UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

Cesar de Oliveira Ferreira Silva

Fusão de dados para melhoria da modelagem da produtividade e zonas de manejo em cafés especiais

Data fusion for improving the modeling of yield and management zones in specialty coffees

Campinas
2024

Cesar de Oliveira Ferreira Silva

Data fusion for improving the modeling of yield and management zones in specialty coffees

Fusão de dados para melhoria da modelagem da produtividade e zonas de manejo em cafés especiais

Tese apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Doutor em Engenharia Agrícola, na área de Agricultura Digital.

Thesis presented to the Faculty of Agricultural Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Agricultural Engineering, in the area of Digital Agriculture.

Supervisor/Orientador: Prof. Dr. Stanley Robson de Medeiros Oliveira
Co-supervisor/Coorientador: Prof. Dr. Rodrigo Lilla Manzione

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO CESAR DE OLIVEIRA FERREIRA SILVA E ORIENTADA PELO PROF. DR. STANLEY ROBSON DE MEDEIROS.

Campinas
2024

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Si38d

Silva, Cesar de Oliveira Ferreira, 1993-
Data fusion for improving the modeling of yield and management zones in specialty coffees / Cesar de Oliveira Ferreira Silva. – Campinas, SP : [s.n.], 2024.

Orientador: Stanley Robson de Medeiros Oliveira.
Coorientador: Rodrigo Lilla Manzione.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Café - Cultivo. 2. Agricultura de precisão. 3. Coffea arabica L.. 4. Krigagem. 5. Mapeamento digital. I. Oliveira, Stanley Robson de Medeiros, 1967-. II. Manzione, Rodrigo Lilla. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações Complementares

Título em outro idioma: Fusão de dados para melhoria da modelagem da produtividade e zonas de manejo em cafés especiais
Palavras-chave em inglês:
Coffee - Crop
Precision agriculture
Coffea arabica L.
Kriging
Digital mapping
Área de concentração: Agricultura Digital
Titulação: Doutor em Engenharia Agrícola
Banca examinadora:
Stanley Robson de Medeiros Oliveira [Orientador]
Guilherme Vieira Nunes Ludwig
Priscila Pereira Coltri
Flavia Rodrigues Alves Patrício
Julieta Bramorski
Data de defesa: 12-04-2024
Programa de Pós-Graduação: Engenharia Agrícola

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-5152-6497
- Currículo Lattes do autor: http://lattes.cnpq.br/1913546832402107

Este exemplar corresponde à redação final da **Tese de Doutorado** defendida por **Cesar de Oliveira Ferreira Silva**, aprovada pela Comissão Julgadora em 12 de abril de 2024, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador (FEAGRI/UNICAMP)

Prof. Dr. Guilherme Vieira Nunes Ludwig – Membro Titular (IMECC/UNICAMP)

Dra. Priscila Pereira Coltri – Membro Titular (CEPAGRI/UNICAMP)

Dra. Flávia Rodrigues Alves Patrício –Membro Titular (Instituto Biológico)

Prof. Dra. Julieta Bramorski – Membro Titular (UNIFAP)

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Tese e na Secretaria do Programa da Unidade.

Acknowledgements

To my supervisor, Prof. Dr. *Stanley Robson de Medeiros Oliveira*, for his guidance, trust, and attention to this research.

To my co-supervisor, Prof. Dr. *Rodrigo Lilla Manzione*, for his partnership and companionship since my master's days at UNESP, sharing his practical, and theoretical knowledge of geostatistics, hydrology and life.

To the researchers from Embrapa Digital Agriculture: *Célia Regina Grego*, for her kindness and thoughtfulness in sharing data, knowledge and ideas throughout this research; to *Luciano Vieira Koenigkan* for providing an UAV image for the study area; to *Gustavo Costa Rodrigues* and *Cristina Aparecida Goncalves Rodrigues* for the technical discussions inside this project.

A special thanks to *Josiane Moraes*, the farm manager, and her team, who provided the experimental area and all the necessary support and infrastructure for the execution of the field activities.

I would also like to thank the Consórcio Pesquisa Café for funding the research project, coordinated by Embrapa Digital Agriculture, under the code SEG 10 18 20 01200000.

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Funding Code 001.

My gratitude to the Faculty of Agricultural Engineering (FEAGRI) at Unicamp, for the opportunity to participate in the PhD's program in the period from 2021 to 2024.

My acknowledgments to Embrapa Digital Agriculture for the opportunity to develop part of my PhD's program at its facilities.

Resumo

As zonas homogêneas de manejo (ZHMs) referem-se à subdivisão de um talhão em algumas zonas homogêneas contíguas para orientar aplicações de taxa variável. A delimitação das ZHMs pode ser baseada em abordagens geoestatísticas ou em clusterização. Aqui, ambas as técnicas são usadas conjuntamente. O problema de pesquisa é modelar a dependência espacial dos fatores que influenciam a produtividade de cafés especiais de forma acurada, uma vez que se trata de uma cultura assíncrona que tende a apresentar outliers quando é pouco amostrada. A hipótese da pesquisa é que o uso de técnicas de fusão de dados integradas à geoestatística multivariada fornece modelos capazes de delinear satisfatoriamente ZHMs sobre lavouras de cafés especiais, além de ajudar a aumentar a precisão da interpolação da produtividade. Portanto, o principal objetivo da pesquisa é delinear ZHMs para lavouras de cafés especiais e avaliar se a incorporação de dados fusionados tem um impacto positivo (melhoria) no mapeamento da produtividade do café na área de estudo. Compõem a metodologia duas tarefas principais: (1) comparar diferentes procedimentos para a criação de zonas de manejo e (2) determinar a relação das ZHMs delineadas com i) mapas de produtividade do café e ii) a capacidade de sintetização de cada método com relação às variáveis de entrada dentro das ZHMs delineadas. As técnicas comparadas para resumir os dados espaciais foram: (1) sintetizar as variáveis em um índice de fertilidade do solo (SFI), (2) a técnica MULTISPATI-PCA e (3) a abordagem multivariada por fatores de autocorrelação Min/Max (MAF). Em seguida, foram aplicados métodos de clusterização para realizar a divisão de talhão em ZHMs binárias (agrupando os valores mais baixos e mais altos das variáveis de entrada). A abordagem MAF obteve a melhor divisão de talhão em termos de métricas de agrupamento (teste de McNemar, coeficiente de silhueta e redução de variância). Neste artigo, não usamos a produtividade como uma variável de entrada, mas como uma métrica de avaliação. Neste trabalho, 40 amostras de variáveis químicas do solo selecionadas foram analisadas em conjunto com dados de sensoriamento remoto para o delineamento de ZHMs em dois anos (2022-23). O estudo de caso foi realizado em uma lavoura de café arábica de 4 ha localizada em Minas Gerais, no sudeste do Brasil. O MAF também foi o melhor para diferenciar áreas de baixa e alta produtividade nas ZHMs. Essa abordagem flexível pode orientar o manejo preciso de nutrientes em áreas com baixa amostragem, permitindo o uso conjunto de ferramentas de ciência de dados e conhecimento agronômico para delinear estratégias de aplicação de taxa variável.
Palavras-chave: Café - Cultivo. Agricultura de precisão. *Coffea arabica L.*. Krigagem. Mapeamento digital.

Abstract

Homogeneous management zones (HMZs) are the subdivision of a field into a few contiguous homogeneous zones to guide variable-rate application. Delineating HMZs can be based on geostatistical or clustering approaches. The research problem is to model the spatial dependence of the factors influencing the coffee yield of specialty coffees in an accurate way, since it is an asynchronous crop that tends to have outliers when is low sampled. The research hypothesis is that the use of data fusion techniques integrated with multivariate geostatistics provides models capable of successfully delineating HMZs over specialty coffee crops, as well as helping to increase the accuracy of yield interpolation. Therefore, the main objective of the research is to delineate HMZs for specialty coffee crops and assess whether the incorporation of fused data has a positive impact (improvement) on mapping the coffee yields of the areas under study. Here, both techniques are joint-used. There are two main tasks in the methodology: (1) compare different procedures for creating management zones and (2) determine the relation of the HMZs delineated with i) coffee yield maps and ii) the summarizing power of each method for each input variable inside the HMZs delineated. The techniques compared to summary spatial data were: (1) the soil fertility index (SFI), (2) the MULTISPATI-PCA technique, and (3) the multivariate Min/Max autocorrelation factors (MAF) approach. Then, clustering methods were applied to perform field partition into binary HMZs (grouping lower and higher values of input variables). MAF approach achieved the best field partition in terms of clustering metrics (McNemar's test, Silhouette Score Coefficient, and variance reduction). In this paper we did not use yields as a cluster variable but as a measure of success. In this work, 40 samples of selected soil chemical variables were jointly analyzed with remotely sensed data for HMZ delineation in two years (2022-23). The case study was performed in a 4-ha arabica coffee crop located in Minas Gerais, Southeast Brazil. MAF also was the best one for separating low- from high-yielding areas over the HMZs. This flexible approach can guide precision nutrient management in low sampled areas, allowing the joint use of data science tools and agronomical knowledge to delineate variable rate application strategies.
Keywords: Coffee - Crop. Precision agriculture. *Coffea arabica L.*. Kriging. Digital mapping
.

List of Figures

List of Tables

List of symbols

B	block
$\overline{C}$	point-to-block covariance
C_0	nugget effect
$C+C_0$	sill
$\|B\|$	volume of the block which is called spatial support
Cov	point-to-point covariance
F	binary indicator of fertility status
$H_i(Y)$	Hermite polynomials
M	variance-covariance matrix
N	number of observed values *in-situ*
R	range
Q	orthogonalisation of variance-covariance matrix
y_i	observed value *in-situ* at a position i
$\hat{y}_i$	estimated value at a position i
Y	transformed variable with mean zero and unit standard deviation
Z_i	observed value at location i
$P(B)$	predicted values at a block
λ	(block) kriging weight
Λ	diagonal matrix eigenvalue of matrix B
$\gamma(h)$	experimental variogram
μ	Lagrange multiplier
Ψ_i	Hermite coefficients
Φ	Gaussian anamorphosis transformation
$\mathbf{W}$	row-sum standardized connectivity matrix

List of Acronyms

AI	artificial intelligence
AL	average linkage
B	boron
BCOK	block cokriging
CEC	cation exchange capacity
CEL	centroid linkage
CLA	clustering large applications
COK	cokriging
COL	complete linkage
CONAB	Companhia Nacional de Abastecimento
CPA	coffee precision agriculture
FAC	fuzzy analysis clustering
FCM	fuzzy c-means
FCS	fuzzy c-shells
Fe	iron
FK	factorial kriging
HCL	hard competitive learning
HMZ	homogeneous management zones
ISPAG	International Society of Precision Agriculture
K	potassium
KED	kriging with external drift
KME	k-means
KRME	kriged reduced mean error
KRMSE	kriged reduced mean squared error

LMC	linear coregionalization model
MAF	minimum/maximum autocorrelation factors
MCA	McQuitty
ME	mean error
Mg^{2+}	magnesium
ML	machine learning
ML	median linkage
MN	McNemar's test
Mn	manganese
MSE	mean squared error
Na	sodium
NDVI	normalized difference vegetation index
NG	neural gas
OK	ordinary kriging
OM	organic matter
P	phosphorus
PA	precision agriculture
PAM	partitioning around medoids
S	sulfur
SCAA	Specialty Coffee Association
SFI	Soil fertility index
SKM	spherical k-means
SL	single linkage
SSC	Silhouette Score Coefficient
UAV	unmanned aerial vehicle
UFCL	unsupervised fuzzy competitive learning
V	base saturation
VR	variance reduction
WAR	Ward
WLS	weighted least-squares
Zn	zinc

Contents

Chapter 1

Introduction

This thesis addresses issues related to precision agriculture (PA), particularly coffee precision agriculture (CPA), with a focus on the contributions of geostatistical modeling, under different approaches, to the applicability of information technologies in improving the management of specialty coffee crops.

1.1 Context

Coffee is part of the Rubiaceae family, in the genus *Coffea*, in which more than 90 species have already been described. Of these, around 25 are commercially exploited, of which only four are of significant importance on the world market: *Coffea arabica*, known as arabica coffee; *Coffea canephora*, known as robusta coffee or conilon coffee, and to a lesser extent: *Coffea liberica* and *Coffea dewevrei*, which produce libérica coffee and excelsa coffee, respectively.

Arabica coffee (*Coffea arabica L.*) is the most important species of the genus *Coffea* and accounts for around 70% of the coffee sold worldwide. It is native to the highlands of Ethiopia, formerly Abyssinia, and is currently grown on the American continent, in Africa and Asia. It is a superior quality beverage, with a striking aroma and a sweet taste, and is widely distributed around the world, consumed pure or in blends with other types of coffee (MOREIRA SILVA; ALVES, 2013).

Robusta or conilon coffee is the name used for varieties of the species *Coffea canephora* Pierre ex Froehner. It is native to the lowland forests of Equatorial Africa, in the Congo River basin, and is currently cultivated in some Central and West African countries, Southeast Asia and South America. It is most commonly used in the preparation of blends, in which it is mixed

with arabica coffee and can make up to 30% of the final product. Because it has a higher soluble solids content than arabica coffee and a higher yield after the roasting process, robusta coffee is an essential component of soluble coffees. The *C. canephora* beans have a high caffeine content, are less aromatic and produce a differentiated drink when roasted (VIEIRA, 2017).

Coffee is one of the most important crops of the Brazilian economy. According to PAM/IBGE (2023), Brazil is the major coffee producer of the world and accounts for 35.7% of the world production. The total physical volume of Brazilian coffee exports in 60-kg bags between April 2023 and March 2024 amounted to 42.80 million bags. The average price of each bag of coffee was 204.10 dollars, which made it possible to collect 8.73 billion dollars in foreign exchange revenue, which converted into the country's currency was equivalent to 43.12 billion Brazilian reals at the time. Of this volume exported, 39.16 million bags were green coffee, which corresponded to 91.5% of the overall total purchased by importers, with 32.83 million bags of arabica coffee, which represented 83.8% of the total green coffee sold, as well as 6.32 million bags of robusta+conilon coffee, which corresponded to 16.2% of the total green coffee exported (OIC, 2024).

As coffee is such an important crop in Brazil, it is necessary to study the factors involved in its production to reduce costs and increase yield. Coffee yield is affected by climate (VIEIRA, 2017), the occurrence of pests (ALVES et al., 2009), plant physiology, tillage system, plant density and population (VIEIRA, 2017), slope and topography (GUIMARÃES et al., 1999) and other factors (FERRAZ et al., 2012b). As a result of the diversity of factors that affect coffee yield, uniform field management based on assumed homogeneity of the total area can decrease farmers' profits.

According to the Specialty Coffee Association (SCAA) (SCA, 2021) definitions, specialty coffee "refers to the highest quality green coffee beans roasted to their greatest flavor potential by true craftspeople and then properly brewed to well-established SCAA developed standards." These standards include scoring higher than 80 points on the quality scale and excellent or outstanding quality in fragrance, aroma, flavor, aftertaste, acidity, body, uniformity, balance, clean cup, sweetness, and overall better taste (SCA, 2021). Specialty coffee refers to a modern demand for exceptional quality coffee, both farmed and brewed to a significantly higher than average standard, and is related to the farmers and the brewer in what is known as the third wave of coffee (ALMEIDA; SPERS, 2019). Over the last 20 years, the specialty coffee segment has

seen a significant increase in demand worldwide, averaging 12% per year, while traditional (commodity) coffees have grown at an annual rate of 2% (SILVA et al., 2021).

From this perspective, spatial analysis can maximize the economic returns by making farm management more efficient and the PA principles meet the reality of managing coffee crops. PA is not a single technology but a toolkit from which farmers choose what they need (LOWENBERG-DEBOER; ERICKSON, 2019). Pierce (1999) considered precision agriculture (commonly also known as "precision farming" or "site-specific management") as a win–win solution both for improving crop yield and environmental quality of agriculture. Farmers usually manage their crop productivity and soil fertility using fixed-rate applications, through which fixed amounts of fertilizers, amendments, and water are applied. This approach considers an area homogeneous, disregarding soil, relief, and plant variations. From the farmers' point of view, it is a practical and convenient method to manage the farm, as it simplifies the management process. However, this approach may overestimate the fertilizer input over the crops, causing negative environmental impacts as erosion and lixiviation, while underestimate them in other areas, leading to suboptimal production (LOWENBERG-DEBOER; ERICKSON, 2019).

The site-specific management of coffee crops requires continuous maps of land properties. Consequently, soil sampling is required to better understand the variations across the field for PA applications (TRANGMAR; YOST; UEHARA, 1986; GOOVAERTS, 1999). However, soil properties present a continuum in their spatial variations, it is difficult to categorize soil samples without introducing errors or over-simplifications (CASTRIGNANÒ; BUTTAFUOCO, 2020). Also, soil properties have been traditionally measured with costly soil sampling and laboratory analyses (SILVA; MANZIONE; OLIVEIRA, 2023; SILVA et al., 2024b), the requested information at high spatial and temporal resolution can often be a limitation to adopt PA practices. These maps usually are produced using spatial interpolation, especially by using geostatistical approaches (CASTRIGNANO; BUTTAFUOCO, 2004; CASTRIGNANÒ; BUTTAFUOCO, 2020; SILVA et al., 2024b; SILVA; MANZIONE; OLIVEIRA, 2023).

Geostatistics is defined as a field of statistics focused on analyzing and interpreting the spatial dependence within a certain area, and the difference between the values of a particular property can be expressed as a function of the distance of separation between the sampled points, which is called semivariance modeling (GOOVAERTS, 1997; GOOVAERTS, 1999). It involves the estimation and modeling of spatial correlation taking into account the heterogeneity and spatial variability. To perform a reliable semivariance modeling, some authors recommend

different minimum sample points, from at least 30 pairs of points (YOST; UEHARA; FOX, 1982; LEGENDRE; FORTIN, 1989) to 100–140 sampled points (WEBSTER; OLIVER, 2007).

In general, the better the semivariance model, the better the quality of the kriging interpolation (PEREIRA et al., 2022). The presence of a single influential outlier can distort the variogram estimates (CRESSIE, 1985). An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. An outlier may be due to a variability in the measurement, an indication of novel data, or it may be the result of experimental error (CHILES; DELFINER, 2012). Low-sampled areas may present outliers, that make spatial modeling with ordinary kriging (OK) problematic. Therefore, it is common practice to immediately exclude any outlying observations, assuming that they are influential (DRIEMEIER et al., 2016; SANCHES et al., 2018). This practice ignores that outliers are not necessarily influential observations(KUTNER et al., 2004) and might result from a secondary, either deterministic or random, process of soil spatial variation (LARK, 2000).

Delineation of homogeneous management zones (HMZ) can aid site-specific applications of farm inputs. HMZs are defined as "the sub-regions within the same piece of land showing similar yield influencing factors within which different crop management practices are carried out at the right time and place to optimize crop productivity and minimize adverse environmental impact" (KHOSLA et al., 2002). Variable-rate application of fertilizers can follow the HMZs to improve its management, since fertilizer use efficiency is quite low under uniform crop management practices (BASSO et al., 2011; SU; ZHAO; DONG, 2018). According to Schepers et al. (2004), HMZs should follow the same yield trend across years and different cultivated crops, for this reason HMZ boundaries may change over the time.

Separating the contribution of each soil property from soil fertility is very difficult, for this reason, delineating HMZs is challenging (MOHARANA et al., 2020). Diverse techniques to delineate HMZs are proposed in the literature, mostly based on geostatistical approaches (CASTRIGNANÒ et al., 2018; CASTRIGNANÒ et al., 2009; CASTRIGNANÒ; BUTTAFUOCO, 2020; BUTTAFUOCO et al., 2010; BUTTAFUOCO et al., 2017, 2021; CASTRIGNANÒ et al., 2000; AGGELOPOOULOU et al., 2013) or clustering based approaches (GAVIOLI et al., 2019; MARTÍNEZ-CASASNOVAS; ESCOLÀ; ARNÓ, 2018; GEORGI et al., 2018; OHANA-LEVI et al., 2019; SCUDIERO et al., 2018; ZERAATPISHEH et al., 2020; MIAO; MULLA; ROBERT, 2018). However, joint using these two approaches is not usual. Most of the re-

search focuses on using OK to interpolate variables to use them as input for clustering such as fuzzy k-means cluster analysis (GILI et al., 2017; METWALLY et al., 2019), fuzzy c-means clustering algorithm (BANSOD; PANDEY, 2013; MARTÍNEZ-CASASNOVAS; ESCOLÀ; ARNÓ, 2018), or apply a dimensionality reduction tool over these interpolated maps as the principal components analysis (PCA) (ORTEGA; SANTIBANEZ, 2007; JIANG et al., 2012; METWALLY et al., 2019), spatially-weighting PCA (ARROUAYS et al., 2011; GAVIOLI et al., 2019), or factorial kriging (CASTRIGNANÒ et al., 2018, 2019; CASTRIGNANÒ et al., 2009).

We consider that these approaches can be successful for field partition into HMZs, however, a lack of interpretability and applicability can happen, since the final decision to delineate HMZs is: what is supposed to explain this field partition? HMZs has already being delineated by using the principal components (PCs) scores to perform the clustering analysis based on usual PCA (LI et al., 2007) or spatially-weighted (ARROUAYS et al., 2011; GAVIOLI et al., 2019). Following the idea of soft computing as the combination of diverse methodologies to exploit tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost, we consider delineating HMZs as a challenging to be dealt with combining dimensionality reduction and unsupervised learning (clustering) in different ways, guided by the needs and possibilities on each study case but with the same “toolkit”.

1.2 Research problem

Geostatistics refers to the statistical analysis of phenomena that change in a continuous spatial manner. It can be defined as the tools that study and predict the spatial structure of georeferenced variables. Geostatistics has been widely applied in agricultural science to solve the problem of estimating soil and plant properties in unsampled locations from sample data (GOOVAERTS, 1997; GOOVAERTS, 1999). Measurement techniques hardly work on the same scale as the process of interest. Therefore, some small-scale variability may be lost because sampling at a lower scale is necessary and can rarely be achieved. However, when a field is low-sampled, outliers may occur, making the performance of univariate kriging techniques problematic (CRESSIE, 1985).

In some situations, the information is multivariate: samples are collected from several locations, and several measurements are taken for each one. The tools used in multivariate geosta-

tistical analysis are analogous to those in univariate analysis and include intrinsic hypothesis, covariance, and cokriging (OLIVER; WEBSTER, 2015). In addition, multivariate geostatistical methods are suited for Big Data applications, since this allows for the use of auxiliary datasets for improving interpolation over unsampled areas, especially when dealing with the presence of outliers and irregular grids (SILVA; MANZIONE; OLIVEIRA, 2023). More sophisticated geostatistical models, like cokriging (WEBSTER; OLIVER, 2007), can include auxiliary data.

Big Data are massive volumes of unstructured and structured datasets considered difficult to process, analyze, and manage using traditional data-processing techniques (RHIF et al., 2020; HU et al., 2022). The increasing number of remotely sensed data provided by sensors coupled on orbital satellites at various spatial and temporal resolutions, the number of data generated has grown exponentially, making multispectral imagery for calculating vegetation indices (RHIF et al., 2020; RHIF; ABBES; FARAH, 2019) and SAR satellite imagery (ZHANG et al., 2022; LIU et al., 2022; ROZNIK; BOYD; PORTH, 2022) significant sources of Big Data (HU et al., 2022).

The **research problem** is to model the spatial dependence of the factors influencing the coffee yield of specialty coffees in an accurate way, since it is an asynchronous crop that tends to have outliers when is low sampled.

1.3 Research hypothesis

The **research hypothesis** is that the use of data fusion techniques integrated with multivariate geostatistics provides models capable of successfully delineating management zones over specialty coffee crops, as well as helping increase the accuracy of yield interpolation.

1.4 Main objective

The **main objective** of the research is to delineate management zones for specialty coffee crops and assess whether the incorporation of fused data has a positive impact (improvement) on mapping the coffee yields of the areas under study.

1.5 Specific objectives

- To understand the state-of-art of the use of geostatistics in precision agriculture in the Brazilian scientific community;
- To evaluate the existence of spatial dependence in the yield of specialty coffees when there is a low amount of data (less than 50 samples) and, consequently, outliers;
- To evaluate the potential for combining spectral response via vegetation index and soil chemistry attributes in the process of delineating management zones and interpolating yield;
- To develop a model combining data fusion and geostatistics capable of delineating management zones consistent with the yield map obtained by interpolation.

1.6 Journal articles published from this research

The bibliometric study presented in chapter 2 of this thesis was published in the journal **Precision Agriculture** (with an impact factor of 6.2) under the title *"Exploring 20-year applications of geostatistics in precision agriculture in Brazil: what's next?"* (SILVA; MANZIONE; OLIVEIRA, 2023).

The soil fertility index (SFI) approach for management zones delineation was published in the journal **Smart Agricultural Technology** (with an CiteScore of 2.6) using data from other coffee crop in the same farm from the present thesis. The article title is *"Summarizing soil chemical variables into homogeneous management zones – case study in a specialty coffee crop"* (SILVA et al., 2024c).

The block cokriging interpolation of coffee yield is presented in a detailed manned under the title *"Improving coffee yield interpolation in the presence of outliers using multivariate geostatistics and satellite data"* (SILVA et al., 2024b) in the journal **AgriEngineering** (with an impact factor of 2.8).

The management zones delineation using the three approaches (SFI, MULTISPATI-PCA, and MAF) *under review* the journal **Precision Agriculture** (with an impact factor of 6.2) under the title *"Combining geostatistical and clustering modeling strategies for delineating specialty coffee management zones under low sampling"* (SILVA et al., 2024a).

1.7 Thesis layout

This thesis is structured as follows:

a) Chapter 2 presents an in-depth literature review about specialty coffees, PA, and geostatistics based on bibliometric techniques;

b) Chapter 3 presents the characterization of the study area and datasets, the methodology with data fusion, dimensionality reduction, HMZs delineation by clustering, and its validation.

c) Chapter 4 presents the analysis of the results (sections 4.1 to 4.5) and the discussion connecting it with state-of-art (sections 4.6 to 4.8)

d) Chapter 5 presents the conclusions of this thesis, recommendations, and suggestions for future research.

Chapter 2

Literature review

2.1 Coffee farming in Brazil

Coffee is an exotic plant that has significant socioeconomic importance for Brazil, due to its significant capacity for generating jobs and distributing income in rural areas. The coffee plant was introduced to Brazil in 1727, in the city of Belém, Pará State, brought from French Guiana by Sergeant Major Francisco de Mello Palheta, at the request of the governor of Maranhão and Grão-Pará. Even then, coffee had great commercial value (FRAGA, 1963; DIAS; SILVA, 2015). From then on, with increased demand from consumer markets in Europe and the United States, coffee growing expanded to the south of the country, where it benefited from the climate and soil conducive to growing the Coffea arabica species. Coffee soon became the leading agricultural export product, initially accounting for more than 70% of Brazil's export revenues. Currently, coffee accounts for around 4% of Brazilian agricultural exports, but even with this smaller share in the generation of foreign exchange in the balance of trade, coffee is an extremely important crop both economically and socially.

Total world coffee production for season 2023-2024 has been estimated at 178 million bags. The world crop of *Coffea arabica* will be 102.2 million bags (57.4%) and *Coffea canephora* 75.8 million bags (42.6%) in the period from October 2023 to September 2024 (OIC, 2024). According to data from the Brazilian National Supply Company or Companhia Nacional de Abastecimento (CONAB) (COMPANHIA NACIONAL DE ABASTECIMENTO, 2023), 2023 production of the Brazilian crop is estimated at 62.3 million processed bags of coffee, including Arabica and Conilon (CONAB, 2023). Of these, arabica would be responsible for the production of 48.77 million bags (79%), while robusta would reach 13.53 million bags (21%). The

total area planted with coffee in Brazil is 2,161,942 ha, with 1,759,906 ha (81.4%) of arabica and 402,036 ha (18.6%) of robusta. Brazil's overall average yield for the 2020/2021 harvest was estimated in 33.48 bags ha^{-1}. Arabica productivity was 32.18 bags ha^{-1}, while robusta yield was 24.28 bags ha^{-1} (COMPANHIA NACIONAL DE ABASTECIMENTO, 2023).

