GUSTAVO WILLAM PEREIRA

FERRAMENTAS COMPUTACIONAIS PARA SUPORTE À DECISÃO NO MAPEAMENTO DE ATRIBUTOS DO SOLO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

Orientador: Domingos Sárvio M. Valente

Coorientador: Daniel Marçal de Queiroz

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
P436f
2021

Pereira, Gustavo Willam, 1978-
Ferramentas computacionais para suporte à decisão no mapeamento de atributos do solo / Gustavo Willam Pereira. – Viçosa, MG, 2021.
1 tese eletrônica (108 f.): il. (algumas color.).

Orientador: Domingos Sárvio Magalhães Valente.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Engenharia Agrícola, 2021.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2021.179
Modo de acesso: World Wide Web.

1. Agricultura de precisão. 2. Sistemas de informação geográfica. 3. Aprendizado do computador. I. Valente, Domingos Sárvio Magalhães, 1978-. II. Universidade Federal de Viçosa. Departamento de Engenharia Agrícola. Programa de Pós-Graduação em Engenharia Agrícola. III. Título.

CDD 22. ed. 631.3

Bibliotecário(a) responsável: Renata de Fátima Alves CRB6/2578

GUSTAVO WILLAM PEREIRA

FERRAMENTAS COMPUTACIONAIS PARA SUPORTE À DECISÃO NO MAPEAMENTO DE ATRIBUTOS DO SOLO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 30 de julho de 2021.

Assentimento:

Gustavo Willam Pereira
Autor

Domingos Sárvio Magalhães Valente
Orientador

À minha estimada esposa Renata e
aos meus amados filhos Mariana e Lucas pelo apoio,
companheirismo, paciência e amor demonstrados ao
longo desta jornada.
Com imenso carinho,
Dedico

Aos meus amados pais, José Carlos e
Maria de Lourdes, e à toda minha família,
Ofereço

AGRADECIMENTOS

A Deus, sempre presente em minha vida, guiando-me nesta jornada.

A minha amada esposa Renata, pelo convívio, amizade, carinho, cumplicidade, amor da minha vida, por estar sempre ao meu lado e apoiar meus sonhos desde o dia que nos conhecemos, pois o seu apoio tornou o meu fardo suportável.

Aos meus filhos, Mariana e Lucas que chegaram para nos alegrar e nos ensinar ser ainda mais fortes.

Aos meus pais José Carlos Pereira (in memorian) e Maria de Lourdes Pereira.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais, pela oportunidade concedida através do convênio DINTER UFV-IFSUDESTEMG.

À Universidade Federal de Viçosa e ao Departamento de Engenharia Agrícola pela oportunidade de realização do curso e por me deixar fazer parte de sua história.

Ao professor Domingos Sárvio Magalhães Valente, por ter acreditado em meu potencial, pela orientação, confiança, amizade e participação irrestrita na execução desse trabalho, além dos valiosos ensinamentos.

Ao professor Daniel Marçal de Queiroz, pelas valiosas contribuições e sugestões no decorrer do trabalho. Aos professores do Departamento de Engenharia Agrícola pelos conhecimentos adquiridos nas disciplinas.

Aos colegas de trabalho do IFSUDESTE-MG e amigos do DINTER, Alexandre, Demison, Flávia, João, Júlio, Leila, Marine, Michael, Paula, Priscila e Thiago, pelo companheirismo, amizade e bom convívio, proporcionando um ambiente mais ameno e agradável durante o período das aulas em Barbacena-MG. Aos amigos do Laboratório de Mecanização Agrícola, em especial André Coelho e Carolina Tavares, pelas contribuições e companheirismo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, CNPq e FAPEMIG pelo suporte financeiro.

BIOGRAFIA

GUSTAVO WILLAM PEREIRA, filho de José Carlos Pereira e Maria de Lourdes Pereira, nasceu em Carangola, Minas Gerais, no dia 22 de agosto de 1978.

Em fevereiro de 1998, iniciou o curso de Ciência da Computação pela Universidade Federal de Viçosa, em Viçosa, Minas Gerais, concluindo em maio de 2002.

Em maio de 2002, ingressou no Programa de Pós-Graduação em Ciência da Computação, em nível de Mestrado, pelo Departamento de Ciência da Computação da Universidade Federal de Minas Gerais.

Em março de 2004, submeteu-se aos exames de defesa da dissertação, para a obtenção do título de *Magister Scientiae*.

No período entre agosto de 2004 à março de 2011 atuou como docente na Instituição de Ensino Superior FASM – Faculdade Santa Marcelina em Muriaé-MG. Em abril de 2011 ingressou no serviço público federal como docente na Universidade Federal do Espirito Santo (UFES). A partir de julho de 2014 até o presente momento é professor efetivo do Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais – Campus Muriaé-MG.

Em agosto de 2017, ingressou no Programa de Pós-Graduação em Engenharia Agrícola, em nível de Doutorado, área de concentração em Mecanização Agrícola, pelo Departamento de Engenharia Agrícola da Universidade Federal de Viçosa.

Em julho de 2021, submeteu-se aos exames de defesa da tese, para a obtenção do título de *Doctor Scientiae.*

RESUMO

PEREIRA, Gustavo Willam, D.Sc., Universidade Federal de Viçosa, julho de 2021. **Ferramentas computacionais para suporte à decisão no mapeamento de atributos do solo**. Orientador: Domingos Sárvio Magalhães Valente. Coorientador: Daniel Marçal de Queiroz.

A agricultura de precisão (AP) é uma técnica de gestão agrícola baseada na observação, medição e resposta às variabilidades espaciais e temporais que ocorrem nas áreas de produção agrícola e de pastagens. A sua adoção pode ser feita por meio de sistemas de apoio à decisão que tem por objetivo maximizar os lucros e aumentar a eficiência no uso de insumos. Em AP, o mapeamento das características físicas e químicas do solo permite estimar com maior precisão a variabilidade espacial do solo. O desenvolvimento de sistemas de informação geográfica (SIG) e sua utilização têm aumentado consideravelmente nos últimos anos, sendo, portanto, utilizado em larga escala na AP. Dado um conjunto de pontos amostrados deseja-se obter mapas dos atributos de solo ou inferir valores em lugares específicos em locais não amostrados, para isso técnicas de interpolação são utilizadas. Existem diversos métodos interpoladores, sendo a Krigagem Ordinária (OK) um dos mais usados. Entretanto, desde o início do século XXI tem havido um interesse crescente em utilizar algoritmos baseados em dados para a geração dos mapas. Estes algoritmos são conhecidos sob o nome "M*achine Learning*" e têm-se mostrados eficientes para produzir previsões espaciais. As técnicas de amostragem de solo e a densidade de amostragem são fatores determinantes para a geração de mapas interpolados dos atributos de solo. Entretanto o custo de análise para uma amostragem convencional mais densa do solo torna-se muitas vezes inviável o processo. Sendo assim técnicas alternativas que apresentem menores custos para amostragem de atributos do solo devem ser implementadas. Assim o presente trabalho tem como objetivo desenvolver um software de suporte a tomada de decisão em AP e analisar diferentes estratégias de amostragem para caracterização da variabilidade espacial dos atributos físicos e químicos do solo. O software desenvolvido consistiu de um plugin (complemento) para o QGIS (programa de computador livre e de código aberto para SIG) para a interpolação de atributos do solo por meio do método da Krigagem Ordinária e da técnica de aprendizado de Máquina "*Support Vector Machine*" (*SVM*). Para isso linguagens de programação como Python e ferramentas de código aberto do QGIS

foram utilizadas, dispensado a aquisição de licenças. O método *SVM* foi implementado de tal forma que covariáveis pudessem ser adicionadas ao modelo. Isso possibilitou que variáveis obtidas de forma mais adensada em campo pudessem ser adicionadas na geração dos mapas de atributos do solo. O interpolador Inverso da Distância Ponderada (IDW) foi utilizado para ajustar o número de observações das layers em formato shape e/ou raster do QGIS ao número de observações da tabela de atributos para a geração dos mapas. O plugin foi desenvolvido de forma a permitir que a própria variável fosse adicionada como covariável ao modelo *SVM* por meio do interpolador IDW. Em três estudos de caso a técnica *SVM* apresentou um desempenho superior ao método *OK* com maiores valores de R^2 e menores valores de RMSE. As três áreas de estudo foram amostradas utilizando-se diferentes densidades amostrais. O Índice de Moran foi utilizado para medir a correlação espacial entre os pontos amostrados. No primeiro estudo de caso, em uma área de 20,2 ha com 141 pontos amostrados, o método *OK* apresentou valores de R^2 superior em três dos dez atributos de solo analisados. O Índice de Moran para estes três atributos apresentaram valores superiores a 0,72 e significativos ao nível de 5% de probabilidade medido pelo seu p-valor. No segundo estudo de caso, realizado em uma área de 204 ha com 204 pontos amostrados, *OK* foi superior para cinco atributos em dez analisados, com Índice de Moran variando entre 0,71 e 0,84, todos significativos. Para estes dois estudos de caso o método *SVM* foi utilizado tendo apenas a própria variável interpolada pelo método IDW como covariável. Já no terceiro estudo de caso, em uma área de 90 ha com 181 pontos coletados, foram utilizadas no método *SVM,* além da própria covariável, outras covariáveis de fácil aquisição como condutividade elétrica aparente (ECa) do solo ou atributos que se modificam pouco ao longo do tempo como altitude, matéria orgânica, silte, argila e areia. Neste terceiro estudo de caso as amostras foram divididas em dois conjuntos de treinamento e teste. Para o conjunto de treinamento o método *SVM* com utilização de covariáveis foi superior ao *OK* para todas as variáveis analisadas em três densidades de grids amostrais.

Palavras-chave: Sistemas de informações geográficas. Krigagem ordinária. Aprendizado de máquina. Amostragem de solo. Variabilidade espacial.

ABSTRACT

PEREIRA, Gustavo Willam, D.Sc., Universidade Federal de Viçosa, July, 2021. **Computational tools for decision making in soil attributes mapping.** Adviser: Domingos Sárvio Magalhães Valente. Co-adviser: Daniel Marçal de Queiroz.

Precision agriculture (PA) is an agricultural management technique based on observation, measurement and response to spatial and temporal variations that occur in agricultural production and pasture areas. Its adoption can be done through decision support systems that aim to maximize profits and increase efficiency in the use of inputs. In AP, the mapping of the physical and chemical characteristics of the soil allows for a more accurate estimation of the spatial variability of the soil. The development of geographic information systems (GIS) and their use have increased considerably in recent years, and are therefore used on a large scale in PA. Given a set of sampled points, it is desired to obtain maps of soil attributes or infer values in specific places in non-sampled locations, for that interpolation techniques are used. There are several interpolating methods, being the Ordinary Kriging (OK) one of the most used. However, since the beginning of the 21st century, there has been a growing interest in using data-based algorithms to generate maps. These algorithms are known under the name “Machine Learning” and have been shown to be efficient in producing spatial predictions. Soil sampling techniques and sampling density are determining factors for the generation of interpolated maps of soil attributes. However, the cost of analysis for a denser conventional sampling of the soil often makes the process unfeasible. Therefore, alternative techniques that present lower costs for sampling soil attributes must be implemented. Thus, this work aims to develop a software to support decision making in PA and analyze different sampling strategies to characterize the spatial variability of physical and chemical soil attributes. The developed software consisted of a plugin (complement) for QGIS (free and open source computer program for GIS) for the interpolation of soil attributes through the Ordinary Kriging method and the Machine learning technique “Support Vector Machine” (SVM). For this, programming languages such as Python and QGIS's open source tools were used, with no need to purchase licenses. The SVM method was implemented in such a way that covariates could be added to the model. This made it possible for variables obtained in a more dense way in the field to be added in the generation of maps of soil

attributes. The Inverse Weighted Distance Interpolator (IDW) was used to adjust the number of observations of layers in shape and/or raster format from QGIS to the number of observations in the attribute table for the generation of maps. The plugin was developed in order to allow the variable itself to be added as a covariate to the SVM model through the IDW interpolator. In three case studies, the SVM technique performed better than the OK method with higher values of R2 and lower values of RMSE. The three study areas were sampled using different sample densities. The Moran Index was used to measure the spatial correlation between the sampled points. In the first case study, in an area of 20.2 ha with 141 sampled points, the OK method presented higher R2 values in three of the ten soil attributes analyzed. The Moran Index for these three attributes showed values above 0.72 and significant at the 5% probability level measured by its p-value. In the second case study, carried out in an area of 204 ha with 204 sampled points, OK was superior for five in ten analyzed, with Moran's Index varying between 0.71 and 0.84, all significant. For these two case studies, the SVM method was used with only the variable interpolated by the IDW method as a covariate. In the third case study, in an area of 90 ha with 181 collected points, in addition to the covariate itself, other easily acquired covariates were used in the SVM method, such as apparent electrical conductivity (ECa) of the soil or attributes that change little when over time as altitude, organic matter, silt, clay and sand. In this third case study the samples were divided into two sets of training and testing. For the training set, the SVM method using covariates was superior to OK for all variables analyzed in three sample grid densities.

Keywords: Geographic Information Systems. Ordinary Kriging. Machine learning. Soil sampling. Spatial variability.

SUMÁRIO

1 INTRODUÇÃO GERAL

O problema da fome, que atinge diferentes regiões do mundo, está associado ao elevado crescimento populacional e à ocorrência de determinados fatores socioeconômicos como a falta de alimento disponível para as pessoas, impossibilidade de se ter acesso ou comprar alimentos, dentro outros. Nos últimos 50 anos, a população mundial cresceu de três bilhões para mais de sete bilhões de habitantes, criando uma alta demanda por alimentos. Estima-se que a população mundial irá aumentar em mais de 30% até 2050, ocasionando uma demanda crescente de 70% na produção de alimento. Para atender a essa crescente demanda, vários estudos e iniciativas foram lançados desde a década de 1990. Avanços no manejo das culturas como sistemas globais de navegação por satélite (GNSS), permitiram a localização precisa de medições pontuais no campo, possibilitando assim o manejo das áreas agrícolas de forma espacialmente variável (KAMILARIS et al., 2017).

Desde o surgimento da agricultura, o homem percebeu que as áreas de produção agrícola não eram uniformes, que os atributos do solo, das plantas e as características do relevo variam espacialmente. Entretanto, até os anos 1990s não se tinha tecnologia para tratar essa variabilidade dos campos de produção. Tratar os campos como uniforme aumenta o impacto ambiental provocado pelas práticas agrícola. Muitos insumos são aplicados em regiões que às vezes não precisam do insumo, contaminando assim o meio ambiente. Regiões de maior potencial produtivo às vezes são tratadas com sub dosagens, limitando a produtividade das culturas.

