample⁻¹) designs for a given soil property, indicating that there is a good overall representation of the underlying surface even with only 12 samples. The inline histograms from these two sampling densities differ, though, an expected response due to the reduction of almost 90% of samples. Note that pH and OM presented a smaller variance than P and K for both sampling designs.

















The means and medians for the validation samples are lower for K, P, and OM than for both sampling designs, whereas they are slightly higher for soil pH. The SDs of the validation samples are lower than that of the grid sampling for K and slightly higher for P, pH, and OM. Except for pH, the inline histograms for the validation samples are similar to the 0.4 ha·sample⁻¹ design. Despite these slight differences, the validation samples capture a similar variability as the grid samples, an important characteristic to guarantee the validity of further analysis using this dataset.

Similarly, Table 5.3 presents the descriptive statistics for one of the variables for each PSS before and after the general filtering procedure. The filtering procedure consistently reduced the SDs and differences between the means and medians of each data source. The comparison of the inline histograms only indicated minor changes in the data distribution, suggesting a successful removal of outliers while maintaining the integrity of the data distribution.

A lower SD and mean ratio indicate a lower elevation and EC_a variance than other variables. The evaluation of maps from these two data sources (not included in this paper) still showed that these variables represented this field's known spatial variability. Such observation highlights the importance of standardizing the dataset to zero mean and unit variance, as ML algorithms can be sensitive to the magnitude and variance of the data, which could result in reduced importance of the two variables mentioned above.

Representative training and validation datasets are essential to guarantee the model's validity and analysis of the results. The descriptive statistics for the same variables from the different data sources are presented for the surface resulting from the IDW interpolation of the data sources and for the 3.5 ha·sample⁻¹ training dataset. Despite some slight differences, the descriptive statistics for the training dataset and interpolated surfaces (Table 5.4) are similar to the ones from the "Filtered data" from Table 3 ("raw data" for elevation). The similarity among the descriptive statistics from these datasets indicates that no substantial changes were induced through the co-location and interpolation procedures, and reliable datasets were generated to achieve the objectives of this study.

Table 5.2 Descriptive statistics for soil testing results for plant-available potassium (K) and phosphorous (P), pH, and soil organic matter (OM) from two sampling densities and validation samples. (Modified from Karp et al., 2024)

Density	Grid Samples						Validation Samples				
	n ^a	Variable	Mean	Standard Deviation	Median	Histogram	n ^a	Mean	Standard Deviation	Median	Histogram
0.4 ha·sample ⁻¹	108	K (ppm)	133.9	37.6	130.5		20	128.0	31.5	122.5	
		P (ppm)	34.0	14.8	33.5			29.3	18.4	24.0	
		pH (-)	7.37	0.49	7.30			7.58	0.79	7.75	
		OM (%)	7.24	0.69	7.20			6.72	0.70	6.75	
3.5 ha·sample ⁻¹	12	K (ppm)	127.8	40.7	136.5		20	128.0	31.5	122.5	
		P (ppm)	31.3	15.4	33.0			29.3	18.4	24.0	
		pH (-)	7.39	0.51	7.20			7.58	0.79	7.75	
		OM (%)	7.3	0.64	7.55			6.72	0.70	6.75	

^a n: number of samples

Table 5.3 Example of descriptive statistics for the elevation raster data and one variable from each PSS before (raw) and after applying the filtering procedure steps from Karp et al. (2022) (a hyphen indicates that the filtering procedure was not applied to that dataset)




















Variable	Raw data				Filtered data			
	Mean	Standard Deviation	Median	Histogram	Mean	Standard Deviation	Median	Histogram
Resistance to penetration (kPa) from 0.01-0.10 m	219.2	215.9	172.0		174.6	63.6	168.3	
Electromagnetic Induction EC _a (mS·m ⁻¹) from 0-0.75 m	248.4	4.7	247.2		248.0	3.3	247.2	
Ground Penetrating Radar Instantaneous amplitude (-) 0.00-0.10 m	250531.7	148691.5	214862.9		223260.4	60323.4	209760.1	
γ -ray ⁴⁰ K (count rate)	294.2	148.5	283.5		286.5	35.6	285.2	
Elevation (m)	1022.6	4.2	1022.6		-	-	-	-

Table 5.4 Example of descriptive statistics from the 3.5 ha·sample⁻¹ co-located training dataset and interpolated surface (15-meter raster) for elevation and one variable from each PSS

Variable	Training Dataset (3.5 ha·sample ⁻¹)				Interpolated Surface			
	Mean	Standard Deviation	Median	Histogram	Mean	Standard Deviation	Median	Histogram
Resistance to penetration (kPa) from 0.01-0.10 m	171.4	56.6	167.5		186.8	59.5	187.3	
Electromagnetic Induction EC _a (mS·m ⁻¹) from 0-0.75 m	248.0	3.4	247.0		247.4	2.8	246.3	
Ground Penetrating Radar Instantaneous amplitude (-) 0.00-0.10 m	202015.3	63791.7	190354.6		238203.8	45946.2	226345.0	
γ-ray ⁴⁰ K (count rate)	278.5	33.4	283.1		289.0	22.8	290.5	
Elevation (m)	1023.1	3.5	1022.5		1022.4	4.5	1023.2	

Correlation analysis

The collinearity in the predictor variables and the relationships between predictors and the soil chemical properties were assessed on the dataset with $0.4 \text{ ha} \cdot \text{sample}^{-1}$ density. A visual inspection of the histograms in Table 5.3 allows to see that even after filtering, the distribution for some of the predictors did not look like a normal distribution. Thus, Spearman's correlation coefficient was used instead of Pearson's. The resulting correlations are presented in Figures 5.2 and 5.3.

The correlations among predictors (Figure 5.2) could lead to a lengthy discussion from a geophysical and engineering-focused perspective (Karp et al., 2023b). Thus, in the present study, the content of Figure 5.2 is discussed and interpreted from a modeling perspective.

Signal responses from different soil layers belonging to the same PSS are often correlated, for which subsequent intervals provide higher correlation values (e.g., Penetrometer 0.11 – 0.20 and 0.21 – 0.30 m). Such behavior is ‘as expected’ since spatial correlation is not limited to its most explored dimension (2D). Across data sources, all predictors significantly correlate to one or more predictor variables.

The observed correlation within and across data sources can be seen as a multicollinearity issue, which might result in unstable estimation of coefficients and variance inflation, consequently affecting the models’ predicting capabilities (Allen, 1997). This known limitation in data fusion is commonly handled using feature selection approaches (Ji et al., 2019; Lachgar et al., 2024) to remove variables that do not contribute to or negatively impact the models’ performance. The present study avoided the removal of predictor variables and entirely relied on the potential of some of the chosen ML algorithms to overcome this limitation. For example, PLSR reduces the dataset dimension and correlations among predictors through latent variables that maximize the explanation of the target variable. RF uses random sampling of variables and observations to reduce the overfitting of the model, which can reduce the effect of multicollinearity. Without a process that can minimize collinearity among predictors, SVM is the most vulnerable to this effect.

The correlations between the predictors and soil properties are presented separately in Figure 5.3 to facilitate the visualization and analysis. Except for ^{238}U and slope, all other variables demonstrated a significant correlation with at least one soil property. The 0.21-0.3 m resistance to penetration and γ -ray count rates for ^{40}K and total count rate (CR) presented significant correlation with all soil properties. pH and P correlate to the greatest number of predictors, 28 and 10, respectively. These two soil properties also present absolute correlation values above 0.5. The

highest absolute correlation for pH, P, K, and OM are with "Penetrometer 0.31-0.40" (-0.59), "Penetrometer 0.21-0.30" (0.55), "Elevation" (0.44), and " γ -ray ^{40}K " (-0.32), respectively.

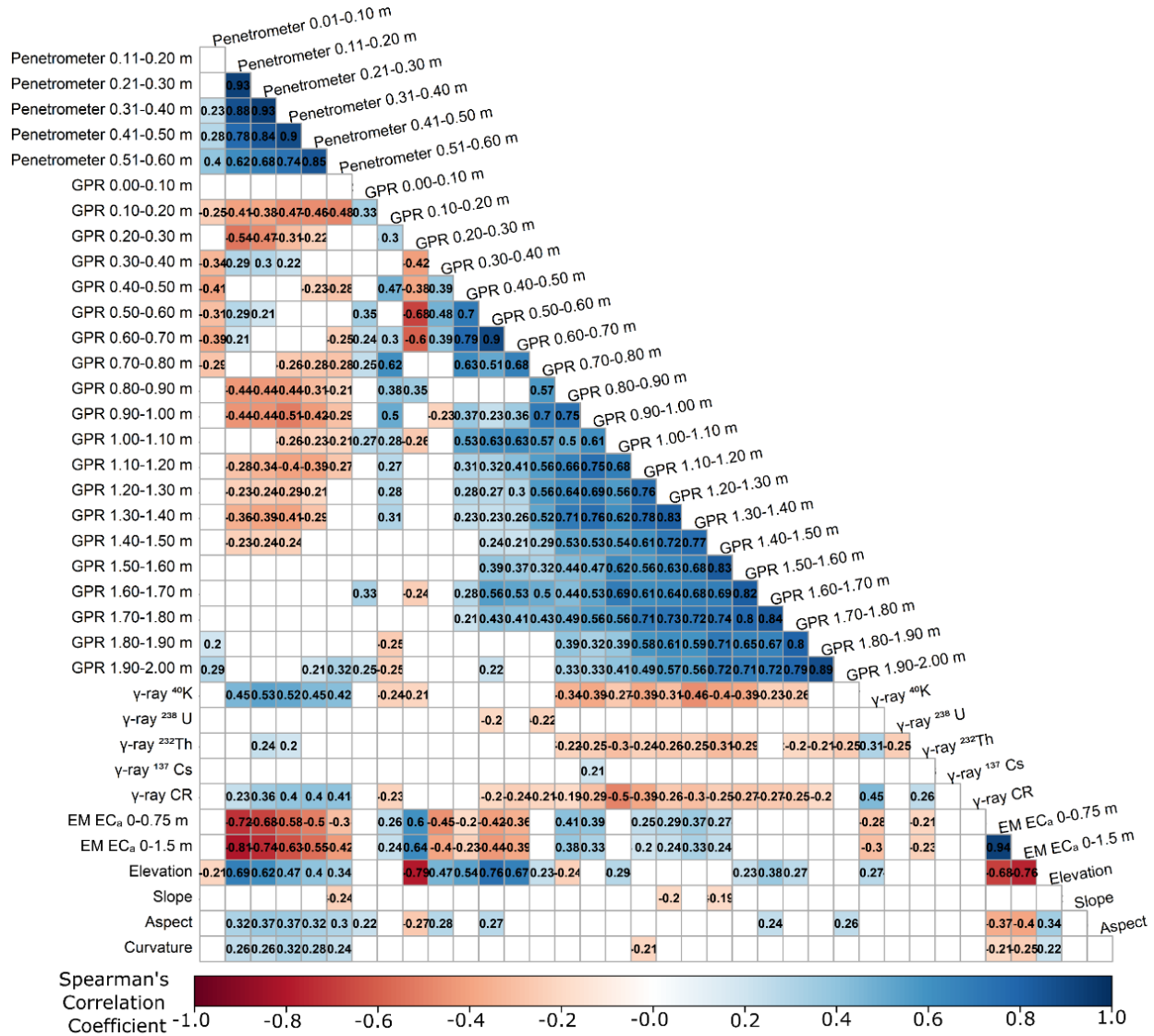


Figure 5.2 Spearman's correlogram for all sensors and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; EC_a- apparent electrical conductivity; EM – Electromagnetic Induction sensor; CR – count rate.

Complete agreement between a predictor variable and soil properties could not be identified (Figure 5.3). Also, each soil property relates to the data sources differently. For instance, pH significantly correlates to multiple variables from GPR, while OM to none. In contrast, all soil

properties correlated to at least three variables from the penetrometer. However, the correlation is stronger (absolute values above 0.5 for some depth intervals) for pH and P and weaker for OM and K (highest absolute value is 0.33). These observations lead to the conclusion that no single sensor can measure or predict all soil properties, a well-known behavior when mapping soil variability with PSS (Adamchuk et al., 2011; Gebbers, 2018).

Upon further inspection of the interpolated surfaces, all predictors clearly defined different aspects of the known variability of the study field, except for ^{238}U and ^{137}Cs , whose maps mainly presented a poorly structured spatial distribution. Since ^{238}U did not show a significant correlation with any soil property, and while analyzing the γ -ray dataset from this field, Karp et al. (2023b) observed pure nugget effect experimental variograms for ^{137}Cs and ^{238}U ; these two nuclides were not used for the model training and predictions.

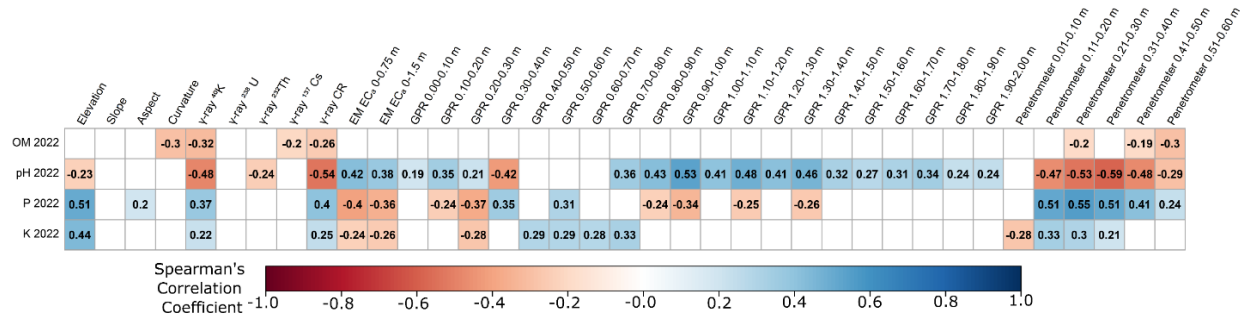


Figure 5.3 Spearman's correlogram matrix for soil properties, all sensors, and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; EC_a- apparent electrical conductivity; EM – Electromagnetic Induction sensor; CR – count rate.

Calibration Results

The model parameters for a given ML algorithm, sampling density, and soil chemical property were tuned, and the best parameters were used to train the calibration models successfully.

Effect and performance of ML algorithms for a given sampling density and soil property

The MSE from the ML algorithms were compared using Levene's test. No statistical significance ($\alpha = 0.05$) was identified for a given soil property and sampling design, meaning homogeneity of variances in the squared errors from the three ML algorithms. A comparison of MSE and R^2 from Figure 5.4 suggests that the multi-objective decision-making approach described

above demonstrated to be an effective approach to select the most robust ML algorithms (red-bordered bars in Figure 5.4). For example, when using the 3.5 ha·sample⁻¹ training dataset for predicting K (Figure 5.4a), RF and SVM resulted in very similar MSE, while RF was selected due to its higher R².

SVM and RF emerged as the most robust algorithms for 3 out of 4 soil properties for the original and extra-low density sampling designs, respectively (Figures 5.4a, c-d). For P, PLSR outperformed the other models for both sampling densities (Figure 5.4b). Ji et al. (2019) compared the performance of ML algorithms (including PLSR, RF, and SVM) to predict soil properties using the fusion of soil γ -ray, reflectance from visible and near-infrared spectra, EC_a from EMI, and elevation. The results presented by Ji et al. (2019) indicate that PLSR was often outperformed by SVM and RF, which aligns with the results observed in Figure 5.4.

The SVM never emerged as the most robust algorithm for the extra-low density sampling design; as previously mentioned, the higher susceptibility of this algorithm to collinearity might be contributing to this result. Thus, future research should evaluate how feature selection could change the observed results.

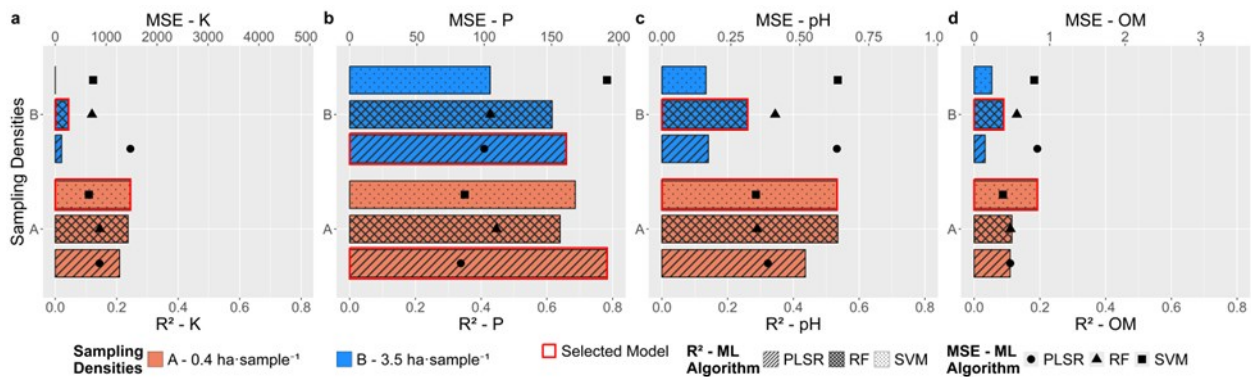


Figure 5.4 Bar-dot plot for the coefficient of determination (R²; bars) and mean squared error (MSE; points) for the 20 validation samples comparing partial least squares (PLSR; hatched bars and circle-shaped points), random forest (RF; crosshatched bars and triangle-shaped points), and support vector machines (SVM; bars with small dots and square-shaped points) as prediction algorithms for plant available potassium (K), phosphorus (P), pH, and soil organic matter (OM) for a given sampling density. A red border around a bar indicates the selected model for that sampling density.

Effect and performance of sampling density on the prediction of the soil properties

Figures 5.5a-d present a focused analysis of the effect of sampling density on predicting soil chemical properties. The R^2 is reported in the boxes on the left side of the panel, while MSE is on the right. Even though Levene's test results in Figures 5.5a-d (uppercase letters following the MSE values) did not reveal significant differences in the variance of squared residuals between the two sampling densities, the performance metrics indicated that calibration models using the original sampling density consistently outperformed the extra-low-density.

According to the results presented in Figure 5.5a-d, only the predictions for P yielded R^2 above 0.5 for both sampling densities (Figure 5.5b). For pH (Figure 5.5c), the R^2 for the 0.4 ha·sample⁻¹ model was 0.53, while 0.26 for the lower density design. These results for pH already suggest a gain in the percentage of the variability that the sensor's fusion can explain when including more samples. This result is 'as expected' since adding more samples improves parameter tuning and model predictions. Also, with the addition of closer samples (higher sampling density), spatial autocorrelation among sampling locations becomes more representative in the training dataset, which is not accounted for in the evaluated ML algorithms but affects the model predictions – the model overfits to the dataset (Hengl et al., 2018). The models for the original and lower density sampling designs accounted for less than 25% and 10%, respectively, of the variability in the validation samples for K (Figure 5.5a) and OM (Figure 5.5d).

The lower R^2 observed for OM (Figure 5.5d) could be attributed to the lower variance of this soil property in this field (Table 5.2). While pH also presented a lower variance (Table 5.2), a stronger relationship between the penetrometer variables and pH was observed (Figure 5.3), which might have contributed to the higher R^2 for this soil property. In contrast to OM and pH, K presented a higher variance (Table 5.2). Thus, the resulting R^2 for K suggests that calibration models trained with the fusion of the five different data sources could not explain more than 24% and 4% of the variability of this soil property in the validation samples when using the 0.4 and 3.5 ha·sample⁻¹ sampling densities, respectively.