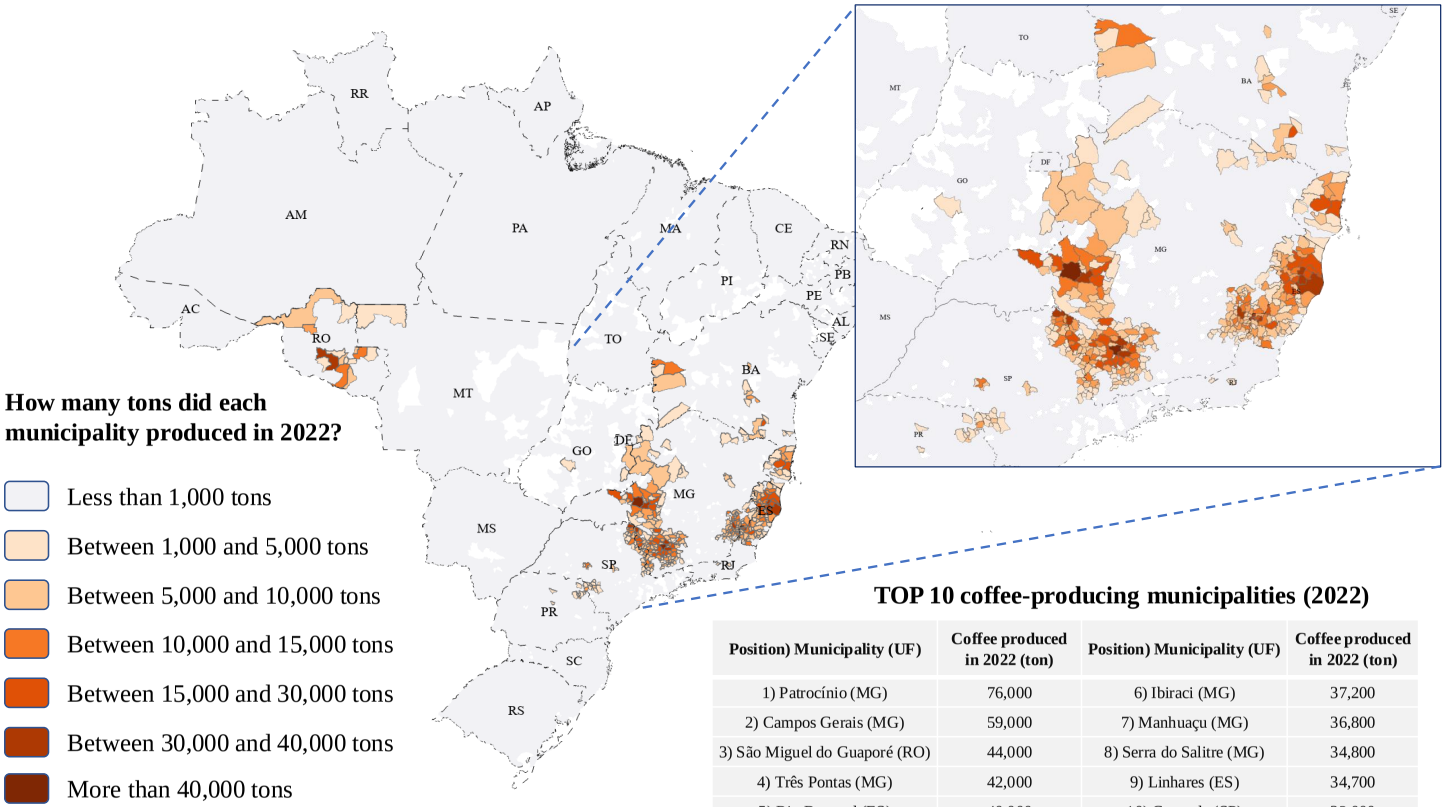
In Brazil, coffee grows in a variety of forms due to the great diversity of climates, altitudes, and types of soil throughout the country (VIEIRA, 2017). Fig. 1 shows the spatial distribution of yearly coffee production in 2022 at Brazilian municipalities (PAM/IBGE, 2023). The state of Minas Gerais is particularly noteworthy, accounting for around 65% of Brazil's production, with the majority of award-winning properties located in the south of the state at altitudes of over 1000 meters, with a predominance of small, semi-mechanized properties or those managed manually (MOREIRA SILVA; ALVES, 2013, p. 20).

Minas Gerais was responsible for 54.9% of national coffee production in 2020, with an estimation of 34.65 million bags (or 60 kg) of processed coffee, corresponding to an increase of 41.1% compared to the previous year's harvest. For Arabica coffee only, Minas Gerais produced 34.34 million bags of processed coffee, which represents 70.4% of Brazil's production of this type of bean. The Matas de Minas region produced 8,589.6 thousand bags of arabica coffee, accounting for approximately 25% of the state's total arabica coffee production. The total area planted with arabica coffee in Minas Gerais is 1,235,477 ha (99%); of these, 1,032,280 ha were occupied by coffee trees in production in 2020. The productivity of this type of coffee was estimated for the 2020/2021 agricultural year in 33.26 bags ha^{-1} in the state (COMPANHIA NACIONAL DE ABASTECIMENTO, 2023).

MapBiomas (2021) mapped land use across the country on an annual basis (1985-2020) and included coffee crops as one of its land use classes. Using geospatial analysis of these mappings, 804,000 hectares of coffee were identified throughout the country. In the state of Minas Gerais, there are 578,000 hectares of coffee, allocated to approximately 102,000 plots, with an average size of 5.5 hectares (MAPBIOMAS, 2021). These plots presented areas ranging from 0.5 to 1453 hectares.

Soil fertility management is crucial to improving coffee yield and quality in the context of specialty coffees, however, uniform management may be unrealistic. For example, Corrêa et al. (1983) proposed nutrient demand amounts based on the expected coffee yield: 4.5 kg N per bag, 0.5 kg P_2O_5 per bag and 4.3 kg of K_2O per bag. This uniform management recommendation is unrealistic as this does not take into account any vegetation variation due to fruit

Figure 1: Coffee-producing municipalities in Brazil according to PAM/IBGE (2023)



TOP 10 coffee-producing municipalities (2022)

Position) Municipality (UF)	Coffee produced in 2022 (ton)	Position) Municipality (UF)	Coffee produced in 2022 (ton)
1) Patrocínio (MG)	76,000	6) Ibiraci (MG)	37,200
2) Campos Gerais (MG)	59,000	7) Manhuaçu (MG)	36,800
3) São Miguel do Guaporé (RO)	44,000	8) Serra do Salitre (MG)	34,800
4) Três Pontas (MG)	42,000	9) Linhares (ES)	34,700
5) Rio Bananal (ES)	40,000	10) Caconde (SP)	32,000

load (SOUZA, 2022). In this sense, including more information on the management decision is crucial. Spatial and temporal variation in soil properties and meteorological conditions may affect coffee growth, grain development, quality, and final yield (ALVES et al., 2009; SILVA et al., 2010; FERRAZ et al., 2017; FERRAZ et al., 2012a; SILVA; LIMA, 2013; FERRAZ et al., 2014; BARROS et al., 2022; ARAÚJO et al., 2017, 2018; LIMA et al., 2016; ANDRADE et al., 2018). To increase farmers' profitability and environmental protection, management practices need to adapt to variable site conditions (FLEMING et al., 2000; COLAÇO; BRAMLEY, 2018). In this manner, specialty coffee management is potentially a precision agriculture application in itself.

2.2 Precision agriculture

The International Society of Precision Agriculture (ISPAG) defined PA as "a management strategy that gathers, processes and analyses temporal, spatial and individual data and combines them with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production" (ISPAG, 2019). According to PA principles, the inherent spatial variability generated by external factors across the croplands needs to be taken into account in such a way that both farmers' profitability and environmental stewardship turn out to be increased (MCBRATNEY et al., 2005; LOWENBERG-DEBOER; ERICKSON, 2019). In this manner, understanding spatially the surface phenomena is fundamental for PA applications.

A relevant topic on PA is the use of HMZs, which are defined as sub-regions of a field and within which the effects on the crop of seasonal differences in weather, soil, management, etc. are expected to be uniform (PERALTA; COSTA, 2013; CASTRIGNANÒ et al., 2018). HMZs delineation is important for the application of PA because farm management decisions are based on it to make decisions on where and how much to apply when scheduling fertilization strategies. For this purpose, it is often useful to define classes from a set of multivariate spatial and temporal data that include properties believed to influence crop yield, such as landscape factors that control water distribution (i.e. elevation and slope) (JACINTHO et al., 2017; DE BENEDETTO et al., 2013), soil physical properties that affect water-holding capacity (i.e. texture and bulk density) (MANZIONE; SILVA; CASTRIGNANÒ, 2020; CID-GARCIA et al., 2013), satellite-based vegetation indices (BASSO et al., 2001; OLDONI et al., 2020) and

soil chemical properties that affect fertility (i.e. pH, electrical conductivity and organic matter) (MOORE et al., 1993; GESSLER et al., 2000; AGGELOPOOULOU et al., 2013; CÓRDOBA et al., 2016).

The delineation of HMZs is based on the spatial variability of the crop in its natural physicochemical characteristics, not taking into account the variability of anthropic actions (such as tillage and fertilization), thus only the effects of this management on the chemical and physical variables sampled can be captured, in other words, in an indirect way (PERALTA; COSTA, 2013; BUTTAFUOCO et al., 2010). The recognition of spatial patterns of soil attributes across the field is essential to the feasibility of applying different and localized management practices in the context of PA. However, the high heterogeneity between different soil chemical variables in the soil makes summarizing it into unique zones a challenge that involves agronomical, mathematical, and computational aspects (CÓRDOBA et al., 2016). Also, developing rational, replicable strategies is crucial for bringing PA to the field (CÓRDOBA et al., 2016).

Synthesizing different chemical variables is a challenge because when adding high variability to a model, the interpretation of the model becomes more complex (GUASTAFERRO et al., 2010). Addressing this challenge leads to a gain of knowledge and applicability of HMZs maps because usually, HMZs have an arbitrary number of zones, decided by the agronomical expert together with the farmer (JACINTHO et al., 2017).

Spatial heterogeneity of soil characteristics is an inevitable problem and represents one of the intrinsic characteristics of soil properties (ELBASIOUNY et al., 2014; SHE et al., 2016). Ignoring the heterogeneity may result in the under application or over application of fertilizers at specific sites (FU; TUNNEY; ZHANG, 2010; BUTTAFUOCO et al., 2010).

Most properties in an agricultural field show spatial dependence at many scales, therefore geostatistics is usually preferred to describe spatial variation (CASTRIGNANO; BUTTAFUOCO, 2004; BOLUWADE; MADRAMOOTOO, 2013; BERNARDI et al., 2017; PARIS et al., 2019). According to the geostatistical paradigm, any soil or crop attribute is considered a random regionalized variable that varies continuously and its variation can be described by a spatial covariance function (MATHERON, 1970). The kriging-based technique can provide the best linear unbiased estimation for the soil properties at unsampled locations (ISAAKS; SRIVASTAVA, 1989; EMADI et al., 2016). Nevertheless, in PA it may be sensible to divide the field into a few practical management zones. To ensure spatial contiguity because of spatially continuous variation (ORTEGA; SANTIBÁÑEZ, 2007; SCUDIERO et al., 2013; CÓRDOBA

et al., 2016), a probability-based approach in a smoothed manner may be used to make HMZs more useful and applicable in the field.

2.3 State-of-art of geostatistics for precision agriculture applications in Brazil

2.3.1 Bibliometric research questions (RQs) for state-of-art analysis

Bibliometric analysis is an approach for analyzing and examining the evolution of literature in a particular field, allowing insights into its articles, authors, subjects, sources, and intra-relationships from a set of documents based on citations rather than content (PALLOTTINO et al., 2018). Bibliometric analysis aptly summarizes the bibliographic materials and provides an efficient quantifiable analysis. According to Donthu et al. (2021), it helps and empowers researchers to get a broad overview, discover knowledge gaps, develop fresh research ideas, position their planned contributions to the field, and promote multidisciplinary research.

In this context, we have attempted to address this bibliometric study by answering the following underlying research questions (RQ):

1. RQ1: What have been the main keywords from peer-reviewed papers on geostatistics for PA applications from Brazilian researchers?

2. RQ2: What are the past perspectives indicated by the peer-reviewed papers on geostatistics for PA applications from Brazilian researchers?

3. RQ3: What are the trends and contexts being followed by the peer-reviewed papers previously analyzed?

We give a historic account of perspectives for geostatistics advances from the 2000s to 2020s connecting it with the use of geostatistics for PA in Brazil. We mention their successes in the past, we identify their merits and weaknesses in the present, and we conjecture on future developments. To address this goal, we have based on the hypothesis that a bibliometric analysis will be able to outline the current state of using geostatistical tools for PA applications in Brazil in the last twenty years (2002-2022) to reveal their current research trends and hotspots.

2.3.2 Bibliometric state-of-art analysis

For the bibliometric analysis with Scopus-indexed peer-reviewed journal articles, we searched for “Precision Agriculture” together with “Geostatistics” under the heading ‘Article title, Abstract, Keywords’ in the Scopus database in early April 2023. 144 peer-reviewed documents were identified. The actual search coding was “(TITLE-ABS-KEY (“precision agriculture”) AND TITLE-ABS-KEY (“geostatistics”)) AND PUBYEAR $>$ 2001 AND PUBYEAR $<$ 2023 AND (LIMIT-TO (AFFILCOUNTRY , “Brazil”)) AND (LIMIT-TO (DOCTYPE , “ar”))”. As a comparison, the same search was performed using “machine learning”, “data science”, and “artificial intelligence” instead of “geostatistics”.

We have chosen this search coding after testing using "Brazil" as a keyword instead of the authors’ country, where we found that the amount of articles is smaller (only 52) and all of them can be captured using the author’s country. The reason is the absence of mentioning the country of the study area in the title nor the abstract or keywords in most of the articles (92 articles).

In terms of descriptive analysis, the study analyzed the author’s keywords in perspective of proceedings papers and peer-reviewed journal articles using word cloud, relevance accounting, and co-occurrence networks.

We used keyword analysis tools to identify the most frequently used author’s keywords. The author’s keywords represent the theme of the research articles and provide a direction to scale the issues involved in the articles (COMERIO; STROZZI, 2019). The co-occurrence network is represented through nodes and links, where the node’s size represents the keyword’s degree of connectiveness with other keywords. For reducing redundancy in these analyses, the terms “geostatistics”, “kriging”, “precision agriculture”, and “brazil” were excluded. Here we unify the terms “semivariogram” and “variogram” for the convenience of organizing the keywords and also because the use of these terms reflects the confusing situation in geostatistical literature (BACHMAIER; BACKES, 2008). Some authors write “variogram” (WACKERNAGEL, 2003), while others write “semivariogram” (OLEA, 1991; CRESSIE, 2015) considering that the “semivariogram” is half the variance of the difference of two random variables, which is the actual “variogram” (ISAAKS; SRIVASTAVA, 1989). Another correct alternative term to "variogram" could be “semivariance", which is quite unusual in agronomical studies.

Among the 144 articles analyzed, 125 recorded a total of 1,479 citations, while the others did not record any citations. All articles have either PhD or MSc as first authors. 93.75% of the corresponding authors were the first author based on a Brazilian institution (135 articles). When

performing the same search using the terms “machine learning”, “data science”, and “artificial intelligence” instead of “geostatistics”, 61, 16, and 23 documents were returned, respectively.

Fig. 2 shows the most cited journals regarding the 144 articles of geostatistics in PA in Brazil retrieved between 2002-2022. “Revista Brasileira de Ciência Solo” (“Brazilian Journal of Soil Science”, in English), “Geoderma”, “Soil Science Society of America Journal”, “Precision Agriculture”, and “Engenharia Agrícola” (“Agricultural Engineering”, in English) contain the articles with the highest number of citations.

Figure 2: Most local cited journals regarding articles of geostatistics in PA in Brazil (2002-2022).

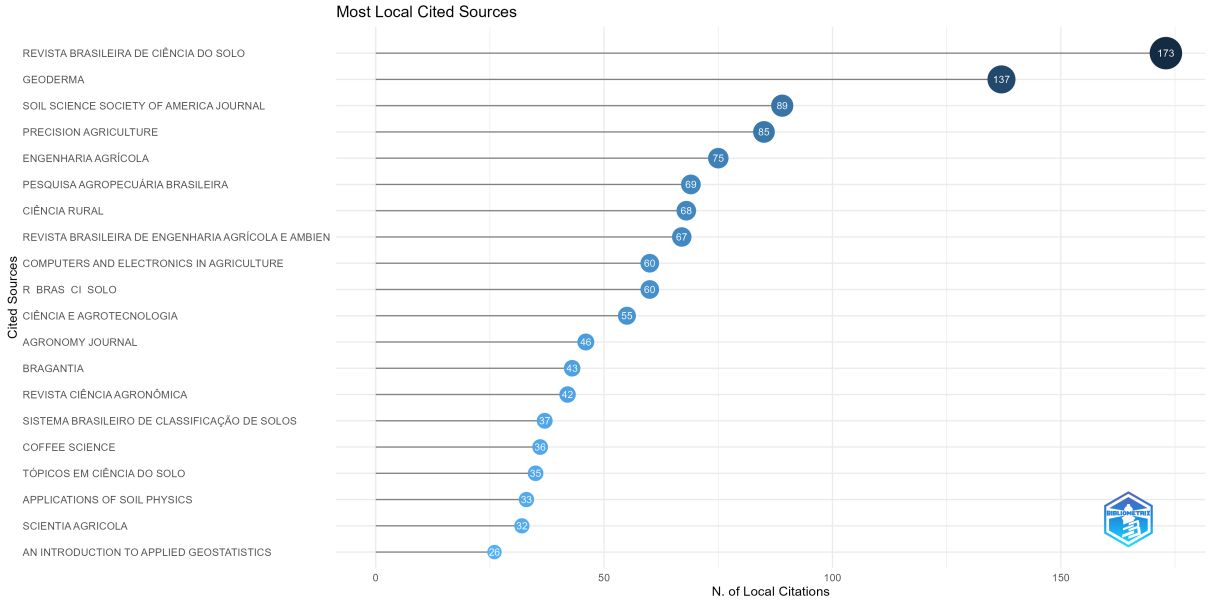


Fig. 3 shows the word cloud of keywords listing the most relevant keywords plot from 144 peer-reviewed articles on geostatistics for PA. Word Cloud is a visual representation of keywords giving greater prominence to the words that appear most frequently. The number of keywords was restricted to 25. We considered “precision agriculture”, “geostatistics”, and "brazil” as redundant. “Spatial variability”, “ordinary kriging”, “soil management”, “semivariogram”, and “soil fertility” are the most frequent keywords. The Trend Topics plot from 20-year peer-reviewed journal articles is shown in Fig. 4, revealing the predominance over time of the most relevant keywords.

The co-occurrence of the author’s keywords plot from articles is shown in Fig. 5. We have chosen five as the minimum number of keyword occurrences to enable the presentation of 25 keywords on the map (see Fig. 5).

Figure 3: Word cloud of author's keywords plot from 20-year peer-reviewed journal articles of geostatistics in PA in Brazil (2002-2022). Terms "geostatistics", "kriging", "precision agriculture", "brazil" were considered redundant and were excluded.

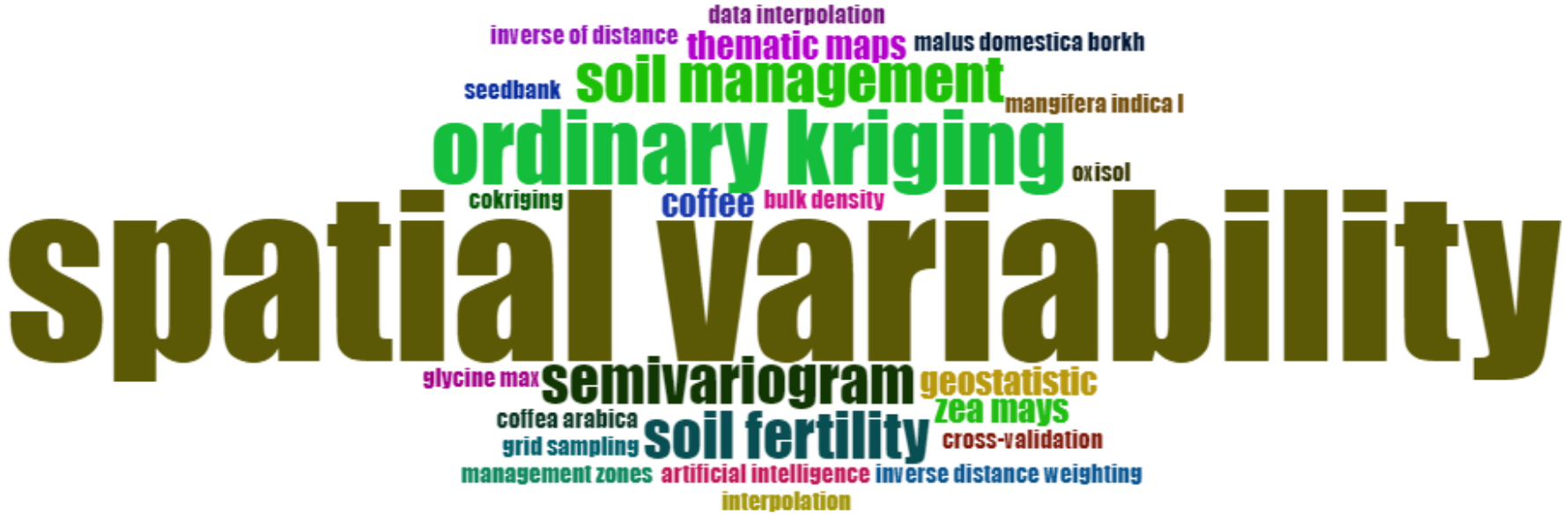
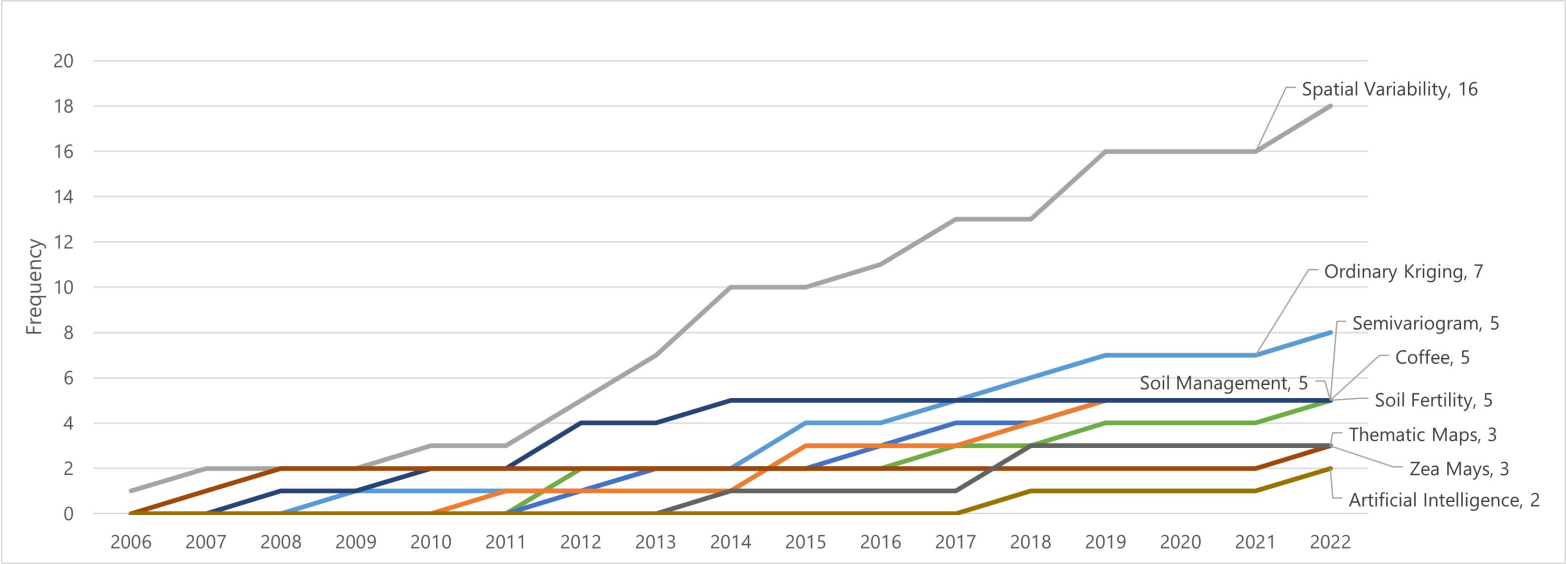


Figure 4: Words' frequency over time from peer-reviewed journal articles about geostatistics in PA in Brazil. The focused period for this analysis is the second part (2007-2022).



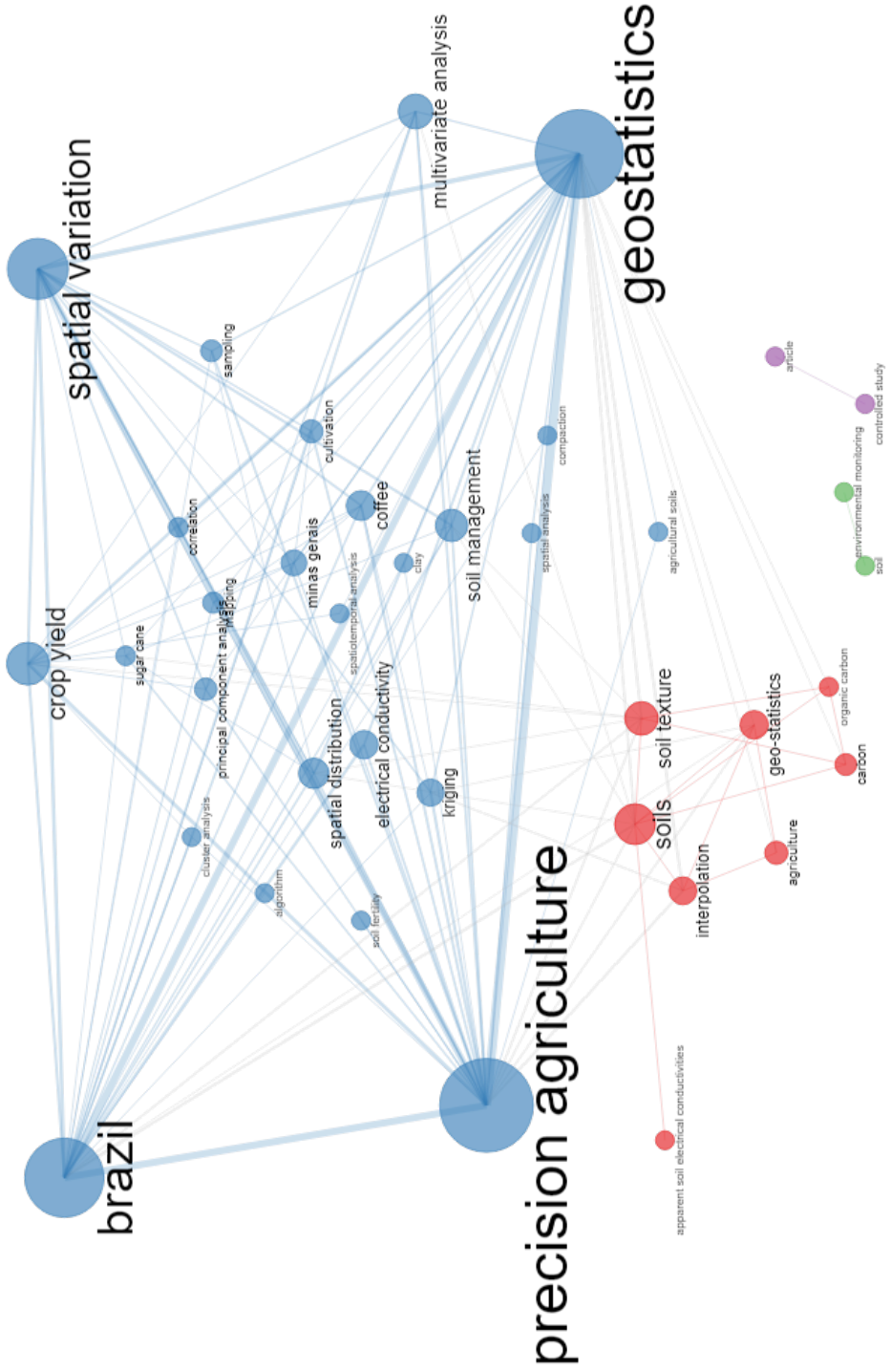


Figure 5: Co-occurrence of author's keywords network from 20-year peer-reviewed journal articles of geostatistics in PA in Brazil (2002-2022)

The first research question (RQ1) aimed to introspect the pattern of main keywords from 144 peer-reviewed papers on geostatistics for PA applications from Brazil-based researchers. Besides redundant terms, the most relevant ones were ‘spatial variability”, “ordinary kriging”, and “soil management”. There is notably a similar predominance of interpolation and mapping studies based on OK for PA. Multivariate geostatistics, data fusion, and disruptive combinations of ML with geostatistics had no participation.