Com o desenvolvimento da eletrônica embarcada em máquinas agrícola, o surgimento das tecnologias de informação e comunicação (TICs) e os sistemas de posicionamento com base em satélites, tornou-se possível o desenvolvimento da agricultura de precisão (FAR e REZAEI-MOGHADDAM, 2018). Na Agricultura de Precisão (AP), o manejo das culturas é definido com base na análise da variabilidade espacial dos fatores que influenciam a produção agropecuária. Isso torna a agricultura mais sustentável tanto econômica como ambientalmente (NAIME et al., 2011). A AP é um conjunto de técnicas que tem por objetivo maximizar o retorno financeiro. Uma dessas técnicas é a aplicação de insumos a taxa variada utilizando ferramentas e

tecnologias para identificar a variabilidade do solo e das culturas buscando melhorar as práticas agrícolas e otimizar o uso de insumos agronômicos (KHANAL, 2017).

A AP trabalha com a premissa de cada ponto da área de produção tem um potencial produtivo e os insumos devem ser aplicados considerando esse potencial. Portanto, na AP as dosagens não são uniformes, praticando o que é denominado de manejo a sítio específico. Para definir esse manejo localizado, quatro conjuntos de tecnologias de computação são utilizados baseados em componentes: sistemas de informações geográficas (SIG), sistema global de navegação por satélite (GNSS), sistemas de aplicação de taxa variável (VAR) e tecnologias de detecção (NORTON e SWINTON, 2018). O desenvolvimento de sistemas de informação geográfica (SIG) e sua utilização têm aumentado consideravelmente nos últimos anos. Os SIG's estão sendo utilizados para modelagem complexa, como previsão de produção, monitoramento de colheita, mapeamento do solo, etc... Em algumas situações é necessário converter dados em mapas que mostram como o valor de determinada variável se modifica em uma dada região. Em outras, deseja-se inferir valores dessa variável em lugares específicos que não foram amostrados. Em ambas as situações, é necessário interpolar para estimar esses valores não amostrados (OLIVER e WEBSTER, 1990).

Existem diferentes métodos interpoladores, com o objetivo de predizer valores em locais não amostrados, entre os mais tradicionais podem ser citados: método dos polígonos, método da triangularização, método da média local, método do inverso da distância, método da krigagem. Estes métodos se caracterizam por serem métodos de combinação linear ponderada. A diferença entre a krigagem dos demais é que a krigagem apresenta estimativa não tendenciosa e a mínima variância associada ao valor estimado (YAMAMOTO e LANDIM, 2013). A krigagem consiste em um método de interpolação geoestatístico, que aplica o conceito da teoria das variáveis regionalizadas, a partir da dependência espacial entre as amostras vizinhas a partir do semivariograma, com o objetivo de estimar valores em locais não amostrados (GIACOMIN, 2014). A krigagem e suas muitas variantes têm sido usadas como a técnica de melhor predição linear para pontos espaciais desde os anos 60. O número de aplicações publicadas em kriging aumentou desde 1980 e a técnica tem sido utilizada em uma variedade de campos (HENGEL, et. al, 2018).

Desde o início do século XXI, no entanto, tem havido um interesse crescente em usar algoritmos computacionalmente mais intensivos e principalmente baseados em dados. Estas técnicas são também conhecidas sob o nome Aprendizado de Máquina, e são aplicáveis a vários problemas de mineração de dados, reconhecimento de padrões, regressão e classificação. O aprendizado de máquina ou "M*achine Learning*" (ML) pode ser amplamente definido como métodos computacionais que usam a experiência para melhorar o desempenho ou para fazer previsões precisas. Consiste em projetar algoritmos de previsão eficientes e precisos. Como o sucesso de um algoritmo de aprendizado depende dos dados usados, o aprendizado de máquina é inerentemente relacionado à análise de dados e estatísticas. Em termos mais gerais, as técnicas de aprendizagem são métodos baseados em dados que combinam conceitos fundamentais em ciência da computação com ideias de estatísticas, probabilidade e otimização (MOHRI et. al, 2012).

Na AP é necessário analisar a dependência espacial das variáveis. Para isso, a amostragem deve ser realizada considerando tanto o número de pontos quanto a sua distribuição na área estudada (YAMAMOTO & LANDIM, 2013). Um sistema de amostragem com alta densidade de pontos representa, muitas vezes, um longo tempo desprendido para aquisição das amostras, além do alto custo de análises laboratoriais. Todavia, a diminuição da densidade de amostragem pode levar a erros na estimativa da variabilidade espacial dos nutrientes do solo, podendo ocasionar erros de recomendação. Um sistema ótimo de amostragem deve proporcionar uma estimativa com menor custo de amostragem sem, no entanto, deixar de representar a variabilidade existente no campo de produção. Uma das técnicas para reduzir o número de amostras para análise de solo é definir zonas de manejo (YAN et al., 2007; VALENTE et al., 2012). A zona de manejo é uma área que apresenta características semelhantes entre os fatores que limitam a produtividade e/ou qualidade do produto, podendo, por isso, ser tratada com a mesma dosagem de insumo a ser aplicado (VALENTE, 2010).

Para definir as zonas de manejos diferentes variáveis podem ser utilizadas. Dependendo das variáveis escolhidas, o resultado da geração das zonas de manejo pode variar. Dois critérios são importantes na escolha das variáveis, o primeiro é que as variáveis devem ser estáveis em relação ao tempo, e o segundo é que a aquisição

dessas variáveis deve ter custo acessível. Dentre as variáveis mais utilizadas para definição de zonas de manejo em AP, encontra-se a condutividade elétrica aparente do solo (ECa), pois ela atende a esses dois critérios citados. A ECa do solo pode ser definida como a capacidade de conduzir ou transmitir corrente elétrica. Neste caso, o solo torna-se um potencial indicador de condutividade elétrica por conta dos eletrólitos dissolvidos em sua solução (MOLIN, et al., 2013). Diversos trabalhos têm utilizado a condutividade elétrica para a caracterização da variabilidade do solo (CORWIN & LESH, 2005; VALENTE et al., 2012; NASCIMENTO et al., 2013; SILVA et al., 2018). A medição da ECa do solo é uma técnica de baixo custo, fácil, confiável e de rápida medição.

1.1 Objetivos

Diante do exposto, esse trabalho foi realizado com o objetivo de desenvolver um sistema de suporte para a AP utilizando técnicas de interpolação por OK e por ML. Os objetivos específicos do trabalho foram:

- Implementar e comparar a técnica de interpolação por Krigagem Ordinária com a técnica de "*Machine Learning*" Máquina de Vetores de Suporte;
- Desenvolver um plugin (complemento) QGIS de código aberto e disponibilizá-lo para utilização de forma gratuita através das plataformas GitHub e repositório de plugins do QGIS;
- Comparar a performance entre as estratégias de amostragem por Célula, Convencional (valor médio dos pontos amostrados na área), por ZM's definidas por ECa do solo e pelo método de amostragem em grid aplicados em duas densidades amostrais utilizando como método interpolador a Krigagem Ordinária.

1.2 Organização da tese

A presente tese foi organizada em cinco capítulos. No primeiro capítulo, foi apresentado uma introdução geral, justificando o desenvolvimento deste trabalho. No segundo capítulo foi implementado a técnica de Aprendizado de Máquina, Máquina de Vetores de Suporte e comparado com a Krigagem Ordinária. No terceiro capítulo o plugin QGIS "Smart-Map" foi desenvolvido e um estudo de caso foi implementado para demostrar sua aplicabilidade, no quarto capítulo foram comparadas as

estratégias de amostragem. No último capítulo são apresentadas as considerações finais da tese.

1.3 Referências

CORWIN, D.L.; LESH, S.M. Apparent soil eletrical conductivity measurements in agricultural. **Computers and Eletronics in Agriculture**. v. 46, p. 11-43, 2005.

FAR, S. T.; REZAEI-MOGHADDAM, K. Impacts of the precision agricultural technologies in Iran: An analysis experts' perception e their determinants. **Information processing in agriculture**, v. 5, n. 1, p. 173-184, 2018.

GIACOMIN, G. et al. Análise comparativa entre métodos interpoladores de modelos de superfícies. **Revista Brasileira de Cartografia**, v. 6, n. 66/6, 2014.

HENGL, T. et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

KAMILARIS, A.; KARTAKOULLIS, A.; PRENAFETA-BOLDÚ, F. X. A review on the practice of big data analysis in agriculture. **Computers and Electronics in Agriculture**, v. 143, p. 23-37, 2017.

KHANAL, S.; FULTON, J.; SHEARER, S. An overview of current and potential applications of thermal remote sensing in precision agriculture. **Computers and Electronics in Agriculture**, v. 139, p. 22-32, 2017.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning.** Massachusetts Institute of Technology press, 2012.

MOLIN, J. P.; FAULIN, G. D. C. Spatial and temporal variability of soil electrical conductivity related to soil moisture. **Scientia Agrícola**, v. 70, p.1-5, 2013.

NASCIMENTO, E. F.; RABELLO, L. M.; BASSOI, L. H. Definição da malha de amostragem da condutividade elétrica do solo para obtenção de zonas de manejo em pomar de videira. **Embrapa Instrumentação-Capítulo em livro científico. Agricultura de precisão para culturas perenes e semi-perenes.** p. 413-420, 2013.

NAIME, J. M.; CAMARGO N. J.; VAZ, C. M. P. Avaliação geral, resultados, perspectivas e uso de ferramentas de agricultura de precisão. **Embrapa Instrumentação-Capítulo em livro científico. Agricultura de precisão para culturas perenes e semi-perenes.** pag. 69-72, 2011.

NORTON, G. W.; SWINTON, S. M. Precision agriculture: global prospects and environmental implications. In: **Tomorrow's Agriculture: Incentives, Institutions, Infrastructure and Innovations-Proceedings of the Twenty-fouth International Conference of Agricultural Economists: Incentives, Institutions, Infrastructure and Innovations-Proceedings of the Twenty-fouth International Conference of Agricultural Economists**. Routledge, p. 269. 2018.

OLIVER, M. A.; WEBSTER, R. Kriging: a method of interpolation for geographical information systems. **International Journal of Geographical Information System**, v. 4, n. 3, p. 313-332, 1990.

SILVA, T. S. M. et al. Condutividade elétrica da solução de solo em função da condutividade elétrica aparente e da umidade do solo sob aplicação de cloreto de potássio com uso da reflectometria no domínio do tempo. **Irriga**, v.10, n.2, p. 174-183, 2018.

VALENTE, D. S. M. Desenvolvimento de um sistema de apoio à decisão para definir zonas de manejo em cafeicultura de precisão. **Tese (Doutorado em Engenharia Agrícola)** – Universidade Federal de Viçosa. Viçosa, p. 120, 2010.

VALENTE, D. S. M.; QUEIROZ, D. M.; PINTO, F. A. C.; SANTOS, N. T.; SANTOS, F. L. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. **Scientia Agrícola**, Piracicaba, v. 69, n. 3, p.173-179, 2012.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística - Conceitos e Aplicações.** Oficina de textos, 1 ed. 216 p., São Paulo, 2013.

YAN, L.; ZHOU, S.; CI-FANG, W.; FANG, L., HONG-YI, L. Optimised spatial sampling scheme for soil electrical conductivity based on variance quad-treee (VQT) method. **Agricultural Sciences in China**, v.6, n.12, p.163-1471, 2007.

2 MAPEAMENTO DIGITAL DO SOLO UTILIZANDO MÁQUINA DE VETORES DE SUPORTE COMBINADO COM INVERSO DA DISTÂNCIA PONDERADA [1]

2.1 Resumo

O mapeamento do solo frequentemente envolve a interpolação de valores desconhecidos a partir de pontos vizinhos que foram amostrados. A Krigagem tem se mostrado o melhor interpolador para predizer valores em locais não amostrados. Entretanto, o método da Krigagem requer um grande número de pontos amostrados para gerar mapas precisos. Técnicas de Aprendizado de Máquina (ML) tem se mostrado como uma alternativa potencial para produzir mapas de solo com um menor número de pontos de amostragem. Portanto, o objetivo deste estudo foi implementar um método híbrido baseado nas técnicas Inverso da Distância Ponderada (IDW) e ML, e compará-lo com os métodos IDW e a Krigagem Ordinária (OK). O método híbrido consistiu em utilizar o método IDW como um método preditor dos atributos de solo e, em seguida, usar o método ML para ajustar a predição. O algoritmo de ML utilizado foi o *Support Vector Machine* (*SVM*). O método criado foi denominado IDW-SVM. Um programa de computador foi desenvolvido utilizando a linguagem Python 3.7 com base nesse método. Bibliotecas do Python foram utilizadas para realizar a predição e mapeamento de atributos físicos e químicos do solo. Este algoritmo foi testado em duas áreas, uma de 20,2 ha cultivada com café (AREA1), e uma outra com 204 ha cultivada com soja (AREA2). O desempenho do método IDW-SVM foi comparado com os métodos tradicionais de interpolação, o IDW e a Krigagem Ordinária (OK). O Índice de Correlação Espacial de Moran (ICM) foi utilizado para medir a dependência espacial. O Coeficiente de Determinação (R^2) e a Raiz Quadrada do Erro Quadrático Médio (RMSE) foram utilizados para avaliar o desempenho dos métodos, com base na validação cruzada *leave-one-out cross-validation* (*LOOCV*). A validação dos modelos foi realizada utilizando diferentes números de pontos amostrais variando de 25 a 141 pontos na AREA1 e de 36 a 204 pontos na AREA2. Os resultados demostraram que, para os grids amostrais menos densos com 25 e 50 pontos (AREA1) e 36 e 72 pontos (AREA2) o IDW foi superior ao OK. Para grids com maior densidade o OK obteve melhores resultados que o IDW. Independentemente do

[1] Este capítulo se refere a uma versão em português do artigo "Digital soil mapping using support vector machine combined with inverse distance weighting" submetido no periódico *Precision Agriculture*. ISSN 1573-1618.

número de amostras, o IDW-SVM apresentou um desempenho superior ao IDW e OK para predição e mapeamento dos atributos do solo de acordo com os critérios analisados.

Palavras-chave: Agricultura de precisão. Amostragem de solo. Krigagem ordinária. Interpolação.