The above observations regarding the modes' performance can also be observed in the predicted surfaces. Figure 5.6 compares thematic maps from the data fusion calibration models with the ordinary kriging (OK) interpolation of the 0.4 ha·sample⁻¹ sampling density. A visual comparison of the surfaces for P and pH revealed that maps originated from ML models for both sampling densities (P – Figures 5.6g and i; pH – Figures 5.6l and n) agree strongly with their

respective OK maps (P – Figure 5.6f; pH – Figure 5.6k). In contrast, such an agreement is weaker for K and OM, with the original sampling design models predicting surfaces closer to OK rather than for the extra-low sampling design models. Overall, the observed agreements between the ML predicted surfaces with OK interpolation suggests that the co-location procedures for the training and prediction dataset were effective.

Considering that the fused dataset contained 37 variables, some might negatively affect the model's predictability, as could the different collection dates for the data sources. Again, this indicates that future work should explore feature selection approaches to reduce the number or better select predictors or data sources.

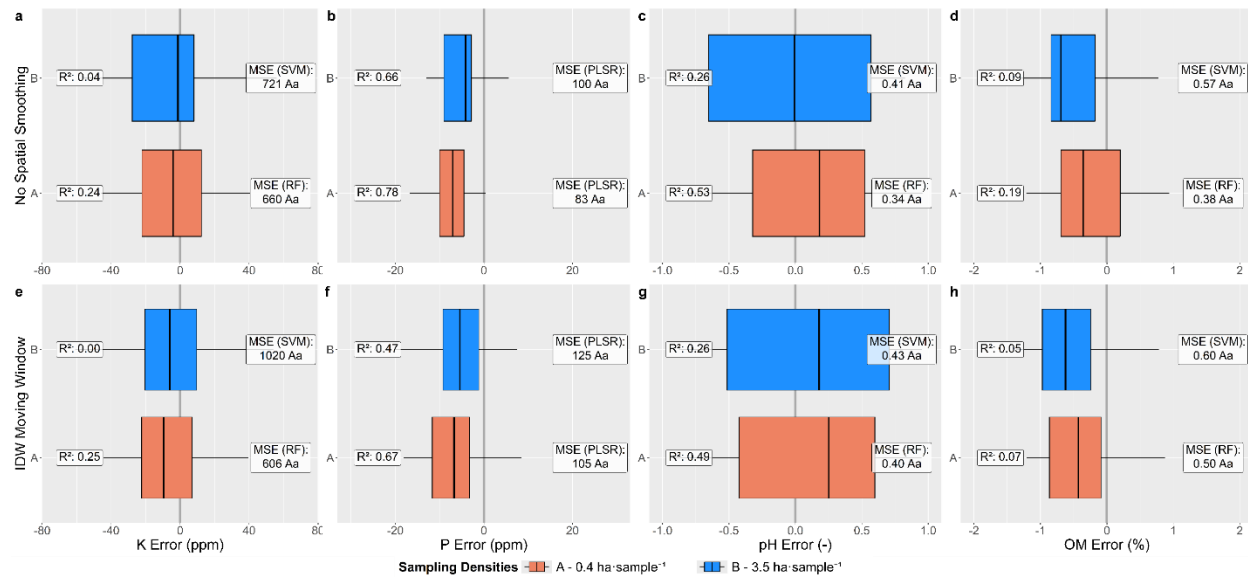


Figure 5.5 Box plots for the prediction errors for plant-available potassium (K), phosphorus (P), pH, and soil organic matter (OM). Results from the predictions based on the fusion of all data sources for two sampling densities, 0.4 and 3.5 ha·sample⁻¹, are indicated by a capital letter and a color. While partial least squares (PLSR), random forest (RF), and support vector machines (SVM) were tested for each scenario, only the best-performing algorithms are presented (refer to Figure 5.4 for all algorithms). MSE followed by different uppercase letters differ significantly at $\alpha = 0.05$ within the panel (between sampling densities), and different lowercase letters for the same sampling density but between "No Spatial Smoothing" and "IDW Moving Window"

Effect of an IDW smoothing approach in the surface predicted by a non-spatial ML algorithm

Although research results have supported that adding coordinates, sampling distances, and neighboring observations as covariates can improve the prediction capability of the ML algorithms

(Hengl et al., 2018; Pereira et al., 2022; Sekulić et al., 2020; Talebi et al., 2022), such approaches were not explored in the current study, as a focus was given to calibrating the fused dataset.

None of the evaluated ML algorithms accounted for the spatial component in the data; the predicted surfaces can present spatial outliers. This behavior can be observed in the maps for “3.5 ha·sample⁻¹” and “0.4 ha·sample⁻¹” in Figure 5.6. From a practical perspective, such spatial inconsistency in the maps might affect prescription maps. Therefore, an IDW smoothing approach was evaluated, and the results are presented in Figures 5.5e-h. Since none of the MSE values were followed by a different lowercase letter, Levene’s test did not indicate a significant difference ($\alpha = 0.05$) in the MSE after applying the IDW Smoothing Window for a given sampling density. However, the smoothing approach often worsens the performance metrics (Figures 5.5e-h) compared to the standalone model predictions (Figures 5.5a-d). A visual comparison of the smoothed maps and standalone predictions (Figure 5.6) suggests that the proposed smoothing approach reduced spatial inconsistencies.

For a reduction of 90% in the number of samples, using ML algorithms and the fusion of PSS and topography data presented some potential to predict the spatial variability of P and pH when using extra-low sampling density. From the perspective of a PA practitioner, such results might be more appealing than those obtained for the 0.4 ha·sample⁻¹ design. It is important to note that these results are applicable for the specific experimental site. The effect of sampling density on the model performance will vary for different fields and soil properties, as it depends on the spatial structure of the specific site and dataset. Thus, the results presented above should be generalized with care. Also, there are still limitations on the prediction of OM and K, which are also crucial for fertility management purposes. Therefore, further evaluations should be performed using different combinations of data sources and adding additional data sources that can help predict these two chemical properties.

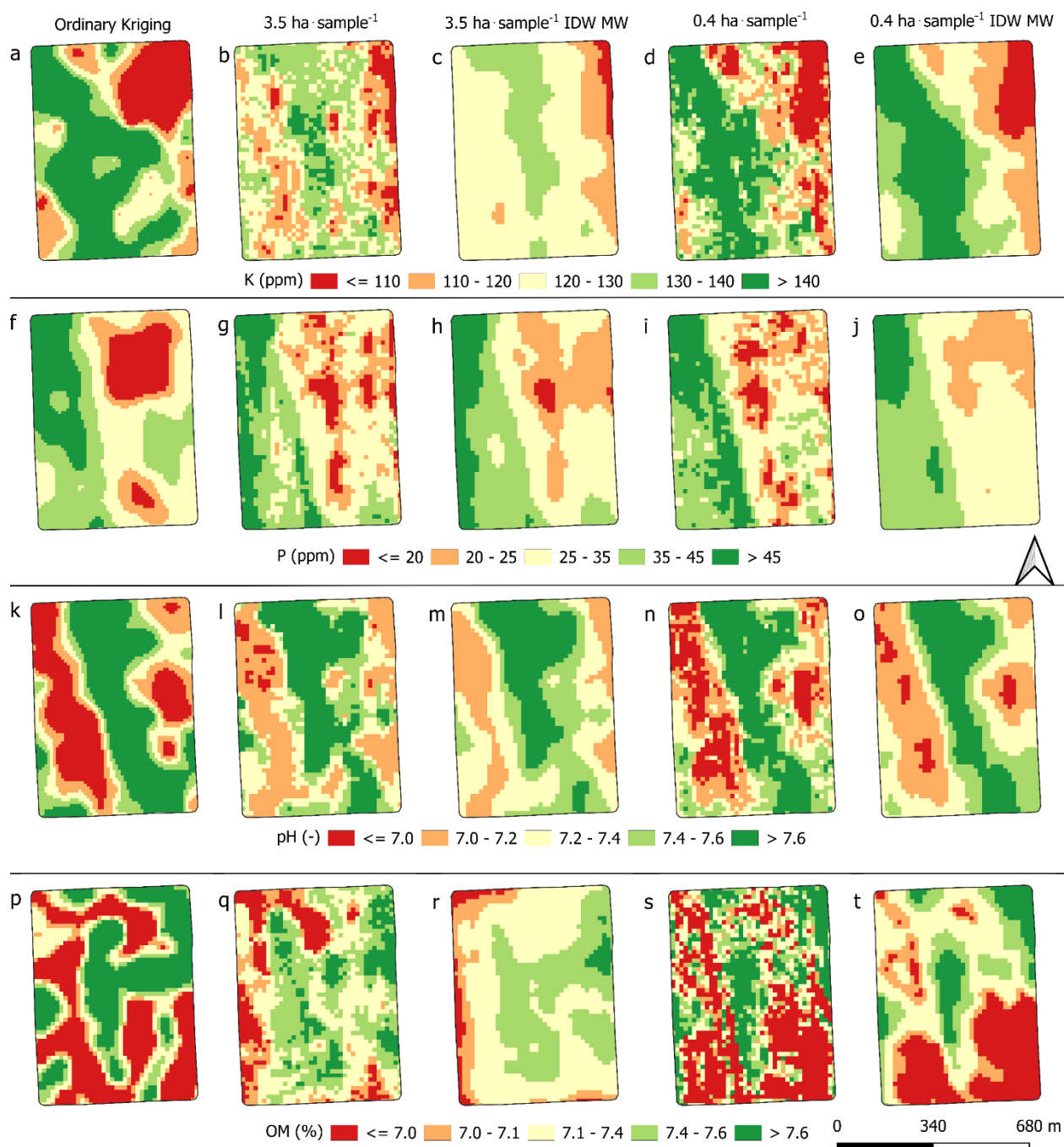


Figure 5.6 Thematic maps for plant-available potassium (K; a-e), phosphorus (P; f-j), pH (k-o), and soil organic matter (OM; p-t) from the interpolation of the 0.4 ha · sample⁻¹ using ordinary kriging (first column of maps), and calibration model predictions for the 3.5 ha · sample⁻¹ (second and third columns) and 0.4 ha · sample⁻¹ (fourth and fifth columns) before and after the “IDW Moving Window” (IDW MW) was applied

5.4 Conclusion

Although significant correlations between soil chemical properties and PSS and topography data were observed, no complete agreement was observed. The complex relationship between the properties measured by the sensors and the soil variables, as previously reported by other researchers, is a possible explanation for this behavior. These results indicate a potential benefit of fusing the data sources.

The machine learning algorithms evaluated did not present statistically significant differences when the squared residuals were compared using Levene's test. However, the use of a multi-objective decision making logic highlighted some differences between the models and presented an effective approach to select the best predictor for the two training datasets (0.4 and 3.5 ha·sample⁻¹).

No statistical difference was observed when comparing the residuals of the most robust ML algorithms from the two sampling densities evaluated. However, the performance metrics indicated that the higher-density dataset provided better predictions. The models for P and pH provided better results than those for OM and K, which models (especially when using the lower-density design) did not account for more than 25% of the variability in the validation dataset. Overall, a visual agreement between some of the predicted surfaces and OK interpolation (more evident for P and pH) was observed suggesting that the co-location procedures adopted for training and prediction were effective.

The ML algorithms evaluated did not consider the spatial component in the data, creating spatial outliers in the predicted surfaces. To overcome this limitation, an IDW-based moving window was evaluated, but while it reduced spatial inconsistencies, it slightly worsened the model performance metrics.

The results presented the potential of calibrating PSS and topography data fusion to predict soil chemical properties. However, this needs to be further explored, especially regarding the different combinations of the data sources.

5.5 Acknowledgments

We thank the Canadian Agri-Food Automation and Intelligence Network (CAAIN), Mitacs, and Telus Agriculture for funding the project and data collection. We also thank SoilOptix, Inc. (Tavistock, ON, Canada) for providing the complete gamma-ray data. This research is part of the

project 'Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services' funded by Mitacs through the Mitacs Accelerate program.

5.6 References

- Adamchuk, V. I., Viscarra Rossel, R. A., Sudduth, K. A., & Lammers, P. S. (2011). Sensor Fusion for Precision Agriculture. In *Sensor Fusion - Foundation and Applications*. InTech. <https://doi.org/10.5772/19983>
- Allen, M. P. (1997). The problem of multicollinearity. In *Understanding Regression Analysis* (pp. 176–180). Boston, MA: Springer US. https://doi.org/10.1007/978-0-585-25657-3_37
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Erickson, B., & Lowenberg-DeBoer, J. (2023). 2023 Precision Agriculture Dealership Survey. Department of Agronomy and Agricultural Economics, Purdue University. https://ag.purdue.edu/digitalag/_media/croplife-purdue-precision-dealer-report-2023.pdf
- Evans, J. S., & Murphy, M. A. (2021). spatialEco. <https://github.com/jeffrejevans/spatialEco>
- GDAL/OGR contributors. (2024). GDAL/OGR Geospatial Data Abstraction software Library. <https://doi.org/10.5281/zenodo.5884351>
- Gebbers, R. (2018). Proximal soil surveying and monitoring techniques (pp. 29–78). <https://doi.org/10.19103/AS.2017.0032.01>
- Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 2018(8). <https://doi.org/10.7717/peerj.5518>
- Ji, W., Adamchuk, V. I., Chen, S., Mat Su, A. S., Ismail, A., Gan, Q., et al. (2019). Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. *Geoderma*, 341(July 2017), 111–128. <https://doi.org/10.1016/j.geoderma.2019.01.006>
- Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2023a). Comparative study of interpolation methods for low-density sampling. In *Precision agriculture '23* (Vol. 34, pp. 563–569). The Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-947-3_71

- Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2024). Comparative study of interpolation methods for low-density sampling. *Precision Agriculture*. <https://doi.org/10.1007/s11119-024-10141-0>
- Karp, F. H. S., Adamchuk, V. I., Melnitchouck, A., Allred, B., Dutilleul, P., & Martinez, L. R. (2023b). Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing. *Journal of Environmental and Engineering Geophysics*, 28(1), 45–61. <https://doi.org/10.32389/JEEG22-018>
- Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2022). Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes. In *Proceedings of the 15th International Conference on Precision Agriculture* (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture. <https://www.ispag.org/proceedings/?action=abstract&id=8792&title=Optimization+of+Batch+Processing+of+High-density+Anisotropic+Distributed+Proximal+Soil+Sensing+Data+for+Precision+Agriculture+Purposes>
- Lachgar, A., Mulla, D. J., & Adamchuk, V. (2024). Implementation of Proximal and Remote Soil Sensing, Data Fusion and Machine Learning to Improve Phosphorus Spatial Prediction for Farms in Ontario, Canada. *Agronomy*, 14(4). <https://doi.org/10.3390/agronomy14040693>
- Nesbitt, I., Simon, F.-X., Hoffmann, F., Paulin, T., & teshaw. (2022). readgssi: an open-source tool to read and plot GSSI ground-penetrating radar data. <https://doi.org/10.5281/ZENODO.5932420>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <http://arxiv.org/abs/1201.0490>
- Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Santos, N. T., & Fernandes-Filho, E. I. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*, 23(4), 1189–1204. <https://doi.org/10.1007/s11119-022-09880-9>

- Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in Large-Margin Classifiers*. Cambridge: MIT Press.
- R Core Team. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria. <https://www.r-project.org/>
- Saifuzzaman, M., Adamchuk, V., Biswas, A., & Rabe, N. (2021). High-density proximal soil sensing data and topographic derivatives to characterise field variability. *Biosystems Engineering*, 211, 19–34. <https://doi.org/10.1016/j.biosystemseng.2021.08.018>
- Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, 12(10), 1–29. <https://doi.org/10.3390/rs12101687>
- Talebi, H., Peeters, L. J. M., Otto, A., & Tolosana-Delgado, R. (2022). A Truly Spatial Random Forests Algorithm for Geoscience Data Analysis and Modelling. *Mathematical Geosciences*, 54(1), 1–22. <https://doi.org/10.1007/s11004-021-09946-w>
- Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method (pp. 286–293). <https://doi.org/10.1007/BFb0062108>

Connecting Text to Chapter 6

Chapter 5 proposed a methodology and evaluated the potential of PSS and topography data fusion for predicting soil chemical characteristics through supervised learning. **Chapter 6**, on the other hand, focused on using an unsupervised spatial clustering algorithm to fuse PSS and topography data to propose improvements and validate the delineation of soil types within a field. **Chapter 6** is the last manuscript of this thesis, ensuring a linear progression of the topics. However, it was the first to be developed during the author's PhD program. The work presented in **Chapter 6** was essential to delineate some of the specific objectives for the thesis. For example, during the development of **Chapter 6**, the lack of a framework to batch-process PSS data was identified, leading to the development of **Chapter 3**.

The results in Chapter 6 might differ slightly from those in **Chapter 5**. When processing GPR data in **Chapter 6**, the time-zero was manually set, while it was automatically calculated in **Chapter 5**. Also, **Chapter 6** was prepared and published before a PSS batch-processing framework was developed; thus, the pre-processing of PSS in **Chapter 6** differs slightly from the framework defined in **Chapter 3**. A few other differences between **Chapters 5** and **6** must be highlighted: elevation from RTK-enabled GNSS in **Chapter 6**, while from LiDAR survey in **Chapter 5**; electromagnetic induction from Fall 2018 in **Chapter 6**, while Fall 2021 in **Chapter 5**; and the use of the ordinary kriged surfaces for the correlation analysis in **Chapter 6**, while the co-located dataset was used in **Chapter 5**.

The preliminary results from **Chapter 6** were first presented at the EEGS-WG-PSS-SEG Symposium: Application of Proximal And Remote Sensing Technologies For Soil Investigations held online, to which the author received the second best presentation award. An expanded version was also presented at the Canadian Society for Bioengineering Annual General Meeting and Technical Conference 2022 in Charlottetown, Prince Edward Island, Canada. Lastly, the content in **Chapter 6** was published in the *Journal of Environmental and Engineering Geophysics*.

Publication:

Karp, F. H. S., Adamchuk, V. I., Melnitchouck, A., Allred, B., Dutilleul, P., & Martinez, L. R. (2023). Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing. *Journal of Environmental and Engineering Geophysics*, 28(1), 45–61.
<https://doi.org/10.32389/JEEG22-018>

Abbreviations for Chapter 6

Abbreviation	Definition
¹³⁷ Cs	Caesium-137 isotope
²³² Th	Thorium-232 isotope
²³⁸ U	Uranium-238 isotope
⁴⁰ K	Potassium-40 isotope
AGRASID	Agricultural Region of Alberta Soil Inventory Database
ANOVA	one-way analysis of variance
ATL	Alberta soil code Antler
CR	Count Rate
DEM	Digital Elevation Model
EC _a	Apparent Electrical Conductivity
EM	Electromagnetic Induction
GC	Galvanic Contact
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPR	Ground Penetrating Radar
Hist.	Histogram
IQR	Inter Quartile Range
LiDAR	Light Detection and Ranging
ML	Machine Learning
PSS	Proximal Soil Sensing
RTK	Real-Time Kinematic
SFCM	Spatial Weighted Fuzzy c-means
WGS-84	Current version of the World Geodetic System

Chapter 6: Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing

Felippe H. S. Karp, Viacheslav I. Adamchuk, Alex Melnitchouck, Barry Allred, Pierre Dutilleul and Luis R. Martinez

Abstract

There is potential use of proximal soil sensors (PSS) to contribute to soil surveys and improve their results, and this study focused on the evaluation of this potential. An analysis using a high-resolution soil survey (1:5000), terrain data, and an ensemble of PSS (gamma ray emission, ground penetrating radar – GPR, apparent electrical conductivity from electromagnetic induction, and galvanic contact) was conducted. First, a geostatistical analysis was performed to characterize the spatial variability of each variable for each sensor and interpolate the data to a common support. The GPR data presented well-delineated groups of depths with similar spatial structure. These groups matched the field soil horizon depths, thus representing the potential for this sensor in soil characterization. A significant correlation was found between most of the variables from each sensor. However, no complete agreement was observed among the data from different PSS. In addition, a visual comparison of the maps showed that each PSS captured the soil spatial variability of the field and delineated regions distinctively. To validate the soil separation provided by the high-resolution soil survey and evaluate the capability of the PSS to distinguish the different soils, an analysis of variance was performed. Although none of the sensors could differentiate all the soils in the field, maps containing an overlay between sensors and soil models provided an important insight: overall, the soils were located correctly but the boundaries needed to be adjusted. Spatial clustering was used to perform a multivariate analysis of the data. A final map containing well-delimited homogenous PSS-based zones was obtained. Accordingly, it is possible to conclude that this approach and the resulting maps can be used to improve the delineation of boundaries between different soil types.