The second research question (RQ2) dealt with the past perspectives detected. The successes of geostatistics for PA applications from Brazil-based researchers were their predominance over other methodologies (machine learning (ML), data science, or artificial intelligence (AI)) while their main weaknesses in the present are the lack of innovation since most of the studies present the same basis (univariate OK for soil variables interpolation) over the last 20 years.

The third research question (RQ3) addressed the associations we can make by combining the insights from the previous research questions. Summing up, studies present similar trends in terms of what kind of geostatistical methodology they are using: OK for interpolation of univariate data. Based on RQ3, we conjecture on future developments, especially regarding geostatistics for PA in Brazil, more exploration on data fusion and multivariate geostatistics is highly recommended.

The next sections of this chapter presents the conjectures developed from the RQ3.

2.4 Geostatistics theory

From the bibliometric analysis, it is notable there is a predominance of interpolation and mapping studies based on OK for PA on both undergraduate and graduate studies detected by analyzing proceeding papers and peer-reviewed articles, respectively. This predominant approach considers only a little portion of the potential approaches under the geostatistical umbrella. There is a wide range of geostatistical methods. For univariate spatial modeling, the most usual method is OK, while for multivariate spatial modeling, the most usual methods are cokriging (COK) (WEBSTER; OLIVER, 2007), and kriging with external drift (KED) (XU et al., 1992).

Geostatistics is widely used for PA applications because of the need for spatially continuous data of agricultural variables (BASSO et al., 2001; CASTRIGNANO; BUTTAFUOCO, 2004). Geostatistics is defined as a field of statistics focused on analyzing and interpreting the spatial dependence (i.e. spatial covariance) within a certain area, and the difference between the values

of a particular property can be expressed as a function of the distance and direction in 2D of separation between the sampled points, which is called semivariance or variogram modeling (GOOVAERTS, 1997; GOOVAERTS, 1999). It involves the estimation and modeling of spatial correlation taking into account the heterogeneity and spatial variability.

Crop field and soil is a continuum that generally follows Tobler's first law of Geography, namely "Everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). This is supported by several studies that have shown that the variability of soil properties is spatially dependent within a certain area, and the difference between the values of a particular property can be expressed as a function of the distance of separation between the sampled points (JUANG; LEE, 2000; LLOYD; ATKINSON, 2001; CASTRIGNANO; BUTTAFUOCO, 2004; CRESSIE, 2015; WEBSTER; OLIVER, 2007; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009; SILVA; LIMA, 2012; ELBASIOUNY et al., 2014; OLIVER; WEBSTER, 2015; EMADI et al., 2016; CASTRIGNANÒ et al., 2018). Notably, this spatial dependence presents uncertainties (MANZIONE; CASTRIGNANÒ, 2019). For this reason, this spatial dependence could be not stationary, and hence change from "place to place" due to external forcing even in relatively small areas or into zones inside the study area (GOOVAERTS, 1998, 2001; BOGUNOVIC et al., 2018).

Choosing the most suitable method for a given dataset has been summarized in Fig. 6. Notably, OK is not the only geostatistical option available. Kriging methods rely on the notion of spatial autocorrelation while ML and other types of modeling do not explicitly (splines, triangulated irregular network, TIN, natural neighbor and inverse distance weighting, IDW). Autocorrelation is a function of distance. In classical statistics, observations are assumed independent, that is, there is no correlation between observations.

In geostatistics, the information on spatial locations allows you to compute distances between observations and to model autocorrelation as a function dependent on the distance and direction of separation in 2D. For this reason, most of the papers use the keywords "spatial variability" and "spatial dependence" to name the results from the kriging analysis. In addition, data transformation is highly recommended for geostatistics analysis. According to Oliver (2015), the geostatistical analysis does not require data to follow a normal distribution. However, variograms comprise sequences of variances, and these can be unstable where data are strongly skewed and contain outliers. If data do not have a near-normal distribution and have a skewness coefficient outside the limits ± 1, because of a long tail, data transformation should

be considered in this case. Unfortunately, data transformation was rarely used in the collected papers (14 articles).

Also, it is important to highlight geostatistics analysis needs well-structured variograms which are not always achieved, even when some spatial autocorrelation is found. A variogram measures the mean degree of dissimilarity (semivariance) between samples separated by a given distance and directions, and thus can describe autocorrelation among observations at specified distances (OLIVER; WEBSTER, 2015).

In general, the semivariance increases with increasing distance between observations until an upper bound is achieved. The value of semivariance at which the variogram plateaus is called the sill, the distance at which the sill is reached is called range, and the semivariance at zero distance is called nugget. The nugget is an estimate of the residual error or spatially uncorrelated error. The observations located within the range are spatially dependent or autocorrelated, while observations away from this distance are not. Sill, nugget, and range are crucial parameters to be met before applying any type of geostatistical method.

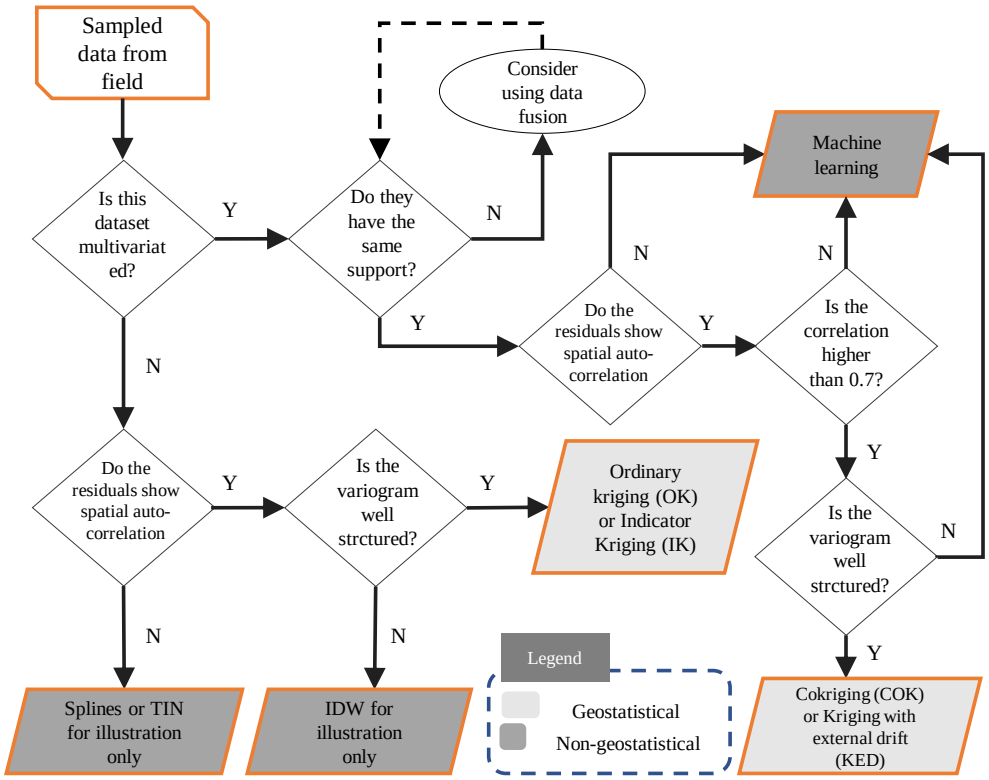
Fig. 6 highlights the need to take into account if the dataset is multivariate or not since bi-or-multivariate methods (COK and KDE) are available. Also, the support of data needs to be checked, and applying data fusion should be considered. Fig. 6 outlines a basic initial framework for strategizing applications of geostatistics for spatial analysis. Applications that also use explicit temporal components are possible (MANZIONE et al., 2019; TAKAFUJI; ROCHA; MANZIONE, 2020, 2019; VAROUCHAKIS et al., 2021; BIVAND et al., 2008; SEKULIĆ et al., 2020; GRÄLER; PEBESMA; HEUVELINK, 2016; HENGL et al., 2012; HIEMSTRA et al., 2009; DE IACO; POSA, 2016; DE IACO; MYERS; POSA, 2002) but are not covered by this flowchart.

2.5 Moving from interpolation to uncertainty modeling

Interpolation (spatial prediction) is the process of estimating a target variable at unsampled locations and can be realized by applying a wide range of models. For an unsampled point in a spatial position, the closer it is to the sampled point, the more likely the attribute value is similar, and this is the most basic assumption of spatial interpolation methods.

Reliably performing a semivariance analysis is highly dependent on the sampling data available. Yost et al. (1982) and Legendre (1989) recommend at least 30 pairs of points to perform

Figure 6: Flowchart for choosing the most suitable method for spatially interpolating a given dataset



a reliable semivariance modeling, while Webster (2007) sets 100–140 sampled points as a minimum sample amount. Regarding the bibliometric analysis, Table 1 summarizes the sample number ranges on which the articles based their geostatistical analysis.

Over the 144 peer-reviewed articles retrived by bibliometric analysis, sampling was done under regular (80 articles), quasi-regular (35 articles) or irregular (29 articles) grids. Notably, most of the research was based on low-sampled studies. This is potentially a weakness because empirically semivariance seems to be better structured on normal or quasi-normal datasets without outliers. Little datasets are less able to well-capture heterogeneity, especially when the difference between the average and extreme values are higher than the standard deviation, trending to be a sparse dataset. Nevertheless, geostatistical methods still achieve reliable results even under low sampling, which is not achieved in machine learning methods, which require datasets with at least hundreds of points (HASTIE et al., 2009).

Considering the amount of research with a small number of sample points, the occurrence of data sets with very low or no spatial autocorrelation can occur, but even in these cases attempts are made to obtain maps from these samplings, hence splines, TIN or IDW have been used in these cases (BERNARDI et al., 2017; BERNARDI et al., 2018). This approach is not justified

Table 1: Ranges of sample amounts from the collected peer-reviewed papers considering its spacing and the study area size

Number of sample	Spacing (m)	Average area (ha)	Number of articles
≤ 50	0-5	1-5	24
	5-10	1-5	4
		5-10	8
	10-25	5-10	6
51-100	0-5	1-5	30
	5-10	5-20	8
		20-50	12
	10-50	20-100	5
101-200	0-5	1-20	18
	5-10	20-100	12
	10-35	100-500	4
201-1000	5-10	100-500	11
≥ 1000	0-10	200-250	2

in terms of spatial statistics strictly, since there is no effective gain of information with this interpolation based on weights without spatial correlation. In practical terms, this approach just creates a poor visualization without spatial dependence.

The challenge of modeling for PA, especially under low-sampled situations, is an opportunity for moving from interpolation studies to uncertainty-based analysis. Uncertainty modeling is a sophisticated statistical approach to data analytics that enables managers to identify key parameters associated with data generation to reduce the uncertainty around the predictive value of that data. A simple interpolation of such relatively sparse spatial data always involves large uncertainties (FOUEDJIO; KLUMP, 2019). Assessing the uncertainty around the predictive value at target locations, and incorporating this assessment to support decision-making is becoming increasingly important (AGGELOPOOULOU et al., 2013; CASTRIGNANÒ et al., 2017; COULSTON et al., 2016; VAYSSE; LAGACHERIE, 2017), however, still very little attention has been paid to their ability to provide reliable prediction uncertainties for PA applications in Brazil.

Uncertainty can be modeled by a probability distribution of an unknown value based on the available related information (MANZIONE; CASTRIGNANÒ, 2019). Geostatistical models have been used widely because of their capacity to provide unbiased estimation of a spatial variable and its uncertainty involved (MANZIONE; CASTRIGNANÒ, 2019; BUTTAFUOCO

et al., 2021; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009), even though technically complex and based on strict assumptions. One advantage of using geostatistics for heterogeneous datasets is the possibility to treat different support data inside a common prediction model (CASTRIGNANÒ; BUTTAFUOCO, 2020). The challenge is to apply the integrated approach to a variety of practical situations, in which it is required to combine spatial data in many forms, from sparse sampling to exhaustive remote/proximal sensing, but crucially providing a measure of prediction uncertainty (MANZIONE et al., 2019).

2.6 Data fusion

To increase the effectiveness of PA in the field it is necessary to improve the accuracy of mapping, interpolation, and estimating the relevant agronomic variables, which could be achieved by intensifying sampling. However, due to the high costs, sampling can never be exhaustive and so a large portion of the field will remain unexplored. For example, Buttafuoco et al. (2017) argue that only using direct soil sampling cannot perform an effective delineation of a field into management zones at the scale required by PA. In this way, combining datasets with different sampling configurations and different supports can be useful (ROSSEL et al., 2011; CASTRIGNANÒ et al., 2017; CASTRIGNANÒ et al., 2018; RODRIGUES et al., 2021). However, the relationship between these different datasets often is not direct nor linear, requiring complex manipulations (CASTRIGNANÒ et al., 2017; CASTRIGNANÒ et al., 2018; RODRIGUES et al., 2021).

In this context, data fusion may be crucial for improving mapping for PA. Data fusion refers to the set of algorithms, processes, and protocols that combine different datasets into a single model that provides complementary views of the same phenomenon. Correlating and fusing information from multiple sources allows more accurate and complete inferences than those that are derived from any single source alone (HALL; MCMULLEN, 2004). Rigorously, data fusion may assume different meanings such as information fusion, sensor fusion, or image fusion. Information fusion is the process of merging information from different sources (ROGOVA; NIMIER, 2004); sensor fusion is the combination of data from different sensors (SASIADEK, 2002) and image fusion is the fusion of two or more images into one, which should be a more useful image (ZHANG, 2004).

Regarding the sensor data fusion, multi-source data can be collected for a variety of spatial scales more often than one of interest and most environmental variables display spatial structures, such as gradients, patches, trends, and complex spatial autocorrelation, which may exist at many scales (ANASTASIOU et al., 2019; CAO; YOO; WANG, 2014; CHANG; BAI, 2018; MANZIONE; CASTRIGNANÒ, 2019; CASTRIGNANÒ et al., 2018; CASTRIGNANÒ et al., 2017, 2021, 2020; MANZIONE; SILVA; CASTRIGNANÒ, 2020; RODRIGUES et al., 2021).

Moreover, the assessment of spatial variations strongly depends on the size of the sampling unit or measurement unit of the sensor. Therefore, the size of the sampling/measurement unit is quite important in the process of investigation because it can critically influence our perception of environmental phenomena (CASTRIGNANÒ et al., 2021; RODRIGUES et al., 2021). Summing up, together with the opportunity of using data from more than one source, there is a challenge of dealing with the difference between them looking for gaining information while understanding the uncertainty involved when combining multi-sensor data. For this reason, sensor data fusion and support analysis are an always-together challenge for applying it in PA.

Data fusion is still little used by Brazil-based researchers in PA. Only one peer-reviewed article dealing with this subject (VASQUES et al., 2020) discuss this topic using multiple linear regression of kriged maps in a case study located in Brazil, the other one presented a case study using factorial kriging (FK) but from a field from Italy (RODRIGUES et al., 2021). FK is a widely used geostatistical approach for fusing multivariate data (BOCCHI et al., 2000; MA et al., 2014; LV et al., 2013; GOOVAERTS, 1992), but still not widely used in Brazil. FK utilizes linear coregionalization model (LMC) fitting, i.e. all experimental simple and cross-variograms are modeled with a linear combination of basic variogram functions to examine the spatial relationship of given variables at multiple scales, and provides a partition of the total variation into various spatial components separately. This partition is provided by principal component analysis of LMC matrices. These components are named “regionalized factors”, and reflect the main features of the multivariate data for each spatial scale and whose scores are interpolated by cokriging. These factors (or loading coefficients) correspond to the covariances between the variables and principal components. Usually, the cokriged regionalized factors are related to short and long-range variation. Regionalized factors can be used as a field partition to guide site-specific management (BUTTAFUOCO et al., 2021, 2017; CASTRIGNANÒ et al., 2018; CASTRIGNANÒ et al., 2017; BUTTAFUOCO et al., 2015; BUTTAFUOCO et al., 2010).

2.7 The problem of change of support

The need for more and better information in PA can be met by wider use of multi-sensor platforms, both ground- and air-based, effectively integrated (BUTTAFUOCO et al., 2017). Furthermore, the data may have different spatial resolutions (support sizes), shapes, and configurations, and their combination results in the problem of change of support (WEBSTER, 1991; RIVOIRARD, 1994; CRESSIE, 1996; GELFAND; ZHU; CARLIN, 2001; EMERY, 2007). Not considering the problem of change of support when combining data from different sources might result in misdiagnosis and misclassification (EMERY, 2007).

Merging data from different sources and sensors are often done by using geographic information systems (GIS) software like ArcGIS (ESRI, 2022) and QGIS (QGIS DEVELOPMENT TEAM, 2023). For example, this software performs union, intersection, zonal averaging, and pixel-by-pixel computations between two raster images with different supports. Despite being fast and scalable, these operations do not treat the change of support problem. Consequently, there is ambiguity about the support of the output and there is no measure of uncertainty associated with input or prediction (NGUYEN et al., 2014).

One of the possible approaches to the problem of change of support from a geostatistical point of view is the block cokriging (BCOK) (BURGESS; WEBSTER, 1980; CRESSIE, 2006; CHILÈS; DELFINER, 2012). BCOK is a kriging method in which the average expected value in an area around an unsampled point is generated rather than the estimated exact value of an unsampled point. BCOK can be used to provide better variance estimates and smooth interpolated results (BORÉM et al., 2021) to effectively upscale point-scale observations to the "block" scale.

Often, a data transformation is needed before applying BK. The Gaussian Anamorphosis approach has been widely used for it (CHILÈS; DELFINER, 2012; BUTTAFUOCO et al., 2017; CASTRIGNANÒ et al., 2020). In this method, each variable is transformed into a normally distributed variable beforehand if it shows a large departure from the Gaussian distribution. If the predictions are produced on a block instead of a point support, a coefficient is needed to obtain an anamorphosis on a block support (CHILÈS; DELFINER, 2012), which will be later used to back-transform block Gaussian estimates into the original distribution from raw data. A support correction coefficient r is determined from the variance of blocks, and the punctual variance is calculated as the sill of the variogram assumed stationary. The variogram based on a point support is calculated on the smallest support, whereas a variogram on a given block support

requires a process of regularization, consisting of discretizing the blocks into equal cells after which a pseudo-experimental variogram is calculated in the fictitious cell centers, and then the point variograms are averaged over the block. Summing up, the BCOK can be used to predict the average values of a multi-variate process at a larger scale, accounting for the size, shape, and orientation of the blocks (CASTRIGNANÒ et al., 2017; RODRIGUES et al., 2021).

Support change and uncertainty go together and special attention is needed when applying BCOK (KANG et al., 2016). Following Tobler's first law of geography, sampling points located closer to a predicted location contribute more strongly than more remote points. However, if the nearer samples have larger uncertainties, samples with smaller uncertainties but at greater distances from the prediction block may be more beneficial for spatial inference (CRESSIE, 2015). Researchers have been working on decomposing the kriging coefficient matrix to incorporate uncertainty. Among the first ones in this task, Watson et al.(1984) modified the observation covariance matrix by adding heterogeneous error terms to the diagonal. This modification was used to predict groundwater levels (SAVELYEVA et al., 2010) and soil salinity (HAMZEHPOUR et al., 2013). However, Christensen et al. (2011) proposed another modification, based on the "right-hand side" vector of the kriging equations using the error terms arguing the modification proposed by Watson et al.(1984) is appropriate for point-to-point estimation, but they cannot be used to upscale a spatial variable to the pixel scale in remote sensing, which can be an example of "block" scale. In this respect, Kang et al. (2016) proposed a BCOK modification using homogeneous or heterogeneous error terms to upscale soil moisture data for interpolation, achieving a decrease in prediction uncertainty by considering heterogeneous error terms.

Chapter 3

Material and Methods

3.1 Description of the study site and agronomic practices

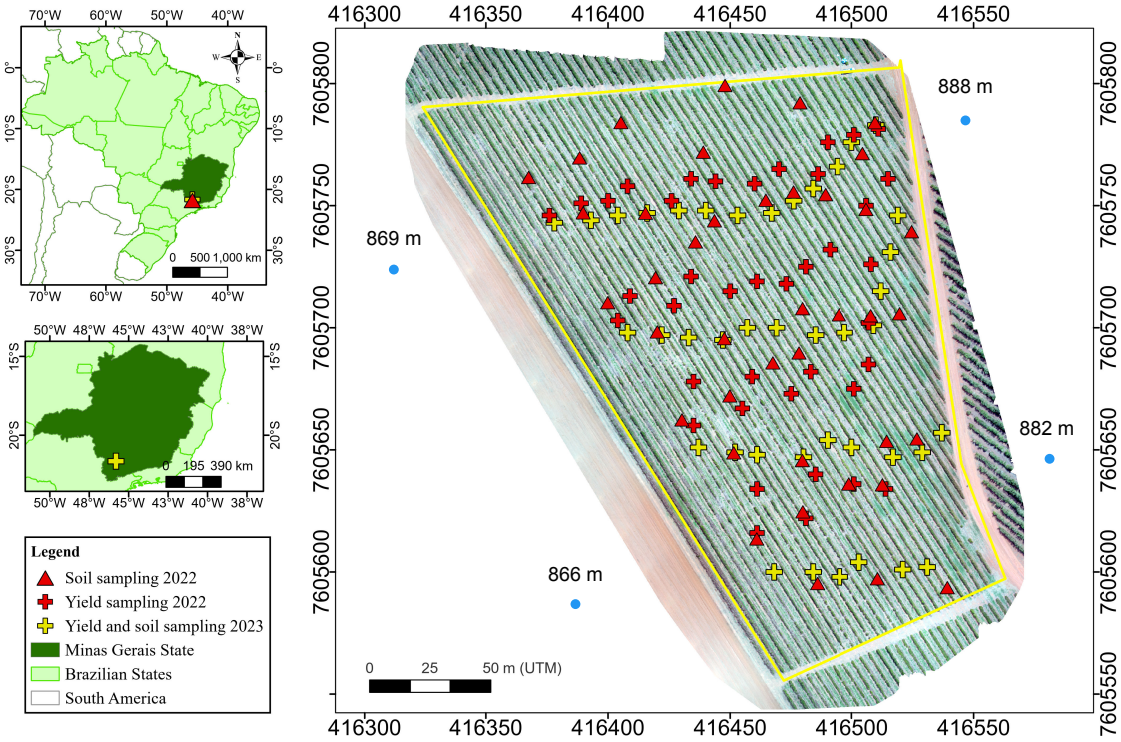
This study was conducted in a specialty coffee crop in the municipality of Paraguaçu, southern Minas Gerais, Brazil, where most of Brazil's coffee crops are concentrated. The coffee crop plot consisted of 3500 trees (Fig. 7). The coffee cultivation (*Coffea arabica L.*), cultivar Catucai Amarelo SL 134, was transplanted in 2012, at a spacing of 3.8 m between rows and 0.75 m between plants. The maximum altitude of this area is 894.3 m. Fig. 7 shows the study area and soil and yield sampling distribution in May 2022 and May 2023, overlapping an image taken on September 29 2021 by an original non-interchangeable camera onboard an unmanned aerial vehicle (UAV) model DJI Mavic 2 Pro, using an interface software "DJI Ground Station Pro". The mosaic building software was the OpenDroneMap (OPENDRONEMAP, 2020) with the flight at 40 m height. The image has a ground sample distance of 0.94 cm/pixel^{-1}.

Uniform fertilization over the entire coffee plot was done directly on the soil in 2021 by applying 42, 10, and 42 kg.ha^{-1} of N, P, K and in 2022 and 2023 by using organic compost produced by aerial composting in an area of the farm near to the coffee plots.

In this area, soil is classified as Argisols. This is characterized as a soil with higher natural fertility (eutrophic), good physical conditions and more gentle terrain have greater potential for agricultural use. Their limitations are more related to their low fertility, acidity, high aluminum content and susceptibility to erosion processes, especially when they occur on rougher terrain (SANTOS et al., 2006).

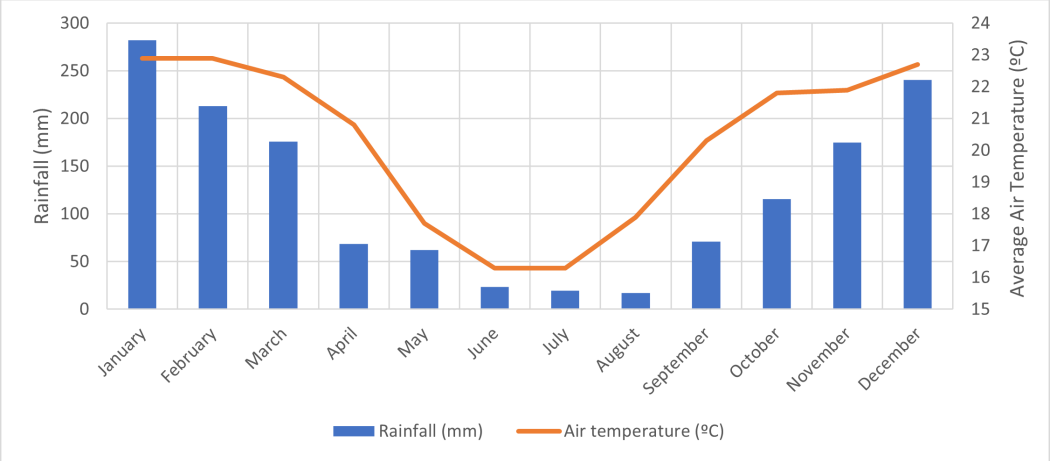
Climatic conditions of Paraguaçu municipality are shown in Fig. 8 in terms of monthly accumulated rainfall and average monthly air temperature. According to data from the National

Figure 7: UAV image of the study area with soil and yield sampling points in May 2022 and May 2023



Meteorological Institute (INMET) (MACHADO et al., 2019), since 1961 the absolute minimum temperature recorded was on June 9, 1985, with a minimum of -1.8 ^{o}C, followed by -0.8 °C on July 21, 1981 and -0.6 ^{o}C on July 18, 2000. The historical maximum is 37.1 ^{o}C on October 3, 2020, with the previous record being October 2014, on the 14th and 15th, when the maximum reached 37 ^{o}C. The record for accumulated rainfall in 24 hours was 140 millimeters (mm) on November 8, 1970. The region is characterized by a mild, tropical altitude climate, with moderate temperatures, and a hot and rainy summer, classified as Cwa according to Köppen's classification (REBOITA et al., 2015).