2.2 Abstract

Soil mapping often involves interpolation of unknown values from neighboring points that were sampled. Kriging has been shown to be the best interpolator to predict values at unsampled points. However, Kriging method requires a high number of sampling points to generate accurate maps. Machine Learning (ML) techniques have potential to produce soil maps with a lower number of sampling points. Therefore, the objective of this study was to implement a hybrid method based on the Inverse Distance Weighting (IDW) and ML techniques and compare it to the IDW alone and to the Ordinary Kriging (OK) methods. The hybrid method consisted of using the IDW method as a predictor method of the soil attribute and then using a ML method to correct the prediction. The ML algorithm used was the *Support Vector Machine* (*SVM*). The created method was called IDW-SVM method. A software based on this IDW-SVM method was developed using the Python language and its libraries to map the soil physical and chemical attributes. This software was tested in two areas, one that had 20,2 ha cultivated with coffee (AREA1) and the other one that had 204 ha cultivated with soybean (AREA2). The performance of the IDW-SVM method was compared to traditional interpolation methods, the IDW and the Ordinary Kriging (OK). The Moran's Spatial Correlation Index was used to measure spatial dependence. Coefficient of Determination (R^2) and Root-Mean-Square Error (RMSE) were used to assess the performance of the methods, based on the leave-one-out cross-validation (LOOCV). Data validation was performed using different numbers of sampling points, ranging from 25 to 141 points in AREA1 and from 36 to 204 points in AREA2. General results showed that, for the less dense sampling, 25 and 50 points in AREA1 and 36 and 72 points in AREA2, IDW was superior to OK. For dense sampling, OK performed better than IDW. Regardless of the number of sampling points, the IDW-SVM method performed better than IDW and OK, for prediction and mapping of soil attributes according to the analyzed criteria.

Keywords: Precision agriculture. Soil sampling. Ordinary kriging. Interpolation.

2.3 Introdução

O mapeamento digital do solo tornou-se uma etapa importante no manejo da cultura desde o desenvolvimento da agricultura de precisão (PA). O mapeamento do solo frequentemente envolve a interpolação de valores desconhecidos de pontos vizinhos que foram amostrados (HEDGE et al., 2017; HENGL; HEUVELINK; STEIN, 2004). Existem diversos métodos interpoladores com o objetivo de predizer valores em pontos não amostrados. Inverso da Distância Ponderada (IDW) e Krigagem Ordinária (OK) tem sido as técnicas de interpolação espacial mais populares utilizadas em AP (SEKULIĆ et al., 2020). IDW é um método determinístico baseado na média ponderada da distância dos vizinhos ao ponto a ser interpolado.

Krigagem supera muitas das deficiências dos métodos tradicionais de interpolação, sendo um interpolador ideal com estimativas não-viesadas (OLIVER; WEBSTER, 1990). A Krigagem pressupõe que as variáveis a serem interpoladas sejam estacionárias e aplica o conceito da teoria das variáveis regionalizadas (CRESSIE, 1990; TRANGMAR; YOST; UEHARA, 1985). Vários estudos têm sido conduzidos utilizando o método da Krigagem para predição e mapeamento de atributos do solo (COELHO et al., 2018; HENGL; HEUVELINK; STEIN, 2004; VALENTE et al., 2012). Entretanto, além das pressuposições estatísticas citadas, o método de krigagem necessita de uma quantidade suficiente de amostras para permitir o ajuste de um modelo teórico de semivariância (GIACOMIN et al., 2014; WEBSTER; OLIVER, 1992). Para se realizar uma modelagem confiável da semivariância, alguns autores recomendam um mínimo de 140 pontos amostrados (WEBSTER; OLIVER, 2001). Outros autores recomendam pelo menos 30 pares de pontos para ter uma boa estimativa da semivariância para cada distância específica *h* (POULADI et al., 2019). Além disso, para modelar a semivariância de maneira adequada é necessário ter conhecimento dos parâmetros geoestatísticos. Em geral, quanto melhor o modelo de semivariância, melhor é a qualidade da interpolação por Krigagem.

Uma das dificuldades associadas para gerar mapas de solo é o grande número de pontos de amostragem necessários para produzir mapas precisos de atributos físicos e químicos do solo. Uma maneira de resolver esse problema é desenvolver métodos que requeiram um número menor de pontos de amostragem para produzir mapas precisos. Esses novos métodos devem ser flexíveis para usar uma

amostragem de alta densidade como covariáveis, por exemplo, produtividade da cultura, índice de vegetação, condutividade elétrica aparente do solo etc. Recentemente, algoritmos de aprendizado de máquina (ML) provaram ser uma técnica eficiente no campo da predição e mapeamento de atributos do solo (KHALEDIAN; MILLER, 2020).

O termo ML pode ser entendido como o processo automatizado dos algoritmos aprenderem com base em conjuntos de dados. Esses algoritmos apresentam eficiência para trabalhar com grandes volumes de dados. Dessa forma, um modelo de ML poderá ser utilizado em mineração de dados, reconhecimento de padrões, regressão e classificação (HEUNG et al., 2016; KHALEDIAN; MILLER, 2020; LIAKOS et al., 2018; MOHRI; ROSTAMIZADEH, 2018; PARMLEY et al., 2019). No campo da predição e distribuição espacial de propriedades do solo, vários trabalhos de ML têm sido desenvolvidos (GUO et al., 2015; HENGL et al., 2015, 2018; HEUNG et al., 2016; POULADI et al., 2019).

Dentre os algoritmos de ML, pode-se destacar: *Cubist*, *K-Nearest Neighbors* (*KNN*), *Support Vector Machine* (*SVM*), *Decision Tree* (DT), *Random Forest (*RF*), LightGBM*, *XGBoost*, *Artificial Neural Networks* (ANNs). O algoritmo *SVM* pode ser utilizado para classificação, regressão e agrupamento. O *SVM* usa funções do kernel para projetar os dados em um novo hiperespaço em que padrões não lineares complexos podem ser simplesmente representados (WERE et al., 2015). Os parâmetros do *SVM,* que são geralmente ajustados pelo usuário (hiperparâmetros) como *Kernel*, *C* e *gamma (γ)* podem ser otimizados com um método sistemático de busca em grade (XU et al., 2018), permitindo assim um ajuste automatizado sem a interferência do usuário. Dentre as vantagens do *SVM* destaca-se: número reduzido de hiperparâmetros para serem ajustados e baixo custo computacional para treinamento do modelo (KHALEDIAN; MILLER, 2020). Dentre as desvantagens podemos destacar não ser adequado para grandes conjuntos de dados.

Embora o método de Krigagem seja um dos melhores interpoladores, ele requer alta densidade de amostragem e conhecimento do usuário para modelar o semivariograma. Por causa disso, esse método pode se tornar caro e não adequado para certas aplicações. Por outro lado, os algoritmos de ML não precisam de um modelo de semivariograma, permitem o uso de várias covariáveis e podem ser automatizados. Além disso, o método ML pode ser combinado com outros métodos

de interpolação, criando um método híbrido para melhorar sua eficiência de predição. No entanto, quão bom é o desempenho de um método híbrido para interpolação de atributos do solo em comparação com o método de Krigagem para atributos do solo com diferentes dependências espaciais? Assim, o objetivo deste trabalho foi implementar um método híbrido que utiliza como gerardor de *features* o Inverso da Distância Ponderada (IDW) para o modelo de ML e compará-lo com os métodos de interpolação IDW e OK.

2.4 Material e métodos

Área de Estudo

Duas áreas, denominadas ÁREA1 e ÁREA2, foram utilizadas no presente estudo. A AREA1, localizada no município de Araponga (20º42'32" S e 42º34'17" W), região sudeste do estado de Minas Gerais, Brasil (Figura 2.1). Essa área possui aproximadamente 20,2 ha e é cultivada com café (*Coffea Arábica L.*), altitude média 904 m, relevo montanhoso e o solo predominantemente classificado como Latossolo Vermelho-Amarelo Distróficos (LVAd) de acordo com a classificação atualizada da Embrapa Solos (SANTOS et al., 2018).

A AREA2, localizada no município de São Desidério (12°25'12'' S e 45°29'46" W), altitude média de 493 m, região oeste do estado da Bahia, Brasil, possui aproximadamente 204 ha (Figura 2.2). Nessa área cultiva-se soja (Glycine max) desde 2014. O solo é predominantemente classificado como Latossolo Amarelo Distróficos (LAd), de acordo com a classificação atualizada da Embrapa Solos (SANTOS et al., 2018). A área de produção é plana e o clima da região apresenta inverno seco e verão chuvoso, caracterizando-se como um clima tropical úmido, de acordo com a classificação atualizada da Köppen-Geiger (KOTTEK et al., 2006).

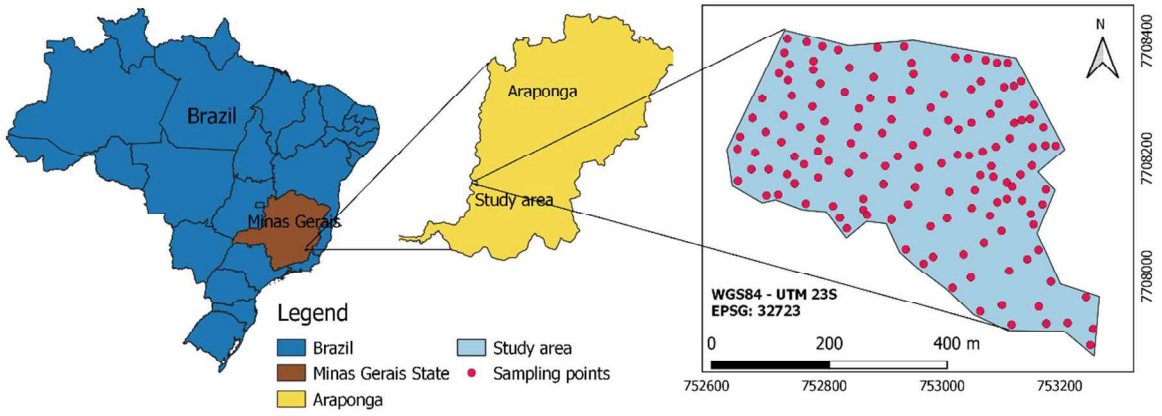


Figura 2.1: Localização geográfica da área de estudo AREA1 e distribuição dos pontos amostrais em Araponga, Sudeste de Minas Gerais, Brasil.

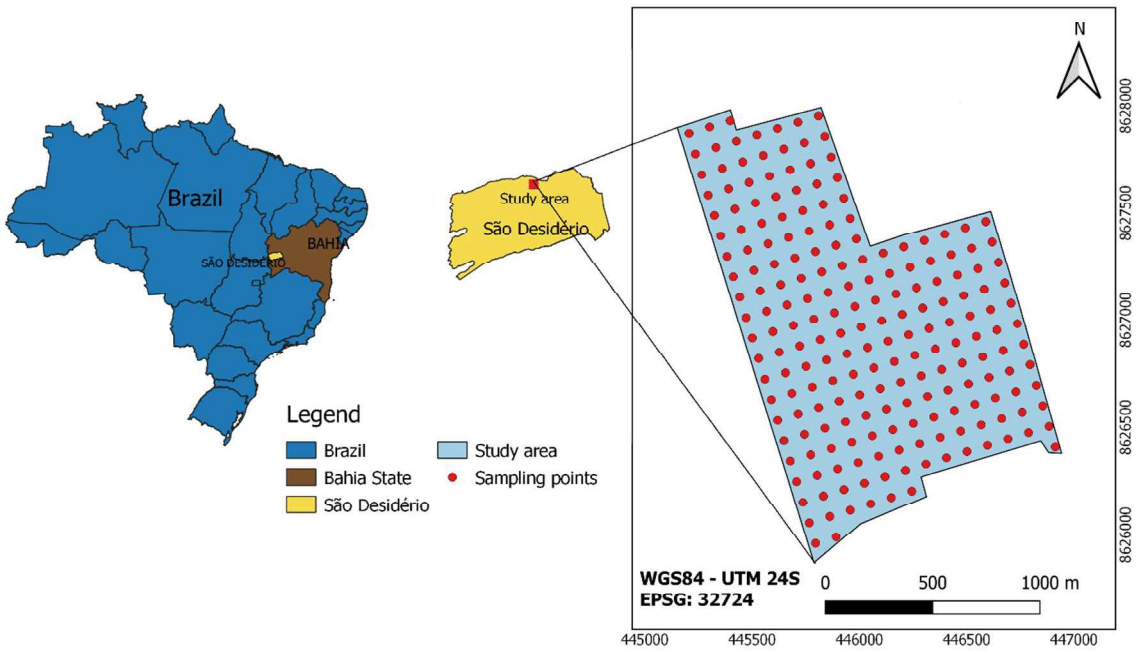


Figura 2.2: Localização geográfica da área de estudo AREA2 e distribuição dos pontos amostrais em São Desidério, Oeste da Bahia, Brasil.

Na AREA1 foram coletados 141 pontos amostrais e na AREA2 foram coletados 204 pontos amostrais. No presente trabalho, para efeito de análise da performance dos métodos de interpolação, foram utilizados os seguintes atributos de solos: Areia Fina (FS), Areia Grossa (CS), Argila (CLA), Cálcio (Ca^{+2}), Fósforo (P), Magnésio (Mg^{+2}), Matéria Orgânica (OM), Potássio (K^{+}), Potencial Hidrogeniônico (pH) e Silte (SIL), totalizando assim 10 atributos de solos. As Tabelas 2.1 e 2.2 apresentam as estatísticas descritivas das duas áreas. Os dados da AREA1 foram disponibilizados

por (VALENTE et al., 2012), enquanto que os dados da AREA2 foram disponibilizados por (MARTINS et al., 2020).

Tabela 2.1: Estatística descritiva dos atributos químicos e físicos do solo na AREA1 (VALENTE et al., 2012).

Atributo	Unidade	Média	Desvio Padrão	Valor Mínimo	Valor Máximo	CV (%)
pH (1)		5,76	0,50	4,83	7,74	8,63
P (2)	(mg dm^{-3})	4,36	17,20	1,10	22,10	57,90
K^+ (3)	(mg dm^{-3})	89,21	32,06	17,00	165,00	35,94
Ca^{2+} (4)	(cmolc dm^{-3})	2,70	1,20	0,43	9,22	44,66
Mg^{2+}(5)	(cmolc dm^{-3})	1,21	0,48	0,21	3,71	39,52
OM (6)	(g dm^{-3})	5,42	1,03	2,50	7,96	18,93
CS (7)	(g kg^{-1})	27,00	4,39	17,00	41,00	16,41
FS (8)	(g kg^{-1})	12,16	1,90	7,00	20,00	15,59
SIL(9)	(g kg^{-1})	7,49	2,04	4,00	15,00	27,30
CLA (10)	(g kg^{-1})	53,57	4,90	39,00	66,00	9,14

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte; 10/ CLA, Argila.

Tabela 2.2: Estatística descritiva dos atributos químicos e físicos do solo na AREA2 (MARTINS et al., 2020).