6.1 Introduction

The application of geophysics in agriculture has become more common with the increased adoption of precision agriculture (Gebbers et al., 2009). Using proximal and remote soil sensors in agriculture allows the evaluation of the spatial heterogeneity of fields that might be related to

soil and landscape properties. These results differ from those obtained from traditional soil surveys (low density/resolution). Using Proximal Soil Sensors (PSS), the field variability can be characterized precisely and accurately as a result of a higher collection density and lower costs. Allred et al. (2008) presented multiple case studies of geophysical techniques applied to agriculture, such as electromagnetic induction (EM), electrical resistivity, ground penetrating radar (GPR) and gamma ray (γ -ray) attenuation. The results of PSS surveys are mostly used to understand the variability within fields and delineate management zones (De Benedetto et al., 2013; Dhawale et al., 2014; Castrignanò et al., 2017; Castrignanò et al., 2018; Sanches et al., 2022).

Besides the use of PSS data, soil type boundaries obtained through field soil surveys provide valuable information when mapping the variability of an agricultural field. This data layer not only provides the distribution of soil types but also describes each specific soil which may be useful information for determining agricultural management practices. However, high resolution surveys are expensive due to the need for a large number of samples, borings and trenches. In contrast, most nationwide available soil surveys are low resolution and do not provide much insight on within-field variability. For example, Canada's detailed soil surveys are mostly between 1:50,000 and 1:100,000 (Agriculture and Agri-food Canada, 2021). When higher resolution surveys are available, most of the soil type boundaries are predicted, and thus, they are of doubtful accuracy. James et al. (2003) compared a soil classification with EM data and observed some inconsistencies, whereupon they decided to resample the field. As a result, the authors determined that the previous soil boundaries needed to be adjusted.

Since it is known that there is a relationship between different PSS data and soil surveys (Inman et al., 2002; Simeoni et al., 2009; André et al., 2012), these two techniques could be used together to reduce costs and increase the accuracy of soil surveys. This concept is not new, and others have pointed out the benefits of using this combination. Doolittle (1987) stated that the conventional methods of soil surveys were slow and tedious to perform, and they generated incomplete data. The same author also mentioned that using GPR during surveys could reduce costs by 70% by decreasing the number of samples and crew size while improving the quality of the maps. However, more research is needed to validate, compare, or propose improvements to the delineation of soil type boundaries using proximal soil sensors, especially focusing on the comparison and use of sensors' arrays and data fusion.

Therefore, this study aimed to evaluate the potential of using an ensemble of proximal soil sensors to validate and propose a potential strategy to improve soil type delineation obtained during soil surveys.

6.2 Materials and Methods

Site and Data Description

A 43-ha field located in Olds, Alberta, Canada (51°46'10.69"N, 114° 5'20.04"W; WGS84) was used for this study. According to the Canadian Climate Normals (Environment Canada, 2019), yearly precipitation at this location is 492.4 mm, while maximum and minimum temperatures averaged 9.6 °C and -2.5 °C. Regional soils originated from parent material characterized by glacial and glaciolacustrine veneer overlaying till (Bowser et al., 1951; Pawluk and Bayrock, 1969; Walker and Mcneil, 2004). The field is cultivated with annual crops in a rotation system of canola, wheat, and barley.

In 2003, a detailed soil survey (scale of 1:5000) took place on the farm where this field is located (Walker and Mcneil, 2004). During the soil survey, 20-25 soil inspection sites were located every 65 ha (160 ac – a quarter of a section based on the Alberta Township System). The survey classified the soils following the guidelines from The Canadian System of Soil Classification (Soil Classification Working Group, 1998) and the distribution of soils (soil mapping units) on the farm was described using soil models. Soil models are obtained by classifying the soil at a series level and are determined based on the most dominant soil series in that specific location. For each soil model, a symbol composed of three or four letters (the dominant soil series) plus a number (identifies soil patterns that repeat in the landscape) is defined. The survey classification at soil series and soil model definition was based on the Agricultural Region of Alberta Soil Inventory Database (AGRASID) (ASIC, 2001) model. Table 6.1 presents the soil models, soil subgroup classification, texture, and a summary of important observations about the soil models present in the study area. Figure 6.1 shows the spatial delineation of the data described in Table 6.1.

Table 6.1 Field AGRASID soil models classified during the soil survey performed in 2003. Information retrieved from Walker and Mcneil (2004)

Soil symbol	Soil model name	Dominant soil subgroup	Texture	Observation
ATL4	Antler 4	Orthic Black Chernozemic	Loam to Clay Loam	Less glaciolacustrine veneer than ATL4; Present at steeper slopes
ATLP4	Antler-Lonepine 14	Orthic Black Chernozemic	Loam	More glaciolacustrine veneer than ATL4; Present at mid and upper slopes
DRNI16	Deadrick-Niobe 16	Gleyed Black Solonetz	Clay Loam	Soil from Solonetzic Order where B horizon has significant amount of exchangeable sodium; Present at Concave slopes; Has greater Post-Glacial erosion from glaciolacustrine
MRNI4	Minaret-Niobe 4	Black Solod/Gleyed Black Chernozemic	Silt Loam	Soil affected by subsurface discharge
MRT4	Minaret 4	Gleyed Black/Gleyed Calcareous Black Chernozems	Silt Loam	Imperfectly drained
NTOL1	Netook-Olds 1	Calcareous Black Chernozems	Loam	Soil not affected by groundwater discharge
NTOL2	Netook-Olds 2	Calcareous Black Chernozems	Loam	Has significant imperfectly drained Gleyed Black Chernozemic soils - not present in NTOL 1

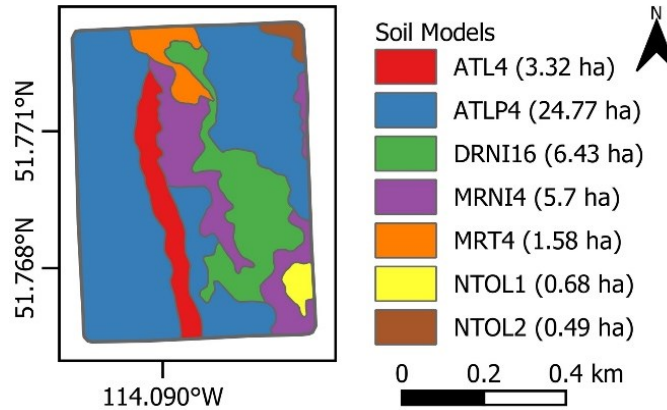


Figure 6.1 Spatial distribution of the soil models within the study field – refer to Table 6.1 for details on each soil model.

To evaluate the soil spatial variability and validate the delineation for the soil models generated during the 2003 soil survey, a collection of a series of high-resolution proximal soil sensor data was carried out. Table 6.2 presents the different sensors used and mapped variables, and Figures 6.2(a)-(d) shows the characteristics of the data collection (sampling locations, swath, distance between consecutive points, density, and direction of travel). Since the setup used for three of the sensors (Veris 3100, EM-38-MK2, and SIR-4000) required pulling or pushing the data acquisition unit, the data collection followed the same direction as the crop seeding operation (North-South). By following a different direction (e.g., East-West), the plants' stubbles would impose difficulties during data acquisition or add noise to the data. For the SoilOptix, as the sensor was attached to the front of an all-terrain vehicle, it was not subjected to the same limitation as the other sensors, so data was acquired in the East-West direction.

The data was not collected during the same season but as follows: Veris 3100 in Spring 2017, EM38 in Fall 2018, SoilOptix in Spring 2019, and Ground Penetrating Radar (GPR) in Summer 2020. All data were geo-referenced using global navigation satellite system (GNSS) receivers. SoilOptix data was collected using a Real-Time Kinematic (RTK) GNSS receiver, which provides a high-level accuracy (<4 cm). Thus, its elevation was used to obtain the digital elevation model (DEM) and terrain slope. GPR data was collected using a 400 MHz antenna with a trace increment of 0.02 m, a range of 60 ns and a total of 512 samples per scan.

All PSS except GPR already output a file containing the sensor readings and GNSS location. Therefore, they could be directly processed as a spatial dataset. For GPR, a specific processing strategy was adopted, and it is discussed in detail under the GPR Data Processing section.

Table 6.2 Description of proximal soil sensors, most significant measured physical property, and mapped variables. EC_a – apparent electrical conductivity, RTK – Real-Time Kinematic.

Sensor model	Manufacturer/provider	Technique	Measured Physical Property	Variables
Veris 3100	Veris Technologies (Salinas, Kansas-U.S.)	EC _a - Galvanic Contact (GC)	Electrical resistivity/conductivity	EC _a Shallow (0-0.3 m), EC _a Deep (0-0.9 m)
EM38-MK2	Geonic (Mississauga, Ontario-CA)	EC _a - Electromagnetic Induction (EM)	Electrical resistivity/conductivity	EC _a Shallow (0-0.75 m), EC _a Deep (0-1.5 m)
SoilOptix	SoilOptix (Tavistock, Ontario-CA)	Passive Gamma-Ray (γ -ray)	Gamma radiation	RTK Elevation, Count Rate, and Activity per mass for: ¹³⁷ Cs, ²³² Th, ²³⁸ U, ⁴⁰ K
SIR-4000 (400 MHz Antenna)	GSSI (Nashua, New Hampshire-U.S.)	Ground Penetrating Radar (GPR)	Dielectric permittivity	Soil Profile Amplitude

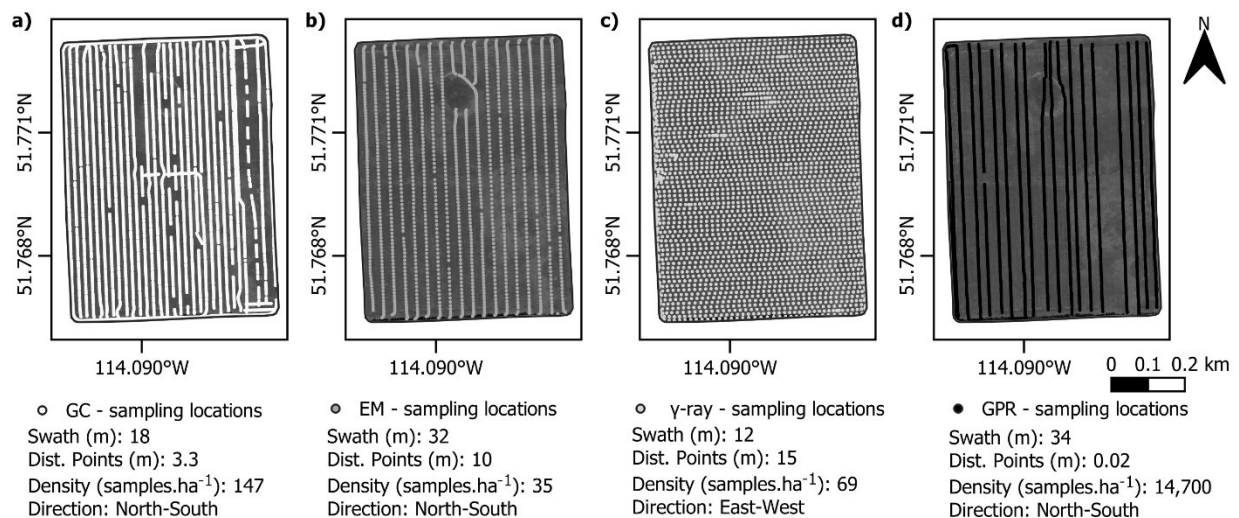


Figure 6.2 Proximal soil sensors data acquisition characterization for (a) Galvanic Contact (GC) – Veris 3100, (b) Electromagnetic Induction (EM) – EM38-MK2, (c) Passive Gamma-Ray (γ -ray) – SoilOptix, and (d) Ground Penetrating Radar (GPR) – SIR-4000. Dist. Points – distance between consecutive points.

GPR Data Processing

GSSI proprietary file formats (.DZT and .DZG containing the GPR and GNSS readings, respectively) were opened and processed using a custom script developed under the python language using the library readgssi (Nesbitt et al., 2022). The processing of each GPR transect started with (1) setting the time zero – a static procedure to remove the direct wave from the sensor, (2) the employment of a “dewow” filter to remove the low frequency signal that is diffused into the ground – the filter was developed following the definition described by Sensors & Software Inc. (2003), (3) removal of background noise by applying a horizontal filter with a window size equals to the length of the transect – this filter was applied carefully (analyzing the generated radargrams) in order not to remove the desired reflectors, interface between soil horizons, (4) execution of Hilbert transformation to calculate the signal envelope (instantaneous amplitude) which is proportional to the square root of the signal energy at a given time – an adimensional attribute that involves the analysis of the complex part of the signal (Claerbout, 1976), (5) conversion from signal travel time to relative depth by using an estimated dielectric constant (relative permittivity). The dielectric constant obtained during the GPR survey was 12.79, which is within the range of values (10-20) reported in the literature for wet loamy soils (Daniels, 1996). It is important to mention that this represents an estimate of the relative permittivity value of the soil, which was calculated in the field using the hyperbola fitting method; however, this might change for different soil depths and throughout the field. The procedure adopted in this paper uses the same value for the entire field, a common practice when processing GPR surveys; however, this might impact the accuracy of the calculated depths.

The frequency of GNSS data collected during the survey was lower than the GPR collection frequency. Therefore, linear interpolation was applied. To process the GPR as a spatial dataset, the recorded matrix was transposed (traces in columns were now placed in rows), and the latitude and longitude from the GNSS were associated with each trace. Since the GPR readings were very dense within the transect (one trace every 0.02 m – Table 2), horizontal stacking was performed in which a boxcar median was applied with a window of 250 traces. Thus, the final distance between points from the same transect became 5 m. This process reduced the number of points (reducing computational power for processes applied during the Data Ensemble section), worked as a smoothing filter and would potentially mitigate the effect of anisotropy during the data interpolation. Also, to reduce and smooth the vertical resolution of the data (depth), a boxcar

median with a 0.1 m size was applied to calculate the median of the instantaneous amplitude for every layer. The maximum depth for processing was set to 2 m. Thus, it was possible to obtain a final dataset with 20 depth layers from 0-2 m in 0.1 m intervals. The size of the depth intervals was determined based on the analysis of the radargrams and the publication from Castrignanò et al. (2018), which used a similar GPR data processing strategy and data arrangement.

Dataset Ensemble

As shown in Figure 2, different resolutions and positioning of transect were used for each sensor. Therefore, all the PSS readings must be brought into the same scale to compare the data and use it as a tool for validating and improving the delineation of the soil boundaries. The rescaling process is described below.

It is known that outliers are present during sensor data collection for environmental sciences applications, and their identification and removal is an important step in improving data quality (Spekken et al., 2013; Lyle et al., 2014; Maldaner et al., 2022). Thus, a global filter was applied to every sensor variable. The implemented filter included removing within-field maneuvers, abrupt variations in speed, and an Inter Quartile Range (IQR) filter. The latter was applied by setting lower and upper limits equivalent to 1.5 times the 25% and 75% quartiles. In addition, a local filter was applied to the data. Thus, a similar strategy as the one described by Maldaner et al. (2022) was used in which an anisotropic (removal outliers within its neighborhood but in the same transect) and an isotropic (removal of omnidirectional outliers) filter are applied. Both filters take as input a user-specified distance that determines the search neighborhood for the outliers. Then lower and upper limits are determined based on a percentage of the median. For the dataset used in this research, a distance equal to 1.5 times the data collection swath, and 15% of the median, were used.

After removing the global and local outliers, ordinary kriging was used to interpolate the data. Before interpolating, the variables were tested for normality, and when rejected, they were subjected to Box-Cox transformation (Box and Cox, 1964) to improve data normality. In addition, the data was standardized to zero mean and unit variance to account for the differences in magnitude on different PSS' variable readings. In sequence, each variable was interpolated using block kriging instead of point kriging to reduce the problem of change of support (Chilès and Delfiner, 2012; Castrignanò et al., 2018). After interpolating the RTK Elevation collected during

the SoilOptix sensor survey, the terrain slope was calculated and added to the dataset. A grid size of 15 m was established and used to interpolate all the variables.

Data were read, filtered and interpolated with custom scripts built under the language R and using the libraries *sf* (Pebesma, 2018), *raster* (Hijmans, 2021), *gstat* (Pebesma, 2004; Gräler et al., 2016), *dplyr* (Wickham et al., 2021) and *bestNormalize* (Peterson, 2021).

Validation of Soil Delineation

Using the interpolated dataset, a correlation analysis was performed to understand the interactions among the different sensors and variables. Afterwards, based on the overlap between the spatial location of the interpolated dataset and the 2003 soil survey polygons, each interpolated observation (grid cell) received a categorical variable representing the soil model it belonged to. Then, a one-way analysis of variance (ANOVA) took place by using the defined soil models as grouping factors. When significant differences were found, Tukey's honest significant difference test was performed as an ad-hoc test to investigate which means of the soil models would differ from the others. This step was performed by using the function *HSD.test* from the R package *agricolae* (de Mendiburu, 2021). It is important to notice that due to the different areas covered by each soil model, the number of observations was unbalanced; to account for this, *agricolae*'s function uses the Tukey-Kramer test. This analysis aimed to evaluate the different PSS variables' capability to differentiate the distinct soil models.

In sequence, a multivariate analysis was implemented to evaluate the potential of using proximal soil sensing data to improve the delineation of the soils' boundaries, in which PSS and terrain data variables were used as inputs to a spatial clustering tool. A custom script developed under the language R used the library *geocmeans* (Gelb and Apparicio, 2021) to perform a spatial weighted fuzzy c-means (SFCM) algorithm. The classical fuzzy c-mean would calculate the probability of an observation to belong to a group considering only the original data membership matrix (without considering the data spatial component). While the SFCM also considers a modified spatial version of the original matrix (Cai et al., 2007). This process avoids the creation of non-spatial consistent groups, which is important when trying to delineate the boundaries for homogeneous zones, such as for soil distribution in space.

The above-mentioned modification of the original fuzzy c-means is obtained by creating a spatial weighting matrix that considers the neighbors of each observation (Getis, 2009) and uses it to calculate a local average for the observations, composing the spatially modified version of the

original matrix (\underline{x}). In addition, a coefficient (α) is used to determine the weight of this spatially modified matrix in the process of updating the membership matrix and cluster centers. If α is 0, classical fuzzy c-means is applied. If $\alpha=1$, the original and the spatially weighted matrices have the same weight, while if $\alpha=2$, the spatially weighted matrix will have twice as much weight as the original matrix. This approach has been adopted in medical studies involving the clustering of brain images (Cai *et al.*, 2007; Zhao *et al.*, 2013) and has recently been applied to environmental sciences (Gelb and Apparicio, 2021). Equations 6.1 and 6.2 present the matrix and centers update functions, respectively.

$$u_{ik} = \frac{\left(\|x_k - v_i\|^2 + \alpha\|\underline{x}_k - v_i\|^2\right)^{\left(-\frac{1}{m-1}\right)}}{\sum_{j=1}^c \left(\|x_k - v_j\|^2 + \alpha\|\underline{x}_k - v_j\|^2\right)^{\left(-\frac{1}{m-1}\right)}} \quad (6.1)$$

$$v_i = \frac{\sum_{k=1}^N u_{ik}^m (x_k + \alpha \underline{x}_k)}{(1 + \alpha) \sum_{k=1}^N u_{ik}^m} \quad (6.2)$$











where U is the membership matrix in which u_{ik} represents the probability that the observation k belongs to cluster i , v_i is the center for cluster i (which is updated with matrix U), v_j is the center for cluster j , c is the total number of clusters (defined by the user beforehand), m is the fuzziness degree (the higher the m , the fuzzier the classification is), N is the number of observations, $\|x_k - v_i\|$ is the Euclidean distance between the observation k at the original matrix and the center of cluster v_i , $\|\underline{x}_k - v_i\|$ is the Euclidean distance between the spatially modified observation k and the center of cluster v_i .