Figure 8: Hitorical climatic conditions (Rainfall and average air temperature) of Paraguaçu municipality (Minas Gerais state, southeast Brazil)



3.2 Soil and yield sampling and remotely-sensed covariates

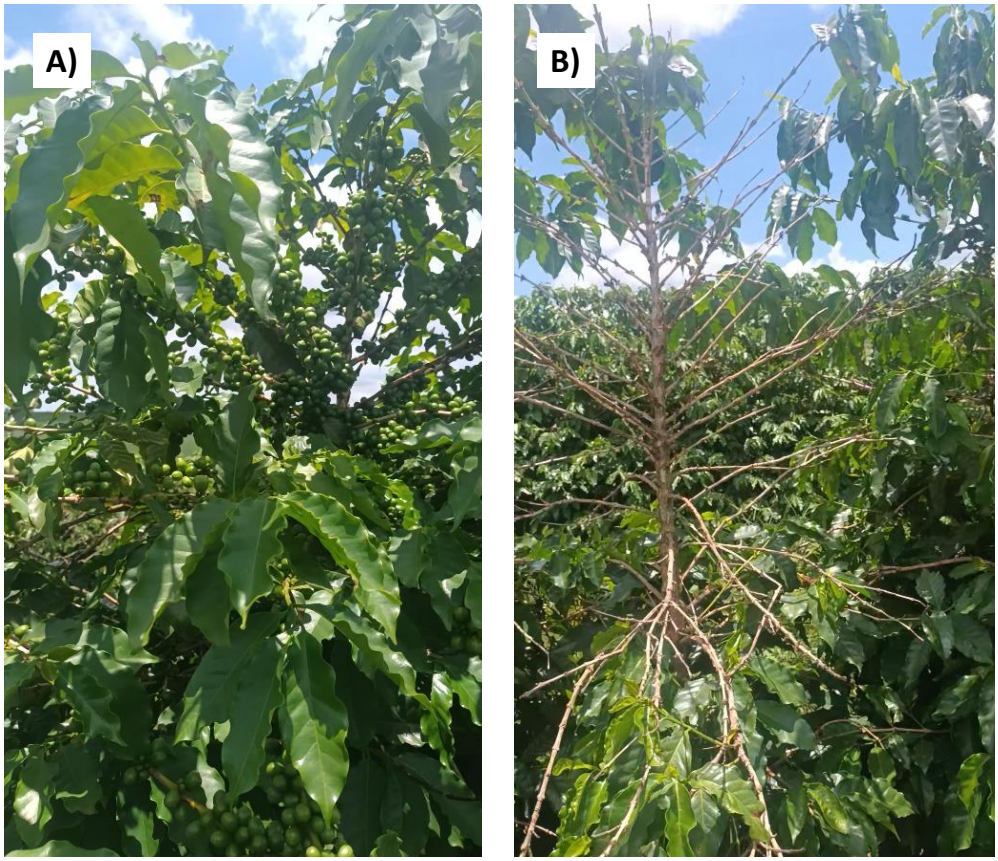
Forty composite topsoil samples (0–0.2 m) were collected from each grid point using a 5 × 5 m grid map and handheld GNSS unit (GARMIN GPS Map 62s, USA). The composite soil samples were collected in polythene bags and transported to the laboratory. Soil samples were then air-dried, thoroughly mixed, ground gently by a wooden mortar, and finally passed through a 2-mm sieve and stored in plastic bottles for soil analysis. In both years, 40 point samples were performed.

Soil samplings were performed by collecting subsamples under the crown projection in the layer of 0-20 cm, using a Dutch auger, in each plant composing the sampling point. These subsamples were homogenized to form a composite sample representative of the point in question and sent to the Laboratory of Soil Analysis. The following soil chemical attributes were evaluated: pH ($CaCl_2$ extractor), availability of phosphorus (P) (Mehlich), availability of potassium (K) (Mehlich 1 extractor), availability of sodium (Na) (Mehlich 1 extractor), availability of iron (Fe) (Mehlich 1 extractor), availability of manganese (Mn) (Mehlich 1 extractor), availability of zinc (Zn) (Mehlich 1 extractor), availability of boron (B) (wet extractor), exchangeable calcium (Ca^{2+}) (1 mol L^{-1} KCL extractor), exchangeable magnesium (Mg^{2+}) (1 mol L^{-1} KCL extractor), sulfur (S) (phosphate extractor), cation exchange capacity (CEC), base saturation (V), and organic matter (OM), following the methodology described by (ALVAREZ et al., 1999).

Samples of coffee yield were collected in May 2022 and May 2023 by getting subsamples around in each group of 2 coffee trees composing the sampling point following the grid shown in Fig. 7. The coffee plot presents higher and lower yield levels in different locations over

the plot in alternated years, a characteristic named 'biennial yield' (CAMARGO; CAMARGO, 2001). As an example of this biennality of coffee crops, Fig. 9A shows a coffee tree full of fruit, while Fig. 9B shows an empty one separated by 5 meters. Therefore, there is a high short-range spatial variability of coffee yield characterized by the bienniality of coffee yield. This scenario highlights the challenge involved in the spatial modeling of coffee yield since there is potentially an inescapable occurrence of outliers.

Figure 9: Biennial yield highlighted by neighboring coffee trees full of fruit (A) and empty (B) separated by 5 meters in December 2023



The data sampling was conducted by the Embrapa researcher, Célia Regina Grego, as part of the project "*Environmental characterization of specialty coffee production systems as a function of spatial variability and its relationship with production and quality in regions of southern Minas Gerais*" within the Coffee Research Consortium (ConCafé).

The normalized difference vegetation index (NDVI) from Sentinel-2 satellite imagery as the auxiliary variable. NDVI was calculated using Eq. 3.1:

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

where *NIR* is the percent near infrared reflectance (0.83 to 0.88 μm) and *red* is percent red reflectance (0.64 to 0.67 μm). Both are bands from Sentinel-2 satellite imagery.

NDVI values range from -1 to 1, and areas occupied by denser vegetation tend to present NDVI close to 1 (higher vegetative vigor). NDVI has been applied to detect seasonality effects, phenological stage of vegetation, length of the growing season, peak greenness, and physiological variations of leaves. Higher values correspond to healthy vegetation, with a higher density of green biomass (between 0.10 and 1). In exposed soil or less dense vegetation, positive values close to zero are obtained because under this condition there is higher absorption of radiation in the near-infrared band, which explains the low values of NDVI.

Slope over the coffee crops were retrieved from ALOS World 3D (AW3D30) with a horizontal resolution of approximately 30 meters (1 arcsec mesh) (TAKAKU; TADONO; TSUTSUI, 2014).

A step-by-step flowchart (Fig. 10) synthesizes the different approaches to delineate HMZs. The Sections which explain each step are highlighted inside this flowchart. Three approaches for HMZ delineation were applied, referencing to the dimensionality reduction of several fused variables into a single final interpolated map. This final interpolated map obtained from each approach is partitioned into two clusters using various methods, and the best one is chosen to represent the two management zones, according to the validation metrics. We chose these three combinations because they have different but comparable premises: in the SFI method, the dimensionality reduction is performed deterministically, in the MULTISPATI-PCA method the dimensionality reduction is performed by a spatially weighted PCA, while in the MAF method, the PCA is decomposed based on a geostatistical analysis.

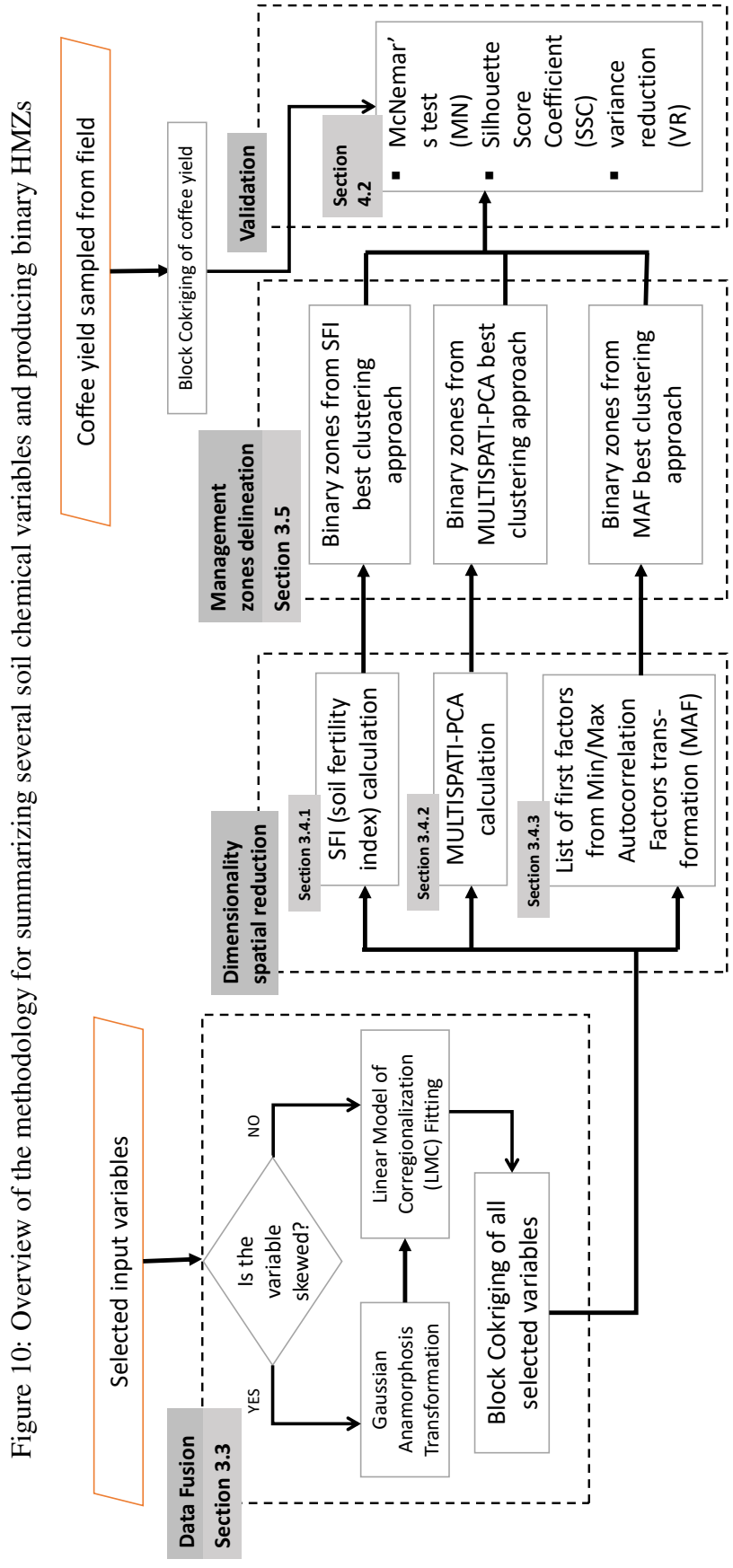


Figure 10: Overview of the methodology for summarizing several soil chemical variables and producing binary HMZs

3.3 Data fusion by multivariate geostatistics modeling

To jointly analyze the heterogeneous data set, including data from the different sensors mostly not collocated and with different support and level of uncertainty, a complex but well integrated approach of multivariate geostatistical procedures was defined. A previous support check is needed (CASTRIGNANÒ; BUTTAFUOCO, 2020), followed by a support regularization if the variables have different supports (CASTRIGNANÒ et al., 2019). Here, we performed descriptive and exploratory statistics over all variables. We want to regularize the variables in a manner that presents mean zero and unit standard deviation. To achieve this distribution, we performed Gaussian anamorphosis transformation (CASTRIGNANO; BUTTAFUOCO, 2004), then fitted the LMC and used it for the BCOK interpolation.

3.3.1 Sampled data migration

To perform multivariate analysis on different sensor data, the raw data from soil sampling and from satellite collocated into the less numerous file containing chemical soil attributes measurements by migrating them to the nearest NDVI sample point up to a maximum distance of 10 m (basically the satellite grid).

3.3.2 Variable selection

Moran's bivariate spatial autocorrelation statistic (CZAPLEWSKI, 1993) and Pearson correlation were calculated among all the variables. Variables were selected after the removal of variables with no significant spatial autocorrelation at 95% significance and the removal of variables that are highly correlated with each other (> 0.7). We consider this procedure to reduce the redundancy between less influential variables. The variables are not assumed as linearly correlated with coffee yield *a priori*. When two or more variables present a high correlation, their direct variogram is analyzed to check if some of them show a nugget effect. In this case, the variables with a more structured variogram will be chosen.

3.3.3 Preprocessing by Gaussian anamorphosis transformation

The Gaussian anamorphosis transformation is used to convert skewed and non-Gaussian statistically distributed variables into a new one with mean zero and unit standard deviation (LAJAUNIE, 1993). The BCOK variance and standard deviation using transformed variables instead

of the original ones are closer to the linear and optimal situation (GOOVAERTS, 1997). This transformation is based on the fitting of a polynomial expansion, as defined in Eq. 3.2, named Hermite polynomials:

$$\Phi = \sum \Psi_i H_i(Y) \tag{3.2}$$

where $H_i(Z)$ are the Hermite polynomials, Ψ_i are the Hermite coefficients. In practice, this polynomial expansion is stopped to a given order. Instead of being strictly increasing, the function consequently shows maxima and minima outside an interval of interest, that is for very low probability of Y. The modeling of the anamorphosis starts with the discrete version of the curve on the true data set;

The function Φ is reversible and able to convert the non-Gaussian variable into a new variable with mean zero and unit standard deviation in Eq. 3.3:

$$Y = \Phi^{-1}(Z) \tag{3.3}$$

Then, we performed the geostatistical analysis with the new standardized variables. After that, we back-transformed the predictions into the raw distribution by using the same reversible anamorphosis function. In practice, an anamorphosis will be fitted to the data and a check will have to be made whether at least the bivariate distribution of the Gaussian transform is bi-Gaussian for all spatial distance dasses. Wackernagel (2003, p.248-253) demonstrates the bijectivity of Gaussian anamorphosis, thus the function Φ is reversible.

3.3.4 Fitting of linear model of coregionalization (LMC)

LMC is an unified model which considers the direct and cross experimental variograms of all the n variables followed by a weighted least-squares (WLS) over the pairs of samples at each lag (JOURNEL; HUIJBREGTS, 1976). To well perform WLS the variables need to be highly correlated. The $n\ (n+1)/2$ direct and cross experimental variograms of all the n variables are fitted by a linear combination of the N_S standardized variograms of unit sill, $g^u(\mathbf{u})$. Under a matrix notation, the LCM follows Eq. 3.4:

$$\boldsymbol{\Gamma}(\boldsymbol{h}) = \sum_{u=1}^{N_S} \boldsymbol{B}^u g^u(\boldsymbol{h}) \tag{3.4}$$

where $\boldsymbol{\Gamma(h)} = [\gamma_{ij}(\boldsymbol{h})]$ is a symmetric matrix (order $n \times n$) where diagonal elements contains direct variograms and out-of-diagonal elements contains cross variograms; $\boldsymbol{B}^u = [b_{ij}^u]$ (the coregionalization matrix) is a symmetrical semi-definite matrix (order $n \times n$) containing the sampled values b_{ij}^u at spatial support u (CASTRIGNANÒ et al., 2000; CASTRIGNANÒ et al., 2017).

3.3.5 Block cokriging (BCOK)

The basis for a geostatistical modeling is the variogram (GOOVAERTS, 1997; WEBSTER; OLIVER, 2007; OLIVER; WEBSTER, 2015; GOOVAERTS, 1999). This is the mathematical description of the spatial autocorrelation (or spatial dependence) between a sampled value and its neighboring sampled values. Eq. 3.5 shows the empirical variogram, $\gamma(h)$, which is a discrete variation based on the difference between sampled values separated by a distance h.

$$P(B) = \sum_{i=1}^{N} \lambda_i Z_i \tag{3.5}$$

where Z_i is the observed value at location i, $P(B)$ is the predicted values at a block, N is the number of pairs of observations, B is the block, and λ are the weights.

A variogram shows the spatial structure and variability of a variable over an area. High spatial dependence means that spatial similarity can be found by analyzing the sampled values by Eq. 3.6:

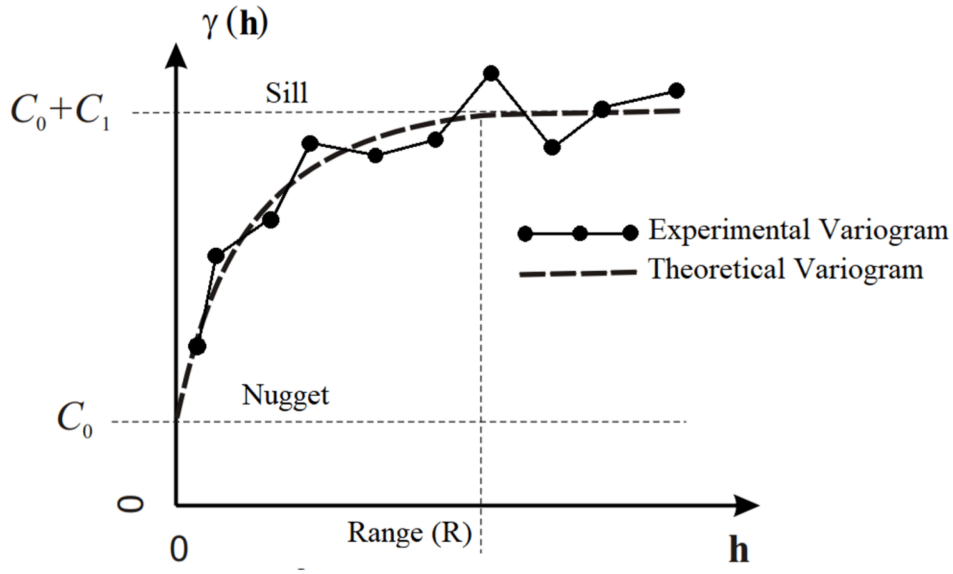
$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(B) - Z(B+h)]^2 \tag{3.6}$$

where $\gamma(h)$ is the experimental variogram; $Z(B)$ and $Z(B+h)$ are the observed values at blocks B and $B+h$; $N(h)$ is the number of observation pairs separated by distance h.

Variograms have three main parameters to consider when evaluating the spatial structure of samples: nugget (C_0), sill ($C+C_0$), and range (R) (Fig. 11). The variance increases with distance and stabilizes at a constant value ($C+C_0$) at a given separation distance, the so-called range of spatial dependence (or range, R). The sill approximates the variance of the samples for stationary data. Samples separated by distances greater than the range are not spatially autocorrelated, because the variance is equal to a random variation with no spatial correlation. If the variogram reaches a plateau (sill) at a distance, the variable is stationary. If the variance increases continuously, without reaching a plateau, it indicates the presence of trend effects and

non-stationarity. Ideally, the experimental variogram should pass through the origin and then the variation is zero. However, many soil properties have non-zero variance when h tends to zero. This discontinuity at the origin is called the nugget effect and is represented by unexplained spatial variation (microvariability at a shorter distance than the shortest sampling distance) or purely random variance (such as measurement or sampling error).

Figure 11: Example of a variogram model



The experimental variogram must be calculated over different angles to check the existence of anisotropy. If there is no sign of anisotropy (different behaviors over different directions), an "omnidirectional" empirical variogram is calculated (usually over the angle 0^o) (GOOVAERTS, 1997). Then, a theoretical continuous model of the variogram is fitted over the discrete empirical variogram. The most common models are the spherical, Gaussian, exponential, power-law, and linear functions (BERNARDI et al., 2017).

When using BCOK, the main difference consists in the calculation of the point-to-block covariance (CASTRIGNANÒ et al., 2017), following Eq. 3.7:

$$\overline{Cov}(B,\boldsymbol{x_i}) = cov(Z(B),Z(\boldsymbol{x_i})) = \int_B \frac{Cov(v,\boldsymbol{x_i})}{|B|} du' \tag{3.7}$$

where $\overline{Cov}$ is point-to-block covariance, $|B|$ is the volume of the block which is called spatial support and Cov is point-to-point covariance (CASTRIGNANÒ et al., 2017).

The punctual LMC for different combinations of variables requires regularization over the same block support that in this study was 10 m by 10 m by 0.2 m by applying block cokriging (BCOK) over the selected block grid. BCOK method can be understood as the summation of

points into the block grid, for this reason the coarser pixel size will be the final spatial resolution. We considered the depths (0.2 m) of soil samples when designing the block grid. In other words, applying BCOK is a solution for the problem of change of support (SILVA; MANZIONE; OLIVEIRA, 2023; CHILES; DELFINER, 2012; ARMSTRONG, 1998; JOURNEL; HUIJBREGTS, 1976).

3.4 Dimensionality spatial reduction

3.4.1 Soil fertility index (SFI) technique

Each soil variable observed value was converted into a binary new variable F_i, which is equal to 0 if this soil variable value indicates bad conditions of soil, otherwise is equal to 1, indicating good conditions of soil.

$$F = \left\{ \begin{array}{ll} 0, & if\ Z\ is\ classified\ as\ \text{“}bad\text{”} \\ 1, & if\ Z\ is\ classified\ as\ \text{“}good\text{”} \end{array} \right\} \tag{3.8}$$

where F is the binary indicator of fertility status and Z is the observed values of each variable.

This binary conversion is based on cutoff thresholds according to Table 2. These cutoffs are consistent with arabica coffee, can be considered in the context of specialty coffee and the references are indicated beside each cutoff.

Table 2: Cutoff thresholds for soil chemical attributes under coffee cultivation

Attributes	Condition	Attributes	Condition
V (%)	> 50.00 [a]	m (%)	< 50.00 [a]
pH ($CaCl_2$)	> 5.40 [a]	OM (g dm^{-3})	> 4.00 [a]
CEC	> 80.00 [b]	TOC (g dm^{-3})	> 15.00 [c]
H (mmolc dm^{-3})	> 36.00 [b]	Zn (mg dm^{-3})	> 3.00 [c]
P (mmolc dm^{-3})	> 50.00 [b]	Mn (mg dm^{-3})	> 20.00 [c]
K (mmolc dm^{-3})	> 4.00 [b]	Fe (mg dm^{-3})	> 1.50 [c]
Ca (mmolc dm^{-3})	> 30.00 [b]	B (mg dm^{-3})	> 1.00 [c]
Mg (mmolc dm^{-3})	> 10.00 [b]	S (mmolc dm^{-3})	> 10.00 [c]
Na (mmolc dm^{-3})	> 0.50 [c]	NDVI	> 0.50 [b]

[a] Alvarez et al. (1999)
[b] Vieira (2017)
[c] Guimarães et al. (1999)

After the conditional cutoff coding for the soil chemical variables, the converted binary values were inserted into the Soil fertility index (SFI), the higher this percentage, the higher the number of variables classified as "higher fertilization need", following Eq. 3.9:

$$SFI = 100 \sum_{i=1}^{N} \frac{F_i}{N} \tag{3.9}$$

where F_i is the binary indicator of fertilization needs for each soil variable observed value, N is the number of variables.

3.4.2 MULTISPATI-PCA technique

Given a matrix X $(n \times p)$ of data containing several (p) measurements at each of n data points, the MULTISPATI-PCA algorithm introduces a spatial weighting matrix $\mathbf{W}$ in the PCA of the standardized $\mathbf{X}$. The matrix $\mathbf{W}$ is a row-sum standardized connectivity matrix. If $\mathbf{W} = [c_{ij}]$ is the connectivity matrix indicating the strength of interactions between points i and j, then $\mathbf{W} = [c_{ij}/\Sigma c_{ij}]$.

By extension of the lag vector, a lag matrix $\overrightarrow{\mathrm{X}} = \mathbf{WX}$ can be defined. The two tables $\overrightarrow{\mathrm{X}}$ and $\mathbf{X}$ are fully matched, i.e., they have the same columns (variables) and rows (observations). MULTISPATI-PCA aims to identify multivariate spatial structures by studying the link between $\overrightarrow{\mathrm{X}}$ and $\mathbf{X}$ using the coinertia analysis (DRAY; SAID; DEBIAS, 2008).

The lag matrix $\overrightarrow{\mathrm{X}}$ is composed of the averages of neighboring values weighted by the spatial connection matrix (i.e., that only the neighboring points are taken into account). The row scores of this analysis maximize the scalar product between a linear combination of the original variables and a linear combination of the lagged variables. (ARROUAYS et al., 2011) evidenced that MULTISPATI-PCA overcomes PCA to detect and map trends in the multivariate distribution of topsoil characteristics. MULTISPATI PCs are therefore "smooth" and show strong spatial structures on the first few axes, while PCA scores can be rough, smooth, or mixed and can show spatial structures on any axis (even distant ones).

3.4.3 Multivariate Min/Max autocorrelation factors (MAF)

The method of minimum/maximum autocorrelation factors (MAF) is a multivariate transformation based on PCA that spatially orthogonalizes the attributes into uncorrelated factors for all lag spacings. MAF is similar to FK approach (DESBARATS; DIMITRAKOPOULOS, 2000).

PCA technique is performed on correlated attributes and provides uncorrelated factors by orthogonalisation of variance-covariance matrix M following Eq. 3.10:

$$M = Q^T \Lambda Q \tag{3.10}$$

Matrix Q is orthogonal and describes eigenvector of matrix M. Also Λ is a diagonal matrix and describes eigenvalue of matrix M. The PCA decorrelated factors are just defined in lag equal to zero (h = 0), following Eq. 3.11:

$$Y_{PCA}(u) = \sqrt{\Lambda}\, Q\, Z(u) = A\, Z(u) \tag{3.11}$$

The distribution of factors is normal because the assumption of normal distribution of multivariate and multiplication by $\sqrt{\Lambda}$, so the variogram matrix will be given by Eq. 3.12:

$$\Gamma_{Y_{PCA}}(h) = A\, \Gamma_Z(h)\, A^T = \sum_{s=1}^{s} A\, M_s\, A^T\, \gamma_s(t) \tag{3.12}$$

Then, decorrelated components are achieved to have fewer dimensions. Based on Eq. 3.11, A is achieved by multiplying eigenvector by inverse square root of eigenvalue since, $A^{-1} = Q^T \sqrt{\Lambda}$.

MAF can then be modeled and simulated independently, thus avoiding the inherent difficulties associated with joint simulation. The simulated MAF scores are back-transformed to the simulations of the original attributes. The MAF transformation technique was developed by Switzer and Green (1985). Their method is a data-based approach and was originally used for multivariate spatial imaging but was later applied within a geostatistical context by Desbarats and Dimitrakopoulos (2000), which is the main difference between this method and MULTISPATI-PCA. Once the theoretical covariance function is known and is modeled by a two-structure LMC, the MAF can be derived from the model coregionalization matrices.

The main steps of MAF methods are i) performing a normal score transformation of the attributes (sphering transformation) and then ii) calculating the MAF transformation coefficients assuming a two-structure LMC of the normal scores and two successive principal component decompositions. These coefficients are then used to transform transform the normal scores into MAF scores.

The correlation between MAF and PCA factors as shown in Eq. 3.13 is Q_1 matrix.

$$F_{MAF}(u) = Q_1\, Y_{PCA}(u) = Q_1\, \sqrt{\Lambda}\, Q\, Z(u) = M\, Z(u) \tag{3.13}$$

The matrix of eigenvalues L is diagonal hence the elements $Y(u)$ and $Y(u+h)$ are orthogonal at lag to lag 0, regardless of the coregionalization model. Eq. 3.14 shows that more eigenvalue results in less correlation between $Y(u)$ and $Y(u+h)$ than the uncorrelated factors specified accordingly. It is interesting to note that MAF factors which are more decorrelated, have larger eigenvalues in comparison with those factors which have smaller one. MAF factors associated with larger eigenvalues are chosen to simulate and make data reduction to continue (DESBARATS; DIMITRAKOPOULOS, 2000). For the two-structure LMC for $Z(u)$, it is readily shown that orthogonality has been ensured at all lags.

$$corr[Y(u), Y(u+h)] = I + \frac{\Lambda}{2} \tag{3.14}$$

Finally, based on LMC and decomposition of covariance-variance B, matrix $V = AB_1A^T$ is achieved. Then, in a second rotation, orthogonally diagonalized matrix $V = Q_1^T\ \Lambda_1\ Q_1$ is achieved. In this step MAF factors, which are calculated by eigenvector matrix and reduced the dimension by eigenvalue matrix whereas the database dose not lose.

3.5 Homogeneous zones delineation by clustering

Clustering methods can provide field partition based on the evaluation of similarity between spatial points, grouping them into more similar points, therefore splitting the field into groups with more similarity. Clustering methods are more complex than the empirical methods and enable greater differentiation between classes using less-subjective criteria (GAVIOLI et al., 2019).

Clustering can be a solution for the key task of HMZ delineation in PA: given a data set of georeferenced data records with high spatial resolution, we would like to discover spatially mostly contiguous zones on the field which exhibit similar characteristics within the zones and different characteristics between zones (RUSS; KRUSE, 2011).