Atributo	Unidade	Média	Desvio Padrão	Valor Mínimo	Valor Máximo	CV (%)
pH (1)		5,53	0,29	4,88	6,29	5,25
P (2)	(mg dm^{-3})	13,67	12,64	4,80	45,00	3,08
K^+ (3)	(mg dm^{-3})	66,78	20,06	28,00	127,00	30,12
Ca^{2+} (4)	(cmolc dm^{-3})	1,66	0,35	1,01	2,78	21,24
Mg^{2+}(5)	(cmolc dm^{-3})	0,65	0,18	0,21	1,24	28,40
OM (6)	(g dm^{-3})	0,77	0,45	0,00	2,44	58,86
CS (7)	(g kg^{-1})	27,29	7,14	11,80	49,70	26,17
FS (8)	(g kg^{-1})	47,29	7,39	26,70	62,40	15,63
SIL(9)	(g kg^{-1})	5,17	1,62	2,58	11,09	31,48
CLA (10)	(g kg^{-1})	20,25	2,03	15,41	28,75	10,02

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte; 10/ CLA, Argila

Geração dos cenários para avaliação dos métodos de interpolação

Para se realizar a interpolação em diferentes condições de dependência espacial e densidade amostral foram gerados diferentes grids de amostragens a partir dos grids originais na AREA1 e na AREA2. Na AREA1 foram definidos os grids de amostragem com densidades de: 25, 50, 75, 100 pontos, a partir do grid original de

referência de 141 pontos. Na AREA2 utilizou-se grids amostrais de 36, 72, 108, 144 a partir do grid original de referência de 204 pontos.

Para gerar os grids de amostragem em diferentes densidades a partir do grid original foi utilizado o algoritmo *Hipercubo latino* (*LHS*) proposto por McKay et al. (1979). O *LHS* é um método estatístico para gerar uma distribuição de pontos a partir de uma distribuição multidimensional. O LHS recebe como entrada o grid original e gera como saída um grid com menor densidade, a partir dos pontos já amostrados melhor distribuídos na área. Este método de amostragem é muito utilizado para construir experimentos em computador (OLSSON; SANDBERG; DAHLBLOM, 2003). Vários trabalhos têm utilizado o algoritmo *LHS* como técnica para determinação dos pontos amostrais desde os anos 80 (CARRÉ; MCBRATNEY; MINASNY, 2007; MINASNY; MCBRATNEY, 2007a; MULDER; DE BRUIN; SCHAEPMAN, 2012; OLSSON; SANDBERG, 2002; SHIELDS; ZHANG, 2016).

A partir dos grids de amostragem gerados pelo algoritmo *LHS* para a AREA1 e AREA2, realizou-se a interpolação para cada um dos 10 atributos do solo. Dessa forma, obteve-se os mapas interpolados para cada atributo, grid amostral, em cada uma das duas áreas. Os mapas interpolados foram gerados utilizando os métodos IDW, OK e um híbrido IDW-SVM. O método híbrido consistiu em usar o método IDW como método preditor do atributo do solo e, em seguida, usar o método *SVM* para corrigir a predição. O método criado foi denominado IDW-SVM.

Na AREA1 foi definido um grid com células de tamanho 5 x 5 m, totalizando 7860 pontos interpolados. Na AREA2 o grid foi definido com dimensões de 15 x 15 m, totalizando 8705 pontos interpolados. Um algoritmo baseado no método IDW-SVM foi desenvolvimento utilizado a linguagem Python versão 3.7 (criado por Guido Van Rossum e gerenciado pela Python Software Foundation, Delaware, EUA). A linguagem Python foi escolhida, devido a sua facilidade de adoção e por possuir diversas bibliotecas e pacotes disponíveis. Linguagens de programação de alto nível, como R e Python, têm aumentado a acessibilidade aos algoritmos de ML (KHALEDIAN; MILLER, 2020).

Critérios de performance para comparação entre os métodos de interpolação

A dependência espacial dos 10 atributos de solo analisados em cada área foi calculada de acordo com o Índice de Moran univariado. O Índice de Moran é um indicador útil para analisar a autocorrelação espacial de variável ambiental. Seu valor varia entre −1 (autocorrelação espacial negativa perfeita) e +1 (autocorrelação espacial positiva perfeita), e o valor 0 indica que não há autocorrelação (GUO et al., 2015). O Índice de Moran foi calculado de acordo com a Equação 2.1 (LEGENDRE; FORTIN, 1989).

$$I = \frac{n}{W}.\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x}).(x_j - \bar{x})}{\sum_{i}^{n}(x_i - \bar{x})^2} \quad (2.1)$$

em que: n é o número de pontos amostrados; x_i, x_j representam os valores observados dos atributos de solo nos pontos i, j; $\bar{x}$ a média dos *n* pontos observados do atributo de solo; w_{ij} é a matriz de pesos espaciais com valor 0 na diagonal (w_{ii} = 0); e W representa a soma de todos os w_{ij}.. A matriz de peso espacial (W) foi calculada usando a biblioteca Python PySAL (Rey e Anselin, 2010). A biblioteca PySAL implementa uma função "*kernel*" que calcula os pesos dos vizinhos com base na distância de separação.

Para se conhecer a dependência espacial dos atributos de solo analisados em grids com diferentes densidades amostrais, foram geradas 100 combinações de pontos amostrais a partir de 100 execuções do algoritmo *LHS* para cada atributo do solo e densidade do grid. A partir das 100 execuções do algoritmo *LHS* para cada densidade amostral foram calculados os valores médios dos Índices de Moran e o pseudo p-valor para cada atributo e densidade do grid.

Os critérios de comparação utilizados para medir o desempenho entre os métodos foram o Coeficiente de Determinação (R^2) e a Raiz Quadrada do Erro Quadrático Médio (RMSE) da validação cruzada *leave-one-out cross-validation (LOOCV)* (CELISSE; ROBIN, 2008). A partir das 100 combinações geradas pelo algoritmo *LHS* para cada atributo de solo e densidade do grid, escolheu-se uma combinação ao acaso para comparar a performance dos modelos. O método *LOOCV* foi aplicado para os três métodos de interpolação, IDW, OK e IDW-SVM. Após

predição de todos os pontos do grid amostral pelo método *LOOCV*, foram calculados o R^2 e o RMSE entre os valores preditos e os valores observados.

O R^2 e o RMSE foram calculados de acordo com as Equações 2.2 e 2.3 respectivamente. O R^2 é um valor relativo que indica a quantidade de variação explicada pelo modelo, um valor próximo de 1 indica um modelo perfeito, isto é, 100% da variação tem sido explicada pelo modelo (HENGL et al., 2018). O RMSE é um valor absoluto que está na mesma unidade do atributo interpolado, descreve a precisão do modelo, o que significa quão próximos os valores previstos estão dos valores reais, para isso utiliza a raiz quadrada da soma dos quadrados dos resíduos.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \widehat{x_i})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2} \tag{2.3}$$

em que: $\widehat{x_i}$ representa o valor estimado do atributo de solo no ponto i; $\bar{x}$ a média dos *n* pontos amostrados do atributo de solo; x_i o valor observado do atributo de solo no ponto i; e n, o número de pontos amostrados.

Interpolação pelo método IDW

O método IDW é popular entre os usuários de sistemas de informações geográficas (*GIS*). Ele fornece uma boa estimativa da variável espacial (WONG, 2017), além de ser relativamente simples de entender e usar. Um algoritmo em Python foi implementado para calcular a interpolação por IDW, considerando cada atributo de solo e densidade do grid amostral. O IDW foi calculado de acordo com a Equação 2.4.

$$\hat{v} = \frac{\sum_{i=1}^{n}\frac{1}{d_i^p}.v_i}{\sum_{i=1}^{n}\frac{1}{d_i^p}} \tag{2.4}$$

em que: n é o número de pontos vizinhos em relação ao ponto a ser interpolado; d_i representa a distância do i-ésimo ponto ao ponto a ser interpolado; v_i representa o valor observado no i-ésimo ponto; p representa a potência a ser utilizada para o IDW.

Nas duas áreas para os 10 atributos de solo e densidades de grids amostrais deferentes foi considerado o valor p = 1 e n = 16, considerando-se assim os 16 vizinhos amostrados mais próximos ao ponto a ser estimado.

Interpolação pelo método OK

A análise geoestatística e a construção dos mapas interpolados foram realizadas no software GS+, versão 5.0 (GAMA DESIGN SOFTWARE, 2000). A estrutura espacial de cada atributo de solo foi caracterizada e quantificada por semivariogramas utilizando modelos teóricos (linear, linear com patamar, esférico, exponencial e gaussiano). O ajuste do modelo de semivariância teórico foi baseado na menor soma residual dos quadrados (RSS) e no maior coeficiente de determinação (R^2). Além disso, foram definidos os seguintes parâmetros para cada modelo: efeito pepita (C0), contribuição (C1), patamar (C0 + C1) e alcance (A). O número de vizinhos considerados para a obtenção dos mapas interpolados por OK foi igual a 16, o mesmo número usado quando aplicado o método IDW.

Como foram interpolados 10 atributos do solo em cada área com 5 diferentes grids de amostragem, foram necessários ajustar 50 semivariogramas para a AREA1 e 50 semivariogramas para a AREA2, totalizando assim 100 semivariogramas ajustados.

Interpolação pelo método IDW-SVM

Para treinar o modelo de interpolação pelo método IDW-SVM definiu-se uma matriz de treinamento (*X*). Na matriz *X* as coordenadas (*coordX* e *coordY*) do ponto e o valor predito do atributo de solo alvo (*idwA*) obtido pelo método IDW, calculado com base nos valores dos 16 vizinhos mais próximos do ponto a ser estimado, foram considerados como variáveis independentes (*features*). O valor conhecido no próprio ponto não entrou no cálculo do atributo *idwA*. Foi gerado assim uma matriz *X*, em que as *n* linhas representavam os *n* pontos amostrais e três colunas como as variáveis independentes (*coordX*, *coordY*, *idwA*). O valor conhecido do atributo de solo no ponto amostrado (*target_A*) foi considerado como variável dependente (*y*) do modelo. A Figura 2.3 representa o conjunto de treinamento do método IDW-SVM dividido em variáveis independentes (matriz *X*) e variável dependente (vetor *y*).

Especificamente em modelos de ML para predição e mapeamento de atributos de solo podem ser utilizadas como *features* do modelo: coordenadas georreferenciadas, atributos de solos, valores de reflectância obtidos por imagens de satélite, dados de sensores, dentre outras informações que contribuem para a construção do modelo de ML. No entanto, no presente estudo, optou-se por utilizar como *features*, apenas o IDW do próprio atributo de solo a ser interpolado e as informações georreferenciadas dos pontos. A variável dependente (*y*) representa o atributo de solo observado, para o qual se quer predizer seus valores em locais não amostrados. A matriz *X* e o vetor *y* foram as entradas do conjunto de treinamento do método IDW-SVM. Os dados de entrada foram padronizados em média zero e desvio padrão um.

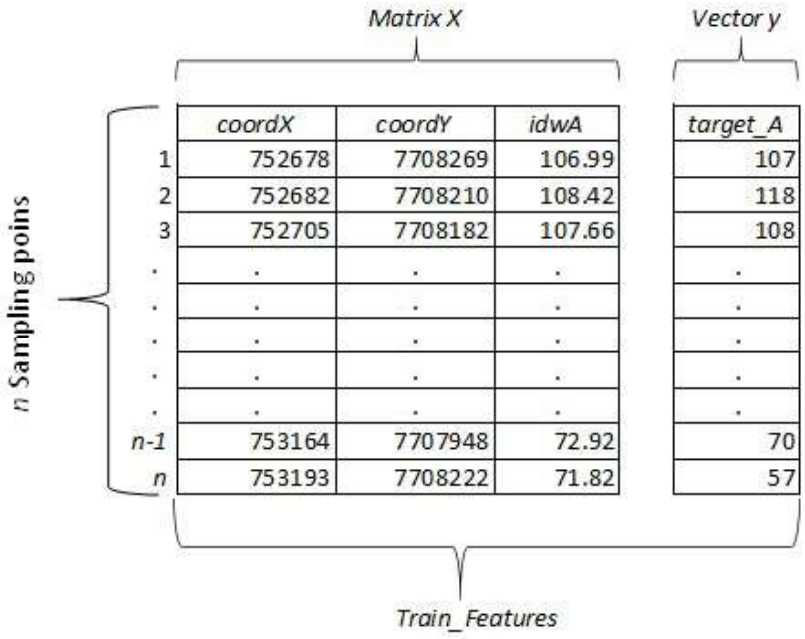


Figura 2.3: Definição do conjunto de features (Matriz X) e target (Vetor y) de treinamento do algoritmo IDW-SVM. Na Matriz X, coordX e coordY são as coordenadas x e y do ponto amostrado, respectivamente; idwA representa o cálculo do IDW dos 16 vizinhos mais próximo do ponto amostrado. No Vetor y, target_A representa os valores amostrados do atributo a ser interpolado. Cada linha dos dados de treinamento representa uma amostra.

Para construção do modelo IDW-SVM foi utilizado o *kernel RBF* (*Radial Basis Function*). Dois hiperparâmetros *C* e *gamma,* possuem um número infinito de combinações por seus valores; portanto, o desafio é a identificação de uma combinação "ideal" de valores para esses hiperparâmetros (HEUNG et al., 2016). Com o objetivo de diminuir o tempo na parametrização destes hiperparâmetros um método sistemático de busca em grade foi utilizado para otimização (XU et al., 2018). A validação cruzada *k-fold* foi utilizada para se obter os valores ótimos destes

hiperparâmetros. Nesse método o conjunto de dados foi dividido aleatoriamente em cinco subconjuntos (*5-folds*) - quatro das partições, que representam 80% dos dados, foram usadas para treinar o modelo e os 20% restantes foram usados para validação. Este processo foi repetido 5 vezes, usando cada *fold* para validação uma única vez. Para cada modelo, uma faixa de valores dos hiperparâmetros foi testada e as previsões finais foram feitas com base na combinação de valores de parâmetros que produziram as menores taxas de erro média do procedimento de validação cruzada (CV). Procedimentos semelhantes foram implementados em (BRUNGARD et al., 2015; HEUNG et al., 2016; HEUNG; BULMER; SCHMIDT, 2014). A Figura 2.4 apresenta a validação cruzada 5-*fold* implementada para otimização dos hiperparâmetros.