6.3 Results and Discussion

Analysis of proximal soil sensors data

The GPR data processing described in the previous section generated a dataset with 5 m between consecutive points (approx. 64 samples·ha⁻¹) and median instantaneous amplitude (envelope of GPR signal) for every 0.1 m from the relative depths of 0 to 2 m. In sequence, all the sensors were subject to the outlier removal process. Examples of the data distribution before and after applying this procedure for each sensor are presented in Table 6.3. In general, the adopted procedure removed abnormal values improving the data distribution (mean closer to the median) without compromising the spatial patterns delineated by the data.

Table 6.3 Comparison of the data distribution before and after the global and local outlier removal process and percentage of data removed. GPR – Ground Penetrating Radar, Med. Inst. Ampl. – Median Instantaneous Amplitude, EM – Electromagnetic Induction, GC - Galvanic Contact, EC_a – apparent electrical conductivity, Hist. – Histogram.

Variable	Removed (%)	Before Outlier Removal			After Outlier Removal		
		Mean	Median	Hist.	Mean	Median	Hist.
GPR 0.0-0.1 m (Med. Inst. Ampl.)	29.1	63286.3	56300.2		54920.3	53439.4	
EM EC _a 0-0.75 m (mS·m ⁻¹)	17.4	32.7	30.1		31.1	29.8	
GC EC _a 0- 0.3 m (mS·m ⁻¹)	30.3	18.3	15.2		17.6	15.1	
γ-ray ¹³⁷ Cs (Count Rate)	26.3	7.0	6.5		6.5	6.3	
Elevation (m)	17.2	1037.9	1037.7		1037.8	1037.3	

Although the removal of outliers improved the data distribution, the in-line histograms displayed in Table 6.3 seemed to present a non-normal shape. That could be attributed to some natural anomalies in the field, such as high values of EC_a caused by a poorly drained area. Thus, the normality of distribution was assessed for each of the sensors and their variables. The null hypothesis (data arising from a normal distribution) was rejected at a significance level of 0.05 for all variables. Therefore, the data were subjected to a Box-Cox transformation as required, variable by variable. The results presented in the variogram analysis are for the transformed data, standardized to a zero mean and a unit variance. For the maps, a back-transformation was applied using the same parameter estimates (Box-Cox lambda, mean and variance). The best variogram model was selected for each variable based on the error sum of squares for a weighted least-squares fitting.

Ground Penetrating Radar (GPR). Initial values for the variogram parameters were determined for each GPR depth and then fitted to a model. Table 6.4 presents the selected model, nugget, sill, and range for each depth. From this table, it is possible to observe that the GPR readings seem to have four major groups of similar spatial structure. This same pattern can be observed in the spatial

variability (relative information between high and low values) presented by the maps of these variables (Figure 6.4). The first group from the relative depths of 0 to 0.4 m (Figures 6.4(a)-(d)), the second from 0.4 m to 0.7 m (Figs 6.4(e)-(g)), the third from 0.7 m to 1.3 m (Figures 6.4(h)-(m)), and fourth from 1.3 m to 2 m (Figures 6.4(n)-(o)). This pattern presented by the maps and geostatistical analysis agrees with most of the depths for the soil horizons described for the profiles during the soil survey.

An example of an Alberta soil code Antler (ATL) pedon is presented in Figure 6.3 (Walker and Mcneil, 2004). This soil code is a major component of 2 soil models identified during the survey (ATL4 and ATLP4 – Table 1) covering approximately 27 ha of the field. The pedon presented in Figure 6.3 does not show the entire depth of this soil profile which, when described in the soil survey report, reaches the bedrock at approximately 1 m. Thus, based on Figure 6.3 and the description of the soil model, the first group of relative depths from GPR would represent horizons A and B, while the second and third groups would represent two stages of development of the C horizon, and lastly, the fourth group would represent the bedrock formation.

However, it is known that with the travel of the GPR waves through the soil, an attenuation in the strength of the signal is observed that can affect the accuracy of the amplitude readings. Thus, in deeper layers, the differences in the spatial structure of the data observed through the variogram analysis and maps might be due to the attenuation of the GPR signal. Consequently, based on the maximum depth of 1 m for the soil described in the survey and this known effect of the signal attenuation of the GPR, the data used for further analysis in this research was limited to 1 m relative depth.

Also, even though the GPR relative depths for the groups of similar spatial structures align with the soil horizons presented by the pedon from Figure 6.3, these depths were not confirmed by taking a soil core. Moreover, the same dielectric constant was used to process the data for the whole field. While this is a common practice, this property changes with differences in soil types and conditions. Therefore, there could be some uncertainty regarding the depth calculation and layering definition. Thus, future work should consider taking a few common mid-point data to estimate the permittivity, evaluate its spatial variability in the area and use this information to process the data accordingly.

Besides the fact that a process to mitigate the anisotropy on the GPR data was performed (stacking of 250 traces), a few of the maps presented in Figure 6.4 show some evidence of a

directional variability which could still be an artifact of the higher density of points in the data acquisition direction. Since this anisotropy is caused by a characteristic of the data collection process and is not necessarily related to the spatial variability of the studied field, a future improvement to the data processing could be the increase in the number of traces used during the stacking process. A suggestion would be to set the distance between consecutive points to be the same as the swath distance. A similar approach could also be used in processing any sensor data containing a higher sampling density in the direction of travel (e.g. GC and EM data).

Still, the results obtained by analyzing the GPR data spatial structure and comparing it to the pedon presented in Figure 6.3 highlight the potential of this data for the delineation of different soil types and is in accordance with similar approaches adopted by other authors (Simeoni et al., 2009; André et al., 2012; Novakova et al., 2013; Zhang et al., 2014) who also found GPR to be useful for the analysis of soil types, profiles, and horizons' identification.

Table 6.4 Variogram model parameters for Ground Penetrating Radar (GPR) median instantaneous amplitude (Med. Inst. Ampl.) for every 0.1 m depth interval.

Variable	Model	Nugget	Sill	Range (m)
GPR 0.0-0.1 m (Med. Inst. Ampl.)	Exponential	0.24	0.89	32.7
GPR 0.1-0.2 m (Med. Inst. Ampl.)	Exponential	0.20	0.90	74.8
GPR 0.2-0.3 m (Med. Inst. Ampl.)	Exponential	0.21	0.82	61.5
GPR 0.3-0.4 m (Med. Inst. Ampl.)	Exponential	0.25	0.94	38.0
GPR 0.4-0.5 m (Med. Inst. Ampl.)	Exponential	0.23	1.35	257.7
GPR 0.5-0.6 m (Med. Inst. Ampl.)	Exponential	0.20	2.90	945.2
GPR 0.6-0.7 m (Med. Inst. Ampl.)	Exponential	0.19	1.38	269.7
GPR 0.7-0.8 m (Med. Inst. Ampl.)	Exponential	0.08	0.80	32.8
GPR 0.8-0.9 m (Med. Inst. Ampl.)	Exponential	0.00	0.91	15.6
GPR 0.9-1.0 m (Med. Inst. Ampl.)	Exponential	0.09	0.75	22.6
GPR 1.0-1.1 m (Med. Inst. Ampl.)	Exponential	0.09	0.70	35.4
GPR 1.1-1.2 m (Med. Inst. Ampl.)	Exponential	0.18	0.99	49.4
GPR 1.2-1.3 m (Med. Inst. Ampl.)	Exponential	0.17	1.03	43.9
GPR 1.3-1.4 m (Med. Inst. Ampl.)	Exponential	0.00	0.88	14.3
GPR 1.4-1.5 m (Med. Inst. Ampl.)	Exponential	0.00	0.90	24.2
GPR 1.5-1.6 m (Med. Inst. Ampl.)	Exponential	0.00	1.00	28.1
GPR 1.6-1.7 m (Med. Inst. Ampl.)	Exponential	0.00	0.79	24.4
GPR 1.7-1.8 m (Med. Inst. Ampl.)	Exponential	0.00	1.07	35.7
GPR 1.8-1.9 m (Med. Inst. Ampl.)	Exponential	0.00	1.05	35.4
GPR 1.9-2.0 m (Med. Inst. Ampl.)	Exponential	0.00	1.01	38.7

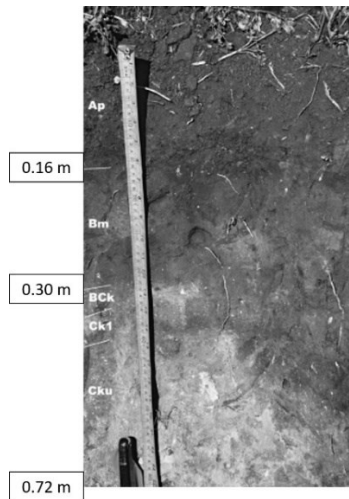


Figure 6.3 Alberta soil code Antler (ATL) sampled soil pedon and horizons depths observed during soil survey (Walker and Mcneil, 2004).

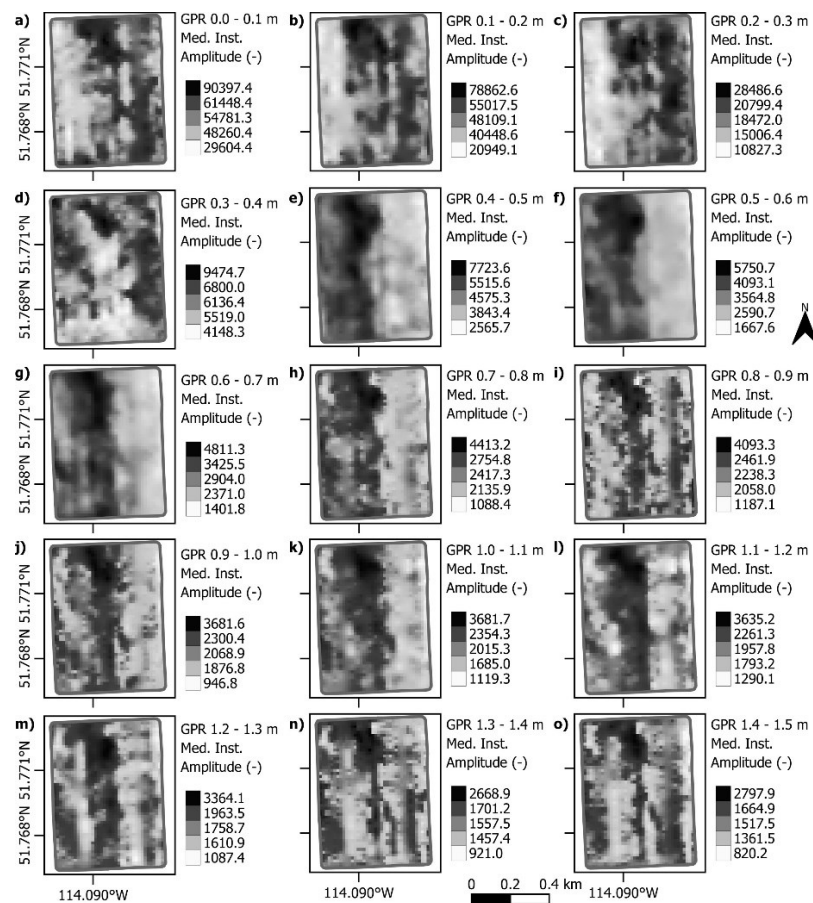


Figure 6.4 Ground penetrating radar (GPR) median instantaneous amplitude (Med. Inst. Amplitude) interpolated maps for every 0.1 m depths from 0 to 1.5 m. Each map has a different legend to highlight the spatial patterns represented by each depth layer.

Electromagnetic Induction (EM) EC_a. The EM apparent electrical conductivity data collected using the EM38-MK2 sensor was used to generate an experimental variogram. A spherical model was selected for both scanning depth layers during the model fitting. The resulting model parameters are presented in Table 6.5. The spatial variability modeled by the shallower layer presented a lower range than the one for the deeper layer, meaning that the latter has a larger scale variability while less detail is presented over smaller distances.

Table 6.5 Variogram model parameters for electromagnetic induction apparent electrical conductivity (EC_a) for shallow (0-0.75m) and deep (0-1.50m) layers.

Variable	Model	Nugget	Sill	Range (m)
Electromagnetic Induction EC _a 0-0.75 m	Spherical	0.16	1.13	263.6
Electromagnetic Induction EC _a 0-1.50 m	Spherical	0.14	1.28	463.0

The difference between the scale of the variability delineated by the two different layers from the EM sensor can also be observed through the maps in Figure 6.5. A likely reason for this observation is an EC_a response from the reflection of the electromagnetic waves over the bedrock material, which could vary less in space than the other soil horizons, attenuating the variability at shorter distances and increasing the range of the model. A comparison with the GPR variogram models for relative depth layers below 0.7 m presents a different behavior, with a reduction in the range, possibly because of its signal attenuation. This phenomenon has less effect on the response provided by the EM38-MK2 since its operational frequency (14.5 kHz) is lower than the one from the antenna used for GPR (400 MHz). As observed above, the maps from the two different scanning depths obtained by the EM38 are slightly different; however, they still present some similar regions (Figs 6.5(a) and (b)). The data highlights two regions with high apparent electrical conductivity, one at the central-north section of the field and another at the southeast. At the same time, a low EC_a is observed on a large section on the west side of the map.

The comparison between EM (Figure 6.5) and GPR (Figure 6.4) maps also presents some similarities. The regions with high (central north) and low (west) EC_a are well delineated on most GPR maps. In contrast, for the high EC_a at the southeast region, only the GPR median instantaneous amplitudes for 0-0.3m could capture this variability.

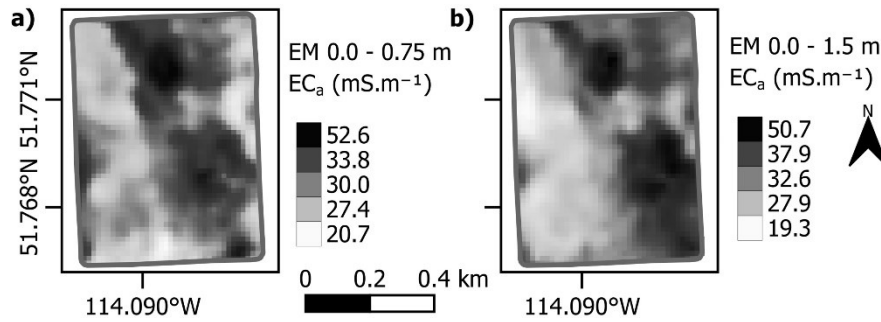


Figure 6.5 Electromagnetic induction (EM) apparent electrical conductivity (EC_a) interpolated maps for (a) shallow and (b) deep layers. Each map has a different legend to highlight the spatial patterns represented by each depth layer.

Galvanic Contact (GC) EC_a Data. The variogram modeling for the galvanic contact EC_a data from Veris 3100 presented a better fit using a spherical model. The values for nugget, sill and range (Table 6.6) are very similar for both scanning depths collected by this sensor, indicating that the spatial variability structure captured and modeled for the two depths might be similar. An analysis of the interpolated maps confirms this assumption by presenting two very similar maps (Figures 6.6(a) and (b)). The two maps present low EC_a zones across the west side of the field and two higher response zones at the north and southeast of the field, similar to the regions presented by the EM maps (Figure 6.5)

Table 6.6 Variogram model parameters for galvanic contact apparent electrical conductivity (EC_a) for shallow (0-0.3 m) and deep (0-0.9 m) layers.

Variable	Model	Nugget	Sill	Range (m)
Galvanic Contact EC_a 0-0.3 m	Spherical	0.06	1.16	244.5
Galvanic Contact EC_a 0-0.9 m	Spherical	0.06	1.15	250.9

The agreement mentioned above between the two sensors was expected since both use different methods to measure the same property, EC_a . It is important to observe that the data presented by both sensors do not completely agree with each other. This difference could be attributed to the different scanning depths provided by each instrument and the data collection time frame, GC in Spring 2017 and EM in Fall 2018.

A comparison of the maps for GC and GPR leads to similar conclusions as those drawn for the EM, where a similar spatial variability is observed between GC maps and the maps for the different

GPR layers. The overall similarities among EM and GC EC_a with GPR were also expected since the instantaneous amplitude values obtained during the data processing for the latter depend on the attenuation characteristics of the soil, which in turn depends on the soil electrical conductivity of the different soil layers.

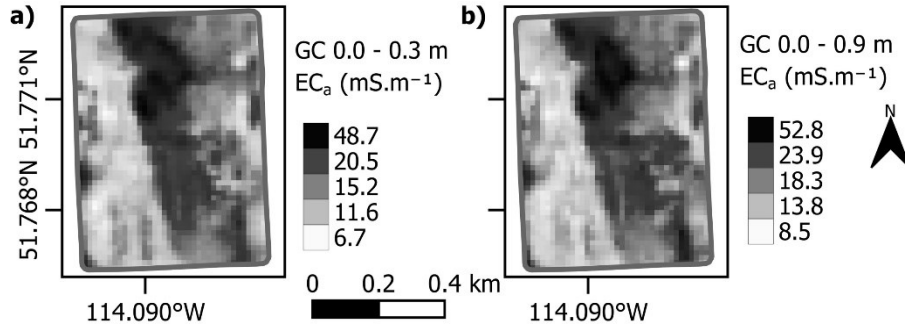


Figure 6.6 Galvanic contact (GC) apparent electrical conductivity (EC_a) interpolated maps for (a) shallow and (b) deep layers. Each map has a different legend to highlight the spatial patterns represented by each depth layer.

Gamma Ray Data. The experimental variograms for the passive gamma ray data from SoilOptix sensor presented a pure nugget effect for ¹³⁷Cs and ²³⁸U. Thus, these two variables were removed from the analysis. The other three variables presented an experimental variogram with a noticeable spatial structure; thus, a model was fitted. For ⁴⁰K and ²³²Th, an exponential model was a better fit than the spherical, while a spherical model returned a lower sum of square errors for the Count Rate.

Table 6.7 Variogram model parameters for SoilOptix passive gamma ray for ¹³⁷Cs, ²³²Th, ²³⁸U, ⁴⁰K, and Count Rate.

Variable	Model	Nugget	Sill	Range (m)
γ-ray ¹³⁷ Cs	-	-	-	-
γ-ray ²³⁸ U	-	-	-	-
γ-ray ²³² Th	Exponential	0.7	1.01	36.83
γ-ray ⁴⁰ K	Exponential	0.67	1.08	74.64
γ-ray Count Rate	Spherical	0.45	1.09	188.17

The resulting parameters for each of the fitted models are presented in Table 6.7, while the interpolated maps are in Figures 6.7(a)-(c). All the variables provided by the SoilOptix system presented a higher nugget effect than the ones obtained for the data from the three sensors

previously discussed. That could be due to different reasons, such as a random variability at short distances (random error) or spatial structure that happens at a smaller scale than the sampling distances (Chilès and Delfiner, 2012); however, more precise explanation could be determined if samples at a shorter distance were available.