Several clustering methods were applied to each summarizing approach to find the best field partition for each approach. We followed the list of clustering methods suggested by Gavioli et al. (2019). The clustering methods used can be divided into hierarchical methods

and partitioning methods. Hierarchical clustering methods split the dataset into several groups in two or more steps. They define a series of nested batches starting with a group with all n values to form *n* groups with one value in each. The starting point is named divisive hierarchical clustering, while if the clustering procedure starts in the other extreme it is named agglomerated hierarchical clustering (JAIN; DUBES, 1988).

According to Jain and Dubes (1988), agglomerated hierarchical methods are more common because it has a lower computation cost. Partitioning clustering methods split the dataset into several n groups without a hierarchical structure but grouping by the similarity between the values. These methods only split the dataset seeking to find groups naturally present in this dataset (JAIN; DUBES, 1988). These methods optimize the partition evaluation function, i.e., they seek to organize a set of n elements into k groups, G_1, ..., G_k, while maximizing or minimizing a pre-established evaluation function. 7 hierarchical clustering algorithms and 10 non-hierarchical clustering algorithms (partitioning methods) were implemented and evaluated (Table 3).

We clustered the PCs from MULTISPATI-PCA and MAF methods with eigenvalues higher than one and accumulated variance higher than 80%. We have not tested simply performing the clustering with the regularized variable maps since we consider the innovation of the present methodology is adding spatial autocorrelation indices into before clustering the maps.

Table 3: Clustering algorithms for the delineation of HMZs

Method (Acronym)	References
average linkage (AL) [a]	Jain and Dubes (1988); Seifoddini et al. (1989)
centroid linkage (CEL) [a]	Jain and Dubes (1988); Dixit and Naskar (2019)
complete linkage (COL) [a]	Jain and Dubes (1988); Hansen and Delattre (1978)
median linkage (ML) [a]	Jain and Dubes (1988); Klastorin (1985)
McQuitty (MCA) [a]	McQuitty (1966); (1964)
Ward (WAR) [a]	Ward (1963)
single linkage (SL) [a]	Jain and Dubes (1988); Gower and Ross (1969)
clustering large applications (CLA) [b]	Kaufman and Rousseeuw (2009); Rai and Singh (2010)
fuzzy analysis clustering (FAC) [b]	Kaufman and Rousseeuw (2009); Gosain and Dahiya (2016)
fuzzy c-means (FCM) [b]	Bezdek (2013); Suganya and Shanthi (2012); and Nayak, Naik, and Behera (2015)
fuzzy c-shells (FCS) [b]	Dave and Bhaswan (1992)
hard competitive learning (HCL) [b]	Xu and Wunsch (2010); Cheung (2002)
k-means (KME) [b]	MacQueen et al. (1967)
neural gas (NG) [b]	Martinetz, Berkovich, and Schulten (1993); Qin and Suganthan (2004)
partitioning around medoids (PAM) [b]	Kaufman and Rousseeuw (2009); Van der Laan, Pollard, and Bryan (2003)
spherical k-means (SKM) [b]	Dhillon and Modha (2001); Hornik, Feinerer, Kober, and Buchta (2012)
unsupervised fuzzy competitive learning (UFCL) [b]	Pal, Bezdek, and Hathaway (1996); Chung and Lee (1994)

[a] for hierarchical method
[b] for partitioning method

3.6 Evaluation and analysis of results

3.6.1 Performance evaluation

Different variogram models were evaluated: mean error (ME) (Eq. 3.15), mean squared error (MSE) (Eq. 3.16), kriged reduced mean error (KRME) (Eq. 3.17), and kriged reduced mean squared error (KRMSE) (Eq. 3.18):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(Z(x_i) - Z(B,\boldsymbol{x_i})) \tag{3.15}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Z(x_i) - Z(B,\boldsymbol{x_i}))^2 \tag{3.16}$$

$$KRME = \frac{1}{N}\sum_{i=1}^{N}\frac{Z(x_i) - Z(B,\boldsymbol{x_i})}{s} \tag{3.17}$$

$$KRMSE = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Z(x_i) - Z(B,\boldsymbol{x_i}))}{s}\right]^2 \tag{3.18}$$

where $Z(x_i)$ is the sampled value at location i, $Z*(x_i)$ is the predicted value at location i, N is the number of pairs of sampled and predicted values, and s is the standard deviation of the sampled values.

The ME and KRME values close to zero indicates a good model performance (CASTRIGNANÒ; BUTTAFUOCO, 2020). MSE indicates good model performance when their values is lower than the variance of the sample values (ADHIKARY et al., 2011). KRMSE should inside the range $1 \pm (2\sqrt{2})$/N (ADHIKARY et al., 2011).

3.6.2 Validation of management zones

To assess whether our method of using soil and crop attributes to delineate HMZs could characterize the spatial variation in crop yield, the map of clustering from MULTISPATI-PCA, SFI, and MAF method were compared with the available yield maps from two seasons by computing tests and measures of agreement in two-way contingency tables. The tests and measures of the agreement include:

- McNemar's test (MN) (MCNEMAR, 1947), which is appropriate when the data come from matched pairs of grid nodes with a dichotomous response. It tests the null hypothesis of marginal homogeneity and is an asymptotic chi-square test with one degree of freedom;
- the Silhouette Score Coefficient (SSC), which is derived from the silhouette coefficient, an evaluation criterion that measures the quality of the internal formation and the external separation between groups;
- the variance reduction (VR), which is the sum of the variances of the variables within each HMZ. The expectation is that the sum of the variances of the sub-areas will be less than the original variance of the area. Therefore, the higher the VR value, the better the HMZs have been defined in terms of variance reduction;

Descriptive statistics of the soil chemical variable samples in each HMZ were generated to evaluate each variable individually. It could indicate the ability of the methodology to summarize and generalize several soil chemical variables without losing local information, crucial for the actual application of fertilizers.

Boxplots of coffee yield in each HMZ was generated to assess the ability of the methodology to significantly separate higher from lower yields following the soil fertility.

3.7 Analysis tools

All clustering analysis were performed with the R software (R DEVELOPMENT CORE TEAM, 2022) using the following packages: the package 'cluster' (MAECHLER et al., 2013) for performing PAM, AL, CEL, CLA, COL, HCL, and calculating SSC; the package 'e1071' (DIMITRIADOU et al., 2006) for performing BCL, FCM, FCS, and UFCL; the package 'cclust' (DIMITRIADOU; DIMITRIADOU, 2007) for performing FAC, FCM, FCS, NG, and HCL; the package 'skmeans' (HORNIK et al., 2017) for performing SKM; the package 'fastcluster' (MÜLLNER, 2013) for the other clustering methods. Also the MULTISPATI-PCA calculation was performed with R by using the package 'adespatial' (DRAY et al., 2018). The geostatistical analyses were performed with Isatis.neo version 2023.08 (GÉOVARIANCES, 2023). Remotely-sensed covariates were retrieved by using Google Earth Engine.

Chapter 4

Results and discussion

4.1 Plot characterization and variable selection for data fusion LMC regularization

The descriptive statistical parameters of all the analyzed variables are presented in Table 4. The higher the CV, the more heterogeneous the data set. Soil P, K, Na, B, Zn, Fe, and yield presented high variability, with a coefficient of variation (CV) higher than 30% in 2022. On the other hand, in 2023 only Soil P, K, S, B, Zn, and yield CV were high. Soil pH was low (less than 10%) in both years. All other soil chemical variables had medium variability, with a coefficient of variation (CV) between 10% and 30%.

According to Ronquim et al. (2010), pH is an indicator of the biological-physical-chemical condition of the soil. An excessively acidic soil (pH very low) or excessively alkaline soil (pH very high) is less favorable for agriculture because there will be less oxygen, less organic matter, less water retention and infiltration, and more toxic ions. High soil acidity can generate high levels of Ca while Mg deficiency, which affects the full development of plants and the achievement of high yields since the low pH reduces the availability of some nutrients and increases the toxic effect of aluminum on plants (VIEIRA, 2017; RONQUIM, 2010).

Table 4: Descriptive statistics of soil chemical variables in the coffee crop (n = 40 in both years)

	May 2022					May 2023				
Attributes	Min	Mean ± SD	Max	CV	Skewness	Min	Mean ± SD	Max	CV	Skewness
pH ($CaCl_2$)	4.80	5.71 ± 0.40	6.80	6.97	0.22	5.00	5.71 ± 0.25	6.20	4.38	-0.27
CEC	9.00	89.51 ± 15.47	121.50	11.75	-0.37	53.90	91.63 ± 15.64	121.90	17.07	-0.35
TOC	51.80	15.37 ± 1.81	19.00	17.28	-0.65	11.00	15.20 ± 2.82	23.00	18.56	0.97
P (mmolc dm^{-3})	2.00	27.95 ± 21.32	90.00	76.29	1.07	6.00	52.1 ± 32.07	138.00	61.54	0.94
K (mmolc dm^{-3})	2.00	4.46 ± 5.37	37.00	119.06	5.54	1.00	2.53 ± 0.79	4.50	31.01	0.48
Ca (mmolc dm^{-3})	25.00	48.88 ± 11.58	75.00	23.7	-0.24	25.00	48.33 ± 10.62	73.00	21.98	0.081
Mg (mmolc dm^{-3})	6.00	13.9 ± 3.63	21.00	26.13	-0.31	9.00	16.83 ± 4.08	27.00	24.22	0.29
S (mmolc dm^{-3})	6.00	14.85 ± 3.52	22.00	23.69	-0.19	3.00	5.10 ± 2.46	14.00	48.19	2.42
Na (mmolc dm^{-3})	2.00	6.25 ± 5.44	30.00	87.03	2.42	0.10	0.28 ± 0.09	0.50	34.69	0.67
V (%)	55.50	74.74 ± 7.55	85.50	10.11	-0.59	56.00	73.93 ± 6.21	85.00	8.40	-0.34
OM (g dm^{-3})	25.00	35.98 ± 6.85	56.00	19.04	1.21	19.00	26.15 ± 4.77	39.00	18.24	0.89
B (mg dm^{-3})	0.18	0.64 ± 0.25	1.54	39.61	0.69	0.56	1.12 ± 0.35	2.06	31.70	0.96
Zn (mg dm^{-3})	1.50	9.22 ± 5.02	31.90	54.38	2.04	2.60	8.59 ± 4.11	18.00	47.77	0.32
Mn (mg dm^{-3})	6.00	16.43 ± 6.29	37.00	38.29	0.95	6.00	15.38 ± 4.43	25.00	28.81	-0.19
Fe (mg dm^{-3})	23.00	53.30 ± 24.95	124.00	46.81	1.11	15.00	19.80 ± 3.13	26.00	15.81	0.62
Yield (kg . $tree^{-1}$)	0.81	5.18 ± 2.57	11.575	49.57	0.29	9.56	4.919 ± 1.83	0.41	37.12	-0.71

Min – Minimum value; Max – Maximum value; CEC – cation exchange capacity; SD – Standard deviation; CV – Coefficient of variation; TOC – Total Organic Carbon; OM – organic matter; m – Aluminum saturation; V – Base saturation.

V represents the percentage of CEC occupied by bases (Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}) in relation to the exchange capacity determined at pH 7. At pH 7 soils were considered 100% base-saturated and had zero base saturation at pH 4. Soils with V equal to or greater than 50% are denominated eutrophic soils (tending to present higher fertility). Soil with base saturation values less than 50% are denominated dystrophic soils (tending to present lower fertility). Base saturation (V) can indicate the amount of cations, such as Ca, Mg, K, and identify if the soil is acidic at a level that is harmful to the crop. The soil OM contributes to an increase in soil CEC which can serve to retain and increase the reserve of soil cations and improves soil structure physics and soil water relations. Soils with higher OM content are associated with increased population and diversity of microorganisms (NÚÑEZ et al., 2011).

Liming is important for reducing soil acidity, increasing the Ca and Mg content, and neutralizing Al^{3+} (VIEIRA, 2017). However, it should be done according to the interpretation of the soil analysis and before fertilization, because in excess, liming can cause a deficiency of B, Zn, Fe, and Mn (ALVAREZ et al., 1999; PABON et al., 2020).

pH values showed a trend in soil acidity, presenting an average pH of 5.71 in both years. Considering the high variability of almost all soil chemical variables, the study area can be considered suitable for receiving different and localized management practices as long the soil samples show different chemical behaviors in different spatial positions of the coffee crop. As there are heterogeneous variables that generate variability in crop yields, the most used techniques in data analysis for developing HMZs are clustering and geostatistical models (BAZZI et al., 2013).

Eight variables were chosen: pH, Fe, Ca, Na, K, P, Mg, and NDVI for both years. These macro- and micronutrients showed high correlation with other chemical variables and indicators, such as the CEC, base saturation (V), sum of bases (BS), organic matter (OM), total organic carbon (TOC), S, B, Mn, and Zn. Heat maps with the correlation matrix over the variables in 2022 and 2023 are shown in Fig 12. Also, OM, TOC, S, B, and Zn showed pure nugget effects on their variograms. All these variables showed Moran's Index equal to - 0.05 ± 0.0002 (p-value ≤ 0.05), indicating there is spatial autocorrelation, even in the presence of outliers and under low sampling.

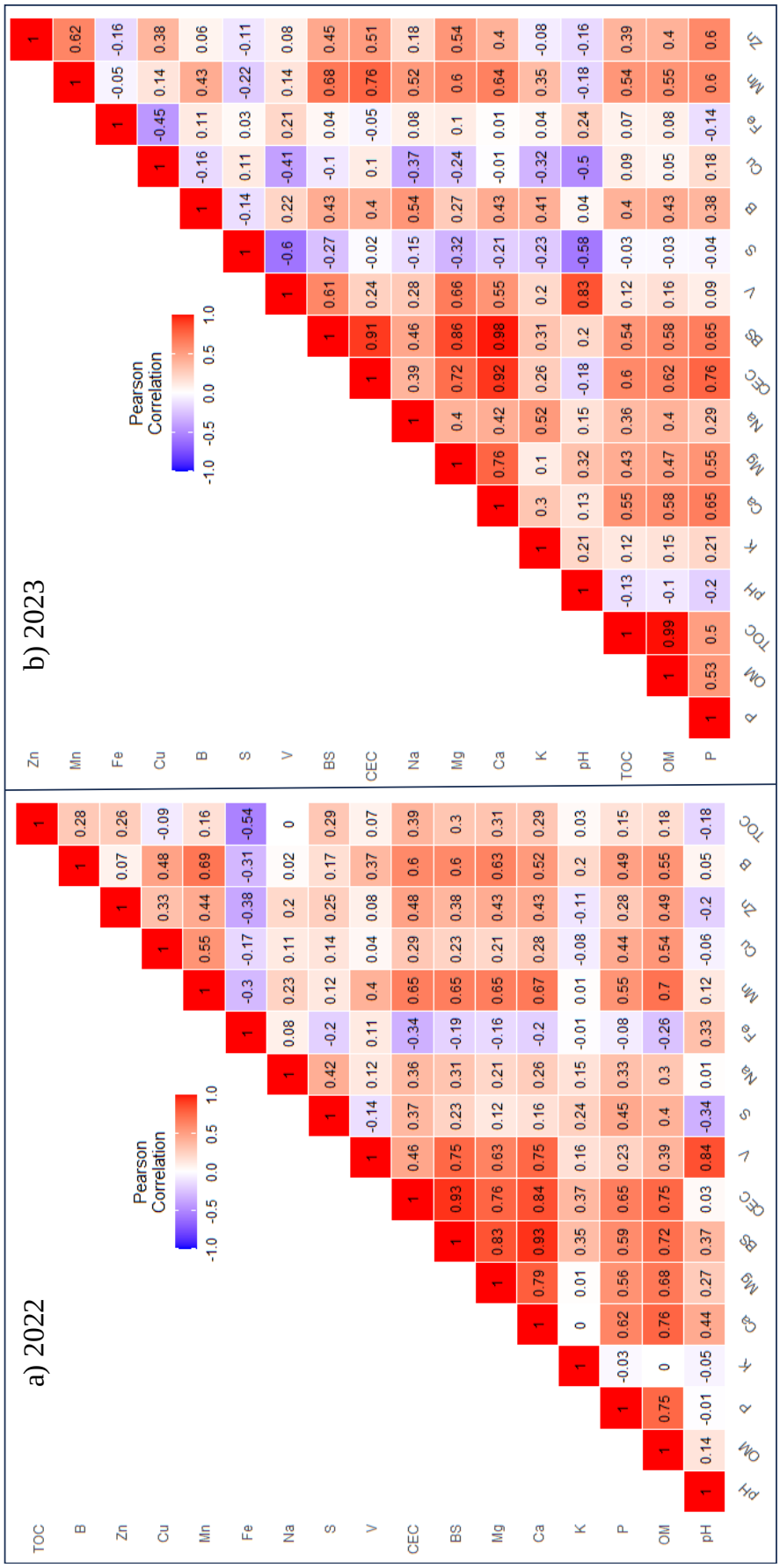


Figure 12: Heatmap of correlation between soil chemical variables in 2022 and 2023

Fig. 13 shows the transformed values by Gaussian Anamorphosis (x-axis) against original values (y-axis). The larger the dataset, the lower the sinuosity of the Hermite Polynomials on the extreme sides. Skewed distributions could be more controlled after this data transformation. One can see how different distributions could be squeezed into a Gaussian distribution centered on zero. Of course, this transformation was performed before calculating experimental variograms and fitting its theoretical models and it was back-transformed after finishing the geostatistical modeling. Considering that several authors showed how data transformation improves the variogram modeling (CASTRIGNANÒ et al., 2000; CASTRIGNANÒ; BUTTAFUOCO, 2020; MANZIONE; CASTRIGNANÒ, 2019; MANZIONE; SILVA; CASTRIGNANÒ, 2020; CASTRIGNANÒ et al., 2018; BUTTAFUOCO et al., 2010; SHADDAD; BUTTAFUOCO; CASTRIGNANÒ, 2020; CASTRIGNANÒ et al., 2019; BUTTAFUOCO et al., 2021), we did not test without data transformation.

Anisotropic variograms were simultaneously calculated for four directions with 45^o (not shown) angular increments and ± 22.5^o angular tolerance. No sign of relevant anisotropy was found in the observations. A relevant anisotropy sign could be found when the sill, range, and nugget are different in different directions.

For 2022, an isotropic LMC was fitted to all experimental variograms (considering the eight selected variables) including the following spatial structures: nugget effect = 0.02, cubic model with range = 32.3 m, and spherical model with range = 78.8 m. For 2023, the isotropic LMC was fitted to all experimental variograms including the following spatial structures: nugget effect = 0.02, cubic model with range = 28.19 m, and spherical model with range = 88.10 m. Long ranges were achieved indicating a good spatial variability explainability by using the isotropic LMC for regularization.

Selected variable's regularized maps are shown in Fig. 14A-P. Because of the outliers, we plotted each map with an individual legend instead of using a single legend for different maps of the same variable. Slope map (Fig. 14T) were retrieved by using OK (nugget effect = 0.05, spherical model with range = 53.7 m) to help the distinction between intrinsic soil and terrain effects, since it was not effective for the regularizing LMC (not shown here), for this reason it was interpolated separately.

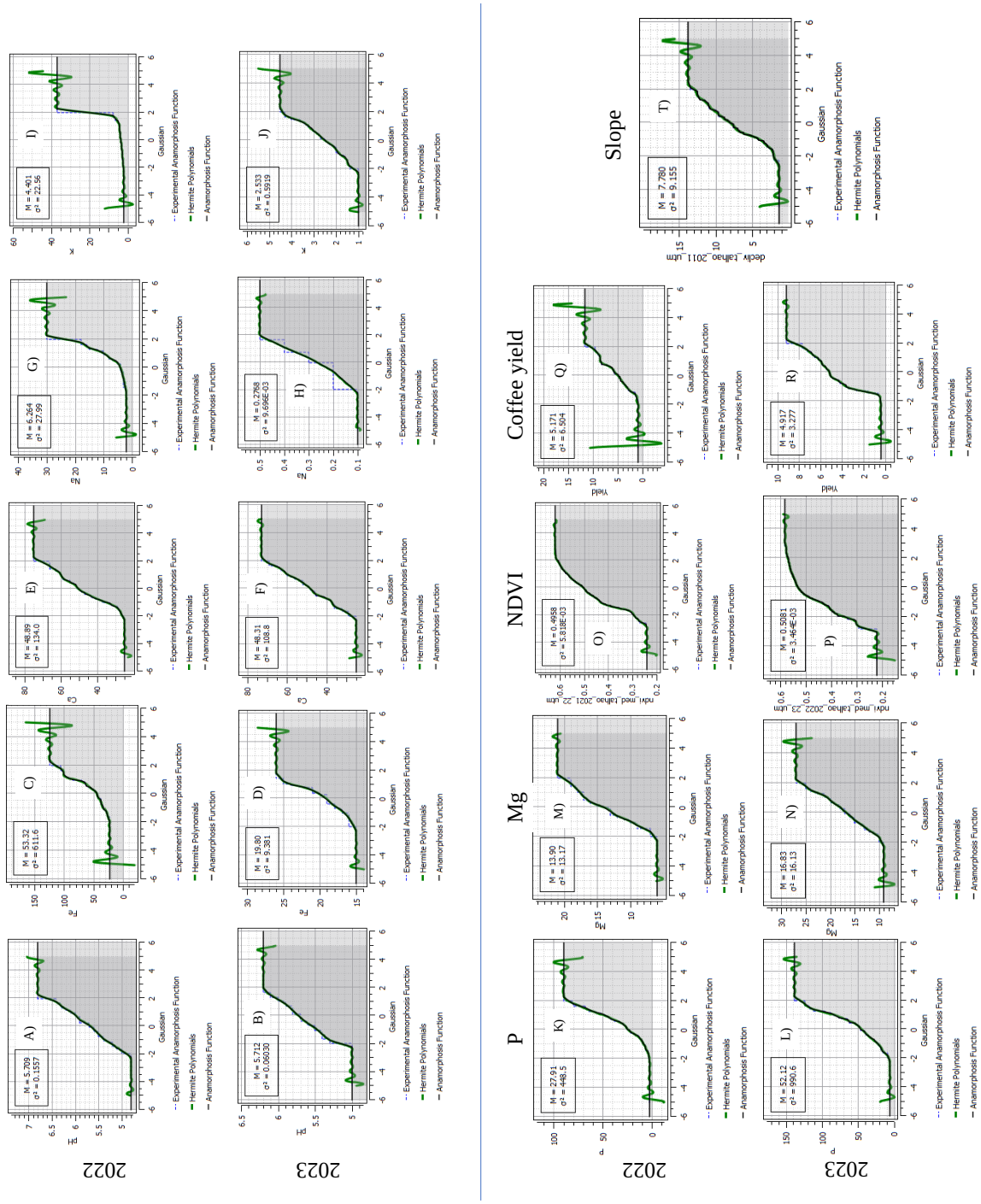


Figure 13: Gaussian anamorphosis of selected input variables

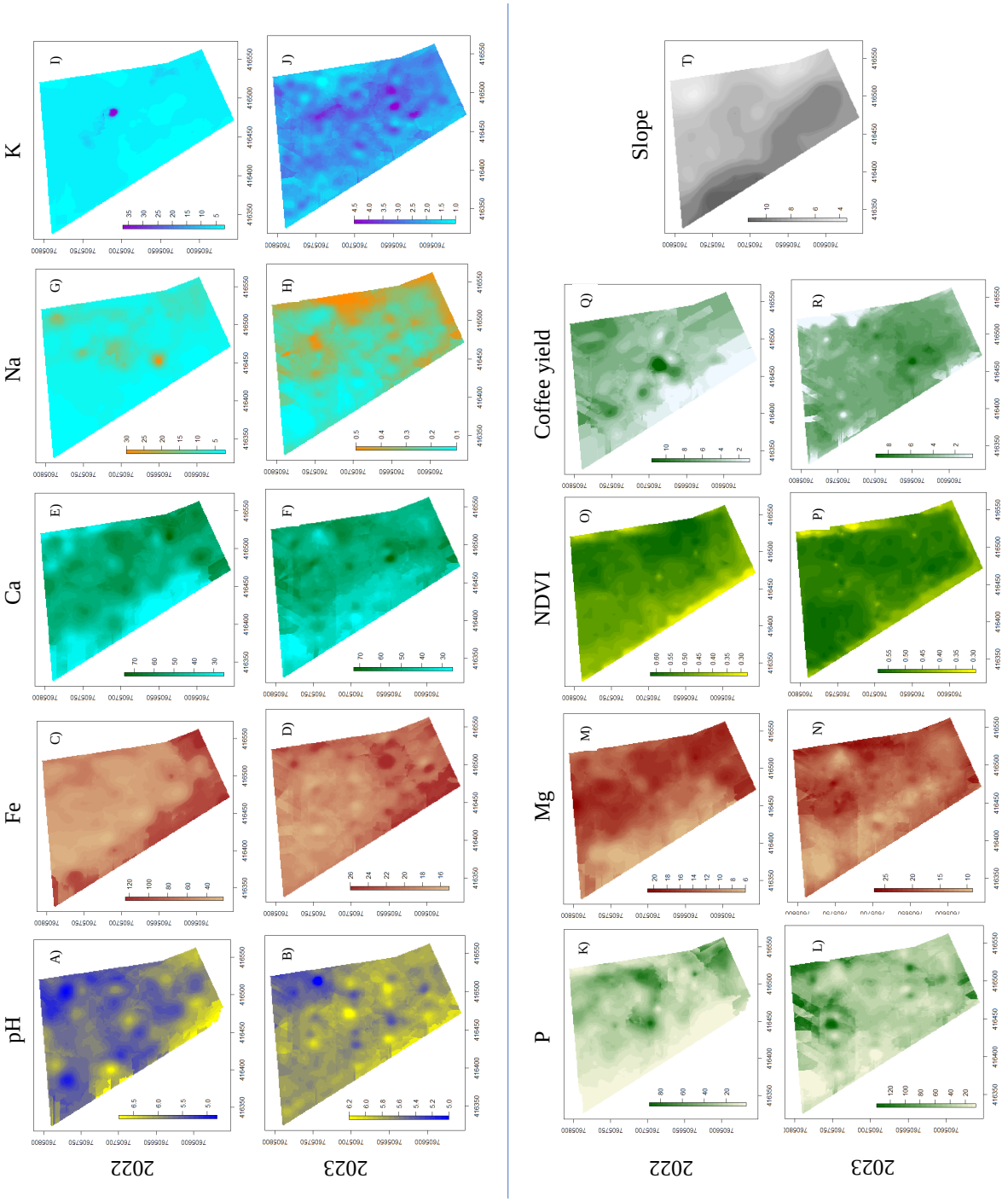


Figure 14: Maps of selected input variables after BCOK regularization

The yield maps over the 2 years are presented in Fig. 14Q,R. They look quite different and show low-yielding zones near the plot borders while the high-yielding zones are in the central area of the plot around outliers. In this paper we did not use yields as a cluster variable but as a measure of success.

The error statistics such as ME, MSE, KRME, and KRMSE were estimated and presented in Table 5. As expected, the maps with outliers in their input datasets showed the worst metrics, as Fe, Na, K, and coffee yield in 2022, and P in 2023. However, these error statistics values are satisfactory since ME and KRME values were close to zero, while MSE values were lower than SD of the sample values and KRMSE were inside the range $1 \pm (2\sqrt{2})/N$. SD of the sample values can be found in Table 3. All MSE values are lower than SD, except for pH.