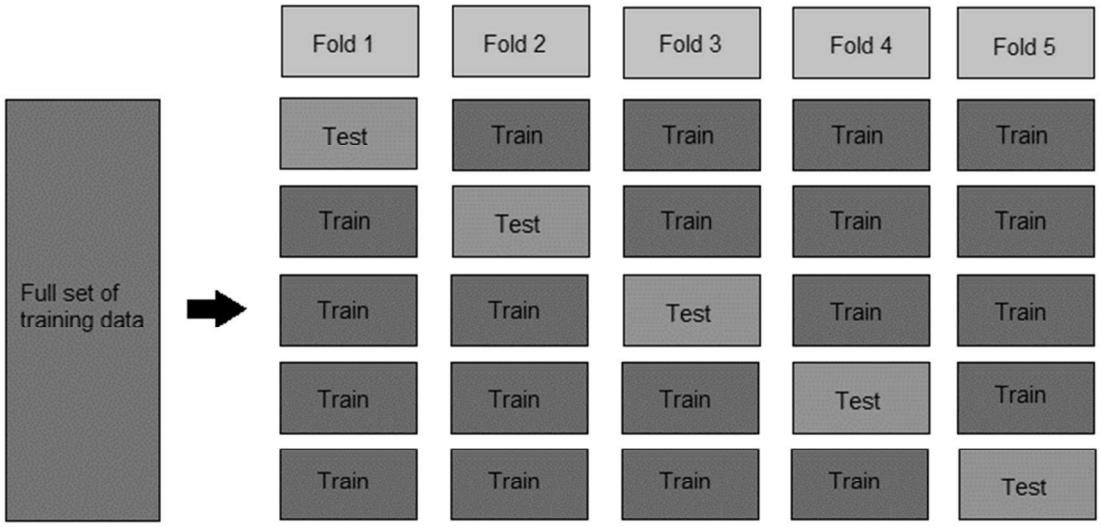


Figura 2.4: Divisão dos dados em conjunto de treino e teste para realizar validação cruzada 5-fold para otimizar os parâmetros C e gamma do SVM.

Os valores ótimos de *C* e *gamma* foram determinados para cada atributo e densidade do grid amostral. Após definido o modelo IDW-SVM foi então realizada uma validação cruzada *leave-one-out* (*LOOCV*), da mesma forma que foi realizada com os métodos de interpolação IDW e OK. A validação cruzada *LOOCV* consiste em utilizar todos os dados de treinamento deixando um de fora. O ponto de fora então é interpolado por um dos métodos de interpolação. Essa estratégia é aplicada para todos as amostras do conjunto de treinamento. Como os valores reais do conjunto de treinamento são conhecidos, calcula-se dessa forma os valores de R^2 e RMSE da validação cruzada *LOOCV*.

O conjunto de teste foi obtido a partir do grid de células do mapa a serem interpolados (Figura 2.5). Dessa forma, na AREA1 foram gerados 7860 pontos (números de células de tamanho 5 x 5 m na AREA1) e 8705 pontos (números de células de tamanho 15 x 15 m na AREA2). A interpolação (predição) de cada ponto do conjunto de teste foi obtida a partir do modelo IDW-SVM ajustado a partir do conjunto de treinamento.

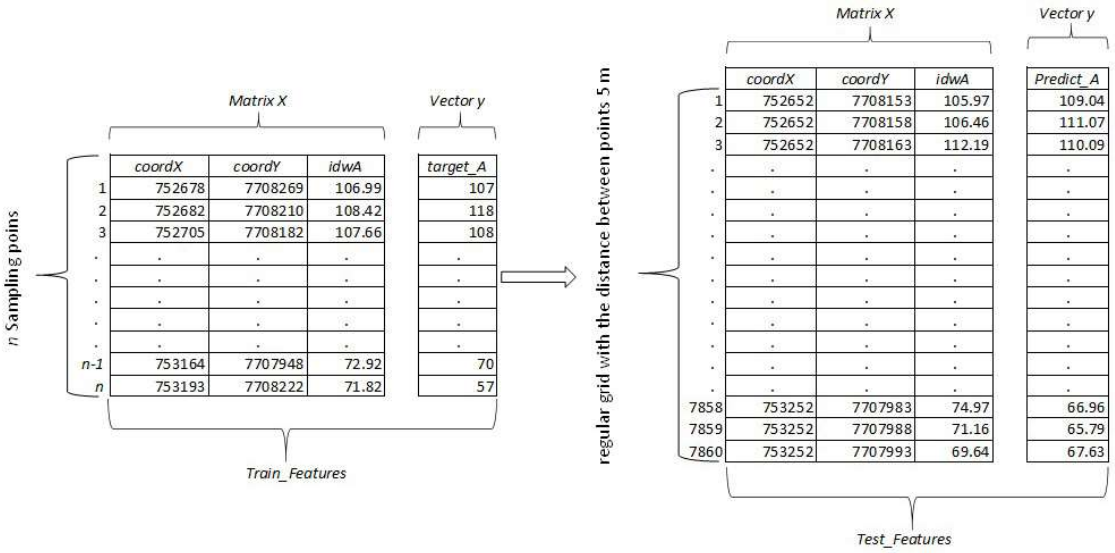


Figura 2.5: Definição do conjunto de teste (grid de células a serem interpolados) para geração dos mapas interpolados pelo método IDW-SVM.

As colunas *coordX* e *coordY* do conjunto de teste formam as coordenadas (X, Y) do ponto central do grid de células 5m x 5m na AREA1 e 15m x 15m na AREA2. A coluna *idwA* no conjunto de teste foi obtido pelo IDW do atributo de solo alvo *(target_A*), calculado (de acordo com a Equação 4) com base nos valores dos 16 vizinhos mais próximos (obtidos no conjunto de treinamento) ao ponto a ser estimado no conjunto de teste (mapa interpolado).

Após a definição das *features* no conjunto de teste, realizou-se o ajuste no modelo IDW-SVM utilizando para os hiperparâmetros *C* e *gamma* os valores encontrados na otimização durante a definição do conjunto de treinamento. Com o modelo ajustado obteve-se os valores preditos (*Predict_A*) para o conjunto de teste. Estes valores preditos pelo método IDW-SVM para o conjunto de teste foram utilizados para gerar o mapa interpolado para cada atributo e densidade do grid.

2.5 Resultados e discussão

Autocorrelação espacial com Índice de Moran

A partir das 100 execuções do algoritmo *LHS* foi possível calcular o valor médio do Índice de Moran para cada atributo e densidade do grid (número de amostras). Na AREA1 foram considerados densidades amostrais de 7 a 141 pontos e na AREA2 de 7 a 204 pontos. Na Figura 2.6 é possível observar o valor médio do Índice de Moran para cada atributo de solo e densidade do grid. A linha AVG representa o valor médio entre os 10 atributos analisados para cada densidade do grid. Nas duas áreas, à medida que se aumentou a densidade do grid amostral, percebeu-se uma tendência de aumento do Índice de Moran.

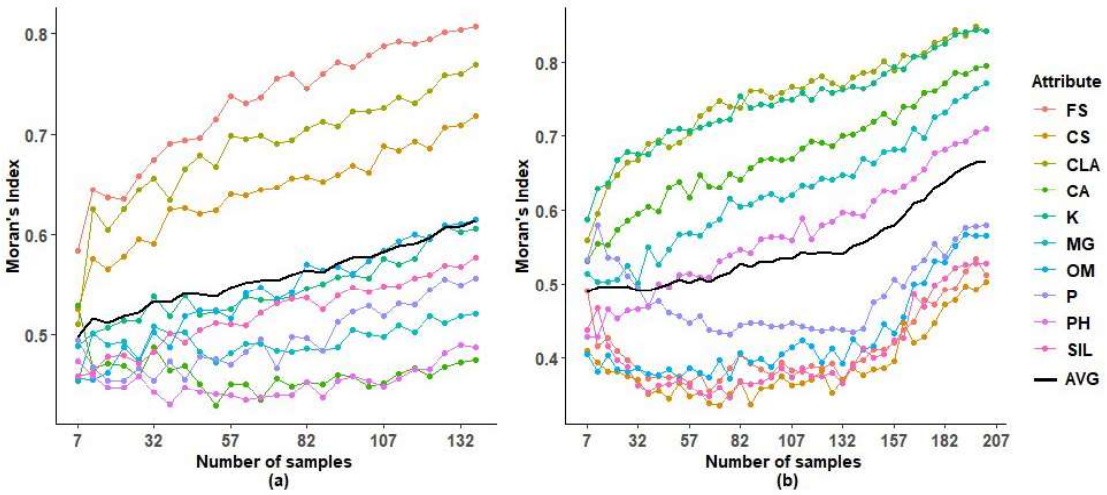


Figura 2.6: Índice de Moran médio calculado para os atributos de solo analisados: (a) AREA1, (b) AREA2. FS, Areia Fina; CS, Areia Grossa; CLA, Argila; CA, Cálcio; K, Potássio; MG, Magnésio; OM, Matéria Orgânica; P, Fósforo; PH, Acidez Ativa em água; SIL, Silte; AVG, Média dos atributos analisados.

Na AREA1 (Figura 2.6a) pode-se observar um crescimento da média do Índice de Moran à medida que se aumenta o número de pontos. Isso provavelmente ocorreu porque os atributos do solo apresentavam, em média, um equilíbrio em termos de comportamento de dependência espacial. Além disso, a AREA1 é 10 vezes menor que a AREA2. Os atributos de solo FS, CLA e CS apresentaram um Índice de Moran médio superior aos outros atributos e com um crescimento constante entre as combinações de 7 a 141 pontos. Os demais atributos também apresentaram um crescimento constante, porém, menos acentuado, com valores médios inferiores à média (linha AVG) dos 10 atributos de solo.

Na AREA2 (Figura 2.6b) os atributos CLA, K^+, Ca^{2+}, Mg^{2+} e pH apresentaram uma tendência crescente para o Índice de Moran, com valores superiores à média geral entre os 10 atributos analisados. Enquanto que P, OM, SIL, FS e CS, mantiveram-se constante até 132 pontos. A partir de 132 pontos, esses atributos apresentaram um crescimento consistente e mais acentuado do Índice de Moran. Possivelmente a partir dessa densidade amostral houve um aumento de dependência espacial para a maioria dos atributos do solo dessa área.

Comparação entre os métodos IDW, OK e IDW-SVM

Para realizar a comparação entre os métodos IDW, OK e IDW-SVM escolheu-se ao acaso dentre as 100 combinações geradas pelo algoritmo *LHS*, para cada densidade de grid amostral, uma combinação de pontos amostrais. O Índice de Moran e o pseudo p-valor para os 10 atributos do solo e 5 densidades de grids amostrais da combinação escolhida são apresentados nas Tabela 2.3 e 2.4 para as AREA1 e AREA2 respectivamente. O pseudo p-valor foi obtido a partir de 999 permutações entre os pontos do grid amostral com o objetivo de verificar sua significância ao nível de 5% de probabilidade.

Tabela 2.3: Valores de Índice de Moran e pseudo p-valor calculados para cada conjunto de amostras escolhida, para 5 grids amostrais e 10 atributos de solo na AREA1.

Atributo	25 Amostras		50 Amostras		75 Amostras		100 Amostras		141 Amostras	
	Moran	p-valor	Moran	p-valor	Moran	p-valor	Moran	p-valor	Moran	p-valor
pH (1)	0,30	0,306	0,37	0,006*	0,37	0,254	0,51	0,056	0,49	0,009*
P (2)	0,48	0,001*	0,47	0,011*	0,52	0,019*	0,53	0,002*	0,56	0,001*
K^+ (3)	0,40	0,073	0,50	0,002*	0,51	0,014*	0,58	0,002*	0,61	0,001*
Ca^{2+} (4)	0,33	0,244	0,43	0,030*	0,33	0,073	0,47	0,183	0,48	0,015*
Mg^{2+}(5)	0,31	0,288	0,46	0,014*	0,40	0,427	0,58	0,055	0,52	0,001*
OM (6)	0,35	0,457	0,52	0,001*	0,59	0,001*	0,61	0,001*	0,62	0,001*
CS (7)	0,67	0,001*	0,66	0,001*	0,79	0,001*	0,76	0,001*	0,81	0,001*
FS (8)	0,47	0,023*	0,61	0,001*	0,73	0,001*	0,67	0,001*	0,72	0,001*
SIL(9)	0,43	0,028*	0,46	0,012*	0,52	0,009*	0,58	0,002*	0,57	0,001*
CLA (10)	0,65	0,001*	0,74	0,001*	0,75	0,001*	0,71	0,001*	0,77	0,001*

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte; 10/ CLA, Argila.

Obs: *Significância ao nível de 5% de probabilidade.

Tabela 2.4: Valores de Índice de Moran e pseudo p-valor calculados para cada conjunto de amostras escolhida, para 5 grids amostrais e 10 atributos de solo na AREA2.

Atributo	36 Amostras		72 Amostras		108 Amostras		144 Amostras		204 Amostras	
	Moran	p-valor	Moran	p-valor	Moran	p-valor	Moran	p-valor	Moran	p-valor
pH (1)	0,32	0,491	0,48	0,004*	0,62	0,001*	0,59	0,001*	0,71	0,001*
P (2)	0,45	0,345	0,35	0,002*	0,40	0,310	0,45	0,012*	0,58	0,002*
K^+ (3)	0,73	0,001*	0,71	0,001*	0,77	0,001*	0,82	0,001*	0,84	0,001*
Ca^{2+} (4)	0,55	0,033*	0,53	0,001*	0,77	0,001*	0,71	0,001*	0,80	0,001*
Mg^{2+}(5)	0,51	0,019*	0,55	0,001*	0,73	0,001*	0,70	0,001*	0,77	0,001*
OM (6)	0,34	0,065	0,30	0,113	0,47	0,055	0,40	0,042*	0,57	0,003*
CS (7)	0,43	0,396	0,20	0,230	0,37	0,477	0,38	0,229	0,50	0,240
FS (8)	0,41	0,498	0,27	0,254	0,43	0,094	0,42	0,054	0,52	0,097
SIL(9)	0,50	0,245	0,33	0,062	0,47	0,054	0,41	0,111	0,53	0,108
CLA (10)	0,67	0,001*	0,68	0,001*	0,76	0,001*	0,82	0,001*	0,84	0,001*

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte; 10/ CLA, Argila.

Obs: *Significância ao nível de 5% de probabilidade.

Para o grid de densidade menor, 25 pontos amostrais da AREA1 (Tabela 2.3), aquele com a menor densidade de pontos, os atributos do solo P, CS, FS, SIL e CLA apresentaram dependência espacial significativos. FS, CLA e CS foram os atributos com maior Índice de Moran médio (Figura 2.6a). Para uma densidade de 36 pontos da AREA2 (Tabela 2.4) foram os atributos K^+, Ca^{2+}, Mg^{2+}e CLA que tiveram os maiores Índice de Moran médio (Figura 2.6b).