All three γ -ray variables that presented a spatial structure had a similar nugget and sill values but an increase in range in the following order: ^{232}Th , ^{40}K , and Count Rate. However, it is important to notice that for the variables in which an exponential model was used, a practical range is calculated (95% of the sill). That could explain the more considerable difference between the range for the count rate (spherical) and the other two variables (exponential).

Even though there were some differences in the range for the variogram models, the maps obtained for the radionuclides and count rate all presented a similar variability. Low values were presented in the middle of the field, delineating a region on the central-north section that matched the high EC_a and instantaneous amplitude region at the same spot. Thus, it is possible to conclude this region is constant throughout the different sensors. In fact, there is a poorly drained region (slough) at this spot, which would cause higher readings for EC_a , while low gamma ray values, as water attenuates the gamma ray signals. Also, it is important to observe that the other high EC_a region that can be identified in the southeast of the field is not as well delineated in the radionuclides maps, while on the Count Rate maps, it starts to appear, but it is smaller and less prominent than that observed for EM-38 and Veris 3100 maps.

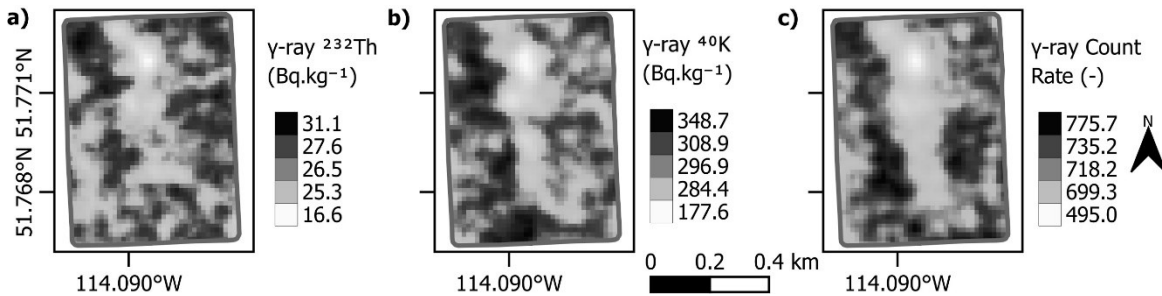


Figure 6.7 SoilOptix passive gamma ray (γ -ray) maps for (a) ^{232}Th , (b) ^{40}K , and (c) Count Rate.

Terrain Data. An experimental variogram was generated for the elevation data from the RTK GNSS receiver used during γ -ray data collection and due to its lower sum of square errors, a Gaussian model was fitted and used to create a DEM. A sill of 2.86 and a nugget of 0.01 were observed. The very low nugget value indicates that even at small distances, the variability could be modeled. Also, a large practical range of 513.41 m was obtained, which was expected since the

field is located on the Canadian prairies and presents a characteristic landscape for the region, with undulated plains, hills, and some plateaus while the changes in the topography are smooth. No variogram analysis was calculated for Slope since this data layer was generated based on the DEM and no interpolation was necessary.

Figure 6.8(a) presents the maps for DEM and Figure 6.8(b) terrain slope. A visual comparison between these two maps, with the data previously presented for PSS, shows some agreement. The highly elevated area on the east side of the field matches the low EC_a readings (Figures 6.5 and 6.6). At the same time, the bottom of the steep slope delineates where the EC_a starts to increase, creating a very notable diagonal line in the maps (especially on GC maps). Those same regions can be observed in some GPR maps and gamma ray readings but are not as prominent as for the EC_a sensors. That might reflect mostly the water flow in the field and spatial variability of soil moisture but could also be related to soil erosion and nutrient runoff.

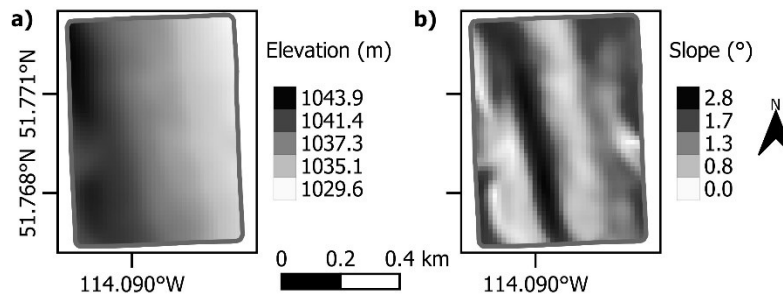


Figure 6.8 Maps for (a) digital elevation model and (b) terrain slope.

Correlation Analysis. Previously the different maps were briefly compared, and it was possible to identify some agreements and draw some conclusions about the reasons for this. Now, to better evaluate and compare the data presented for the PSSs, elevation, and slope, a correlation analysis was performed. The results are presented in Figure 6.9. Four other depths for the GPR (0.0-0.3 m, 0.0-0.9 m, 0.0-0.75 m, and 0.0-1.5 m) were added to this comparison but not used for any further analysis. Those were obtained by averaging the instantaneous amplitude at depths that corresponded the best with the shallow and deep EC_a readings from EM and GC. These four new GPR depths were added to evaluate if they would present a higher correlation with EC_a sensors' readings when compared to the GPR 0.1 m layers. Also, it is important to remember that the sensors' data were collected during different seasons, which adds uncertainty to the comparisons among the sensors, as most likely soil and environmental conditions that influence the data (e.g. moisture and temperature) were different during the data acquisition.

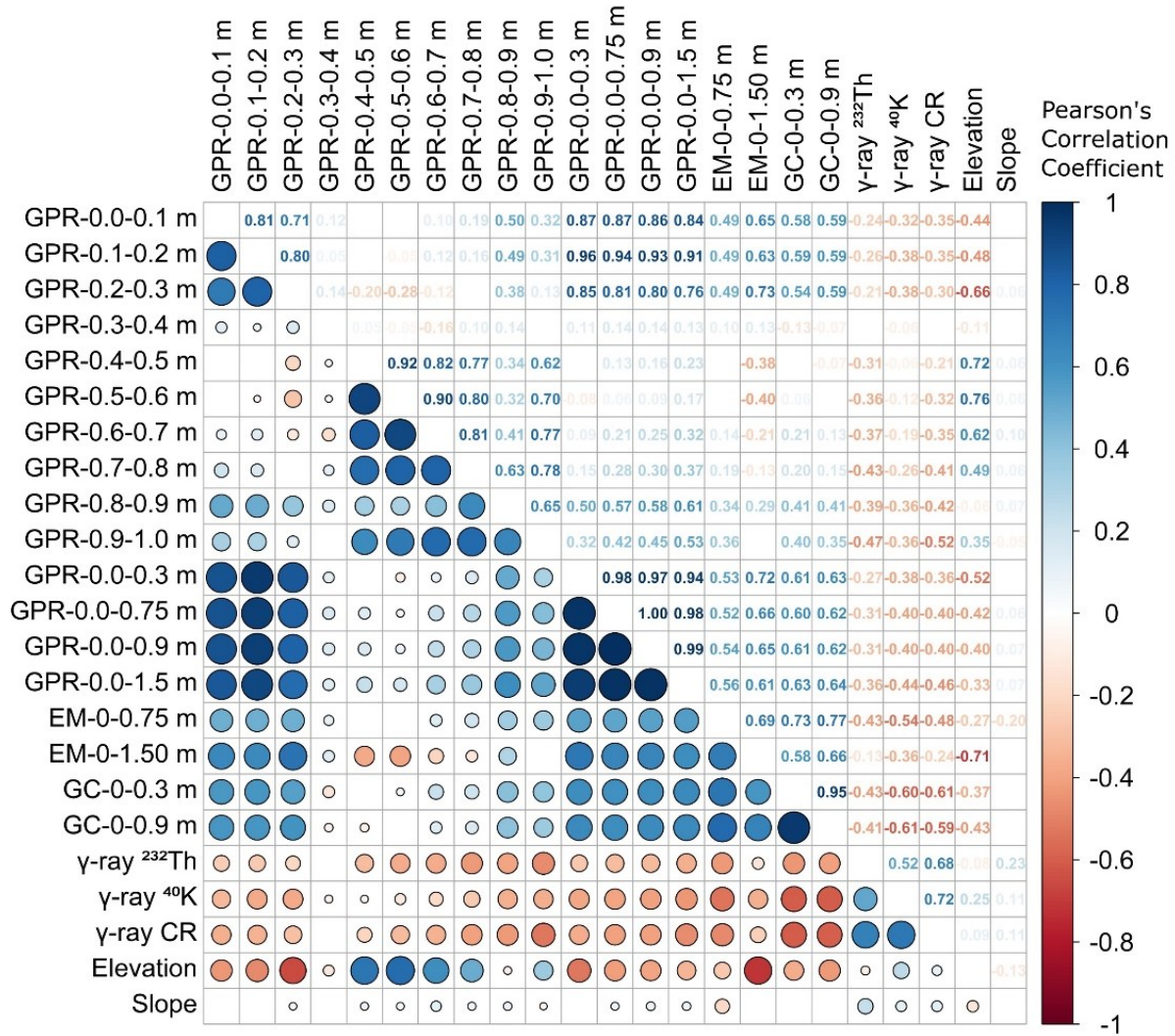


Figure 6.9 Correlation analysis for all proximal soil sensors, digital elevation model, and terrain slope. The upper triangular matrix presents the numerical values for the Person's correlation coefficient (r), while the lower one shows a graphical representation of the correlation – a larger circle radius means a higher absolute r. Circles and numbers are filled with a color scale based on the r values. Empty cells represent the non-significant correlation at a significance level of 0.05. GC – galvanic contact apparent electro conductivity, GPR – ground penetrating radar, EM – electromagnetic induction apparent electro conductivity, CR – count rate.

Comparing the correlation among different depths from GPR leads to the same observation when evaluating the variogram parameters and maps for these variables. At least three very well-delimited groups of distinct spatial variability match the soil horizons within 0 to 1 m depths. Looking at the correlogram, the first group occurs at the relative depth of 0 to 0.3 m, the second

from 0.4 m to 0.8 m, and the third from 0.8 m to 1 m. Those can be easily observed due to the high correlation within each group's depth range, while the correlation with the depths from other groups, in absolute values, is smaller or not significant. The only instantaneous amplitude that does not follow this pattern is 0.9-1 m, which presents a higher correlation with the previous group (0.4-0.8 m). That could be the result of a not complete transition from one soil horizon to the next at that depth.

Examining the GPR maps with the other sensors also presented some significant correlations. Both EC_a sensors show a higher correlation with the GPR readings from the depths of 0 to 0.3 m. Also, when comparing the matched scanning depths of GC and EM, the correlation showed a slight increase, but the correlation pattern did not match what was expected. For example, one would expect that averaging the instantaneous amplitude from GPR at 0-0.3 m would correlate the best with GC shallow EC_a readings. However, the data did not demonstrate this; the highest correlation for GC shallow depth (0-0.3 m) is with GPR-0.0-1.5 m. Also, intuitively GPR and EC_a readings would present a negative correlation since high soil EC_a tends to attenuate the radar signal and reduce the amplitude. These observations could be due to the different data collecting timing but should be further studied in future research; however, it is beyond the scope of the present paper. The correlation between soil gamma ray data and GPR presents an overall negative correlation. In absolute values, higher correlations were observed for depths higher than 0.5 m for ^{232}Th , while for Count Rate and ^{40}K , the higher absolute values happen with GPR depths from 0-0.3 m and 0.7-1 m.

The higher correlation between γ -ray and deeper GPR readings could be attributed to the fact that below 0.5 m the soils in this field mostly reached the C horizon, which is constantly going through soil generation weathering processes, known to produce gamma ray radiation (Dickson et al., 1996). Thus, GPR could be capturing the variability that exists at the parent material or C horizon that affects the radionuclides decay, causing the higher correlation. On the other hand, the highest correlation for ^{40}K with GPR shallower depths could be related to the K fertilizer application due to agricultural practices. Since ^{40}K contributes to Count Rate, the latter also presents a higher absolute correlation to the GPR shallower depths. Finally, the correlation between GPR and the elevation presents high positive numbers for the relative depths from 0.4 to 0.8 m. In contrast, Slope correlations are very low, not only with GPR but among the other PSS. The overall high absolute correlation coefficients between GPR and elevation could be attributed to soil

erosion and material deposition on the field's lower part (east side). The eroded material would change the distance between the sensor and the soil horizons, causing the split of the field in east and west (Figure 6.4), following a line that matches the topographic ridge observed in Figure 6.8. That assumption agrees with what was described by Glaser and Wagner (2019), as they studied and modeled the effect of deposition of material (dynamic terrain condition) on the time-domain EM investigations to detect unexploded ordinance and munitions and explosives of concern. It is also important to note that material deposition can change the dielectric constant, reinforcing the need for future work to map this property and use it during data processing.

EM and GC correlations are relatively higher than the ones among any other combination of sensors because both sensors measure the same attribute. The higher correlation is observed for the GC deep (0-0.9 m) and EM shallow (0-0.75 m) maps. Similar correlations between both sensors and scanning depths have been reported in the literature (Fritz et al., 1999; Sudduth et al., 2003). Also, when comparing the EC_a maps with gamma ray, negative correlations were observed with higher absolute values for the comparison between EC_a shallower depths and gamma ray. Research developed in Canada that used a similar sensor to the EM38 also reported negative correlations between EC_a and gamma ray. At the same time, the highest absolute values were observed between shallower EC_a responses and gamma ray readings (Ji et al., 2019). Others in Australia presented some contrasting results with the correlation between EM38 and gamma ray reading varying from positive to negative in different study sites (Rodrigues et al., 2015).

Overall, through the analysis of the different data layers obtained from terrain and PSS, it is possible to note that there are some similarities among them. However, each of these layers does not completely correlate with each other, and the comparison among the different maps obtained for each variable (Figures 6.4-6.8) highlights how different each sensor captures the variability in the field. Some sensors are more affected by the soil parent material (γ -ray), while others can capture the variability at different depths (GPR). Thus, the data ensemble presented shows high potential in differentiating the distinct soil models for this field.

Proximal soil sensors data and soil models

To evaluate the points presented in the previous paragraph and to answer tentatively one of the research questions regarding the capability of PSS on the definition of soil types/models, the results from the One-Way ANOVA and Tukey-Kramer ad-hoc test are presented in Table 6.8. Based on the statistical theory that the higher significant F-Test values would represent a better

differentiation among the groupings (different soil models), the rows of this table are in decreasing order of the F-test values. In addition, Figures 6.10(a)-(f) present an overlay of the soil model boundaries with a selected variable from each of the PSS, DEM, and slope.

Based on the results in Table 6.8, all the variables rejected the One-Way ANOVA null hypothesis at a significance level of 0.05, meaning that at least one of the averages among the different soil models was significantly different. Thus, the ad-hoc test was run for every variable, and based on the results, none of the variables could completely differentiate all soil models. Based on the F-test value and ad-hoc test results, the data layers GPR-0.5-0.6 m, GPR-0.6-0.7 m, and Elevation presented the highest capability of separating the soil models. However, a visual analysis of Figure 6.10(c) indicates a good agreement between the EM 0.0 – 1.5 m data and the soil's delineation. Therefore, for this field, a single sensor or data layer would not be able to differentiate all the different soil models described by the soil survey performed in 2003.

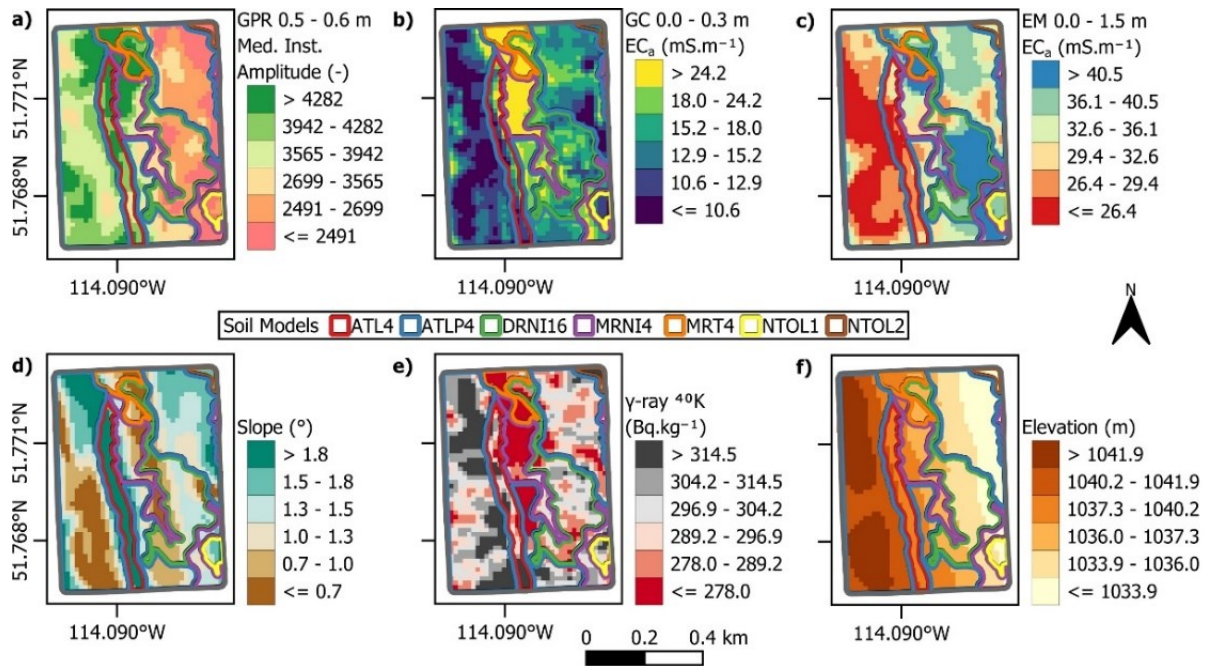


Figure 6.10 Overlay of soil models and (a) Ground penetrating radar (GPR) median instantaneous amplitude (Med. Inst. Amplitude), (b) galvanic contact (GC) apparent electro conductivity (EC_a), (c) electromagnetic induction (EM) apparent electro conductivity, (d) terrain slope, (e) γ -ray for ^{40}K , and (f) elevation (m)

Using proximal soil sensors data to improve the soil survey boundary delineation

A closer analysis of the maps presented in Figure 6.10 leads to the conclusion that most of the soil models' boundaries are in fairly homogenous regions, meaning that their placement seems to be reasonably correct. However, some inconsistencies exist between the PSS data and soil models' boundaries. Considering the previous analysis presented for the sensors data, which regions were constant among the different soil sensing techniques, it is possible to conclude that the sensors provided reliable information about the soil variability. Thus, demonstrate the potential for using this information to better delineate soil boundaries, which has also been reported in the literature (e.g., Adamchuk et al., 2011; James et al., 2003).

For example, the high EC_a and GPR instantaneous amplitudes observed on the north-central area of the field seem to match very well with the definition of the MRT4 soil model, an imperfectly drained soil. However, when comparing the current MRT4 boundaries with the region presented in the PSS maps, there seems to be a mismatch in shape and size. Another example is the extension of the ATL4 soil model, which according to Table 6.1, differs from ATLP4 by the lesser presence of glaciolacustrine veneer that has been eroded by subsurface water due to ATL4 being located at steeper slopes. Based on this description and the slope map in Figure 6.10, ATL4 should extend to the northwest corner of the field.

As noted in the paragraph above and the previous section, information from the different soil sensing techniques is needed to delineate the different soil models. Therefore, a multivariate analysis is needed. Thus, all the PSS and terrain data layers were used in the SFCM algorithm, where 7 clusters were defined. The number of clusters was to match the number of soil models. Figure 6.11(a) presents the clustering results and Figure 6.11(b) an overlay with the soil models' boundaries.