Table 5: Performance evaluation of BCOK regularization interpolation of transformed variables

	2022				2023			
Variable	ME	MSE	KRME	KRMSE	ME	MSE	KRME	KRMSE
pH ($CaCl_2$)	0.83	0.75	0.75	0.984	0.81	0.74	0.72	0.971
Fe (mg dm^{-3})	2.98	2.63	2.49	0.914	1.41	1.28	1.16	0.949
Ca (mmolc dm^{-3})	2.36	1.99	1.74	0.974	2.29	2.18	1.86	0.927
Na (mmolc dm^{-3})	3.71	3.54	3.48	0.877	0.09	0.08	0.08	0.921
K (mmolc dm^{-3})	3.88	3.59	3.18	0.827	1.15	0.88	0.56	0.934
P (mmolc dm^{-3})	3.08	2.91	1.42	0.984	3.51	3.14	3.08	0.894
Mg (mmolc dm^{-3})	1.48	1.26	1.24	0.977	1.31	1.29	1.11	0.931
NDVI	0.09	0.08	0.04	0.984	0.09	0.08	0.07	0.989
Yield (kg . $tree^{-1}$)	1.58	1.49	1.31	0.984	1.34	1.41	1.45	0.987

4.2 Homogeneous management zones from SFI approach

The soil chemical variables were transformed into binary values, if values were below a cutoff threshold, it was coded with a value of 0; if values were above, it was coded with a value of 1. After performing the coding step, the SFI (Eq. 3.9) was calculated. The cutoff thresholds of all variables were presented in Table 2. Here, we applied the binary conversion only for the selected variables (pH, Na, P, Ca, K, Fe, Mg, and NDVI).

Notably, the SFI technique is a binary coding that does not explicitly capture spatial auto-correlation like MULTISPATI-PCA and MAF. Table 6 shows the clustering metrics over SFI maps for field binary partition. SL clustering method was the best one for 2022, achieving a

variance reduction of 10.12% with a statistically significant departure from the hypothesis of full agreement between the two HMZs at p-value < 0.01 according to the the McNemar's test. WAR method was the best one for 2023 with a variance reduction of 9.23% with a statistically significant adherence at p-value < 0.05 according to the the McNemar's test.

Table 6: Clustering metrics for SFI-based HMZ delineation

	2022			2023		
Method	VR	SSC	MN	VR	SSC	MN
AL	3.71	0.298	0.08	3.77	0.335	0.09
CEL	3.55	0.499	0.12	3.92	0.388	0.12
CLA	2.98	0.312	0.10	3.12	0.351	0.09
COL	2.23	0.453	0.18	2.67	0.211	0.16
FAC	3.39	0.518	0.07	4.01	0.498	0.13
FCM	2.99	0.567	0.12	3.34	0.312	0.16
FCS	4.31	0.290	0.11	4.62	0.348	0.07
HCL	5.21	0.298	0.11	5.98	0.367	0.06
KME	8.21	0.431	0.02*	7.98	0.310	0.03*
ML	7.23	0.342	0.13	7.02	0.356	0.15
MCA	3.78	0.441	0.18	3.03	0.237	0.17
WAR	8.91	0.423	0.01*	**9.23**	**0.501**	**0.001***
NG	5.33	0.439	0.09	5.45	0.445	0.09
PAM	5.23	0.387	0.13	4.99	0.227	0.17
SL	**10.12**	**0.512**	**0.002****	7.21	0.450	0.12
SKM	2.82	0.142	0.19	3.12	0.251	0.17
UFCL	3.89	0.234	0.15	4.01	0.256	0.15

** McNemar's Test is significant at $p < 0.01$
* McNemar's Test is significant at $p < 0.05$
Bold method name indicates it was the best one

4.3 Homogeneous management zones from MULTISPATI-PCA approach

The dimensionality reduction for correlated variables and identification of spatially weighted orthogonal linear recombination among variables and PCs were analyzed using the selected variables. To aggregate and summarise the individual maps produced by BCOK into HMZs, the first two sPCs were used to be clustered. They retained 85.44% in 2022 and 83.56% in 2023 of the total variance and all of them presented eigenvalues higher than one.

In the analysis of the coefficients of sPCs, which act as weights for the original variables in those components (Table 7), the first component (sPC1) had higher weighting coefficients, in absolute values, for the Mg, Na, and P for both years, with NDVI and K only for 2022, and with Ca, Fe, and pH only for 2023. This indicates high spatial and temporal mobility of nutrients, showing that adding an autocorrelation component can improve the ability of the HMZ delineation to be more effective in capturing soil fertility. However, loadings with a strong correlation to PCs were rare, indicating a limited capture of spatial variability.

Table 8 shows the clustering metrics over sPC1 and sPC2 maps for field binary partition. WAR clustering method was the best one for 2022, achieving a variance reduction of 10.77% with a statistically significant departure from the hypothesis of full agreement between the two HMZs at p-value < 0.01 according to the the McNemar's test. KME method was the best one for 2023 with a variance reduction of 10.23% with a statistically significant adherence at p-value < 0.01 according to the the McNemar's test.

Table 7: Loadings of soil variables and vegetation index in the first two principal components of MULTISPATI-PCA principal components (PCs)

Year	sPC	NDVI	Mg	Ca	K	Fe	Na	P	pH
2022	sPC1	**0.595**	**0.414**	0.514	**0.652**	0.224	**0.619**	**0.651**	0.300
	sPC2	-0.146	0.157	**-0.590**	0.052	***0.799***	0.232	0.097	***-0.813***
2023	sPC1	-0.319	**0.646**	**0.510**	0.145	**0.566**	**0.583**	**0.660**	**0.578**
	sPC2	**0.620**	-0.375	-0.027	**0.636**	0.411	0.214	-0.215	0.177

Bold method name indicates this loading is medium to highly correlated ($> |0.5|$) with this PC
Italic and bold method name indicates this loading is strongly correlated ($> |0.7|$) with this PC

Table 8: Clustering metrics for MULTISPATI-PCA-based HMZ delineation

	2022			2023		
Method	VR	SSC	MN	VR	SSC	MN
AL	4.16	0.480	0.09	3.99	0.580	0.11
CEL	3.99	0.569	0.09	4.24	0.381	0.09
CLA	6.13	0.465	0.09	5.89	0.401	0.08
COL	4.29	0.580	0.09	4.21	0.520	0.08
FAC	9.65	0.568	0.004*	9.01	0.521	0.001*
FCM	3.39	0.197	0.17	3.67	0.201	0.12
FCS	3.99	0.369	0.11	2.67	0.201	0.16
HCL	9.26	0.576	0.05	8.26	0.523	0.06
KME	10.23	0.580	0.002*	**10.23**	**0.578**	**0.001****
ML	10.23	0.580	0.002*	9.34	0.590	0.01*
MCA	3.39	0.518	0.07	4.01	0.498	0.06
WAR	**10.77**	**0.565**	**0.0002****	10.01	0.580	0.002*
NG	6.33	0.5385	0.09	7.45	0.345	0.10
PAM	6.14	0.469	0.12	7.34	0.378	0.09
SL	10.12	0.718	0.001**	9.01	0.456	0.01*
SKM	5.29	0.476	0.08	6.01	0.473	0.02
UFCL	5.29	0.369	0.09	6.10	0.401	0.10

** McNemar's Test is significant at $p < 0.01$

* McNemar's Test is significant at $p < 0.05$

Bold method name indicates it was the best one

4.4 Homogeneous management zones from MAF approach

To aggregate and summarise the individual maps produced by BCOK into HMZs, the first three PCs were used to be clustered. They retained 91.18% in 2022 and 89.16% in 2023 of the total variance and all of them presented eigenvalues higher than one. Compared with MULTISPATI-PCA loadings (Table 7), there were several loadings with a strong correlation to the PCs, indicating a greater ability to capture spatial variability within the PCs.

In the analysis of the correlation coefficients of MAF (Table 9), the first factor (MAF1) had a high correlation with NDVI for both years, as expected, since this is the most populated input dataset. In 2022, this first factor was more correlated with soil variables (Mg, Ca, Na, and pH) than was in 2023 (K only). In other words, the NDVI was the most relevant variable to explain spatial variability in both years. Variables with outliers tend to be less explanatory and thus are more related to MAF2 and MAF3.

Table 10 shows the clustering metrics over MAF1, MAF2, and MAF3 maps for field binary partition. HCL clustering method was the best one for 2022, achieving a variance reduction of 12.09% with a statistically significant departure from the hypothesis of full agreement between the two HMZs at p-value < 0.01 according to the the McNemar's test. KME method was the best one for 2023 with a variance reduction of 12.67% with a statistically significant adherence at p-value < 0.01 according to the the McNemar's test.

Furthermore, MAF yielded usually the best results in terms of the VR index (Table 10), in other words, this approach identified classes with larger differences between the respective normalized average and lower internal residual values. Differences in the normalized average yield between classes indicate that soil conditions influence the crop response (GAVIOLI et al., 2016).

Table 9: Correlation between the soil variables and vegetation index and the Min/Max Autocorrelation Factors (MAF) based on the Sphering transformed PCs which eigenvalues are higher than one

Year	MAF	NDVI	Mg	Ca	K	Fe	Na	P	pH
2022	MAF1	***-0.78***	***-0.77***	***0.77***	0.13	0.26	**0.21**	0.31	**0.59**
	MAF2	-0.23	0.21	-0.23	-0.11	***0.99***	-0.12	-0.21	0.57
	MAF3	-0.68	-0.67	-0.11	***0.95***	0.09	0.057	**-0.43**	-0.02
2023	MAF1	***-0.71***	0.25	0.61	**0.41**	0.07	0.29	-0.09	0.16
	MAF2	-0.41	0.66	0.50	-0.06	***0.82***	**0.65**	0.64	0.06
	MAF3	0.09	***-0.75***	***-0.91***	-0.58	-0.08	-0.59	***-0.91***	**0.49**

Bold method name indicates this loading is medium to highly correlated ($>$ |0.5|) with this PC
Italic and bold method name indicates this loading is strongly correlated ($>$ |0.7|) with this PC

Table 10: Clustering metrics for MAF-based HMZ delineation

	2022			2023		
Method	VR	SSC	MN	VR	SSC	MN
AL	4.25	0.345	0.10	4.01	0.421	0.09
CEL	4.01	0.290	0.11	4.12	0.378	0.08
CLA	6.12	0.456	0.08	5.77	0.412	0.09
COL	4.12	0.342	0.09	4.23	0.453	0.08
FAC	7.21	0.290	0.01	7.88	0.250	0.02
FCM	3.91	0.298	0.02	3.86	0.345	0.03
FCS	2.90	0.190	0.12	3.01	0.201	0.11
HCL	**12.09**	**0.578**	**0.0002****	9.21	0.451	0.001**
KME	10.32	0.498	0.01*	**12.67**	**0.567**	**0.0002****
ML	8.21	0.431	0.02	7.98	0.310	0.03
MCA	9.21	0.399	0.01*	8.12	0.410	0.02*
WAR	10.34	0.488	0.009**	9.81	0.419	0.01*
NG	7.01	0.478	0.09	6.89	0.456	0.02*
PAM	6.90	0.345	0.12	6.12	0.341	0.05
SL	8.90	0.432	0.09	5.23	0.321	0.04*
SKM	5.03	0.47	0.12	5.67	0.347	0.02*
UFCL	4.98	0.41	0.10	5.12	0.451	0.09

** McNemar's Test is significant at $p < 0.01$
* McNemar's Test is significant at $p < 0.05$
Bold method name indicates it was the best one

4.5 Homogeneous management zones final maps

To obtain a delineation of the field into homogeneous areas, the best clustering application for each dimensionality reduction approach in each year was considered for the final HMZ map, shown in Fig. 15. For SFI, the SL clustering method was used for 2022, and the WAR method for 2023. For MULTISPATI-PCA, the WAR clustering method was used for 2022, and the KME method for 2023. For MAF, the HCL clustering method was used for 2022, and the KME method for 2023. We named the zone which groups the lower values of input variables as "Zone 1" and the other one, with higher values, as "Zone 2".

Summing up, these maps graphically realize the fusion of several soil chemical variables and remotely sensed data into a partition of the field into zones with different chemical properties of the soil which can provide insights about the fertility of the soil and its impacts on coffee yield, even under a biennial behavior. This can be summarized by checking the boxplots of coffee yield over the zones delineated from the different approaches in Fig. 16. We set zone 1 as the low-yielding zone and zone 2 as the high-yielding one even before plotting the boxplot because zone 1 groups the low values of the chemical variables while zone 2 groups the higher ones, as shown in Table 11.

MAF-based HMZ presented the best coffee yield differentiation for both years according to the boxplots (Fig. 16), even in the presence of outliers. By visual inspection, when comparing the slope map (Fig. 14T) with the HMZs (Fig. 15), it is notable that slope controlled the fertility zoning in 2022, while in 2023 this does not happen, mostly because the field is more controlled by agronomical reasons (homogeneous fertilization and crop management) than by the terrain when the spatial autocorrelation is considered. SFI-based HMZs look more related to slope for both years. However, SFI showed the worst results in terms of clustering metrics and yield differentiation in both years. Notably, MAF's zone 2 in 2023 was quite reduced compared with other techniques, however showed the best clustering metrics values and yield differentiation, indicating that using this zoning could reduce the fertilization application in a more precise way.

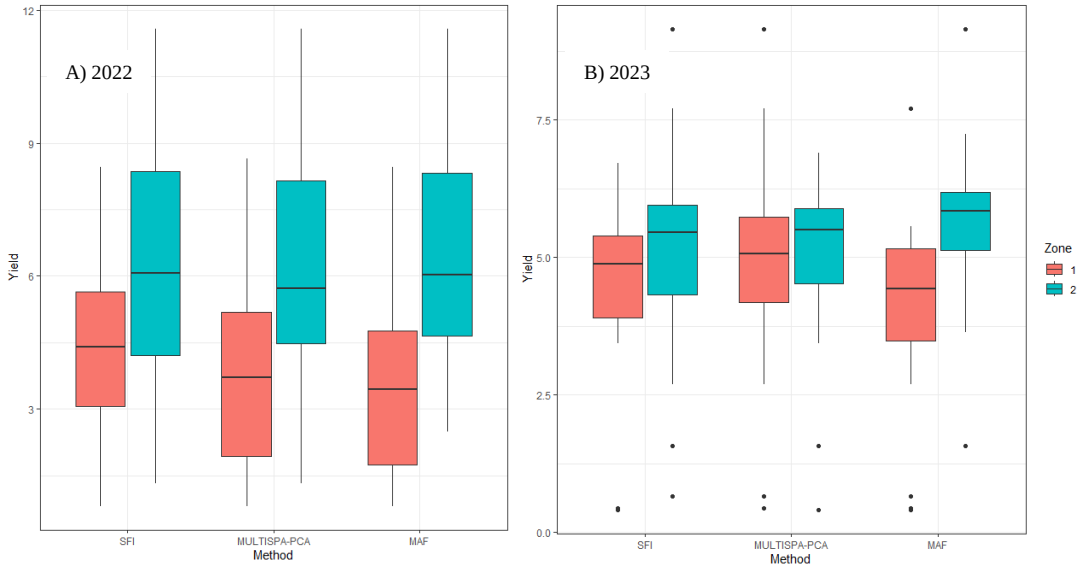
The combined analysis of MN, SSC, and VR results confirms the recommendation of the field partition into two classes. Gavioli et al. (2016) assessed the field partition into 2, 3, and 4 HMZs using MULTISPATI-PCA method and found the split into 2 zones obtained smoother boundaries with the best multidimensional variance reduction. Córdoba et al. (2016) and Peralta et al. (2013) also considered two classes HMZs as the best choice because is a easier field operations choice.

Figure 15: HMZs from different approaches in 2022 and 2023

Table 11: Average values of soil properties, vegetation (NDVI), terrain (slope), and coffee yield (kg . tree^{-1}) over HMZs

	Year	Zone	Mg	Ca	K	Fe	Na	P	pH	NDVI	Slope	Coffee yield
SFI	2022	1	7.21	32.67	9.89	69.54	19.78	21.23	5.12	0.43	9.87	4.35
		2	12.34	48.98	10.23	68.65	17.11	31.34	5.14	0.45	5.78	6.03
	2023	1	12.45	34.90	2.12	18.54	0.31	42.12	5.45	0.46	9.12	4.69
		2	15.45	45.90	2.45	18.99	0.29	42.23	5.20	0.42	5.99	5.45
MULTISPATI-PCA	2022	1	6.45	34.89	9.99	67.54	19.28	20.23	5.15	0.42	7.92	3.89
		2	15.78	47.56	10.13	67.01	17.31	32.34	5.14	0.46	7.15	5.62
	2023	1	12.34	35.78	2.13	19.87	0.34	25.45	5.21	0.39	6.89	5.12
		2	18.98	44.89	2.35	17.13	0.35	58.45	5.44	0.45	7.45	5.39
MAF	2022	1	7.12	33.98	10.01	67.14	16.12	20.83	5.12	0.37	9.65	3.46
		2	14.78	46.89	10.03	67.06	8.34	32.19	5.12	0.48	6.12	6.01
	2023	1	13.78	31.98	2.09	18.98	0.41	24.45	5.02	0.36	7.49	4.35
		2	17.12	47.62	2.61	18.64	0.29	59.45	5.78	0.47	6.15	5.93

Figure 16: Boxplots of coffee yield over HMZs from different methods in 2022 and 2023



4.6 Incorporating soil chemical summarized information into precision agriculture

Research in PA has focused on dividing a field into a few relatively uniform management zones, as a practical and cost-effective approach for variable application of agronomical procedures. In this study, different types of data, including soil, coffee yield, and remotely sensed parameters, were used to delineate HMZs in a specialty coffee crop. Since spatial–temporal variability in specialty coffee production depends on many factors, multivariate approaches were preferred through the effective integration of different informative layers. The achievement of compact classes within a coffee crop, which are both contiguous in geographic space and relatively homogeneous in attribute space, is quite difficult to reach, nevertheless highly desirable in PA.

By identifying fields where the relationship between soil and yields is fundamentally different from others, unique management areas could be delineated. Usually, HMZs are delineated from a single variable, such as altitude (JACINTHO et al., 2017) or soil apparent electrical conductivity (PERALTA; COSTA, 2013; SCUDIERO et al., 2013). However, multivariate approaches for delineating HMZs are needed for a more general and comprehensive application of PA (ORTEGA; SANTIBÁÑEZ, 2007; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009; CÓRDOBA et al., 2016). Also, making HMZ a brief spatial summary of several characteristics from the field is a challenge, because both spatial and intrinsic heterogeneity make HMZs too

general for real applications (CÓRDOBA et al., 2016). Reducing the scope of the HMZs into a group of variables like soil fertility, weather drivers, yield factor, etc, could make multivariate approaches easier to understand in the field and more appropriate for farmers to use in the field.

In the multivariate approach, the most commonly used methodologies are the joint use of dimensionality reduction techniques such as principal component analysis (PCA) and clustering techniques such as c-means or k-means. When applying a PCA, the research objective usually is to find the key loading factors to explain a main independent factor. Peralta et al. (2013) found that soil properties and nutrient concentrations were compared with apparent electrical conductivity (ECa) using principal components (PC)-stepwise regression and ANOVA. The dimensionality reduction in this case means using fewer variables to explain another one with regression modeling. In a second moment, clustering is applied to group the principal components into groups (HMZs).

The results indicate that clustering methods extensively used for delineating HMZs, such as fuzzy methods (FAC, FCS, or FCM) and KME (K-means), may not be adequate to accomplish this task every time. Gustaferro et al. (2010), Dobermann et al. (2003) and Gavioli et al. (2019) already pointed out that FCM performs poorly compared with others clustering methods. Dobermann et al. (2003) also found good performances using WAR (as we found when using SFI and MULTISPATI-PCA for delineating HMZs).

The values of VR, SSC, and MN varied when applying clustering into different dimensionality-reduced maps, showing that a single clustering method cannot be considered as a universal method for any situation. The best algorithms according to the VR index formed groups with high internal similarity and adequate inter-group separation (high external dissimilarity) (GAVIOLI et al., 2019). Notably, not all clustering methods achieved reasonable results. Usually, a single clustering method is used for delineating an HMZ map using several interpolated maps as input (SILVA; MANZIONE; OLIVEIRA, 2023), without applying a dimensionality reduction step as done here. Nowadays, a wide range of clustering approaches are available and also more sophisticated spatial modeling techniques can make the HMZ delineation more accurate. Also, greater computing capacity is available on personal computers or in cloud servers, enabling multiple methodologies to be used and tested to use the one with the best results, making methodologies more agnostic and less dependent on a single technique.

4.7 Applicability of HMZs

In the present research, the methodology is focused on converting the relevant variables on soil fertility based on agronomical knowledge. The results of the present methodology (Fig. 10) showed a good summarizing ability, which produced HMZs that separated several soil chemical variables into two zones and the information was well preserved as shown by the boxplots (Fig. 16). A smoothed zoning as presented here, without fragmented spatial distribution, is highly desirable, because usually highly accurate classifications and clustering methods produce unrealistic maps for site-specific management, with isolated pixels and corners which are ignored in real applications. Also, a smoothed zoning does not need to be redesigned by hand at the end of modeling, just needs a threshold to separate different zones.

The question of the best way to integrate different types of data to derive the most information is a common problem in PA. As long as there is still no universal protocol for delineating HMZs (MARTÍNEZ-CASASNOVAS; ESCOLÀ; ARNÓ, 2018), delineating HMZs may have different approaches and consequently different solutions using different input variables considered for clustering. At present, agricultural and agronomical collections of data often include the outcomes from different sensors, characterised by various spatial resolutions and degrees of uncertainty. This study has demonstrated that combining multivariate geostatistics with clustering approaches can provide useful means for automatically delineating management zones, joint-using data science, and agronomical knowledge to support HMZ interpretation.

4.8 Temporal instability of HMZs and nutrient mobility on soil

Temporal stability of the HMZs is another important issue in site-specific agriculture, which might require the extension of the current methodology for new experiments with the same sampling design over different periods. Several authors have shown the inconsistency over time of spatial variability patterns of important crop and soil properties such as yield, protein content, and plant available N (PERALTA; COSTA, 2013; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009; CÓRDOBA et al., 2016; AGGELOPOOULOU et al., 2013; BERNARDI et al., 2017; PARIS et al., 2019; SILVA; MANZIONE; OLIVEIRA, 2023; ROBERT, 2002; THOMAS, 1970). A common finding is that temporal variability is generally much higher than spatial

variability and the definition of stable low- and high-yielding potential zones is very uncertain. Also, Dobermann et al. (2003) and Gavioli et al. (2019) report the reduction of spatial fragmentation by using PCA-based methods and kriging interpolation (when using the appropriate grid and theoretical model) as advantages of these approaches. Smooth HMZs are better for practical use for site-specific applications.

The instability of spatial variability is due to the mobility of soil nutrients. which are transferred from year to year, and nutrient transport rates depend on local conditions (LAMBERT; LOWENBERG-DEBOER; MALZER, 2007; THOMAS, 1970). The most mobile nutrients usually are phosphorus (P), potassium (K), and nitrogen (N) (THOMAS, 1970). For example, Lambert, Lowenberg-Deboer, and Malzer (2007) found that P transport rates were heterogeneous due to local topographic and chemical variations in the soil.

Another example is the "antagonism" between nutrients in the soil. Fe is considered not very mobile in the plant, however, it is one of the micronutrients most accumulated by the coffee tree, not because of its metabolic demand, but because of the high availability of Fe in soils (PRIMAVESI, 2002). The availability of Fe to the plant is associated with the clay content and OM content of the soil, as clay soils tend to retain Fe, making it unavailable for absorption. On the other hand, adequate OM levels enable plants to better use this nutrient due to its acidifying and reducing characteristics, as well as its ability to form chelates in adverse soil pH conditions (VIEIRA, 2017; MOREIRA SILVA; ALVES, 2013; BORÉM et al., 2021). In soils with high acidity, Fe deficiency can occur due to an excess of Mg, which reduces its absorption through antagonism. This antagonism can be seen in the maps in Fig. 14.

Checking the plot characterization shown in Table 4, it is clear the relevant differences in the chemical soil variables over time. For example, the mobility of Na can be detected by the decreasing of CV from 87.03% to 34.69%. Ca, Na, K, and V had a CV difference between 2022 and 2023 around 15-30%. On the other hand, OM, CEC, and TOC are highly sensible to tillage and fertilization (using single rate application for this case). Consequently, HMZs can change from year to year.

Knowing the mobility of nutrients in the soil is intrinsic to the dynamics of the soil-water-plant interface (THOMAS, 1970) makes the PA paradigm even more important, as it demonstrates the need for periodic mapping, and consequently the possibility of rethinking the recommendations made and changing strategies, as well as detecting more complex problems that will require further studies.

Chapter 5

Conclusions

5.1 Concluding remarks and contributions of this thesis

Firstly, from the bibliometric study for understanding the state-of-art of the use of geostatistics in PA in the Brazilian scientific community, we detected using geostatistics for PA has been limited, mostly for univariate interpolation purposes. In this manner, our proposed methodology is innovative, considering the multivariate geostatistics for PA is still not widely used. From a statistical point of view, disregarding the heterogeneity of data and merging all data together might lead to considerable estimation errors and bias. Geostatistics can produce a satisfactory solution to this common problem by taking the change of support into account.

Then, delineating HMZs and interpolating coffee yield even with the presence of outliers, ranking the effects of using different dimensionality reduction approaches followed by different clustering methods, are proposed to facilitate the evaluation of innovative approaches over hard-to-deal spatial modeling scenarios like a low-sampled field with outliers.

A data fusion approach based on multivariate geostatistics, for combining spatial data of different types from soil sampling and remotely sensed retrieving was described and applied over a specialty coffee plot. In summary, this work reported the following findings:

- There is generally high spatial variability of soil chemical attributes and coffee yield after performing Gaussian anamorphosis transformation and LMC regularization;
- High temporal variability was observed over the 2 years while the spatial patterns remained quite unstable;

- Spatial variability were found and modeled even with the presence of outliers by using BCOK interpolation with LMC regularization;
- Management zones can be delineated by adopting a combination of multivariate geostatistics with different clustering approaches;
- MAF approach provided the best coffee yield differentiation between zones.

In this research, data fusion allowed the use of spatial information from different sources. After detecting the current state-of-art of the use of geostatistics to PA, it was clean the main current use is limited to univariate modeling. In this manner, a data fusion flexible approach is a contribution to the PA community in Brazil.

The joint use of geostatistics and clustering methods (from data science) is another important point, since previous studies use them separated for the same goal, delineating HMZs. Data science and agronomical knowledge supported HMZ delineation. Summing up, a heterodox combination of tools allowed us to deal with heterogeneous datasets for a challenging research problem.

5.2 Future research

Future research might focus on new experiments with the same sampling design along different periods, looking to find more evidence about the robustness of the methodology to ensure the assurance of the farmers in its application. Farmers can focus on managing the variation within the coffee complete cycle (around 2 years), and a better strategy might be to combine the use of HMZs with crop-based in-season remote sensing. The latter information could be incorporated efficiently into a decision support system (DSS) software aimed at supporting farmers in their agricultural management.

Future studies may focus on smallholder farmers, more specifically on farmers' social and economic behavior towards the adoption of PA practices as the use of HMZs and whether the adopters have a competitive advantage compared to non-adopters. Also, a better understanding of how the new generation, according to the level of connectivity in the rural environment.

References

ADHIKARY, P. P.; DASH, C.; BEJ, R.; CHANDRASEKHARAN, H. Indicator and probability kriging methods for delineating Cu, Fe, and Mn contamination in groundwater of Najafgarh Block, Delhi, India. **Environmental Monitoring and Assessment**, Springer, Nova Iorque, v. 176, n. 1, p. 663–676, 2011. DOI: 10.1007/s10661-010-1611-4.

AGGELOPOOULOU, K.; CASTRIGNANÒ, A.; GEMTOS, T.; BENEDETTO, D. Delineation of management zones in an apple orchard in Greece using a multivariate approach. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 90, p. 119–130, 2013. DOI: 10.1016/j.compag.2012.09.009.

ALMEIDA, L. F. de; SPERS, E. E. **Coffee Consumption and Industry Strategies in Brazil: A Volume in the Consumer Science and Strategic Marketing Series**. Londres: Woodhead Publishing, 2019.

ALVAREZ, V. V. H. et al. Interpretação dos resultados das análises de solo. In: RIBEIRO, A. C.; GUIMARÃES, P. T. G.; ALVAREZ, V. V. H. (Eds.). **Recomendações para o uso de corretivos e fertilizantes em Minas Gerais**. Viçosa: Comissão de Fertilidade do Solo do Estado de Minas Gerais, 1999. P. 25–32.