Para o grid de densidade maior, 141 pontos da AREA1 (Tabela 2.3) todos os atributos tiveram dependência espacial significativa. Com densidade de 204 pontos da AREA2 (Tabela 2.4) somente os atributos do solo CS, FS e SIL não apresentaram Índice de Moran significativos. O Índice de Moran para estes 3 atributos não foram significativos para nenhuma densidade de grid amostral. Também foram estes 3 atributos que tiveram os menores Índice de Moran médio (Figura 2.6b). Os atributos de composição textural na AREA1 como CLA, CS e FS e os macronutrientes na AREA2 como K^+, Ca^{2+}, Mg^{2+} foram os atributos com melhores valores de Índice de Moran, como apresentado na Figura 2.6 e Tabelas 2.3 e 2.4.

O R^2 e o RMSE foram utilizados como critérios de performance para comparar os métodos IDW, OK e IDW-SVM, a partir da validação cruzada *LOOCV*. A superioridade do IDW-SVM neste trabalho está relacionada ao fato de o método utilizar os valores interpolados pelo IDW como variável independente no modelo ML.

Neste trabalho a terceira *feature* (*idwA*) da matriz X do conjunto de treinamento, foi obtida a partir da interpolação (utilizou-se o interpolador IDW) dos valores da vizinhança da variável dependente (y). No modelo *SVM* implementado foram utilizados além das coordenadas X, Y, apenas um atributo como mostrado na Figura 2.3. Um modelo com um pequeno número de *features* é mais interpretável e a precisão do modelo pode ser melhorada e evitar o risco de sobre ajuste (GREGORUTTI; MICHEL; SAINT-PIERRE, 2017). Outra característica importante dos algoritmos de *Machine Learning* está relacionada a sua capacidade em reconhecimento de padrões. Os algoritmos de ML em especial o *SVM* com kernel RBF permite que o método produza um plano de separação não linear entre as classes possibilitando a modelagem de fenômenos não lineares que podem ocorrer na distribuição espacial dos atributos de solo (HEUNG et al., 2016; KHALEDIAN; MILLER, 2020; WERE et al., 2015).

Os valores de R^2 e RMSE para os 10 atributos de solo analisados separadamente na AREA1, para os métodos IDW, OK e IDW-SVM podem ser visualizados nas Figuras 2.7 e 2.8. Os atributos que apresentaram os melhores valores de R^2 foram CLA, CS e FS, considerando os 5 grids amostrais com 25, 50, 75, 100 e 141 pontos (Figura 2.7). O Índice de Moran para estes atributos foram significativos, independente da densidade do grid como mostrado na Tabela 2.3, com valores entre 0,47 (FS, 25 amostras) e 0,81 (CS, 141 amostras). Na Figura 2.6a foram estes os atributos de solo que apresentaram os melhores valores médios de Índice de Moran. O método OK apresentou melhores resultados com estes atributos a partir do grid de densidade de 75 pontos, confirmando o fato de que para a krigagem deve-se ter uma quantidade de pontos considerável para conseguir um bom ajuste do semivariograma (GIACOMIN et al., 2014; WEBSTER; OLIVER, 1992) e estes pontos devem ter uma boa correlação espacial. O Índice de Moran aumentou de 0,71 à 0,77 para CLA e de 0,76 à 0,81 para CS para os grids com 100 a 141 pontos amostrais respectivamente, e diminuiu de 0,73 a 0,67 para FS, para os grids com 75 a 100 pontos amostrais, respectivamente. O método IDW-SVM obteve bons resultados tanto para grid com baixa densidade quanto para o grid com uma densidade maior.

O método IDW-SVM produziu os menores erros de previsão para a maioria dos atributos (Figura 2.8). No grid amostral de 25 pontos foram os atributos CLA, CS, FS, K^+, Mg^{2+} e pH que tiveram os menores valores de RMSE para o IDW-SVM. Mais altos

valores de R^2 também foram obtidos para o IDW-SVM quando comparado com os valores do IDW e OK. No grid amostral com 141 pontos o método IDW-SVM obteve os menores valores de RMSE e os maiores valores de R^2 para os atributos OM, SIL, P, Mg^{2+} e Ca^{2+}. À medida que se diminui a densidade do grid amostral o método IDW-SVM tende a ser melhor que o método OK, tanto no R^2 quanto no RMSE. Com o aumento da densidade do grid amostral a diferença entre os dois métodos foi menor, considerando-se os 10 atributos de solo analisados.

A Tabela 2.5 apresenta a área calculada dos polígonos para os dez atributos de solos analisados na AREA1 considerando as cinco densidades amostrais. De acordo com a área pode-se medir o desempenho de cada método para cada atributo de solo. Quanto maior a área, maior é o R^2 e o RMSE. O método OK apresentou os maiores valores de R^2 para os atributos CS, FS e CLA. Na Figura 2.6a estes três atributos apresentaram os maiores valores médios para o Índice de Moran. Para os demais atributos o método IDW-SVM foi superior.

Tabela 2.5: Área calculada para os três métodos de interpolação para os dez atributos de solos analisados na AREA1.

Atributo	Unidade	IDW		OK		IDW-SVM	
		R2	RMSE	R2	RMSE	R2	RMSE
pH (1)		0,006	0,540	0,008	0,533	0,073	0,374
P (2)	(mg dm^{-3})	0,052	9,271	0,028	9,959	0,105	6,816
K^+ (3)	(mg dm^{-3})	0,097	1867,3	0,091	2045,5	0,105	1646,2
Ca^{2+} (4)	(cmolc dm^{-3})	0,003	3,316	0,001	2,552	0,030	1,829
Mg^{2+}(5)	(cmolc dm^{-3})	0,013	0,446	0,016	0,442	0,060	0,260
OM (6)	(g dm^{-3})	0,141	1,971	0,136	1,724	0,187	1,824
CS (7)	(g kg^{-1})	0,599	30,020	0,712	22,594	,0,677	22,924
FS (8)	(g kg^{-1})	0,422	4,452	0,467	4,108	0,452	4,247
SIL(9)	(g kg^{-1})	0,076	5,709	0,089	7,289	0,113	5,532
CLA (10)	(g kg^{-1})	0,663	30,835	0,719	26,230	0,641	29,021

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte; 10/ CLA, Argila.

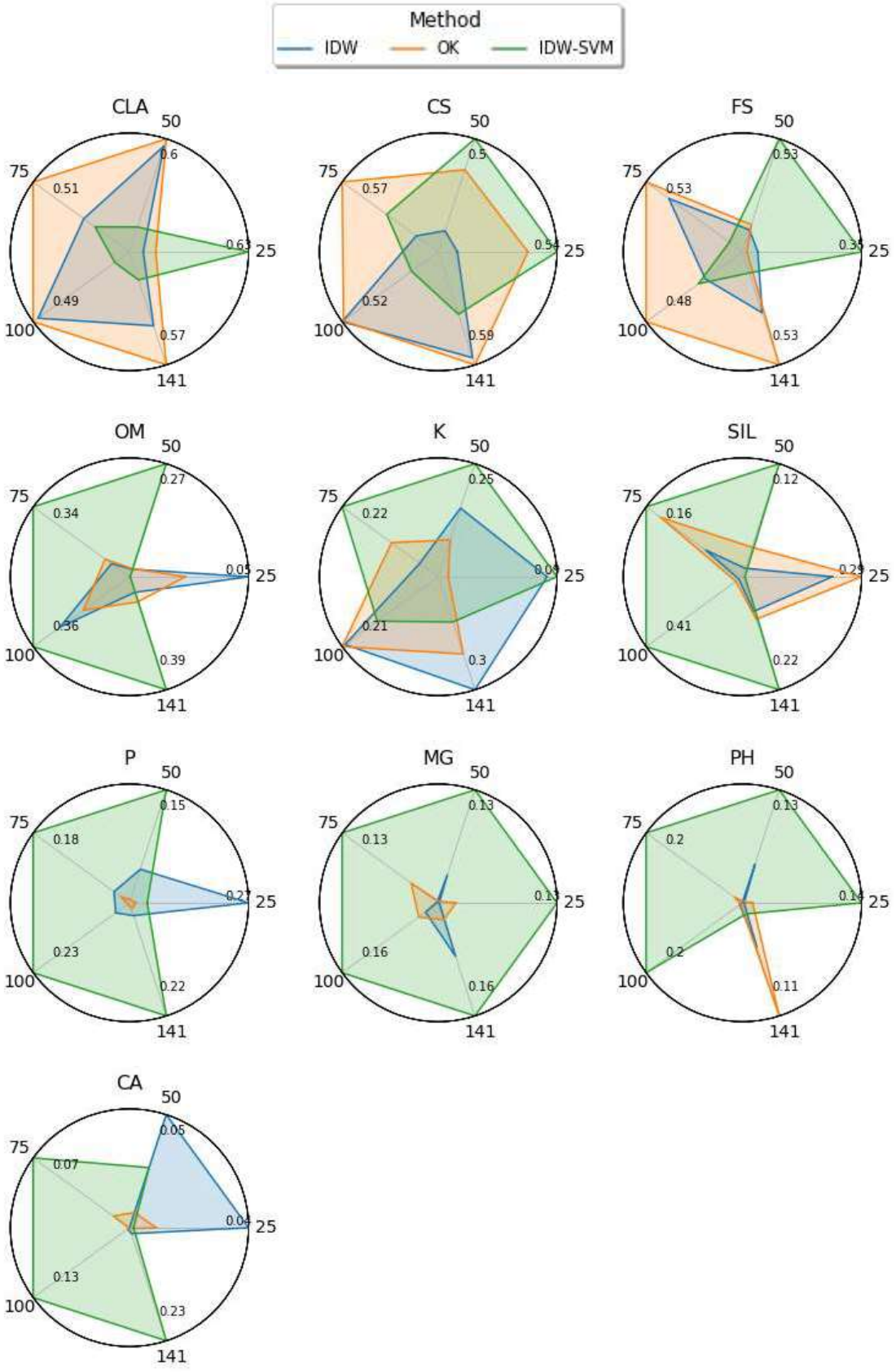


Figura 2.7: R^2 calculado para os 10 atributos de solo analisados na AREA1 para diferentes números de pontos amostrais.

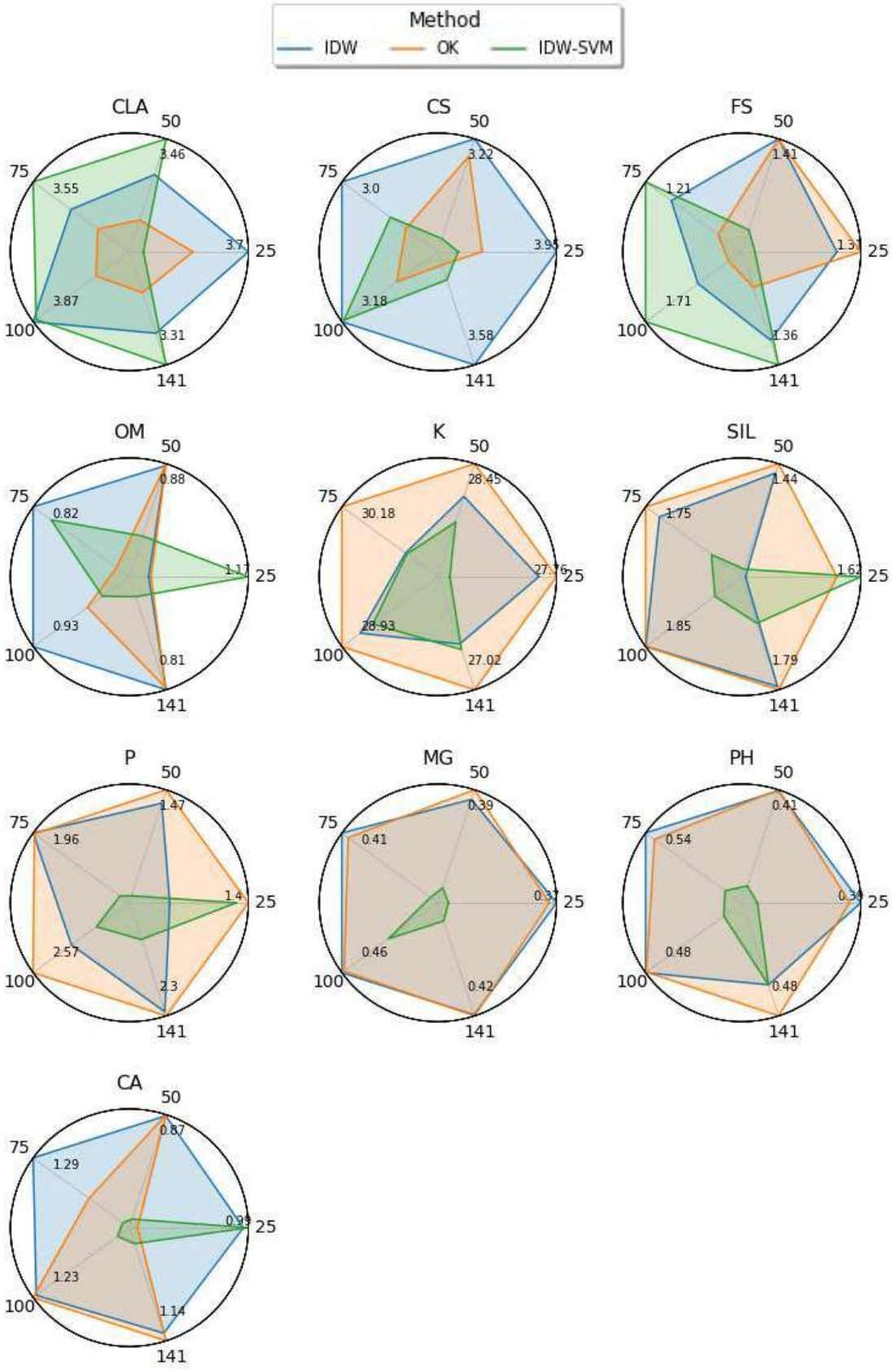


Figura 2.8: RMSE calculado para os 10 atributos de solo analisados na AREA1 para diferentes números de pontos amostrais.

Os valores de R^2 para os 10 atributos de solo analisados separadamente na AREA2, para os métodos IDW, OK e *SVM* são apresentados na Figuras 2.9. Os atributos K^+, CLA, Ca^{2+}, Mg^{2+} apresentaram melhores valores de R^2 (Figura 2.9). Para os grids de 36 e 72 pontos K^+ e CLA obtiveram os maiores valores para o Índice de Moran, entretanto devido ao baixo número de amostras (36 pontos em uma área de 204 ha), R^2 da OK foi inferior ao IDW-SVM para o grid com densidade de 36 pontos. A partir de 72 pontos o método OK foi superior ao IDW-SVM e IDW para K^+ e CLA, e a partir de 108 pontos o OK foi superior para os atributos K^+, CLA, Ca^{2+}, Mg^{2+}. Para os grids com 108, 144 e 204 pontos, o menor Índice Moran para estes 4 atributos foi 0,70 Mg^{2+} (144 pontos) e o maior foi 0,84 (204 pontos) para CLA e K^+.