Table 6.8 One-Way ANOVA and Tukey-Kramer ad-hoc test results for each proximal soil sensor and terrain data. The rows of the table are ordered by the F-Test values, and different lowercase letters represent significant differences in the averages among the different soil models. GPR – ground penetrating radar, GC – galvanic contact, EM – electromagnetic induction, EC_a – apparent electrical conductivity

Variable	OWA p value	F Test	Soil Models						
			ATL 4	ATLP 4	DRNI 16	MRNI 4	NTOL 1	MRT 4	NTOL 2
GPR 0.5-0.6 m	<0.05	1377.0	b	c	d	c	d	a	E
GPR 0.6-0.7 m	<0.05	1303.5	a	b	c	b	d	a	E
Elevation	<0.05	867.7	a	a	c	c	d	b	E
GC EC _a 0-0.3 m	<0.05	671.6	d	de	c	b	e	a	De
GPR 0.4-0.5 m	<0.05	575.7	a	b	c	b	c	a	D
Slope	<0.05	426.5	d	b	a	b	c	a	C
EM EC _a 0-1.50 m	<0.05	380.8	d	d	ab	b	ab	a	C
GPR 0.0-0.1 m	<0.05	374.0	cd	d	b	b	b	a	C
EM EC _a 0-0.75 m	<0.05	338.2	c	c	b	b	c	a	C
GC EC _a 0-0.9 m	<0.05	329.8	d	d	c	b	d	a	D
GPR 0.9-1.0 m	<0.05	288.9	b	cd	d	bc	e	a	F
GPR 0.7-0.8 m	<0.05	235.0	b	c	d	c	d	a	E
GPR 0.8-0.9 m	<0.05	200.4	bc	d	cd	b	bc	a	E
GPR 0.1-0.2 m	<0.05	191.5	b	c	b	b	b	a	C
GPR 0.2-0.3 m	<0.05	140.0	c	d	ab	b	b	a	D
γ-ray 40K	<0.05	92.6	d	cd	c	b	cd	a	E
γ-ray ²³² Th	<0.05	68.3	c	c	c	b	c	a	D
γ-ray Count Rate	<0.05	67.8	c	c	c	b	c	a	C
GPR 0.3-0.4 m	<0.05	62.4	d	c	c	cd	b	a	C

Group 7 presents a delineation of the known slough in the field, corresponding to the MRT4 soil model, which through the analysis of Figure 6.10 it was concluded that some boundary adjustments were needed. However, this multivariate spatial clustering presents a clear boundary for this soil. The same approach can be applied to the other soil models. Another interesting example is the DRNI16 soil model. According to this soil description, one of the major differences among the other soils in the field is the high content of exchangeable sodium in the B horizon. From previous soil sampling taken from this field, it is known that the regions covered by cluster Group 5 contain a high sodium content. Therefore, based on this information and the geolocation of the current soil survey boundary for DRNI16, one could assume that group 5 delineates the boundaries for the DRNI16 soil model. Based on these observations, it is possible to conclude that the clustering boundaries presented in Figure 11 are very clear and present very well-defined regions that are slightly different from the soil types' boundaries from the soil survey. A possible reason for this discrepancy could be attributed to a known limitation of traditional soil surveys, the low density of soil inspection sites. While for the traditional soil survey for this study field, approximately 1 sample per 3 ha was collected, the PSS data presented at least 35 readings per 1 ha. Thus, the latter provides a finer resolution in data, which increases the capability of delineating boundaries with higher accuracy and precision.

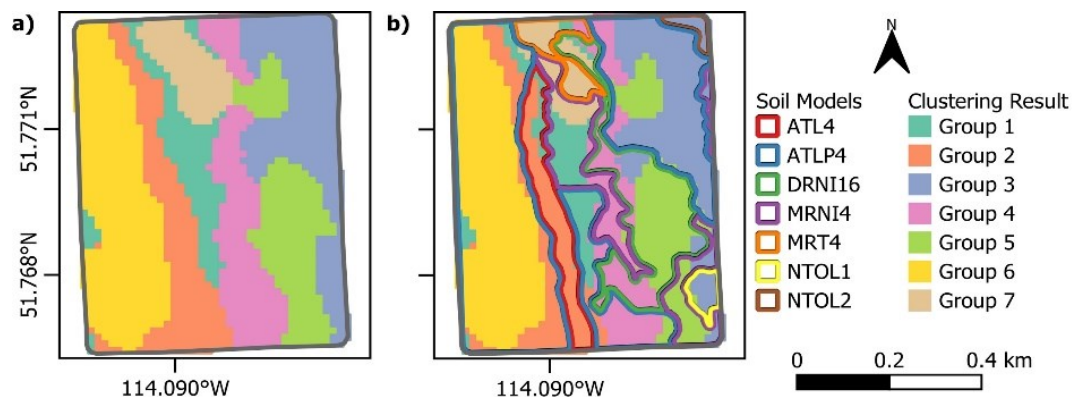


Figure 6.11 Proximal soil sensors and terrain data (a) spatial fuzzy c-means clustering result and (b) overlay with soil models.

As observed through the process described above, using the clusters obtained from the SFCM algorithm could help delineate the boundaries but not necessarily show which soil type/model belongs to each one of these clusters. This because it is an unsupervised clustering method. Thus, the method applied should be used in a collaborative manner between PSS, terrain data and

classical field survey tools (samples, digs, trenches, etc.). Maybe the method presented here could be used prior to the classical field survey, helping with the placement of sampling points, or as used in this paper, where a classical method was employed and now can be adjusted using the clusters.

Another example that reinforces the need for the combination of classical and sensor approaches can be observed in the clustering results for group 4, which in comparison to the current soil models' delineation most likely represents the same soil model as group 3. However, some field survey knowledge would be necessary to confirm this assumption. Another example would be the discussion about the possible existence of NTOL1 and NTOL2 in this field. The clustering algorithm did not create clear boundaries for these two soil models. Once again, field survey information would be necessary to determine their existence in the field.

It is also important to notice that a combination of sensors was needed to produce the results in Figure 6.11 and that using a single sensor would not produce the same results. Tests conducted using one or two sensors provided results that could help improve the soils' boundaries delineation. Still, they were not as spatially constant and sometimes, no clear boundary could be determined. That highlights the benefits of using data from different sensors (data fusion), which have been described in the literature (Adamchuk et al., 2011; Pantazi et al., 2020). Future studies must focus on developing consistent frameworks that would allow the identification of the best set of sensors that would be most beneficial during the survey. In addition, the usage of an acoustic or seismic component in the sensor fusion process could improve the accuracy of the soils' boundary delineation, as it would take into consideration soil density in addition to soil electrical properties, which mainly respond to moisture (Glaser et al., 2021).

From reading this paper, one might think that some coding experience would be necessary to perform the required analysis. However, although customized scripts written in the languages R and python were used, there are multiple tools that, in combination, could generate results similar to those ones presented here. For example, GPR data could be processed using the software provided by the manufacturers (e.g., Sensors and Software Inc., RADAN from GSSI). At the same time, interpolation could be performed using a Geographical Information System (GIS). Finally, for spatial clustering, not all GIS can carry out this analysis; however, there are freely available software that could be used for this matter, such as GeoDa (Anselin et al., 2006) and NSA (Dhawale et al., 2014; Saifuzzaman et al., 2019).

6.4 Conclusion

This work suggests that there is potential in using different proximal soil sensing technologies to improve the delineation of soils within a field. Even though using one proximal sensor can help in this process, a combination of different sensor types and terrain data provided a much better understanding of the spatial variability within the field and, consequently, offered insight into the delineation of soil boundaries, demonstrating the importance of sensor fusion. In addition, by using an unsupervised spatial cluster tool on the data from the sensors and terrain, regions that matched well with the description of soils provided during a soil survey were delineated, thus, offering a possible solution to help improve the accuracy of soil boundary delineation during soil surveys. Since an unsupervised method was used as a grouping mechanism, the user's knowledge is still needed to know which soil belongs to each cluster. In conclusion, there is still a need for the classical survey methods, where one would classify the different soils and later use this information in combination with the generated clusters to delineate the boundaries.

6.5 Acknowledgments

We would like to thank the reviewers for providing constructive comments, which have contributed to the improvement of the manuscript. We thank all Olds College crew and students that helped during the soil survey in 2003 as well as all those involved in collecting the proximal soil sensor data used in this research. We also thank SoilOptix Inc. (Tavistock, ON, Canada) for providing the complete gamma ray analysis. This research is part of the project "Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services" supported by Mitacs through the Mitacs Accelerate program.

6.6 References

- Adamchuk, V.I., Rossel, R.A.V., Sudduth, K.A., and Lammers, P.S., 2011, Sensor Fusion for Precision Agriculture: in *Sensor Fusion - Foundation and Applications*, Thomas, C. (ed.), InTech, Rijeka, 27–40.
- Agriculture and Agri-food Canada, 2021, Detailed Soil Survey (DSS) compilations. Accessed on February 15, 2022. URL <https://sis.agr.gc.ca/cansis/nsdb/dss/v3/index.html>.
- Allred, B., Daniels, J.J., and Ehsani, M.R., 2008, *Handbook of agricultural geophysics*: CRC Press, Boca Raton, 410 pp.
- André, F., van Leeuwen, C., Saussez, S., Van Durmen, R., Bogaert, P., Moghadas, D., de Rességuier, L., Delvaux, B., Vereecken, H., and Lambot, S., 2012, High-resolution imaging

- of a vineyard in south of France using ground-penetrating radar, electromagnetic induction and electrical resistivity tomography: *Journal of Applied Geophysics*, 78, 113–122.
- Anselin, L., Syabri, I., and Kho, Y., 2006, *GeoDa: An introduction to spatial data analysis: Geographical Analysis*, 38, 5–22.
- ASIC (Alberta Soil Information Centre), 2001, *AGRASID 3.0: Agricultural region of Alberta soil inventory database (Version 3.0)*, Revised 3rd ed.: Agriculture and Agri-Food Canada, Research Branch; Alberta Agriculture, Food and Rural Development, Conservation and Development Branch.
- De Benedetto, D., Castrignanò, A., Rinaldi, M., Ruggieri, S., Santoro, F., Figorito, B., Gualano, S., Diacono, M., and Tamborrino, R., 2013, An approach for delineating homogeneous zones by using multi-sensor data: *Geoderma*, 199, 117–127.
- Bowser, W.E., Peters, T.W., and Newton, J.D., 1951, *Soil Survey of Red Deer Sheet.*: Department of Extension, University of Alberta, Edmonton, 86 pp.
- Box, G.E.P., and Cox, D.R., 1964, An Analysis of Transformations: *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243.
- Cai, W., Chen, S., and Zhang, D., 2007, Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation: *Pattern Recognition*, 40(3), 825–838.
- Castrignanò, A., Buttafuoco, G., Quarto, R., Parisi, D., Viscarra Rossel, R.A., Terribile, F., Langella, G., and Venezia, A., 2018, A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture: *Catena*, 167, 293–304.
- Castrignanò, A., Buttafuoco, G., Quarto, R., Vitti, C., Langella, G., Terribile, F., and Venezia, A., 2017, A combined approach of sensor data fusion and multivariate geostatistics for delineation of homogeneous zones in an agricultural field: *Sensors*, 17(12), 1–20.
- Chilès, J.P., and Delfiner, P., 2012, *Geostatistics*: John Wiley & Sons, Inc., Hoboken, 731 pp.
- Claerbout, J.F., 1976, *Fundamentals of Geophysical Data Processing with applications to petroleum prospecting*: Blackwell Scientific Publications, Palo Alto, 274 pp.
- Daniels, D.J., 1996, Surface-penetrating radar: *Electronics & Communication Engineering Journal*, 8(4), 165–182.

- Dhawale, N.M., Adamchuk, V.I., Prasher, S.O., Dutilleul, P.R.L., Ferguson, R.B., 2014, Spatially constrained geospatial data clustering for multilayer sensor-based measurements: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-2, 187–190.
- Dickson, B.L., Fraser, S.J., and Kinsey-Henderson, A., 1996, Interpreting aerial gamma-ray surveys utilising geomorphological and weathering models: *Journal of Geochemical Exploration*, 57(1-3), 75–88.
- Doolittle, J.A., 1987, Using Ground-penetrating Radar to Increase the Quality and Efficiency of Soil Surveys: in *Soil survey techniques*, Reibold, W. U., and Petersen, G. W. (eds.), Soil Science Society of America, Inc., Madison, 11–32.
- Fritz, R.M., Malo, D.D., Schumacher, T.E., Clay, D.E., Carlson, C.G., Ellsbury, M.M., and Dalsted, K.J., 1999, Field Comparison of Two Soil Electrical Conductivity Measurement Systems: in *Proceedings of the Fourth International Conference on Precision Agriculture*, Robert, P.C. Rust, R.H., and Larson, W.E (eds), 4th International Conference on Precision Agriculture, John Wiley & Sons, Ltd , 1211–1217.
- Environment Canada, 2019, Canadian climate normals 1981–2010 station data. Accessed on February 15, 2022. URL https://climate.weather.gc.ca/climate_normals/results_1981_2010_e.html?searchType=stnProv&lstProvince=AB&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=2120&dispBack=0.
- Gebbers, R., Lück, E., Dabas, M., and Domsch, H., 2009, Comparison of instruments for geoelectrical soil mapping at the field scale: *Near Surface Geophysics*, 7, 179–190.
- Gelb, J., and Apparicio, P., 2021, Apport de la classification floue c-means spatiale en géographie: essai de taxinomie socio-résidentielle et environnementale à Lyon: *Cybergeog*, 13, 1–26. [French]
- Getis, A., 2009, Spatial weights matrices: *Geographical Analysis*, 41, 404–410.
- Glaser, D.R., Burch, K., Brinkley, D.L., and Reppert, P., 2021, Localization of deep voids through geophysical signatures of secondary dewatering features: *Geophysics*, 86(3), WA139–WA152.
- Glaser, D.R., and Wagner, A.M., 2019, Dynamic cold regions terrain effects on time-domain electromagnetic induction data: *Cold Regions Science and Technology*, 158, 52–61.

- Gräler, B., Pebesma, E., and Heuvelink, G., 2016, Spatio-temporal interpolation using gstat: The R Journal, 8(1), 204–218.
- Hijmans, R.J., 2021, raster: Geographic Data Analysis and Modeling, R package version 3.4.13, URL <https://github.com/rspatial/raster>.
- Inman, D.J., Freeland, R.S., Ammons, J.T., and Yoder, R.E., 2002, Soil Investigations using Electromagnetic Induction and Ground-Penetrating Radar in Southwest Tennessee: Soil Science Society of America Journal, 66, 206–211.
- James, I., Waine, T., Bradley, R., Taylor, J., and Godwin, R., 2003, Determination of Soil Type Boundaries using Electromagnetic Induction Scanning Techniques: Biosystems Engineering, 86(4), 421–430.
- Ji, W., Adamchuk, V.I., Chen, S., Mat Su, A.S., Ismail, A., Gan, Q., Shi, Z., Biswas, A., 2019, Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study: Geoderma, 341, 111–128.
- Lyle, G., Bryan, B.A., and Ostendorf, B., 2014, Post-processing methods to eliminate erroneous grain yield measurements: Review and directions for future development: Precision Agriculture, 15, 377–402.
- Maldaner, L.F., Molin, J.P., and Spekken, M., 2022, Methodology to filter out outliers in high spatial density data to improve maps reliability: Scientia Agricola, 79(1), 1–7.
- de Mendiburu, F., 2021, agricolae: Statistical Procedures for Agricultural Research. R package version 1.3.5. URL <https://github.com/myaseen208/agricolae>.
- Nesbitt, I., Simon, F.-X., Hoffmann, F., Paulin, T., and Teshaw, 2022, readgssi: an open-source tool to read and plot GSSI ground-penetrating radar data, Python library version 0.0.21, URL <https://github.com/iannesbitt/readgssi>.
- Novakova, E., Karous, M., Zajíček, A., and Karousova, M., 2013, Evaluation of ground penetrating radar and vertical electrical sounding methods to determine soil horizons and bedrock at the locality dehtaře: Soil and Water Research, 8(3), 105–112.
- Pantazi, X.E., Moshou, D., and Bochtis, D., 2020, Utilization of multisensors and data fusion in precision agriculture: in Intelligent Data Mining and Fusion Systems in Agriculture, Pantazi, X. E., Moshou, D., and Bochtis, D (eds), Academic Press, 103–173.
- Pawluk, S., and Bayrock, L., 1969, Some characteristics and physical properties of Alberta tills: Research Council of Alberta, Edmonton, 78 pp.

- Pebesma, E.J., 2004, Multivariable geostatistics in S: The gstat package: Computers and Geosciences, 30(7), 683–691.
- Pebesma, E., 2018, Simple Features for R: Standardized Support for Spatial Vector Data: The R Journal, 10(1), 439.
- Peterson, R.A., 2021, Finding Optimal Normalizing Transformations via bestNormalize: The R Journal, 13(1), 294–313.
- Rodrigues, F.A., Bramley, R.G.V., and Gobbett, D.L., 2015, Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils: Geoderma, 243–244, 183–195.
- Saifuzzaman, M., Adamchuk, V., Buelvas, R., Biswas, A., Prasher, S., Rabe, N., Aspinall, D., and Ji, W., 2019, Clustering Tools for Integration of Satellite Remote Sensing Imagery and Proximal Soil Sensing Data: Remote Sensing, 11(9), 1036.
- Sanches, G.M., Otto, R., Adamchuk, V., and Magalhães, P.S.G., 2022, Spatial variability of soil attributes by an electromagnetic induction sensor: A framework of multiple fields assessment under Brazilian soils: Biosystems Engineering, 216, 229–240.
- Sensors & Software Inc., 2003, EKKO_View Enhanced & EKKO_View Deluxe User's Guide, 132 pp.
- Simeoni, M.A., Galloway, P.D., O'Neil, A.J., and Gilkes, R.J., 2009, A procedure for mapping the depth to the texture contrast horizon of duplex soils in south-western Australia using ground penetrating radar, GPS and kriging: Australian Journal of Soil Research, 47(6), 613–621.
- Soil Classification Working Group, 1998, The Canadian System of Soil Classification: Agriculture and Agri-Food Canada, Revised 3rd ed), Ottawa, 187 pp.
- Spekken, M., Anselmi, A.A., and Molin, J.P., 2013, A simple method for filtering spatial data: in Precision Agriculture '13, Stafford, J.V. (ed), 9th European Conference on Precision Agriculture, Wageningen Academic Publishers, Wageningen, 259–266.
- Sudduth, K.A., Kitchen, N.R., Bollero, G.A., Bullock, D.G., and Wiebold, W.J., 2003, Comparison of electromagnetic induction and direct sensing of soil electrical conductivity: Agronomy Journal, 95(3), 472–482.
- Walker, B.D., and Mcneil, R.L., 2004, Detailed Soil Survey of Olds College Farm, 79 pp.

- Wickham, H., François, R., Henry, L., and Müller, K., 2021, dplyr: A Grammar of Data Manipulation. R package version 1.0.8, URL <https://github.com/tidyverse/dplyr>
- Zhang, J., Lin, H., and Doolittle, J., 2014, Soil layering and preferential flow impacts on seasonal changes of GPR signals in two contrasting soils: *Geoderma*, 213, 560–569.
- Zhao, F., Jiao, L., and Liu, H., 2013, Kernel generalized fuzzy c-means clustering with spatial information for image segmentation: *Digital Signal Processing*, 23(1), 184–199.