ALVES, M. C.; SILVA, F. M.; POZZA, E. A.; OLIVEIRA, M. S. Modeling spatial variability and pattern of rust and brown eye spot in coffee agroecosystem. **Journal of Pest Science**, Springer, Nova Iorque, v. 82, n. 2, p. 137–148, 2009. DOI: 10.1007/s10340-008-0232-y.

ANASTASIOU, E.; CASTRIGNANÒ, A.; ARVANITIS, K.; FOUNTAS, S. 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**, Elsevier, Amsterdã, v. 684, p. 155–163, 2019. DOI: 10.1016/j.scitotenv.2019.05.324.

ANDRADE, A. D. et al. Spatial variability of soil penetration resistance in coffee growing. **Coffee Science**, Universidade Federal de Lavras, Lavras, v. 13, n. 3, p. 341–348, 2018. DOI: 10.25186/cs.v13i3.1456.

ARAÚJO, G. et al. Comparativo entre os atributos químicos do solo amostrados de forma convencional e em malha. **Coffee Science**, Universidade Federal de Lavras, Lavras, v. 12, n. 1, p. 17–29, 2017. DOI: 10.25186/cs.v12i1.1188.

ARAÚJO, G. et al. Plant sampling grid determination in precision agriculture in coffee field. **Coffee Science**, Universidade Federal de Lavras, Lavras, v. 13, n. 1, p. 112–121, 2018. DOI: 10.25186/cs.v13i1.1391.

ARMSTRONG, M. **Basic linear geostatistics**. Nova Iorque: Springer, 1998.

ARROUAYS, D. et al. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. **Geoderma**, Elsevier, Amsterdã, v. 161, n. 3-4, p. 107–114, 2011. DOI: 10.1016/j.geoderma.2010.12.002.

BACHMAIER, M.; BACKES, M. Variogram or semivariogram? Understanding the variances in a variogram. **Precision Agriculture**, Springer, Nova Iorque, v. 9, p. 173–175, 2008. DOI: 10.1007/s11119-008-9056-2.

BANSOD, B. S.; PANDEY, O. An application of PCA and fuzzy C-means to delineate management zones and variability analysis of soil. **Eurasian Soil Science**, Springer, Nova Iorque, v. 46, p. 556–564, 2013. DOI: 10.1134/S1064229313050165.

BARROS, L. S. et al. Dispersão Espacial de Atributos Químicos do Solo de um Açaizeiro na Região Amazônica. **Anuário do Instituto de Geociências**, Universidade Federal do Rio de Janeiro, Rio de Janeiro, v. 45, 2022.

BASSO, B.; RITCHIE, J. T.; CAMMARANO, D.; SARTORI, L. A strategic and tactical management approach to select optimal N fertilizer rates for wheat in a spatially variable field. **European Journal of Agronomy**, Elsevier, Amsterdã, v. 35, n. 4, p. 215–222, 2011. DOI: 10.1016/j.eja.2011.06.004.

BASSO, B. et al. Spatial validation of crop models for precision agriculture. **Agricultural Systems**, Elsevier, Amsterdã, v. 68, n. 2, p. 97–112, 2001. DOI: 10.1016/S0308-521X(00)00063-9.

BAZZI, C. L. et al. Definição de unidades de manejo usando atributos químicos e físicos do solo em uma área de soja. **Engenharia Agrícola**, Associação Brasileira de Engenharia Agrícola, Jaboticabal, v. 33, p. 952–964, 2013. DOI: 10.1590/S0100-69162013000500007.

BERNARDI, A. C. C. et al. Variabilidade espacial de índices de vegetação e propriedades do solo em sistema de integração lavoura-pecuária. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Universidade Federal de Campina Grande, Campina Grande, v. 21, n. 8, p. 513–518, 2017. DOI: 10.1590/1807-1929/agriambi.v21n8p513-518.

BERNARDI, A. C. C. et al. Mapping of yield, economic return, soil electrical conductivity, and management zones of irrigated corn for silage. **Pesquisa Agropecuária Brasileira**, Embrapa, Brasília, v. 53, n. 12, p. 1289–1298, 2018.

BEZDEK, J. C. **Pattern recognition with fuzzy objective function algorithms**. Nova Iorque: Springer, 2013.

BIVAND, R. S.; PEBESMA, E. J.; GOMEZ-RUBIO, V.; PEBESMA, E. J. **Applied spatial data analysis with R**. Nova Iorque: Springer, 2008. v. 747248717.

BOCCHI, S.; CASTRIGNANÒ, A.; FORNARO, F.; MAGGIORE, T. Application of factorial kriging for mapping soil variation at field scale. **European Journal of Agronomy**, Elsevier, Amsterdã, v. 13, n. 4, p. 295–308, 2000. DOI: 10.1016/S1161-0301(00)00061-7.

BOGUNOVIC, I.; TREVISANI, S.; PEREIRA, P.; VUKADINOVIC, V. Mapping soil organic matter in the Baranja region (Croatia): Geological and anthropic forcing parameters. **Science of the Total Environment**, Elsevier, Amsterdã, v. 643, p. 335–345, 2018. DOI: 10.1016/j.scitotenv.2018.06.193.

BOLUWADE, A.; MADRAMOOTOO, C. Assessment of uncertainty in soil test phosphorus using kriging techniques and sequential Gaussian simulation: implications for water quality management in southern Quebec. **Water Quality Research Journal**, International Water Association, Londres, v. 48, n. 4, p. 344–357, 2013. DOI: 10.2166/wqrjc.2013.112.

BORÉM, A.; MARÇAL DE QUEIROZ, D.; SÁRVIO M. VALENTE, D.; ASSIS DE CARVALHO PINTO, F. **Agricultura digital**. São Paulo: Oficina de Textos, 2021.

BURGESS, T.; WEBSTER, R. Optimal interpolation and isarithmic mapping of soil properties: II block kriging. **Journal of Soil Science**, John Wiley & Sons, Nova Iorque, v. 31, n. 2, p. 333–341, 1980. DOI: 10.1111/j.1365-2389.1980.tb02085.x.

BUTTAFUOCO, G. et al. An approach to delineate management zones in a durum wheat field: validation using remote sensing and yield mapping. In: PRECISION agriculture'15. Nova Iorque: Wageningen Academic Publishers, 2015. P. 330. DOI: 10.3920/978-90-8686-814-8_29.

BUTTAFUOCO, G. et al. Geostatistical modelling of within-field soil and yield variability for management zones delineation: a case study in a durum wheat field. **Precision Agriculture**, Springer, Nova Iorque, v. 18, p. 37–58, 2017. DOI: 10.1007/s11119-016-9462-9.

BUTTAFUOCO, G. et al. Taking into account change of support when merging heterogeneous spatial data for field partition. **Precision Agriculture**, Springer, Nova Iorque, v. 22, p. 586–607, 2021. DOI: 10.1007/s11119-020-09781-9.

BUTTAFUOCO, G.; CASTRIGNANÒ, A.; COLECCHIA, A. S.; RICCA, N. Delineation of management zones using soil properties and a multivariate geostatistical approach. **Italian Journal of Agronomy**, PAGEPress, Roma, v. 5, n. 4, p. 323–332, 2010. DOI: 10.4081/ija.2010.323.

CAMARGO, A. P.; CAMARGO, M. B. P. Definition and outline for the phenological phases of arabic coffee under Brazilian tropical conditions. **Bragantia**, Instituto Agronômico, Campinas, v. 60, p. 65–68, 2001.

CAO, G.; YOO, E. H.; WANG, S. A statistical framework of data fusion for spatial prediction of categorical variables. **Stochastic Environmental Research and Risk Assessment**, Springer, Nova Iorque, v. 28, p. 1785–1799, 2014. DOI: 10.1007/s00477-013-0842-7.

CASTRIGNANO, A.; BUTTAFUOCO, G. Geostatistical stochastic simulation of soil water content in a forested area of south Italy. **Biosystems Engineering**, Elsevier, Amsterdã, v. 87, n. 2, p. 257–266, 2004. DOI: 10.1016/j.biosystemseng.2003.11.002.

CASTRIGNANÒ, A.; GIUGLIARINI, L.; RISALITI, R.; MARTINELLI, N. Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics. **Geoderma**, Elsevier, Amsterdã, v. 97, n. 1-2, p. 39–60, 2000. DOI: 10.1016/S0016-7061(00)00025-2.

CASTRIGNANÒ, A.; QUARTO, R.; VENEZIA, A.; BUTTAFUOCO, G. A comparison between mixed support kriging and block cokriging for modelling and combining spatial data with different support. **Precision Agriculture**, Springer, Nova Iorque, v. 20, p. 193–213, 2019. DOI: 10.1007/s11119-018-09630.

CASTRIGNANÒ, A. et al. A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture. **Catena**, Elsevier, Amsterdã, v. 167, p. 293–304, 2018. DOI: 10.1016/j.catena.2018.05.011.

CASTRIGNANÒ, A.; BUTTAFUOCO, G. Data processing. In: AGRICULTURAL Internet of Things and decision support for precision smart farming. Nova Iorque: Academic Press, 2020. P. 139–182. DOI: 10.1016/B978-0-12-818373-1.00003-2.

CASTRIGNANÒ, A.; COSTANTINI, E. A.; BARBETTI, R.; SOLLITTO, D. Accounting for extensive topographic and pedologic secondary information to improve soil mapping. **Catena**, Elsevier, Amsterdã, v. 77, n. 1, p. 28–38, 2009. DOI: 10.1016/j.catena.2008.12.004.

CASTRIGNANÒ, A. et al. A combined approach of sensor data fusion and multivariate geostatistics for delineation of homogeneous zones in an agricultural field. **Sensors**, Multidisciplinary Digital Publishing Institute, Basel, v. 17, n. 12, p. 2794, 2017. DOI: 10.3390/s17122794.

CASTRIGNANÒ, A. et al. A geostatistical fusion approach using UAV data for probabilistic estimation of Xylella fastidiosa subsp. pauca infection in olive trees. **Science of the Total Environment**, Elsevier, Amsterdã, v. 752, p. 141814, 2021. DOI: 10.1016/j.scitotenv.2020.141814.

CASTRIGNANÒ, A. et al. **Agricultural internet of things and decision support for precision smart farming**. Nova Iorque: Academic Press, 2020. DOI: 10.1016/c2018-0-00051-1.

CHANG, N. B.; BAI, K. **Multisensor data fusion and machine learning for environmental remote sensing**. Boca Raton: CRC Press, 2018. DOI: 10.1201/b20703.

CHEUNG, Y.-m. Rival penalization controlled competitive learning for data clustering with unknown cluster number. In: PROCEEDINGS of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02. Xangai: Institute of Electrical and Electronics Engineers, 2002. v. 1, p. 467–471. DOI: 10.1109/ICONIP.2002.1202214.

CHILES, J.-P.; DELFINER, P. **Geostatistics: modeling spatial uncertainty**. Nova Iorque: John Wiley & Sons, 2012. v. 713.

CHILÈS, J. P.; DELFINER, P. **Geostatistics: Modeling Spatial Uncertainty: Second Edition**. Hoboken: John Wiley & Sons, 2012. DOI: 10.1002/9781118136188.

CHRISTENSEN, W. F. Filtered kriging for spatial data with heterogeneous measurement error variances. **Biometrics**, John Wiley & Sons, Nova Iorque, v. 67, n. 3, p. 947–957, 2011. DOI: 10.1111/j.1541-0420.2011.01563.x.

CHUNG, F. L.; LEE, T. Fuzzy competitive learning. **Neural Networks**, v. 7, n. 3, p. 539–551, 1994.

CID-GARCIA, N. M.; ALBORNOZ, V.; RIOS-SOLIS, Y. A.; ORTEGA, R. Rectangular shape management zone delineation using integer linear programming. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 93, p. 1–9, 2013. DOI: 10.1016/j.compag.2013.01.009.

COLAÇO, A. F.; BRAMLEY, R. G. Do crop sensors promote improved nitrogen management in grain crops? **Field Crops Research**, Elsevier, Amsterdã, v. 218, n. 1, p. 126–140, 2018. DOI: 10.1016/j.fcr.2018.01.007.

COMERIO, N.; STROZZI, F. Tourism and its economic impact: A literature review using bibliometric tools. **Tourism economics**, SAGE, Nova Iorque, v. 25, n. 1, p. 109–131, 2019.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira de Café- Primeiro Levantamento**. 2023. Available from: <https://www.conab.gov.br/info-agro/safras/cafe>. Visited on: 3 Nov. 2023.

CÓRDOBA, M. A. et al. Protocol for multivariate homogeneous zone delineation in precision agriculture. **Biosystems Engineering**, Elsevier, Amsterdã, v. 143, p. 95–107, 2016. DOI: 10.1016/j.biosystemseng.2015.12.008.

CORREA, J.; GARCIA, A.; COSTA, P. Extração de nutrientes pelos cafeeiros Mundo Novo e Catuaí. In: 12 Congresso Brasileiro de Pesquisas Cafeeiras. Caxambu: Procafe, 1983.

COULSTON, J. W.; BLINN, C. E.; THOMAS, V. A.; WYNNE, R. H. Approximating prediction uncertainty for random forest regression models. **Photogrammetric Engineering & Remote Sensing**, Elsevier, Amsterdã, v. 82, n. 3, p. 189–197, 2016. DOI: 10.14358/PERS.82.3.189.

CRESSIE, N. Block kriging for lognormal spatial processes. **Mathematical Geology**, Springer, Nova Iorque, v. 38, p. 413–443, 2006. DOI: 10.1007/s11004-005-9022-8.

_____. Fitting variogram models by weighted least squares. **Journal of the international Association for mathematical Geology**, v. 17, n. 5, p. 563–586, 1985. DOI: 10.1007/BF01032109.

_____. **Statistics for spatial data**. Nova Iorque: John Wiley & Sons, 2015.

CRESSIE, N. A. Change of support and the modifiable areal unit problem. **Geographical Systems**, Springer, Nova Iorque, v. 3, n. 2-3, p. 159–180, 1996.

CZAPLEWSKI, R. L. **Expected value and variance of Moran's bivariate spatial autocorrelation statistic for a permutation test**. Nova Iorque: US Department of Agriculture, Forest Service, Rocky Mountain Forest, 1993. v. 309.

DAVE, R. N.; BHASWAN, K. Adaptive fuzzy c-shells clustering and detection of ellipses. **IEEE Transactions on Neural Networks**, Institute of Electrical and Electronics Engineers, Xangai, v. 3, n. 5, p. 643–662, 1992. DOI: 10.1109/72.159055.

DE BENEDETTO, D. et al. An approach for delineating homogeneous zones by using multisensor data. **Geoderma**, Elsevier, Amsterdã, v. 199, p. 117–127, 2013. DOI: 10.1016/j.geoderma.2012.08.028.

DE IACO, S.; MYERS, D.; POSA, D. Space–time variograms and a functional form for total air pollution measurements. **Computational Statistics & Data Analysis**, Springer, Nova Iorque, v. 41, n. 2, p. 311–328, 2002. DOI: 10.1016/S0167-9473(02)00081-6.

DE IACO, S.; POSA, D. Wind velocity prediction through complex kriging: formalism and computational aspects. **Environmental and ecological statistics**, Springer, Nova Iorque, v. 23, p. 115–139, 2016. DOI: 10.1007/s10651-015-0331-x.

DESBARATS, A.; DIMITRAKOPOULOS, R. Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. **Mathematical Geology**, Springer, Nova Iorque, v. 32, p. 919–942, 2000. DOI: 10.1023/A:1007570402430.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. **Machine learning**, Springer, Nova Iorque, v. 42, p. 143–175, 2001. DOI: 10.1023/A:1007612920971.

DIAS, L. O.; SILVA, M. S. Determinantes da demanda internacional por café brasileiro. **Revista de Política Agrícola**, Universidade de São Paulo, Piracicaba, v. 24, n. 1, p. 86–98, 2015.

DIMITRIADOU, E.; DIMITRIADOU, M. E. **The cclust Package**. Viena: Citeseer, 2007.

DIMITRIADOU, E. et al. The e1071 package. **Misc Functions of Department of Statistics (e1071), TU Wien**, p. 297–304, 2006.

DIXIT, R.; NASKAR, R. Region duplication detection in digital images based on Centroid Linkage Clustering of key–points and graph similarity matching. **Multimedia Tools and Applications**, Springer, Nova Iorque, v. 78, p. 13819–13840, 2019. DOI: 10.1007/s11042-018-6666-1.

DOBERMANN, A. et al. Classification of crop yield variability in irrigated production fields. **Agronomy Journal**, Springer, Nova Iorque, v. 95, n. 5, p. 1105–1120, 2003. DOI: 10.2134/agronj2003.1105.

DONTHU, N. et al. How to conduct a bibliometric analysis: An overview and guidelines. **Journal of Business Research**, Elsevier, Amsterdã, v. 133, p. 285–296, 2021. DOI: 10.1016/j.jbusres.2021.04.070.

DRAY, S.; SAID, S.; DEBIAS, F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. **Journal of vegetation science**, John Wiley & Sons, Nova Iorque, v. 19, n. 1, p. 45–56, 2008. DOI: 10.3170/2007-8-18312.

DRAY, S. et al. Package 'adespatial'. **R package**, v. 2018, p. 3–8, 2018.

DRIEMEIER, C. et al. A computational environment to support research in sugarcane agriculture. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 130, p. 13–19, 2016. DOI: 10.1016/j.compag.2016.10.002.

ELBASIOUNY, H.; ABOWALY, M.; ABU ALKHEIR, A.; GAD, A. Spatial variation of soil carbon and nitrogen pools by using ordinary Kriging method in an area of north Nile Delta, Egypt. **Catena**, Elsevier, Amsterdã, v. 113, p. 70–78, 2014. DOI: 10.1016/j.catena.2013.09.008.

EMADI, M. et al. Geostatistics-based spatial distribution of soil moisture and temperature regime classes in Mazandaran province, northern Iran. **Archives of Agronomy and Soil Science**, John Wiley & Sons, Nova Iorque, v. 62, n. 4, p. 502–522, 2016. DOI: 10.1080/03650340.2015.1065607.

EMERY, X. On some consistency conditions for geostatistical change-of-support models. **Mathematical geology**, Springer, Nova Iorque, v. 39, n. 2, p. 205–223, 2007.

ESRI. ArcGIS Pro Advanced 2.8. **Redlands, CA: Environmental Systems Research Institute**, Environmental Systems Research Institute, Redlands, CA, 2022.

FERRAZ, G. A. et al. Variabilidade espacial da força de desprendimento de frutos do cafeeiro. **Engenharia Agrícola**, Universidade de São Paulo, Piracicaba, v. 34, n. 6, p. 1210–1223, 2014. DOI: 10.1590/s0100-69162014000600016.

FERRAZ, G. A. S. et al. Variabilidade espacial e temporal do fósforo, potássio e da produtividade de uma lavoura cafeeira. **Engenharia Agrícola**, Universidade Federal de Lavras, Lavras, v. 32, p. 140–150, 2012.

FERRAZ, G. A. et al. Spatial variability of plant attributes in a coffee plantation. **Revista Ciência Agronômica**, Universidade Federal do Ceará, Fortaleza, v. 48, n. 1, p. 81–91, 2017. DOI: 10.5935/1806-6690.20170009.

FERRAZ, G. A. e. S. et al. Geostatistical analysis of fruit yield and detachment force in coffee. **Precision Agriculture**, Springer, Nova Iorque, v. 13, n. 1, p. 76–89, 2012. DOI: 10.1007/s11119-011-9223-8.

FLEMING, K.; WESTFALL, D.; WIENS, D.; BRODAHL, M. Evaluating farmer defined management zone maps for variable rate fertilizer application. **Precision Agriculture**, Springer, Nova Iorque, v. 2, n. 2, p. 201–215, 2000. DOI: 10.1023/A:1011481832064.

FOUEDJIO, F.; KLUMP, J. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. **Environmental Earth Sciences**, Springer, Nova Iorque, v. 78, n. 1, p. 38, 2019. DOI: 10.1007/s12665-018-8032-z.

FRAGA, C. C. Resenha histórica do café no Brasil. **Agricultura em São Paulo**, v. 10, n. 1, p. 1–21, 1963. DOI: 10.4067/S0718-95162013005000044.

FU, W.; TUNNEY, H.; ZHANG, C. Spatial variation of soil nutrients in a dairy farm and its implications for site-specific fertilizer application. **Soil Tillage Research**, Elsevier, Amsterdã, v. 106, n. 2, p. 185–193, 2010. DOI: 10.1016/j.still.2009.12.001.

GAVIOLI, A. et al. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. **Biosystems engineering**, Elsevier, Amsterdã, v. 181, p. 86–102, 2019. DOI: 10.1016/j.biosystemseng.2019.02.019.

GAVIOLI, A. et al. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 127, p. 302–310, 2016. DOI: 10.1016/j.compag.2016.06.029.

GELFAND, A. E.; ZHU, L.; CARLIN, B. P. On the change of support problem for spatio-temporal data. **Biostatistics**, Springer, Nova Iorque, v. 2, n. 1, p. 31–45, 2001.

GEORGI, C.; SPENGLER, D.; ITZEROTT, S.; KLEINSCHMIT, B. Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. **Precision Agriculture**, Springer, Nova Iorque, v. 19, p. 684–707, 2018. DOI: 10.1007/s11119-017-9549-y.

GÉOVARIANCES. **Isatis.neo Technical Ref. 2023.04.01**. Paris: Geovariances & Ecole Des Mines De Paris Avon Cedex, 2023.

GESSLER, P. et al. Modeling soil–landscape and ecosystem properties using terrain attributes. **Soil Science Society of America Journal**, John Wiley & Sons, Londres, v. 64, n. 6, p. 2046–2056, 2000. DOI: 10.2136/sssaj2000.6462046x.

GILI, A.; ÁLVAREZ, C.; BAGNATO, R.; NOELLEMEYER, E. Comparison of three methods for delineating management zones for site-specific crop management. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 139, p. 213–223, 2017. DOI: 10.1016/j.compag.2017.05.022.

GOOVAERTS, P. **Geostatistics for natural reources evaluation**. Nova Iorque: Oxford University Press, 1997.

GOOVAERTS, P. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. **Journal of soil science**, Elsevier, Amsterdã, v. 43, n. 4, p. 597–619, 1992. DOI: 10.1111/j.1365-2389.1992.tb00163.x.

_____. Geostatistical modelling of uncertainty in soil science. **Geoderma**, Elsevier, Amsterdã, v. 103, n. 1-2, p. 3–26, 2001. DOI: 10.1016/S0016-7061(01)00067-2.

_____. Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. **Biology and Fertility of Soils**, Springer, Nova Iorque, v. 27, p. 315–334, 1998. DOI: 10.1007/s003740050439.

_____. Geostatistics in soil science: state-of-the-art and perspectives. **Geoderma**, Elsevier, Amsterdã, v. 89, n. 1-2, p. 1–45, 1999. DOI: 10.1016/S0016-7061(98)00078-0.

GOSAIN, A.; DAHIYA, S. Performance analysis of various fuzzy clustering algorithms: a review. **Procedia Computer Science**, Elsevier, Amsterdã, v. 79, p. 100–111, 2016. DOI: 10.1016/j.procs.2016.03.014.

GOWER, J. C.; ROSS, G. J. Minimum spanning trees and single linkage cluster analysis. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Journal Storage, Nova Iorque, v. 18, n. 1, p. 54–64, 1969. DOI: 10.2307/2346439.

GRÄLER, B.; PEBESMA, E. J.; HEUVELINK, G. B. Spatio-temporal interpolation using gstat. **R Journal**, R Foundation for Statistical Computing, Viena, v. 8, n. 1, p. 204, 2016.

GUASTAFERRO, F. et al. A comparison of different algorithms for the delineation of management zones. **Precision Agriculture**, Springer, Nova Iorque, v. 11, n. 6, p. 600–620, 2010. DOI: 10.1007/s11119-010-9183-4.

GUIMARÃES, P. T. G. et al. Cafeeiro. In: RIBEIRO, A. C.; GUIMARÃES, P. T. G.; ALVAREZ, V. V. H. (Eds.). **Recomendações para o uso de corretivos e fertilizantes em Minas Gerais**. Viçosa: Comissão de Fertilidade do Solo do Estado de Minas Gerais, 1999. P. 289–302.

HALL, D. L.; MCMULLEN, S. A. H. **Mathematical Techniques in Multisensor Data Fusion**. 2. ed. Londres: Artech House Publishers, 2004.

HAMZEHPOUR, N. et al. Spatial prediction of soil salinity using kriging with measurement errors and probabilistic soft data. **Arid land research and management**, Springer, Nova Iorque, v. 27, n. 2, p. 128–139, 2013. DOI: 10.1007/978-90-481-2322-3_26.

HANSEN, P.; DELATTRE, M. Complete-link cluster analysis by graph coloring. **Journal of the American Statistical Association**, v. 73, n. 362, p. 397–403, 1978.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. Nova Iorque: Springer, 2009. v. 2.

HENGL, T.; HEUVELINK, G. B.; PERČEC TADIĆ, M.; PEBESMA, E. J. Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. **Theoretical and applied climatology**, Springer, Nova Iorque, v. 107, p. 265–277, 2012. DOI: 10.1007/s00704-011-0464-2.

HIEMSTRA, P. H.; PEBESMA, E. J.; TWENHÖFEL, C. J.; HEUVELINK, G. B. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. **Computers & Geosciences**, Elsevier, Amsterdã, v. 35, n. 8, p. 1711–1721, 2009. DOI: 10.1016/j.cageo.2008.10.011.

HORNIK, K.; FEINERER, I.; KOBER, M.; BUCHTA, C. Spherical k-means clustering. **Journal of statistical software**, Elsevier, Amsterdã, v. 50, p. 1–22, 2012.

HORNIK, K.; FEINERER, I.; KOBER, M.; HORNIK, M. K. **Package 'skmeans'**. Viena: R Foundation for Statistical Computing, 2017.

HU, J.; ZHANG, L.; LEE, C.; GUI, R. Advanced big SAR data analytics and applications. **Frontiers in Environmental Science**, Frontiers, Pequim, v. 10, p. 2097, 2022. DOI: 10.3389/fenvs.2022.1063376.

ISAAKS, E. H.; SRIVASTAVA, M. R. **Applied geostatistics**. Nova Iorque: Oxford University Press, 1989.

ISPAG. **Precision Ag Definition**. Nova Iorque: ISPAG, 2019. https://www.ispag.org/about/definition/. Accessed: 2022-21-04.

JACINTHO, J. L.; FERRAZ, G. A.; SILVA, F. M.; SANTOS, S. A. Definição de zonas de manejo para cafeicultura. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Universidade Federal de Campina Grande, Campina Grande, v. 21, n. 2, p. 94–99, 2017. DOI: 10.1590/1807-1929/agriambi.v21n2p94-99.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Nova Iorque: Prentice-Hall, Inc., 1988.

JIANG, H.-L. et al. Delineation of site-specific management zones based on soil properties for a hillside field in central China. **Archives of Agronomy and Soil Science**, Taylor & Francis, Nova Iorque, v. 58, n. 10, p. 1075–1090, 2012. DOI: 10.1080/03650340.2011.570337.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. Nova Iorque: Springer, 1976.

JUANG, K.-W.; LEE, D.-Y. Comparison of three nonparametric kriging methods for delineating heavy-metal contaminated soils. **Journal of Environmental Quality**, Elsevier, Amsterdã, v. 21, n. 1, p. 197–205, 2000. DOI: 10.2134/jeq2000.00472425002900010025x.

KANG, J.; JIN, R.; LI, X.; ZHANG, Y. Block kriging with measurement errors: A case study of the spatial prediction of soil moisture in the middle reaches of Heihe River Basin. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers, Xangai, v. 14, n. 1, p. 87–91, 2016. DOI: 10.1109/LGRS.2016.2628767.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. Nova Iorque: John Wiley & Sons, 2009.

KHOSLA, R. et al. Use of site-specific management zones to improve nitrogen management for precision agriculture. **Journal of Soil and Water Conservation**, Soil and Water Conservation Society, Nova Iorque, v. 57, n. 6, p. 513–518, 2002.