O RMSE dos 10 atributos, considerando os 5 grids amostrais na AREA2 são apresentados na Figura 2.10. Para grids com baixa densidade (36 pontos) o IDW-SVM resultou em menores valores de RMSE, quando comparados com o método OK para todos os atributos de solo, exceto P. Com grids de alta densidade (204 pontos), o RMSE para o método IDW-SVM foi inferior ao método OK para os atributos Mg^{2+}, pH, CS, FS, SIL, OM e P.

A Tabela 2.6 apresenta a área calculada dos polígonos para os dez atributos de solos analisados na AREA2 considerando as cinco densidades amostrais. O método OK apresentou os maiores valores de R^2 para os atributos pH, K^+, Ca^{2+}, Mg^{2+} e CLA. Na Figura 2.6b estes 5 atributos apresentaram os maiores valores médios para o Índice de Moran. Para os demais atributos o método IDW-SVM foi superior.

Tabela 2.6: Área calculada para os três métodos de interpolação para os dez atributos de solos analisados na AREA2.

Atributo	Unidade	IDW		OK		IDW-SVM	
		R2	RMSE	R2	RMSE	R2	RMSE
pH (1)		0,119	0,147	0,136	0,145	0,233	0,121
P (2)	(mg dm^{-3})	0,002	361,8	0,002	388,7	0,006	344,8
K^+ (3)	(mg dm^{-3})	0,810	375,6	0,818	351,4	0,731	361,9
Ca^{2+} (4)	(cmolc dm^{-3})	0,433	0,171	0,538	0,151	0,499	0,158
Mg^{2+}(5)	(cmolc dm^{-3})	0,243	0,050	0,344	0,045	0,279	0,044
OM (6)	(g dm^{-3})	0,004	0,525	0,002	0,538	0,011	0,418
CS (7)	(g kg^{-1})	0,012	128,2	0,010	128,4	0,174	89,22
FS (8)	(g kg^{-1})	0,007	131,0	0,001	138,2	0,077	94,52
SIL(9)	(g kg^{-1})	0,002	6,805	0,002	6,750	0,050	5,511
CLA (10)	(g kg^{-1})	0,675	4,275	0,776	3,813	0,687	4,274

1/ pH, Acidez Ativa em água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CS, Areia Grossa; 8/ FS, Areia Fina; 9/ SIL, Silte;

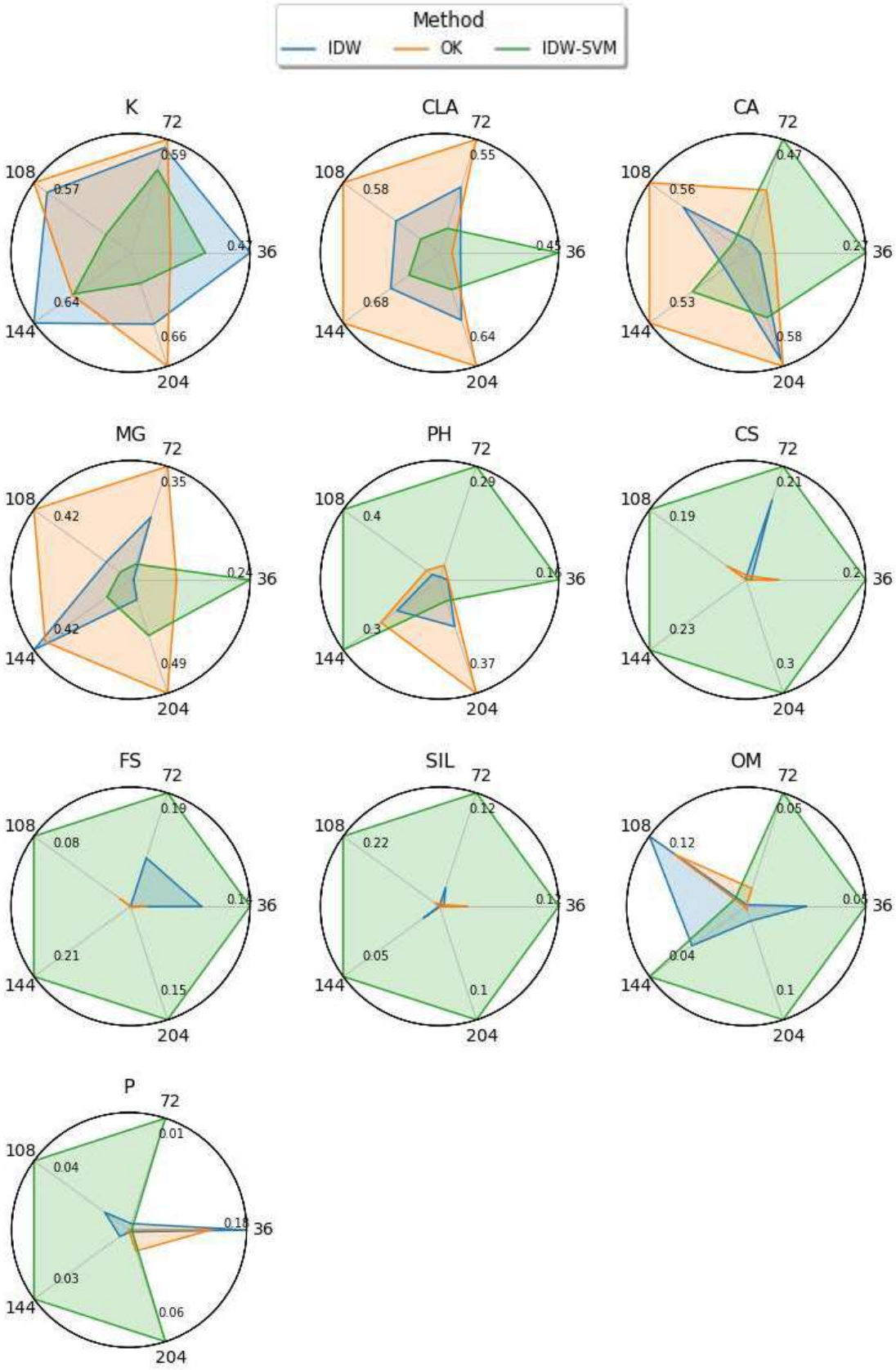


Figura 2.9: R^2 calculado para os 10 atributos de solo analisados na AREA2 para diferentes números de pontos amostrais.

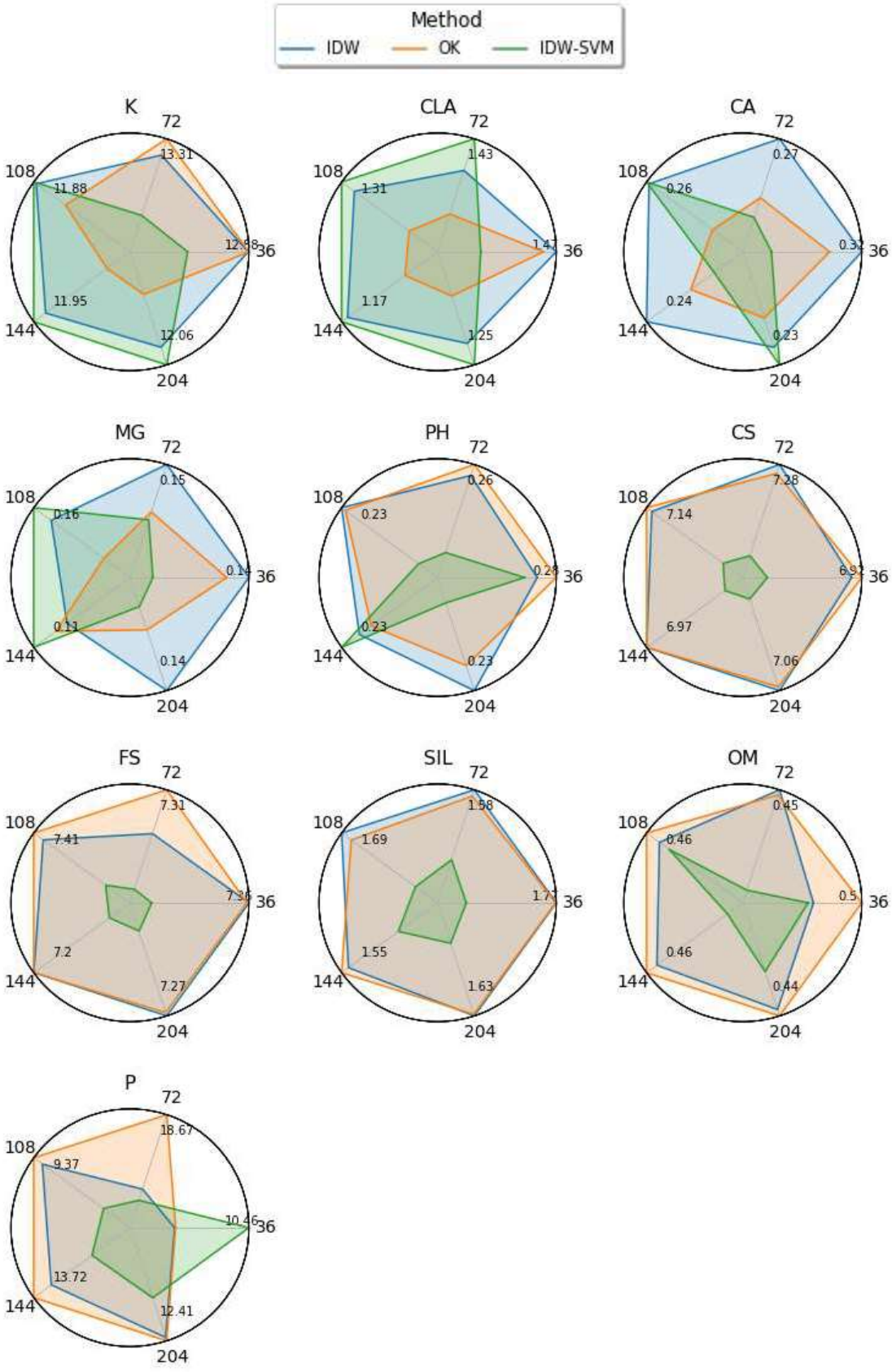


Figura 2.10: RMSE calculado para os 10 atributos de solo analisados na AREA2 para diferentes números de pontos amostrais.

Os atributos que resultaram baixo valor de R^2, na AREA1 (Ca^{2+}, pH, Mg^{2+}, P, SIL) e na AREA2 (P, OM, SIL, FS, CS) são também os atributos que apresentaram baixa correlação espacial pelo Índice de Moran, como mostrado na Figura 2.6 e Tabelas 2.3 e 2.4. Para estes atributos, seus valores do Índice de Moran ficaram abaixo do Índice de Moran médio como apresentado na Figura 2.6.

O método IDW-SVM teve bom desempenho quando comparado ao IDW e OK nas duas áreas. Principalmente para grids com baixa densidade amostral, esse desempenho apresentou-se melhor. Outros trabalhos confirmaram que a *SVM* possui baixa sensibilidade em relação ao tamanho do conjunto de amostras (KESKIN; GRUNWALD; HARRIS, 2019; KHALEDIAN; MILLER, 2020).

Correlação entre Coeficiente de Determinação R^2 e Índice de Moran

As Figuras 2.11 e 2.12 apresentam as correlações (r) entre o R^2 e o Índice de Moran para os métodos IDW, OK e IDW-SVM para a AREA1 e AREA2, respectivamente. Essas correlações são importantes para analisar o desempenho dos métodos IDW, OK e IDW-SVM com grids de diferentes densidades nas duas áreas analisadas.

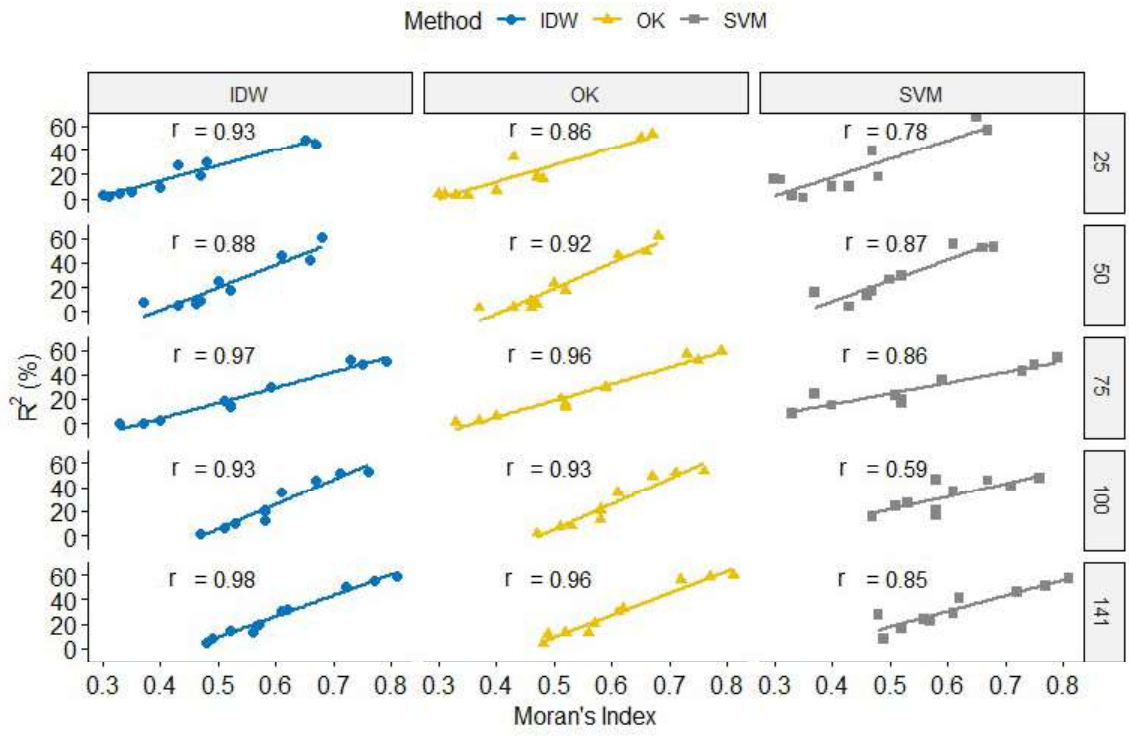


Figura 2.11: Correlação entre o R2 e Índice Moran para os atributos de solos analisados na AREA1.