Abbreviations for Chapter 7

Abbreviation	Definition
CI	Cone-Index
cLHS	conditional Latin Hypercube
CNN	Convolution Neural Network
DEM	Digital Elevation Model
DF	Data Fusion
DSS	Decision Support System
EC _a	Apparent Electrical Conductivity
EMI	Electromagnetic Induction
	Global Navigation Satellite
GNSS	System
GPR	Ground Penetrating Radar
IDW	Inverse Distance Weighting
K	plant-available Potassium
LiDAR	Light Detection and Ranging
ML	Machine Learning
P	plant-available Phosphorus
PA	Precision Agriculture
PLSR	Partial Least Squares
PSS	Proximal Soil Sensing
RF	Random Forest
SVM	Support Vector Machine

Chapter 7: General Discussion

Promising approaches for processing PSS data, interpolating lower sampling densities, and fusion of PSS and topography data for soil characterization were presented in Chapters 3-6. These approaches were carefully selected, proposed, and evaluated to support DSS development for PA soil management strategies.

Chapter 3 was constructed based on a comprehensive literature review of data processing and rasterization strategies used for spatial data. While a well-established procedure is defined for handling data originating from yield monitors, a general standard procedure for PSS is non-existent to the author's knowledge. An abstract from a pedometrics conference (Ji et al., 2017) is one of the few publications defining a framework for processing PSS on-the-go mapping. However, the abstract mentioned above does not provide a detailed description of the adopted procedures.

If PSS data is not processed and interpolated correctly, its value to soil management is minimal. Additionally, as previously highlighted, the 2023 Precision Agriculture Dealership survey (Erickson & Lowenberg-DeBoer, 2023) suggested that expanding the adoption of PA will be limited unless the benefits of using PA strategies surpass the costs; furthermore, well-analyzed and ready-to-use PA information must be provided to practitioners to ease the decision-making process.

If the costs related to data collection are left aside, the development of a DSS with an automated processing procedure has the potential to reduce the costs of extracting valuable information from PSS and provide readily available PSS-based information to PA practitioners. The framework definition for batch-processing of PSS presented in Chapter 3 was essential to further developing this thesis and PA.

The procedures and methods evaluated for the data projection, position offset, and operational filtering in the framework defined in Chapter 3 were demonstrated to be suitable for automation and effective for maximizing the value of the data. In contrast, some limitations were encountered for the methodologies evaluated for the global and local statistical filtering.

The filtering methodology defined by Maldaner et al. (2022) is very sensitive to the required user-defined coefficients and primarily relies on the users' subjective analysis (i.e., visual comparison of thematic maps) to adjust these coefficients. Contrast this with the methodology defined by Leroux et al. (2018) which requires less input from users and is more suitable for automated processing. Still, the results indicated that it tended to remove observations excessively.

Regarding the interpolation of PSS, automatically optimized IDW coefficients produced similar results as kriging-based interpolators. Due to this similarity and the higher computation power required by kriging methods, IDW seems more suitable when developing DSS. On the other hand, the advantage of using kriging is that it provides estimates of the spatial correlation of the data, allowing for an analysis of the spatial structure of the studied variable, an approach used in Chapter 6.

Future research in batch-processing of PSS data should focus on defining optimization criteria to assess and optimize the user-defined coefficients required for Maldaner et al. (2022). Another option would be to explore filtering methodologies that rectify observations considered outliers instead of removing them. Nevertheless, the steps defined in Chapter 3's framework are suitable for DSS development and often improve PSS data quality.

The interpolated maps from PSS obtained from the framework in Chapter 3 provide insights into soil spatial variability in the field. However, as highlighted in multiple sections of this thesis, data originating from a single sensor can represent the interaction of many environmental variables (Adamchuk et al., 2011). Additionally, while some PSS can provide direct measurement of soil properties used for agricultural soil management decisions (e.g., ion-selective electrodes for soil pH; Adamchuk et al., 2005), others offer values that, unless calibrated using soil samples or clustered cannot be directly used for soil management practices.

For example, EC_a can delineate high and low conductivity regions within a field; however, unless more information is gathered and used to analyze what is causing the different conductivity levels within the field, very little could be concluded besides the fact that soil spatial variability exists in this field. Soil EC_a can be correlated to moisture, nutrients, elements (including soil salinity), and other chemical and physical properties (Viscarra Rossel et al., 2011). One could determine the reason for the different conductivity regions using historical information (e.g., yield, satellite imagery, communication with the farmer, etc.) or soil samples for this field. Then, this information can be used to guide soil sampling or to make soil fertility management decisions. It must be noted that this process is not trivial and requires specific skills and time investment.

From a farmer's perspective, the investment in data collection should provide a tangible return (i.e., improvement of their farming practices and increase in profits), while collecting an individual PSS dataset cannot fulfill this expectation. The above remark is one of the reasons Erickson & Lowenberg-DeBoer (2023) listed as the cause of US PA practitioners' low adoption of EC_a .

Without any ancillary dataset, PA soil fertility mapping and subsequent fertilizer prescription rates rely on traditional soil mapping techniques, predominately through the interpolation of lab analysis resulting from a grid pattern georeferenced soil sampling strategy.

As highlighted in Chapter 4, depending on the density and number of samples collected from a field, the uncertainty associated with the kriging model estimates can vary, potentially leading to unreliable management decisions. Therefore, if ordinary kriging using traditional variogram fitting procedures is used as the standard interpolation approach in a DSS, under certain conditions (e.g., low and extra-low density sampling designs), the resulting interpolated surface will be highly uncertain or nonexistent. Figure 4.8 supports this as the traditional variogram fitting procedure only converged for K sampled at $0.8 \text{ ha} \cdot \text{sample}^{-1}$.

Remarkably, evaluating the results from the other nine interpolation approaches leads to a similar conclusion as the traditional ordinary kriging; none of the approaches consistently emerged as the best interpolator. Based on this result, one could decide to average the results from the soil samples and recommend a uniform fertilizer rate. As a farmer who invested in georeferenced soil sampling expecting a variable rate fertilizer prescription for a field, using the average would not be an appealing solution.

To evaluate which, if any, conditions the average would emerge as the best estimate, the ten interpolation approaches were compared to the average. None average generated the most accurate surface for the four fields and soil properties evaluated in Chapter 4. Although generalizing these results to other fields should be done with care, it indicates that an interpolator can often generate a surface that outperforms the field average.

A solution could be implementing all interpolation approaches into a DSS and, for every dataset, automatically evaluating which interpolator produces the most accurate surface. Although possible, this would be computationally expensive. Additionally, a separate validation set of samples would have to be collected to validate the surfaces and select the best interpolator, resulting in a higher cost for farmers, or cross-validations should be used to estimate the interpolation errors.

Another solution would be selecting a robust interpolation method (not necessarily the best, but the one that hardly produces results worse than the field average). Based on the robustness analysis in Chapter 4, for an extra-low density sampling design, at least once, all the ten interpolation approaches were either worse than average or the model-based approaches did not

converge or result in a flat variogram. It must be noted that these results are still limited to a dataset from Central Alberta. To better understand the reliability of these ten interpolators, datasets from various regions around the world with different sampling densities should be used to validate these approaches.

Chapter 4 was an essential study for developing this thesis, as the evaluation of different interpolators determined a baseline with the maximum amount of information that could be extracted from the soil samples alone. The next step was to explore if PSS DF could improve these interpolation results, leading to the development of Chapter 5.

The method proposed in Chapter 5 for PSS DF to predict soil chemical properties utilized the framework from Chapter 3 for pre-processing PSS data and defined automated procedures to co-locate the data, optimize ML models hyperparameter, train the model, and perform predictions. Therefore, it is a well-suited method for DSS development. However, in contrast to the findings reported by Pei et al. (2019), where PLSR emerged as the best and second-best model for six and four of the eleven soil properties analyzed, none of the ML models evaluated in Chapter 5 consistently emerged as the best predictor for all soil properties and sampling densities. It is worth reminding the reader that the differences between the predictions from different ML algorithms were not statistically significant for the dataset used in Chapter 5. Still, Figure 5.4 suggests that RF and SVM emerged as the best predictors for three soil properties for the 0.4 and 3.5 sample·ha⁻¹ sampling densities, respectively. This result indicates that the number of samples might affect the choice of the best predicting model. From a practical point of view, the number of samples from a field will differ, and one would have to evaluate multiple ML algorithms using a validation approach to obtain the best results.

Comparing the results from Figures 4.4 and 5.5 for K, P, and pH provides a few interesting insights. First, when the sampling density of 0.4 ha·sample⁻¹ was used for PSS DF calibration (Figure 5.5 a-c), only the result for K outperformed the best interpolation approach (Figure 4.4 a-c). This result could be associated with the spatial structure of the soil properties. Table 4.4 indicates that the variogram model parameter estimates for K presented a shorter range of spatial correlation and a higher nugget than for P and pH. A plausible assumption is that if the sampling data represents the underlying surface well and the property presents a strong spatial structure, the interpolation of the samples alone should be sufficient and provide an accurate estimate. This assumption is further supported by the 3.5 ha·sample⁻¹ sampling density results in Figure 5.5 a-c.

As the number of samples was reduced by 90%, the spatial information present in the sampling dataset can no longer provide a representative understanding of the underlying surface, which resulted in a better performance of the PSS DF than for the interpolation of the samples alone (Figure 5.5 a-c compared to Figure 4.4 g-i). On the other hand, in the comparison between the interpolation of $0.4 \text{ ha} \cdot \text{sample}^{-1}$ (Figure 4.4 a-c) and the PSS DF when the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ was used for calibration (Figure 5.5 a-c), only the results for K outperformed the best interpolation method when using the higher sampling density.

However, the above-described increase in the accuracy of estimated surfaces for the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ comes with an additional cost of collecting all the PSS data used in the fusion process. Except for CI, all PSS data used in Chapter 5 were collected by contractors, allowing for a simple economic analysis. The cost to collect the data from Chapter 5 per area or sample is as follows: CA\$ 76.43 per soil sample (includes sampling, handling, shipping, and lab analysis), CA\$ $21 \cdot \text{ha}^{-1}$ for the passive γ -ray survey, CA\$ $16.06 \cdot \text{ha}^{-1}$ for the EMI survey, CA\$ $32.64 \cdot \text{ha}^{-1}$ for the GPR, and CA\$ $16.96 \cdot \text{ha}^{-1}$ for the CI (estimated based on the number of hours for the data collection).

For the field in Chapter 5 (43-ha), a quick calculation would indicate that the total cost for the PSS survey would be approximately CA\$ 3720 while sampling at $3.5 \text{ ha} \cdot \text{sample}^{-1}$ would cost approximately CA\$ 920. A total cost of CA\$ 4640, compared to approximately CA\$ 8200 for the samples alone if sampled at $0.4 \text{ ha} \cdot \text{sample}^{-1}$, a cost reduction of 43%. No costs associated with the DEM LiDAR survey used in Chapter 5 were accounted for, as this information could be collected using an RTK-enabled GNSS (as in Chapter 6). Back to the comparison of the interpolation of the $0.4 \text{ ha} \cdot \text{sample}^{-1}$ (Figure 4.4 a-c) and the PSS DF calibration using the $3.5 \text{ ha} \cdot \text{sample}^{-1}$ (Figure 5.5 a-c), the considerable reduction in costs for the latter compared to the former might sound very appealing to a PA practitioner, even with the associated increase in the errors. Also, if no significant soil disturbing operations or drastic management changes (e.g., the addition of an irrigation system) happen in this field, an annual collection of the PSS data would not be necessary, which means the PSS surveying costs would be diluted in multiple years.

The PSS surveys conducted in the field from Chapter 5 were all independent field operations. As highlighted in Chapter 2, multi-sensor platforms have been continuously developed. Multi-sensor platforms, combined with the development and adoption of autonomous equipment (Lowenberg-DeBoer et al., 2020), can further reduce the cost of PSS surveying and improve the scalability of the data collection.

The above remarks are promising but should be generalized with care, as these were obtained for a single field in Central Alberta. Future research should focus on evaluating this technique for multiple fields and different combinations of sensors, including a few other methods (e.g., optical).

As for Chapter 5, Chapter 6 also suggested the potential of using PSS DF for delineating soil boundaries through an unsupervised spatial clustering algorithm. While in Chapter 5, a vector co-location procedure was utilized to create the training dataset, and IDW to interpolate all PSS data and generate the continuous surfaces for the prediction, in Chapter 6, a geostatistical approach similar to the one from Vogel et al. (2022) was employed. This geostatistical approach utilized ordinary block kriging to co-locate all the data from the different PSS to a common support (raster). As mentioned in Chapter 2, there are some advantages related to this co-location procedure, one being the possibility of investigating the spatial structure of the data through the variogram parameter estimates, which is why this geostatistical approach is adopted in Chapter 6. As expected, no individual PSS data could differentiate all the different soils within the analyzed field. At the same time, the DF of all PSS indicated the potential of this technique to improve the delineation of soil boundaries.

In general, the results presented in this thesis agree with the previously reported findings in the literature and expand the knowledge in mapping soil characteristics from a PA perspective. However, further research is necessary to understand better and evaluate the interactions of PSS data and soil properties. Questions that should be answered by future research could be: Would the optimization of PSS combinations improve the results instead of the simple ‘blind’ combination of all sensors? Could a model trained with multiple year/fields be generalized for a region? Could methods that improve the sample distribution within a field, such as conditional Latin Hypercube (cLHS), be used to improve the sampling locations based on PSS DF? Would a cLHS-based sampling result in better predictions than for grid sampling?

Also, from a management perspective, it is crucial to understand that site-specific management of agricultural inputs requires the development of thematic maps, as these would present the spatial variability in the field, allowing for determining the input rates. Thus, scenarios such as low-accuracy thematic maps or fields with smaller spatial variability might still indicate that a uniform rate should be adopted instead of a variable rate approach. Future research should also focus on

developing decision mechanisms and streamlining the decision process based on thematic maps, DF, and PSS in which a PA practitioner would be suggested with the most suitable management practice.

Finally, research should further evaluate the use of more complex ML algorithms that account for the spatial component of the data, such as CNN. Models developed using this CNN often provide high generalization capabilities. Therefore, if a representative dataset is used to train these more complex models, the potential exists to eliminate the need for soil sampling.

Chapter 8: Summary and General Conclusions

This thesis successfully explored the most common and latest advances in mapping soil characteristics. Multiple methodologies were developed and evaluated throughout. Chapter 3 proposed a framework for batch-processing PSS data, which presented the potential to improve the quality of PSS data while automating most of the processes.

In Chapter 4, ten interpolation approaches (including a newly proposed method and a modified kriging-based approach) were evaluated. Although none of the approaches always emerged as the best interpolator, the concept of identifying a robust method was introduced. This concept focused on finding methods that would hardly generate results worse than the field average. The modified kriging approach and IDW with a power value of 1 provided the most robust method.

In Chapters 5 and 6, methodologies and the potential of PSS DF to improve the mapping of soil chemical properties and soil type boundaries were evaluated. In both chapters, the PSS fusion methods employed were presented as suitable for their integration with DSS and potential soil mapping improvements compared to traditional methods.

Finally, Chapter 7 presented a general scholarly discussion, which was followed by a broader discussion of the different chapters and how they integrate. The authors also discussed limitations and improvements to the proposed methods and suggested topics for future research.

The cost and complexity of PA-based soil management technologies result in barriers to the growth and expansion of the adoption of PA; the PA research community must address these barriers. DSS development can reduce costs and facilitate data access and analysis from PA technologies. Overall, the results presented in the current thesis focus on defining methodologies suitable for DSS development. The author's interaction with Olds College of Agriculture & Technology through a Mitacs project has led to his interaction with the Digital Ag Team and support during the development of a DSS web platform (<https://hyperlayer.ag/>) where some of the concepts presented in Chapters 3-6 were integrated.

Chapter 9: Contribution to Knowledge

To the author's knowledge, a processing framework for PSS data has not been defined in the literature, nor has the fusion of a dataset with GPR, GC, EMI, γ -ray, CI, and topography been explored. Therefore, the results presented in the current thesis expanded the knowledge on maximizing the information acquired using PSS and understanding the interactions and potential of the DF of this specific set of sensing techniques.

Throughout this thesis, the author intentionally reminds the reader to observe and analyze the results practically. This was done to focus on methods and discussions that would fulfill the overreaching goal of the thesis (i.e., propose and evaluate techniques to be implemented in DSS for the processing and fusion of spatial data used for soil mapping in agriculture). Through this perspective, this thesis contributed to the definition of statistically sound and scientifically appropriate methods that can be utilized in real-life conditions, supporting further development of DSS, especially the expansion of PA.

More specifically, developing a robust interpolation approach that maximizes the value of low-density sampling density provides PA practitioners with a solution when these sampling conditions are encountered. The unique integration procedure of four PSS techniques combined with ML to evaluate the potential to predict and generate thematic maps for soil chemical properties indicates the potential of such a technique. Lastly, comparing soil series maps against several sensors, a novel approach, has proven to be possible and a procedure that can help improve the delineation of soil boundaries. However, this study was developed in a specific agro-climatic environment, and it is recommended that similar studies be pursued elsewhere.

The methods and results presented are expected to inspire others to develop methods that improve agricultural management, guaranteeing that future generations will have access to the environmental functions we enjoy currently.