KLASTORIN, T. D. The p-median problem for cluster analysis: A comparative test using the mixture model approach. **Management Science**, American Psychological Association, Washington, v. 31, n. 1, p. 84–95, 1985. DOI: 10.1037/a0018535.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; WASSERMAN, W. **Applied linear regression models**. Nova Iorque: McGraw-Hill, 2004. v. 4.

LAJAUNIE, C. L'estimation géostatistique non linéaire. **Cours C-152, Centre de Géostatistique, Ecole des Mines de Paris**, Centre de morphologie mathématique de Fontainebleau, Fontainebleau, 1993.

LAMBERT, D.; LOWENBERG-DEBOER, J.; MALZER, G. Understanding phosphorous in Minnesota soils. **Agricultural Economics**, Elsevier, Amsterdã, v. 37, n. 1, p. 43–53, 2007. DOI: 10.1111/j.1574-0862.2007.00221.x.

LARK, R. A comparison of some robust estimators of the variogram for use in soil survey. **European journal of soil science**, John Wiley & Sons, Nova Iorque, v. 51, n. 1, p. 137–157, 2000. DOI: 10.1046/j.1365-2389.2000.00280.x.

LEGENDRE, P.; FORTIN, M. J. Spatial pattern and ecological analysis. **Vegetatio**, Elsevier, Amsterdã, v. 80, n. 2, p. 107–138, 1989.

LI, Y.; SHI, Z.; LI, F.; LI, H.-Y. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. **Computers and Electronics in Agriculture**, Elsevier, Amstewrdã, v. 56, n. 2, p. 174–186, 2007.

LIMA, J. S. d. S.; SILVA, S. d. A.; OLIVEIRA, R. B. d.; FONSECA, A. S. d. Estimativa da produtividade de café conilon utilizando técnicas de cokrigagem. **Ceres**, Universidade Federal de Viçosa, Viçosa, v. 63, n. 1, p. 54–61, 2016. DOI: 10.1590/0034-737X201663010008.

LIU, H. et al. A Novel Branch and Bound Pure Integer Programming Phase Unwrapping Algorithm for Dual-Baseline InSAR. **Frontiers in Environmental Science**, Frontiers, Pequim, v. 10, p. 890343, 2022. DOI: 10.3389/fenvs.2022.890343.

LLOYD, C.; ATKINSON, P. M. Assessing uncertainty in estimates with ordinary and indicator kriging. **Computer & Geosciences**, Elsevier, Amsterdã, v. 27, n. 8, p. 929–937, 2001. DOI: 10.1016/S0098-3004(00)00132-1.

LOWENBERG-DEBOER, J.; ERICKSON, B. Setting the record straight on precision agriculture adoption. **Agronomy Journal**, Springer, Nova Iorque, v. 111, n. 4, p. 1552–1569, 2019. DOI: 10.2134/agronj2018.12.0779.

LV, J.; LIU, Y.; ZHANG, Z.; DAI, J. Factorial kriging and stepwise regression approach to identify environmental factors influencing spatial multi-scale variability of heavy metals in soils. **Journal of hazardous materials**, Elsevier, Amsterdã, v. 261, p. 387–397, 2013. DOI: 10.1016/j.jhazmat.2013.07.065.

MA, Y. et al. Factorial kriging for multiscale modelling. **Journal of the Southern African Institute of Mining and Metallurgy**, The Southern African Institute of Mining and Metallurgy, v. 114, n. 8, p. 651–659, 2014.

MACHADO, R. D. et al. Generation of 441 typical meteorological year from INMET stations-Brazil. In: PROCEEDINGS of the IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry. Nova Iorque: IEA SHC, 2019.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA, 14. PROCEEDINGS of the fifth Berkeley symposium on mathematical statistics and probability. [S.l.: s.n.], 1967. v. 1, p. 281–297.

MAECHLER, M. et al. Package 'cluster'. **Dosegljivo na**, Citeseer, 2013.

MANZIONE, R. et al. Spatio-temporal Kriging to Predict Water Table Depths from Monitoring Data in a Conservation Area at São Paulo State, Brazil. 7: 1. **Geoinformatics and Geostatistics: An Overview**, John Wiley & Sons, Londres, v. 10, n. 1, p. 2, 2019. DOI: 10.4172/2327-4581.1000205.

MANZIONE, R. L.; CASTRIGNANÒ, A. A geostatistical approach for multi-source data fusion to predict water table depth. **Science of the Total Environment**, Elsevier, Amsterdã, v. 696, p. 133763, 2019. DOI: 10.1016/j.scitotenv.2019.133763.

MANZIONE, R. L.; SILVA, C. O. F.; CASTRIGNANÒ, A. A combined geostatistical approach of data fusion and stochastic simulation for probabilistic assessment of shallow water table depth risk. **Science of the Total Environment**, Elsevier, Amsterdã, v. 765, p. 142743, 2020. DOI: 10.1016/j.scitotenv.2020.142743.

MAPBIOMAS. **Plataforma MapBiomas Brasil**. 2021. Available from: <https://plataforma.brasil.mapbiomas.org/>. Visited on: 21 Jan. 2022.

MARTINETZ, T. M.; BERKOVICH, S. G.; SCHULTEN, K. J. 'Neural-gas' network for vector quantization and its application to time-series prediction. **IEEE transactions on neural networks**, Institute of Electrical and Electronics Engineers, Xangai, v. 4, n. 4, p. 558–569, 1993. DOI: 10.1109/72.238311.

MARTÍNEZ-CASASNOVAS, J. A.; ESCOLÀ, A.; ARNÓ, J. Use of farmer knowledge in the delineation of potential management zones in precision agriculture: A case study in maize (Zea

mays L.) **Agriculture**, Multidisciplinary Digital Publishing Institute, Basel, v. 8, n. 6, p. 84, 2018. DOI: 10.3390/agriculture8060084.

MATHERON, G. La théorie des variables régionalisées et ses applications. **Cahiers du Centre de morphologie mathématique de Fontainebleau**, Centre de morphologie mathématique de Fontainebleau, Fontainebleau, v. 5, p. 1–212, 1970.

MCBRATNEY, A.; WHELAN, B.; ANCEV, T.; BOUMA, J. Future directions of precision agriculture. **Precision Agriculture**, Springer, Nova Iorque, v. 6, n. 1, p. 7–23, 2005. DOI: 10.1007/s11119-005-0681-8.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. **Psychometrika**, Springer, Nova Iorque, v. 12, n. 2, p. 153–157, 1947.

MCQUITTY, L. L. Capabilities and improvements of linkage analysis as a clustering method. **Educational and Psychological Measurement**, Springer, Nova Iorque, v. 24, n. 3, p. 441–456, 1964.

_____. Similarity analysis by reciprocal pairs for discrete and continuous data. **Educational and Psychological measurement**, Springer, Nova Iorque, v. 26, n. 4, p. 825–831, 1966.

METWALLY, M. S. et al. Soil properties spatial variability and delineation of site-specific management zones based on soil fertility using fuzzy clustering in a hilly field in Jianyang, Sichuan, China. **Sustainability**, Multidisciplinary Digital Publishing Institute, Basel, v. 11, n. 24, p. 7084, 2019. DOI: 10.3390/su11247084.

MIAO, Y.; MULLA, D. J.; ROBERT, P. C. An integrated approach to site-specific management zone delineation. **Frontiers of Agricultural Science and Engineering**, Frontiers, Pequim, v. 5, n. 4, p. 432–441, 2018. DOI: 10.15302/J-FASE-2018230.

MOHARANA, P. et al. Geostatistical and fuzzy clustering approach for delineation of site-specific management zones and yield-limiting factors in irrigated hot arid environment of India. **Precision Agriculture**, Springer, Nova Iorque, v. 21, p. 426–448, 2020. DOI: 10.1007/s11119-019-09671-9.

MOORE, I. D.; GESSLER, P. E.; NIELSEN, G.; PETERSON, G. Soil attribute prediction using terrain analysis. **Soil Science Society of America Journal**, John Wiley & Sons, Nova Iorque, v. 57, n. 2, p. 443–452, 1993. DOI: 10.2136/sssaj1993.03615995005700020026x.

MORARI, F.; CASTRIGNANÒ, A.; PAGLIARIN, C. Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 68, n. 1, p. 97–107, 2009. DOI: 10.1016/j.compag.2009.05.003.

MOREIRA SILVA, F.; ALVES, M. d. C. **Cafeicultura de Precisão**. Lavras: Universidade Federal de Lavras, 2013.

MÜLLNER, D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. **Journal of Statistical Software**, Elsevier, Amsterdã, v. 53, p. 1–18, 2013.

NAYAK, J.; NAIK, B.; BEHERA, H. Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In: COMPUTATIONAL Intelligence in Data Mining-Volume 2:

Proceedings of the International Conference on CIDM, 20-21 December 2014. Nova Iorque: Springer, 2015. P. 133–149.

NGUYEN, H.; KATZFUSS, M.; CRESSIE, N.; BRAVERMAN, A. Spatio-temporal data fusion for very large remote sensing datasets. **Technometrics**, Springer, Nova Iorque, v. 56, n. 2, p. 174–185, 2014.

NÚÑEZ, P. et al. Soil fertility evaluation of coffee (Coffea spp.) production systems and management recommendations for the Barahona Province, Dominican Republic. **Journal of Soil Science and Plant Nutrition**, Universidade Federal do Ceará, Fortaleza, v. 11, n. 1, p. 127–140, 2011.

OHANA-LEVI, N. et al. A weighted multivariate spatial clustering model to determine irrigation management zones. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 162, p. 719–731, 2019. DOI: 10.1016/j.compag.2019.05.012.

OIC. **Relatório sobre o mercado de café– fevereiro de 2024**. 2024. Available from: <https://www.consorciopesquisacafe.com.br/images/stories/noticias/2021/2024/Fevereiro/relatorio_oic_fevereiro_2024.pdf/>. Visited on: 13 Apr. 2024.

OLDONI, H. et al. Homogeneous zones of vegetation index for characterizing variability and site-specific management in vineyards. **Engenharia Agrícola**, Universidade de São Paulo, Piracicaba, v. 78, e20190243, 2020. DOI: 10.1590/1678-992X-2019-0243.

OLEA, R. A. **Geostatistical Glossary and Multilingual Dictionary**. Nova Iorque: Oxford University Press, 1991.

OLIVER, M. A.; WEBSTER, R. **Basic steps in geostatistics: the variogram and kriging**. Nova Iorque: Springer, 2015.

OPENDRONEMAP. **OpenDroneMap**. GitHub Repository: eMap, 2020. https://github.com/OpenDroneMap/ODM/. Accessed: 2022-21-04.

ORTEGA, R.; SANTIBÁÑEZ, O. Agronomic evaluation of three zoning methods based on soil fertility in corn crops (Zea maize L.) **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 58, n. 1, p. 49–59, 2007. DOI: 10.1016/j.compag.2006.12.011.

ORTEGA, R. A.; SANTIBANEZ, O. A. Determination of management zones in corn (Zea mays L.) based on soil fertility. **Computers and Electronics in agriculture**, Elsevier, Amsterdã, v. 58, n. 1, p. 49–59, 2007. DOI: 10.1016/j.compag.2006.12.011.

PABON, C. D. R.; SÁNCHEZ-BENITEZ, J.; RUIZ-ROSERO, J.; RAMIREZ-GONZALEZ, G. Coffee crop science metric: A review. **Coffee Science**, Universidade Federal de Lavras, Lavras, v. 15, e151693, 2020. DOI: 10.25186/.v15i.1693.

PAL, N. R.; BEZDEK, J. C.; HATHAWAY, R. J. Sequential competitive learning and the fuzzy c-means clustering algorithms. **Neural Networks**, Elsevier, Amsterdã, v. 9, n. 5, p. 787–796, 1996.

PALLOTTINO, F. et al. Science mapping approach to analyze the research evolution on precision agriculture: World, EU and Italian situation. **Precision Agriculture**, Springer, Nova Iorque, v. 19, n. 6, p. 1011–1026, 2018. DOI: 10.1007/s11119-018-9569-2.

PAM/IBGE. **Produção Agrícola Municipal – Instituto Brasileiro de Geografia e Estatística**. 2023. Available from: <https://sidra.ibge.gov.br/pesquisa/pam/tabelas>. Visited on: 21 Jan. 2022.

PARIS, J. O.; GONTIJO, I.; PARTELLI, F. L.; FACCO, A. G. Variability and spatial correlation of soil micronutrients and organic matter with macadamia nut production. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Universidade Federal de Campina Grande, Campina Grande, v. 24, p. 31–36, 2019. DOI: 10.1590/1807-1929/agriambi.v24n1p31-36.

PERALTA, N. R.; COSTA, J. L. Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 99, p. 218–226, 2013. DOI: 10.1016/j.compag.2013.09.014.

PEREIRA, G. W. et al. Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. **Precision Agriculture**, Springer, Nova Iorque, p. 1–16, 2022. DOI: 10.1007/s11119-022-09880-9.

PIERCE, F. J.; NOWAK, P. Aspects of precision agriculture. **Advances in Agronomy**, Elsevier, Amsterdã, v. 67, p. 1–85, 1999. DOI: 10.1016/S0065-2113(08)60513-1.

PRIMAVESI, A. **Manejo ecológico do solo: a agricultura em regiões tropicais**. São Paulo: Nobel, 2002.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System version 3.28.3**. Viena: Open Source Geospatial Foundation, 2023. Available from: <https://www.qgis.org/en/site/>.

QIN, A. K.; SUGANTHAN, P. N. Kernel neural gas algorithms with application to cluster analysis. In: PROCEEDINGS of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Xangai: ICPR, 2004. v. 4, p. 617–620. DOI: 10.1109/ICPR.2004.1333848.

R DEVELOPMENT CORE TEAM, R. **R: A Language and Environment for Statistical Computing**. Viena: R Foundation for Statistical Computing, 2022.

RAI, P.; SINGH, S. A survey of clustering techniques. **International Journal of Computer Applications**, Springer, Nova Iorque, v. 7, n. 12, p. 1–5, 2010. DOI: 10.5120/1326-1808.

REBOITA, M. S.; RODRIGUES, M.; SILVA, L. F.; ALVES, M. A. Aspectos climáticos do estado de Minas Gerais. **Revista brasileira de Climatologia**, Instituto Agronômico, Campinas, v. 17, 2015. DOI: 10.5380/abclima.v17i0.41493.

RHIF, M.; ABBES, A. B.; FARAH, I. R. A Non-stationary NDVI Time Series with Big Data: A Deep Learning Approach. In: SPRINGER. CONFERENCE of the Arabian Journal of Geosciences. Nova Iorque: [s.n.], 2019. P. 357–359. DOI: 10.1007/978-3-030-72896-0_81.

RHIF, M.; BEN ABBES, A.; MARTINEZ, B.; FARAH, I. R. A deep learning approach for forecasting non-stationary big remote sensing time series. **Arabian Journal of Geosciences**, Springer, Nova Iorque, v. 13, n. 22, p. 1174, 2020. DOI: 10.1007/s12517-020-06140-w.

RIVOIRARD, J. **Introduction to disjunctive kriging and non-linear geostatistics**. Londres: Clarendon Press, 1994.

ROBERT, P. C. Precision agriculture: a challenge for crop nutrition management. In: PROGRESS in Plant Nutrition: Plenary Lectures of the XIV International Plant Nutrition Colloquium. Nova Iorque: Springer, 2002. P. 143–149.

RODRIGUES, M. S. et al. Geostatistics and its potential in Agriculture 4.0. **Revista Ciência Agronômica**, Universidade Federal do Ceará, Fortaleza, v. 51, p. 2, 2021. DOI: 10.5935/1806-6690.20200095.

ROGOVA, G. L.; NIMIER, V. Reliability in information fusion: literature survey. In: PROCEEDINGS of the seventh international conference on information fusion. Xangai: Institute of Electrical and Electronics Engineers, 2004. v. 2, p. 1158–1165.

RONQUIM, C. C. Conceitos de fertilidade do solo e manejo adequado para as regiões tropicais. **Embrapa Territorial-Boletim de Pesquisa e Desenvolvimento (INFOTECA-E)**, Campinas, SP: Embrapa Monitoramento por Satélite, 2010., 2010.

ROSSEL, R. V. et al. Proximal soil sensing: An effective approach for soil measurements in space and time. **Advances in Agronomy**, Elsevier, Amsterdã, v. 113, p. 243–291, 2011. DOI: 10.1016/B978-0-12-386473-4.00005-1.

ROZNIK, M.; BOYD, M.; PORTH, L. Improving crop yield estimation by applying higher resolution satellite NDVI imagery and high-resolution cropland masks. **Remote Sensing Applications: Society and Environment**, Elsevier, Amsterdã, v. 25, p. 100693, 2022. DOI: 10.1016/j.rsase.2022.100693.

RUSS, G.; KRUSE, R. Exploratory hierarchical clustering for management zone delineation in precision agriculture. In: SPRINGER. INDUSTRIAL conference on data mining. [S.l.: s.n.], 2011. P. 161–173. DOI: 10.1007/978-3-642-23184-1_13.

SANCHES, G. M.; MAGALHÃES, P. S.; REMACRE, A. Z.; FRANCO, H. C. Potential of apparent soil electrical conductivity to describe the soil pH and improve lime application in a clayey soil. **Soil and Tillage Research**, Elsevier, Amsterdã, v. 175, p. 217–225, 2018. DOI: 10.1016/j.still.2017.09.010.

SANTOS, H. G. et al. **Sistema brasileiro de classificação de solos**. Rio de Janeiro: Centro Nacional de Pesquisa de Solos, 2006.

SASIADEK, J. Z. Sensor fusion. **Annual Reviews in Control**, Elsevier, Amsterdã, v. 26, n. 2, p. 203–228, 2002.

SAVELYEVA, E.; UTKIN, S.; KAZAKOV, S.; DEMYANOV, V. Modeling spatial uncertainty for locally uncertain data. **geoENV VII–Geostatistics for Environmental Applications**, p. 295–306, 2010. DOI: 10.1007/978-90-481-2322-3_26.

SCA. **What is Specialty Coffee?** Irvine: 505 Technology Drive, 2021. https://sca.coffee/research/what-is-specialty-coffee/. Accessed: 2022-11-06.

SCHEPERS, A. R. et al. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. **Agronomy Journal**, John Wiley & Sons, Hoboken, v. 96, n. 1, p. 195–203, 2004. DOI: 10.2134/agronj2004.1950.

SCUDIERO, E. et al. Delineation of site-specific management units in a saline region at the Venice Lagoon margin, Italy, using soil reflectance and apparent electrical conductivity. **Computers and Electronics in Agriculture**, Elsevier, Amsterdã, v. 99, p. 54–64, 2013. DOI: 10.1016/j.compag.2013.08.023.

SCUDIERO, E. et al. Workflow to establish time-specific zones in precision agriculture by spatiotemporal integration of plant and soil sensing data. **Agronomy**, Multidisciplinary Digital Publishing Institute, Basel, v. 8, n. 11, p. 253, 2018. DOI: 10.3390/agronomy8110253.

SEIFODDINI, H. K. Single linkage versus average linkage clustering in machine cells formation applications. **Computers & Industrial Engineering**, Elsevier, Amsterdã, v. 16, n. 3, p. 419–426, 1989. DOI: 10.1016/0360-8352(89)90160-5.

SEKULIĆ, A. et al. Spatio-temporal regression kriging model of mean daily temperature for Croatia. **Theoretical and Applied Climatology**, Springer, Nova Iorque, v. 140, p. 101–114, 2020. DOI: 10.1007/s00704-019-03077-3.

SHADDAD, S. M.; BUTTAFUOCO, G.; CASTRIGNANÒ, A. Assessment and mapping of soil salinization risk in an Egyptian field using a probabilistic approach. **Agronomy**, Multidisciplinary Digital Publishing Institute, Basel, v. 10, n. 1, p. 85, 2020. DOI: 10.3390/agronomy10010085.

SHE, D.; FEI, Y.; CHEN, Q.; TIMM, L. C. Spatial scaling of soil salinity indices along a temporal coastal reclamation area transect in China using wavelet analysis. **Archives of Agronomy and Soil Science**, John Wiley & Sons, Nova Iorque, v. 62, n. 12, p. 1625–1639, 2016. DOI: 10.1080/03650340.2016.1155698.

SILVA, C. d. O. F.; GREGO, C. R.; MANZIONE, R. L.; OLIVEIRA, S. R. d. M. Combining geostatistical and clustering modeling strategies for delineating specialty coffee management zones under low sampling. **Precision Agriculture**, Elsevier, Amsterdã, in press, 2024.

_____. Improving coffee yield interpolation in the presence of outliers using multivariate geostatistics and satellite data. **AgriEngineering**, Multidisciplinary Digital Publishing Institute, Basel, v. 6, n. 1, p. 81–94, 2024. DOI: 10.3390/agriengineering6010006.

SILVA, C. d. O. F.; MANZIONE, R. L.; OLIVEIRA, S. R. d. M. Exploring 20-year applications of geostatistics in precision agriculture in Brazil: what's next? **Precision Agriculture**, Springer, Nova Iorque, v. 24, p. 2293–2326, 2023. DOI: 10.1007/s11119-023-10041-9.

SILVA, C. d. O. F. et al. Summarizing soil chemical variables into homogeneous management zones – case study in a specialty coffee crop. **Smart Agricultural Technology**, Elsevier, Amsterdã, v. 7, p. 100418, 2024. DOI: 10.1016/j.atech.2024.100418.

SILVA, F. M.; ALVES, M. C.; SOUZA, J. C. S.; OLIVEIRA, M. S. d. Effects of manual harvesting on coffee (coffea arabica L.) crop biannuality in Ijaci, Minas Gerais. **Ciência e Agrotecnologia**, Universidade Federal de Lavras, Lavras, v. 34, p. 625–632, 2010.

SILVA, M. G. da et al. Em busca de sabores, aromas e histórias: uma revisão integrativa acerca dos cafés especiais. **Cadernos de Ciência & Tecnologia**, v. 38, n. 3, p. 26879, 2021. DOI: 10.35977/0104-1096.cct2021.v38.26879.

SILVA, S. A.; LIMA, J. S. S. Multivariate analysis and geostatistics of the fertility of a humic rhodic hapludox under coffee cultivation. **Revista Brasileira de Ciência do Solo**, Sociedade Brasileira de Ciência do Solo, Viçosa, v. 36, n. 2, p. 467–474, 2012.

_____. Relação espacial entre o estoque de nutrientes e a densidade de solo cultivado com cafeeiro. **Pesquisa Agropecuária Tropical**, Embrapa, Brasília, v. 43, n. 4, p. 377–384, 2013.

SOUZA, L. T. d. **Biomass partition and nutrient demand of coffee in different conditions and fertilization according to nutrient demand from simultaneous sinks-fruits and vegetative growth**. 2022. PhD thesis – Universidade de São Paulo.

SU, B.; ZHAO, G.; DONG, C. Spatiotemporal variability of soil nutrients and the responses of growth during growth stages of winter wheat in northern China. **PloS one**, Public Library of Science San Francisco, CA USA, São Francisco, v. 13, n. 12, e0203509, 2018. DOI: 10.1371/journal.pone.0203509.

SUGANYA, R.; SHANTHI, R. Fuzzy c-means algorithm-a review. **International Journal of Scientific and Research Publications**, Springer, Nova Iorque, v. 2, n. 11, p. 1, 2012.

SWITZER, P.; GREEN, A. A. Min/max autocorrelation factors for multivariate spatial imagery. **Computer science and statistics**, Springer, Nova Iorque, v. 32, p. 919–942, 1985.

TAKAFUJI, E. H. d. M.; ROCHA, M. M. da; MANZIONE, R. L. Spatiotemporal forecast with local temporal drift applied to weather patterns in Patagonia. **SN Applied Sciences**, Springer, Nova Iorque, v. 2, n. 6, p. 1001, 2020. DOI: 10.1007/s42452-020-2814-0.

_____. Groundwater level prediction/forecasting and assessment of uncertainty using SGS and ARIMA models: a case study in the Bauru Aquifer System (Brazil). **Natural Resources Research**, Springer, Nova Iorque, v. 28, n. 2, p. 487–503, 2019. DOI: 10.1007/s11053-018-9403-6.

TAKAKU, J.; TADONO, T.; TSUTSUI, K. Generation of high resolution global DSM from ALOS PRISM. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, Elsevier, Amsterdã, v. 40, p. 243–248, 2014. DOI: 10.5194/isprsarchives-XL-4-243-2014.

THOMAS, G. W. Soil and climatic factors which affect nutrient mobility. **Nutrient mobility in soils: accumulation and losses**, John Wiley & Sons, Nova Iorque, v. 4, p. 1–20, 1970.

TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. **Economic geography**, Springer, Nova Iorque, v. 46, sup1, p. 234–240, 1970.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. **Advances in Agronomy**, Elsevier, Amsterdã, v. 38, p. 45–94, 1986.

VAN DER LAAN, M.; POLLARD, K.; BRYAN, J. A new partitioning around medoids algorithm. **Journal of Statistical Computation and Simulation**, Taylor & Francis, Nova Iorque, v. 73, n. 8, p. 575–584, 2003. DOI: 10.1080/0094965031000136012.

VAROUCHAKIS, E. A. et al. Combining geostatistics and remote sensing data to improve spatiotemporal analysis of precipitation. **Sensors**, Multidisciplinary Digital Publishing Institute, Basel, v. 21, n. 9, p. 3132, 2021. DOI: 10.3390/s21093132.

VASQUES, G. M. et al. Field proximal soil sensor fusion for improving high-resolution soil property maps. **Soil Systems**, Elsevier, Amsterdã, v. 4, n. 3, p. 52, 2020.

VAYSSE, K.; LAGACHERIE, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. **Geoderma**, Elsevier, Amsterdã, v. 291, p. 55–64, 2017. DOI: 10.1016/j.geoderma.2016.12.017.

VIEIRA, H. D. **Café Rural**. Rio de Janeiro: Interciência, Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, 2017.

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications**. Nova Iorque: Springer, 2003. DOI: 10.2307/2291758.

WARD JR, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, Taylor & Francis, Nova Iorque, v. 58, n. 301, p. 236–244, 1963. DOI: 10.1080/01621459.1963.10500845.

WATSON, G. S. Smoothing and interpolation by kriging and with splines. **Journal of the International Association for Mathematical Geology**, Springer, Nova Iorque, v. 16, p. 601–615, 1984. DOI: 10.1007/BF01029320.

WEBSTER, R. Local disjunctive kriging of soil properties with change of support. **Journal of Soil Science**, John Wiley & Sons, Nova Iorque, v. 42, n. 2, p. 301–318, 1991.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for environmental scientists**. Nova Iorque: John Wiley & Sons, 2007.

XU, R.; WUNSCH, D. C. Clustering algorithms in biomedical research: a review. **IEEE reviews in biomedical engineering**, Institute of Electrical and Electronics Engineers, Xangai, v. 3, p. 120–154, 2010. DOI: 10.1109/RBME.2010.2083647.

XU, W.; TRAN, T. T.; SRIVASTAVA, R. .; JOURNEL, A. SPE Annual Technical Conference and Exhibition. In: PROCEEDINGS of the IEEE. Washington: OnePetro, 1992. SPE-24742-MS.

YOST, R.; UEHARA, G.; FOX, R. Geostatistical analysis of soil chemical properties of large land areas. II. Kriging. **Soil Science Society of America Journal**, John Wiley & Sons, Nova Iorque, v. 46, n. 5, p. 1033–1037, 1982. DOI: 10.2136/sssaj1982.03615995004600050029x.

ZERAATPISHEH, M. et al. Integration of PCA and fuzzy clustering for delineation of soil management zones and cost-efficiency analysis in a citrus plantation. **Sustainability**, Multidisciplinary Digital Publishing Institute, Basel, v. 12, n. 14, p. 5809, 2020. DOI: 10.3390/su12145809.

ZHANG, L. et al. Adaptive Fusion of Multi-Source Tropospheric Delay Estimates for InSAR Deformation Measurements. **Frontiers in Environmental Science**, Frontiers, Pequim, v. 10, p. 213, 2022. DOI: 10.3389/fenvs.2022.859363.

ZHANG, Y. Understanding image fusion. **Photogrammetric Engineering and Remote Sensing**, Elsevier, Amsterdã, v. 70, n. 6, p. 657–661, 2004.