References

- Adamchuk, V. I., Allred, B., Doolittle, J., Grote, K., & Viscarra Rossel, R. A. (2017). Tools for Proximal Soil Sensing. In C. Ditzler, K. Scheffe, & H. Monger (Eds.), *Soil survey manual, USDA handbook 18* (pp. 355–394). Washington DC: Government Printing Office.
- Adamchuk, V. I., Charles R Hempleman, & Daniel G Jahraus. (2009). On-the-Go Capacitance Sensing of Soil Water Content. In *Mid-Central Conference* (Vol. 0300). St. Joseph, MI: American Society of Agricultural and Biological Engineers. <https://doi.org/10.13031/2013.29481>
- Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44(1), 71–91. <https://doi.org/10.1016/j.compag.2004.03.002>
- Adamchuk, V. I., Lan, J., Abdalla, K., Dias Carlson, P., Debbagh, M., Madramootoo, C., & Kvezereli, B. (2023). Instrumentation for on-the-spot measurement of soil health indicators, 823–829. https://doi.org/10.3920/978-90-8686-947-3_103
- Adamchuk, V. I., Lund, E. D., Sethuramasamyraja, B., Morgan, M. T., Dobermann, A., & Marx, D. B. (2005). Direct measurement of soil chemical properties on-the-go using ion-selective electrodes. *Computers and Electronics in Agriculture*, 48(3), 272–294. <https://doi.org/10.1016/j.compag.2005.05.001>
- Adamchuk, V. I., & Rossel, R. A. V. (2010). Development of On-the-Go Proximal Soil Sensor Systems. In R. A. Viscarra Rossel, A. B. McBratney, & B. Minasny (Eds.), *Proximal Soil Sensing* (pp. 15–28). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-8859-8_2
- Adamchuk, V. I., Viscarra Rossel, R. A., Sudduth, K. A., & Lammers, P. S. (2011). Sensor Fusion for Precision Agriculture. In *Sensor Fusion - Foundation and Applications*. InTech. <https://doi.org/10.5772/19983>
- Adamchuk, V. I., & Wang, C. (2007). Collocating Multiple Self-Generated Data Layers. *GIS Applications in Agriculture*, (1), 185–196. <https://doi.org/10.1201/9781420007718-14>
- Andraski, T. W., Bundy, L. G., & Brye, K. R. (2000). Crop Management and Corn Nitrogen Rate Effects on Nitrate Leaching. *Journal of Environmental Quality*, 29(4), 1095–1103. <https://doi.org/10.2134/jeq2000.00472425002900040009x>

- Ahrends, H. E., & Lajunen, A. (2021). Assessing soil spatial heterogeneity using proximal soil sensing. *Proceedings of IEEE Sensors, 2021-Octob*, 1–4. <https://doi.org/10.1109/SENSORS47087.2021.9639507>
- Akinsunmade, A. (2021). GPR imaging of traffic compaction effects on soil structures. *Acta Geophysica*, 69(2), 643–653. <https://doi.org/10.1007/s11600-020-00530-0>
- Algeo, J., Van Dam, R. L., & Slater, L. (2016). Early-Time GPR: A Method to Monitor Spatial Variations in Soil Water Content during Irrigation in Clay Soils. *Vadose Zone Journal*, 15(11), 1–9. <https://doi.org/10.2136/vzj2016.03.0026>
- Allred, B., Daniels, J. J., & Ehsani, M. R. (2008). *Handbook of agricultural geophysics*. Boca Raton: CRC Press.
- Anastasiou, E., Castrignanò, A., Arvanitis, K., & Fountas, S. (2019). A multi-source data fusion approach to assess spatial-temporal variability and delineate homogeneous zones: A use case in a table grape vineyard in Greece. *Science of the Total Environment*, 684, 155–163. <https://doi.org/10.1016/j.scitotenv.2019.05.324>
- Badewa, E., Unc, A., Cheema, M., Kavanagh, V., & Galagedara, L. (2018). Soil moisture mapping using multi-frequency and multi-coil electromagnetic induction sensors on managed podzols. *Agronomy*, 8(10), 1–16. <https://doi.org/10.3390/agronomy8100224>
- Badua, S., & Sharda, A. (2023). Quantifying real-time opening disk load to assess compaction and potential for planter control. In *Precision agriculture '23* (pp. 43–50). The Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-947-3_3
- Barbedo, J. G. A. (2022). Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps. *Sensors*, 22(6), 2285. <https://doi.org/10.3390/s22062285>
- Behera, S. K., Adamchuk, V. I., Shukla, A. K., Pandey, P. S., Kumar, P., Shukla, V., et al. (2022). The Scope for Using Proximal Soil Sensing by the Farmers of India. *Sustainability*, 14(14), 8561. <https://doi.org/10.3390/su14148561>
- Cao, Y., Yang, W., Li, H., Zhang, H., & Li, M. (2024). Development of a vehicle-mounted soil organic matter detection system based on near-infrared spectroscopy and image information fusion. *Measurement Science and Technology*, 35(4), 045501. <https://doi.org/10.1088/1361-6501/ad179f>
- Castrignanò, A., & Belmonte, A. (2023). *Data Fusion in a Data-Rich Era*. https://doi.org/10.1007/978-3-031-15258-0_7

- Castrignanò, A., Buttafuoco, G., Quarto, R., Parisi, D., Viscarra Rossel, R. A., Terribile, F., et al. (2018). A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture. *Catena*, 167(May), 293–304. <https://doi.org/10.1016/j.catena.2018.05.011>
- Cherubin, M. R., Damian, J. M., Tavares, T. R., Trevisan, R. G., Colaço, A. F., Eitelwein, M. T., et al. (2022). Precision Agriculture in Brazil: The Trajectory of 25 Years of Scientific Research. *Agriculture (Switzerland)*, 12(11), 1–29. <https://doi.org/10.3390/agriculture12111882>
- Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Amado, T. J. C., Simon, D. H., & Damian, J. M. (2015). Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. *Pesquisa Agropecuária Brasileira*, 50(2), 168–177. <https://doi.org/10.1590/S0100-204X2015000200009>
- Chilès, J.-P., & Delfiner, P. (2012). *Geostatistics*. Hoboken, NJ, USA: Wiley. <https://doi.org/10.1002/9781118136188>
- Conway, L. S., Sudduth, K. A., Kitchen, N. R., Anderson, S. H., Veum, K. S., & Myers, D. B. (2022). Soil organic matter prediction with benchtop and implement-mounted optical reflectance sensing approaches. *Soil Science Society of America Journal*, 86(6), 1652–1664. <https://doi.org/10.1002/saj2.20475>
- Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., & Benediktsson, J. A. (2015). Challenges and Opportunities of Multimodality and Data Fusion in Remote Sensing. *Proceedings of the IEEE*, 103(9), 1585–1601. <https://doi.org/10.1109/JPROC.2015.2462751>
- De Benedetto, D., Montemurro, F., & Diacono, M. (2019). Mapping an agricultural field experiment by electromagnetic induction and ground penetrating radar to improve soil water content estimation. *Agronomy*, 9(10). <https://doi.org/10.3390/agronomy9100638>
- Demattê, J. A. M., Dotto, A. C., Bedin, L. G., Sayão, V. M., & Souza, A. B. e. (2019). Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma*, 337(May 2018), 111–121. <https://doi.org/10.1016/j.geoderma.2018.09.010>
- Dhawale, N. M., Adamchuk, V. I., Prasher, S. O., & Rossel, R. A. V. (2021). Evaluating the precision and accuracy of proximal soil vis–nir sensors for estimating soil organic matter and texture. *Soil Systems*, 5(3). <https://doi.org/10.3390/soilsystems5030048>

- Dimitri, C., Effland, A., & Conklin, N. (2005). *The 20th century transformation of U.S. agriculture and farm policy. Economic Information Bulletin* (Vol. 3). <https://ageconsearch.umn.edu/record/59390/>
- Doolittle, J. A. (1987). Using Ground-penetrating Radar to Increase the Quality and Efficiency of Soil Surveys. In W. U. Reybold & G. W. Petersen (Eds.), *Soil Survey Techniques* (pp. 11–32). Madison: Soil Science Society of America, Inc. <https://doi.org/10.2136/sssaspecpub20.c2>
- Dworak, V., Mahns, B., Selbeck, J., Gebbers, R., & Weltzien, C. (2020). Hyperspectral Imaging Tera Hertz System for Soil Analysis: Initial Results. *Sensors*, 20(19), 5660. <https://doi.org/10.3390/s20195660>
- Eitelwein, M. T., Tavares, T. R., Molin, J. P., Trevisan, R. G., de Sousa, R. V., & Demattê, J. A. M. (2022). Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN. *Automation*, 3(1), 116–131. <https://doi.org/10.3390/automation3010006>
- Erickson, B., & Lowenberg-DeBoer, J. (2023). 2023 Precision Agriculture Dealership Survey. Department of Agronomy and Agricultural Economics, Purdue University. https://ag.purdue.edu/digitalag/_media/croplife-purdue-precision-dealer-report-2023.pdf
- FAO. (2009). *How to Feed the World in 2050*. http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf
- Fu, Y., Taneja, P., Lin, S., Ji, W., Adamchuk, V., Daggupati, P., & Biswas, A. (2020). Predicting soil organic matter from cellular phone images under varying soil moisture. *Geoderma*, 361(April 2019), 114020. <https://doi.org/10.1016/j.geoderma.2019.114020>
- Gebbers, R. (2019). Proximal soil surveying and monitoring techniques. In J. V. Stafford (Ed.), *Precision agriculture for sustainability* (1st ed., pp. 29–78). Sawston, Cambridge, UK: Burleigh Dodds Science Publishing. <https://doi.org/10.19103/AS.2017.0032.01>
- Gonçalves, J. R. M. R., Ferraz, G. A. e S., Reynaldo, É. F., Marin, D. B., Ferraz, P. F. P., Pérez-Ruiz, M., et al. (2021). Comparative analysis of soil-sampling methods used in precision agriculture. *Journal of Agricultural Engineering*, 52(1), 83–94. <https://doi.org/10.4081/jae.2021.1117>

- Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press, USA.
- Heggemann, T., Welp, G., Amelung, W., Angst, G., Franz, S. O., Koszinski, S., et al. (2017). Proximal gamma-ray spectrometry for site-independent in situ prediction of soil texture on ten heterogeneous fields in Germany using support vector machines. *Soil and Tillage Research*, 168, 99–109. <https://doi.org/10.1016/j.still.2016.10.008>
- Hubert, M., & Van Der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, 22(3–4), 235–246. <https://doi.org/10.1002/cem.1123>
- International Society for Precision Agriculture (ISPA). (2024). Precision Ag Definition. <https://www.ispag.org/about/definition>. Accessed 5 July 2024
- Jacobsen, J. S., Lorbeer, S. H., Schaff, B. E., & Jones, C. A. (2002). Variation in soil fertility test results from selected northern great plains laboratories. *Communications in Soil Science and Plant Analysis*, 33(3–4), 303–319. <https://doi.org/10.1081/CSS-120002747>
- Ji, W., Adamchuk, V. I., Chen, S., Mat Su, A. S., Ismail, A., Gan, Q., et al. (2019). Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. *Geoderma*, 341(July 2017), 111–128. <https://doi.org/10.1016/j.geoderma.2019.01.006>
- Ji, W., Adamchuk, V., Lauzon, S., Su, Y., & Saifuzzaman, M. (2017). Pre-processing of on-the-go mapping data. In *Pedometrics 2017* (p. 113). Wageningen, Netherlands.
- Journel, A. G., & Huijbregts, C. J. (1976). *Mining geostatistics*. United Kingdom.
- Kassim, A. M., Nawar, S., & Mouazen, A. M. (2021). Potential of on-the-go gamma-ray spectrometry for estimation and management of soil potassium site specifically. *Sustainability (Switzerland)*, 13(2), 1–17. <https://doi.org/10.3390/su13020661>
- Koganti, T., Vigah Adetsu, D., Triantafilis, J., Greve, M. H., & Beucher, A. M. (2023). Mapping peat depth using a portable gamma-ray sensor and terrain attributes. *Geoderma*, 439(July). <https://doi.org/10.1016/j.geoderma.2023.116672>
- Lachgar, A., Mulla, D. J., & Adamchuk, V. (2024). Implementation of Proximal and Remote Soil Sensing, Data Fusion and Machine Learning to Improve Phosphorus Spatial Prediction for Farms in Ontario, Canada. *Agronomy*, 14(4). <https://doi.org/10.3390/agronomy14040693>
- Leroux, C., Jones, H., Clenet, A., Dreux, B., Becu, M., & Tisseyre, B. (2018). A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, 19(5), 789–808. <https://doi.org/10.1007/s11119-017-9555-0>

- Lombardi, F., & Lualdi, M. (2019). Step-frequency ground penetrating radar for agricultural soil morphology characterisation. *Remote Sensing*, 11(9). <https://doi.org/10.3390/rs11091075>
- Lowenberg-DeBoer, J. (2022). *Economics of adoption for digital automated technologies in agriculture. Background paper for The State of Food and Agriculture 2022*. Rome: FAO. <https://doi.org/10.4060/cc2624en>
- Lowenberg-DeBoer, J., & Erickson, B. (2019). Setting the Record Straight on Precision Agriculture Adoption. *Agronomy Journal*, 111(4), 1552–1569. <https://doi.org/10.2134/agronj2018.12.0779>
- Lowenberg-DeBoer, J., Huang, I. Y., Grigoriadis, V., & Blackmore, S. (2020). Economics of robots and automation in field crop production. *Precision Agriculture*, 21(2), 278–299. <https://doi.org/10.1007/s11119-019-09667-5>
- Maldaner, L. F., Molin, J. P., & Spekken, M. (2022). Methodology to filter out outliers in high spatial density data to improve maps reliability. *Scientia Agricola*, 79(1), 1–7. <https://doi.org/10.1590/1678-992x-2020-0178>
- Martini, E., Werban, U., Zacharias, S., Pohle, M., Dietrich, P., & Wollschläger, U. (2017). Repeated electromagnetic induction measurements for mapping soil moisture at the field scale: validation with data from a wireless soil moisture monitoring network. *Hydrology and Earth System Sciences*, 21(1), 495–513. <https://doi.org/10.5194/hess-21-495-2017>
- Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, 57(2), 115–129. <https://doi.org/10.1016/j.inffus.2019.12.001>
- Mitchell, S., Weersink, A., & Erickson, B. (2018). Adoption of precision agriculture technologies in ontario crop production. *Canadian Journal of Plant Science*, 98(6), 1384–1388. <https://doi.org/10.1139/cjps-2017-0342>
- Molin, J. P. (2017). *Agricultura de precisão: números do mercado brasileiro. Agricultura de Precisão: boletim técnico* 3. <http://www.agriculturadeprecisao.org.br/publicacoes/categoria/3>
- Molin, J. P., & Tavares, T. R. (2019). Sensor Systems for Mapping Soil Fertility Attributes: Challenges, Advances, and Perspectives in Brazilian Tropical Soils. *Engenharia Agrícola*, 39(spe), 126–147. <https://doi.org/10.1590/1809-4430-eng.agric.v39nep126-147/2019>
- Pantazi, X. E., Moshou, D., & Bochtis, D. (2020). Utilization of multisensors and data fusion in precision agriculture. In *Intelligent Data Mining and Fusion Systems in Agriculture* (pp. 103–

- 173). Elsevier. <https://doi.org/10.1016/B978-0-12-814391-9.00003-0>
- Pei, X., Sudduth, K. A., Veum, K. S., & Li, M. (2019). Improving in-situ estimation of soil profile properties using a multi-sensor probe. *Sensors (Switzerland)*, 19(5), 1–15. <https://doi.org/10.3390/s19051011>
- Pelletier, B., Dutilleul, P., Larocque, G., & Fyles, J. W. (2009a). Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 1. Estimation of drift and random components. *Environmental and Ecological Statistics*, 16(4), 439–466. <https://doi.org/10.1007/s10651-008-0090-z>
- Pelletier, B., Dutilleul, P., Larocque, G., & Fyles, J. W. (2009b). Coregionalization analysis with a drift for multi-scale assessment of spatial relationships between ecological variables 2. Estimation of correlations and coefficients of determination. *Environmental and Ecological Statistics*, 16(4), 467–494. <https://doi.org/10.1007/s10651-008-0096-6>
- Pentoś, K., Mbah, J. T., Pieczarka, K., Niedbała, G., & Wojciechowski, T. (2022). Evaluation of Multiple Linear Regression and Machine Learning Approaches to Predict Soil Compaction and Shear Stress Based on Electrical Parameters. *Applied Sciences (Switzerland)*, 12(17). <https://doi.org/10.3390/app12178791>
- Rains, G. C., Thomas, D. L., & Vellidis, G. (2001). Soil Sampling Issues for Precision Management of Crop Production. *Applied Engineering in Agriculture*, 17(6), 769–775. <https://doi.org/10.13031/2013.6844>
- Rezaei, A., Karparvarfard, S. H., Naderi-Boldaji, M., Azimi-Nejadian, H., & Tekeste, M. Z. (2022). A new combined penetrometer-dielectric-low frequency acoustic-electrical conductivity sensor for measuring the soil physical characteristics. *Sensors and Actuators A: Physical*, 347(October), 113952. <https://doi.org/10.1016/j.sna.2022.113952>
- Ryazantsev, P. A., Hartemink, A. E., & Bakhmet, O. N. (2022). Delineation and description of soil horizons using ground-penetrating radar for soils under boreal forest in Central Karelia (Russia). *Catena*, 214(January), 106285. <https://doi.org/10.1016/j.catena.2022.106285>
- Schirrmann, M., Gebbers, R., Kramer, E., & Seidel, J. (2011). Soil pH mapping with an on-the-go sensor. *Sensors*, 11(1), 573–598. <https://doi.org/10.3390/s110100573>

- Sheikh, F., Zantah, Y., Abbas, A. A., & Kaiser, T. (2022). See-through Soil Measurements at 300 GHz. In *2022 47th International Conference on Infrared, Millimeter and Terahertz Waves (IRMMW-THz)* (pp. 1–2). Delft, Netherlands: IEEE. <https://doi.org/10.1109/IRMMW-THz50927.2022.9895574>
- Silva, F. C. de S., & Molin, J. P. (2018). On-the-go tropical soil sensing for pH determination using ion-selective electrodes. *Pesquisa Agropecuária Brasileira*, 53(11), 1189–1202. <https://doi.org/10.1590/s0100-204x2018001100001>
- Steele, D. (2017). *Analysis of Precision Agriculture Adoption & Barriers in western Canada*.
- Stepień, M., Gozdowski, D., & Samborski, S. (2013). A case study on the estimation accuracy of soil properties and fertilizer rates for different soil-sampling grids. *Journal of Plant Nutrition and Soil Science*, 176(1), 57–68. <https://doi.org/10.1002/jpln.201100422>
- Sudarsan, B., Ji, W., Biswas, A., & Adamchuk, V. (2016). Microscope-based computer vision to characterize soil texture and soil organic matter. *Biosystems Engineering*, 152, 41–50. <https://doi.org/10.1016/j.biosystemseng.2016.06.006>
- Tavakoli, H., Correa, J., Vogel, S., & Gebbers, R. (2022). RapidMapper – a mobile multi-sensor platform for the assessment of soil fertility in precision agriculture. In *AgEng LAND. TECHNIK 2022* (Vol. 2406, pp. 351–358). Berlin: VDI Verlag. <https://doi.org/10.51202/9783181024065-351>
- Tavares, T. R., Minasny, B., McBratney, A., Cherubin, M. R., Marques, G. T., Ragagnin, M. M., et al. (2023). Estimating plant-available nutrients with XRF sensors: Towards a versatile analysis tool for soil condition assessment. *Geoderma*, 439(August). <https://doi.org/10.1016/j.geoderma.2023.116701>
- Tavares, T. R., Molin, J. P., Hamed Javadi, S., de Carvalho, H. W. P., & Mouazen, A. M. (2021a). Combined use of vis-nir and xrf sensors for tropical soil fertility analysis: Assessing different data fusion approaches. *Sensors (Switzerland)*, 21(1), 1–23. <https://doi.org/10.3390/s21010148>
- Tavares, T. R., Molin, J. P., Nunes, L. C., Wei, M. C. F., Krug, F. J., de Carvalho, H. W. P., & Mouazen, A. M. (2021b). Multi-Sensor Approach for Tropical Soil Fertility Analysis: Comparison of Individual and Combined Performance of VNIR, XRF, and LIBS Spectroscopies. *Agronomy*, 11(6), 1028. <https://doi.org/10.3390/agronomy11061028>
- Tavares, T. R., Mouazen, A. M., Nunes, L. C., dos Santos, F. R., Melquiades, F. L., da Silva, T.

- R., et al. (2022). Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis. *Soil and Tillage Research*, 216(October 2021). <https://doi.org/10.1016/j.still.2021.105250>
- Viscarra Rossel, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., & Lobsey, C. (2011). *Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. Advances in Agronomy* (Vol. 113). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-386473-4.00005-1>
- Viscarra Rossel, R. A., & McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture*, 38(7), 765. <https://doi.org/10.1071/EA97158>
- Vogel, S., Bönecke, E., Kling, C., Kramer, E., Lück, K., Philipp, G., et al. (2022). Direct prediction of site-specific lime requirement of arable fields using the base neutralizing capacity and a multi-sensor platform for on-the-go soil mapping. *Precision Agriculture*, 23(1), 127–149. <https://doi.org/10.1007/s11119-021-09830-x>
- Wadoux, A. M. J. C., Marchant, B. P., & Lark, R. M. (2019). Efficient sampling for geostatistical surveys. *European Journal of Soil Science*, 70(5), 975–989. <https://doi.org/10.1111/ejss.12797>
- Wang, J., Zhao, X., Deuss, K. E., Cohen, D. R., & Triantafilis, J. (2022). Proximal and remote sensor data fusion for 3D imaging of infertile and acidic soil. *Geoderma*, 424(June), 115972. <https://doi.org/10.1016/j.geoderma.2022.115972>
- Webster, R., & Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43(1), 177–192. <https://doi.org/10.1111/j.1365-2389.1992.tb00128.x>
- White, F. E. (1991). Data Fusion Lexicon. Data Fusion Subpanel of the Joint Directors of Laboratories, Technical Panel for C3.
- Xu, Y., Duan, J., Jiang, R., Li, J., & Yang, Z. (2021). Study on the Detection of Soil Water Content Based on the Pulsed Acoustic Wave (PAW) Method. *IEEE Access*, 9, 15731–15743. <https://doi.org/10.1109/ACCESS.2021.3049852>
- Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170(January), 105256. <https://doi.org/10.1016/j.compag.2020.105256>