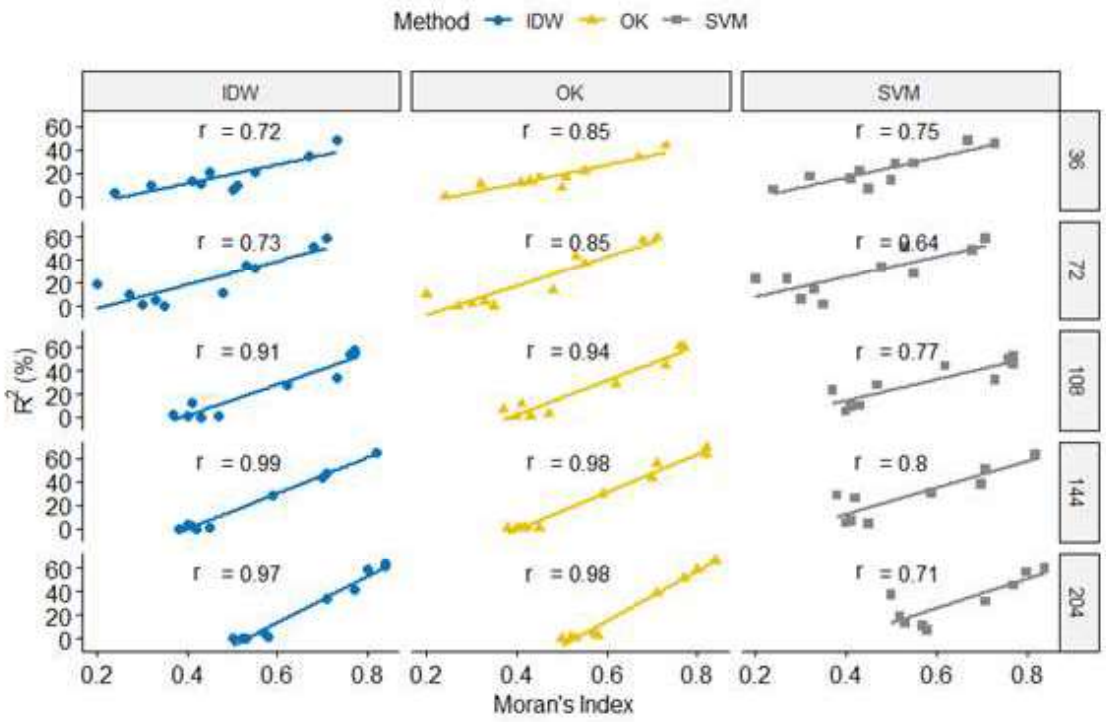


Figura 2.12: Correlação entre o R^2 e Índice Moran para os atributos de solos analisados na AREA2.

Na Figura 2.11, a correlação (r) para o método IDW variou entre 0,88 (50 amostras) e 0,98 (141 amostras), para o método OK a correlação (r) variou de 0,86 (25 amostras) à 0,96 (141 amostras). Estes resultados mostram que a correlação (r) entre o R^2 e Índice de Moran aumentou à medida que se aumentou a densidade do grid amostral para os métodos IDW e OK. Um comportamento similar também foi observado para a AREA2 (Figura 2.12). Para o método IDW a correlação (r) variou entre 0,72 e 0,97 e para o método OK essa correlação ficou entre 0,85 e 0,98 quando 36 e 204 pontos amostrais foram utilizados, respectivamente.

Por outro lado, o método IDW-SVM não apresentou o mesmo padrão de comportamento. Embora inferior essa correlação (r) entre o R^2 e o Índice de Moran, o IDW-SVM produziu melhores valores médios de R^2 e RMSE (Figuras 2.7 e 2.8). Logo o IDW-SVM apresenta uma menor dependência em relação ao Índice de Moran para produzir boas estimativas de atributos do solo, ao contrário do IDW e OK. Uma possível explicação é que o método *SVM* pode capturar padrões diferentes da dependência espacial. Em diferentes locais das áreas de estudo, pode haver gradientes com distribuições diferentes. O *SVM* pode lidar com a adaptabilidade do

comportamento em seus conjuntos de dados por meio do aprendizado supervisionado.

Mapas interpolados

Realizou-se a interpolação de todos os 10 atributos do solo nas duas áreas, obtendo-se assim os mapas interpolados para os 5 grids de densidades amostrais a partir das amostras selecionadas para cada densidade de grid amostral como apresentado nas Tabelas 2.3 e 2.4. Grids com células de tamanho 5 x 5 m e 15 x 15 m foram utilizados para a AREA1 e AREA2, respectivamente.

Os mapas interpolados gerados pelos métodos IDW, OK e IDW-SVM de Potássio (K^+) na AREA1 para os grids de densidade amostrais de 25 e 141 pontos são apresentados na Figura 2.13. As concentrações mais altas de K^+ podem ser visualizadas na parte oeste do mapa. Contudo, também foram encontradas diferenças nos padrões espaciais entre os mapas produzidos pelo IDW-SVM, OK e IDW. As transições entre os limites do K^+ foram mais suaves nos mapas de previsão produzidos pelo método IDW-SVM.

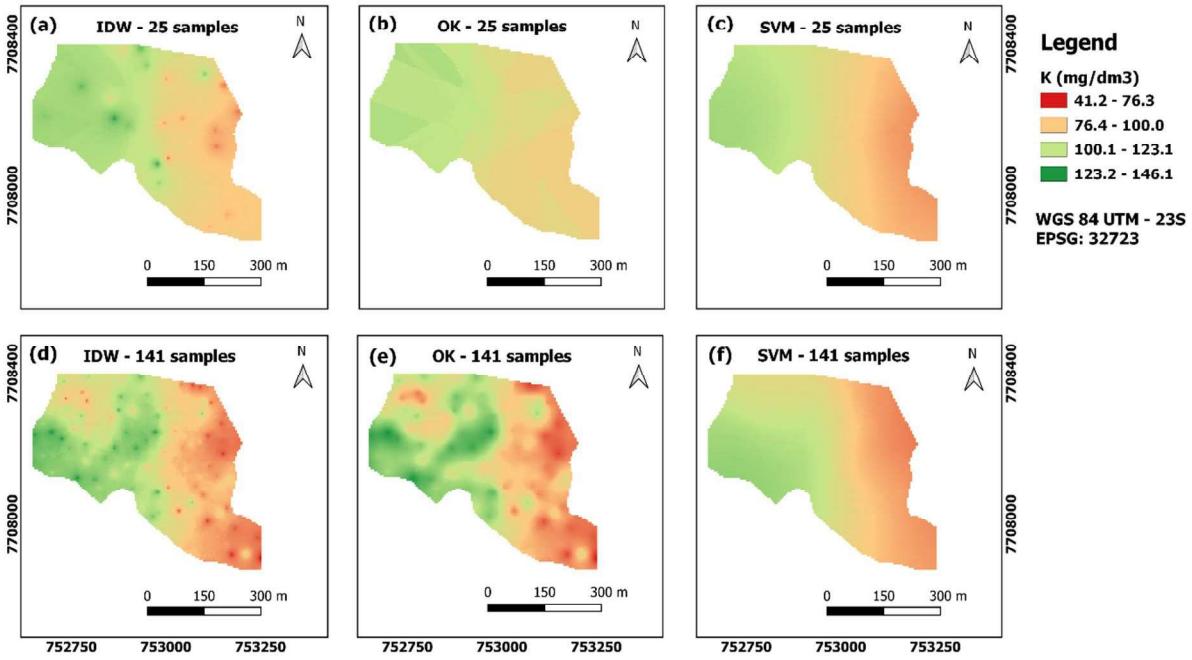


Figura 2.13: Mapa interpolado de K^+ na AREA1 para os métodos IDW, OK, SVM com densidades amostrais de 25 e 141 pontos.

Os mapas interpolados gerados pelos métodos IDW, OK e IDW-SVM de Argila (CLA) na AREA2 para os grids de densidade amostrais de 36 e 204 pontos são apresentados na Figura 2.14. A opção pelo atributo CLA foi por ser o componente de composição textural com maior Índice de Moran para todos os grids amostrais na AREA2 (Tabela 2.4). CLA também apresentou altos valores de R^2 (Figura 2.9) e baixos valores de RMSE (Figura 2.10), para todas as densidades do grid amostral. As concentrações mais baixas de CLA podem ser visualizadas na parte leste do mapa. Essa semelhança entre os mapas possivelmente está associada ao Índice de Moran (Tabela 2.4), R^2 (Figura 2.9) e RMSE (Figura 2.10).

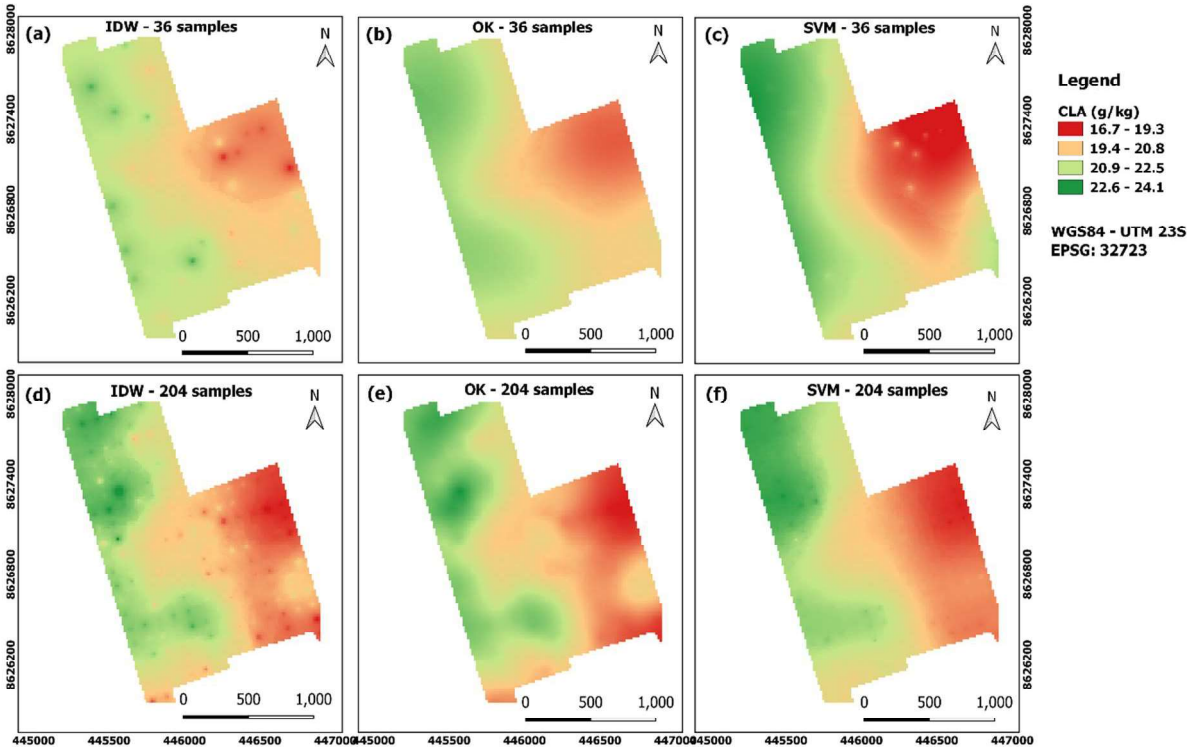


Figura 2.14: Mapa interpolado de CLA na AREA2 para os métodos IDW, OK, IDW-SVM com densidades amostrais de 36 e 204 pontos.

2.6 Conclusões

Neste estudo, 10 atributos do solo foram analisados em duas áreas distintas, em diferentes grades de amostragem, representando diferentes densidades de pontos, com o objetivo de comparar a performance entre o interpolador Inverso da Distância Ponderada (IDW), o método geoestatístico Krigagem Ordinária (OK) e um

método híbrido baseado no IDW com o algoritmo de Aprendizado de Máquina "*Support Vector Machine*" (*SVM*). Através da análise dos 10 atributos de solo, OK obteve melhores resultados que IDW e IDW-SVM para os atributos em que o Índice de Moran apresentou valores significativos e superiores a 0,67. Com baixa densidade de pontos e pouca dependência espacial, o método IDW-SVM apresentou um melhor desempenho.

Comparada com o IDW-SVM, Krigagem Ordinária apresentou uma performance média entre os 10 atributos analisados inferior, independentemente do número de pontos utilizados nas duas áreas. Essa performance se acentua à medida que o número de pontos amostrados diminui, demonstrando, portanto, que a Krigagem Ordinária é influenciada pelo número de pontos amostrais e que estes pontos devem possuir boa correlação espacial para produzir bons resultados.

Para o método IDW-SVM implementado, seus hiperparâmetros foram configurados de forma automática por meio de um algoritmo de otimização em grade, sem a necessidade de interferência do usuário. Apresentou baixa sensibilidade em relação a dependência espacial entre as amostras, se mostrando adaptado tanto a conjunto com grids de baixa densidade de pontos amostrais quanto para grids mais densos. Os resultados neste trabalho confirmaram a viabilidade e aplicabilidade de técnicas de *"Machine Learning",* em especial o algoritmo "*Support Vector Machine*", para predição e mapeamento de atributos físicos e químicos do solo em escala regional.

2.7 Referências

ALBORNOZ, E. M. et al. Development and evaluation of an automatic software for management zone delineation. **Precision Agriculture**, v. 19, n. 3, p. 463–476, 2018.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: the fuzzy c-means clustering algorithm. **Comput. Geosci.**, v. 10, p. 191–203, 1984.

BRUNGARD, C. W. et al. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239, p. 68–83, 2015.

CARRÉ, F.; MCBRATNEY, A. B.; MINASNY, B. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. **Geoderma**, v. 141, n. 1–2, p. 1–14, 2007.

CELISSE, A.; ROBIN, S. Nonparametric density estimation by exact leave-p-out cross-validation. **Computational Statistics and Data Analysis**, v. 52, n. 5, p. 2350–2368, 2008.

CHEN, S. et al. Delineation of management zones and optimization of irrigation scheduling to improve irrigation water productivity and revenue in a farmland of Northwest China**. Precision Agriculture**, v. 21, n. 3, p. 655–677, 2019.

COELHO, A. L. F. et al. An open-source spatial analysis system for embedded systems. **Computers and Electronics in Agriculture**, v. 154, p. 289–295, 2018.

COSTA, M. M. et al. Moisture content effect in the relationship between apparent electrical conductivity and soil attributes. **Acta Scientiarum - Agronomy**, v. 36, n. 4, p. 395–401, 2014.

CRESSIE, N. The Origins of Kriging. **Mathematical Geology**, v. 22, n. 3, p. 239–252, 1990.

DA MATTA CAMPBELL, P. M. et al. Digital mapping of soil attributes using machine learning. **Revista Ciencia Agronomica**, v. 50, n. 4, p. 519–528, 2019.

GAMA DESIGN SOFTWARE. **GS+ User´s Guide Version 5Gamma Design Software, (Michigan: Plainwell)**, 2000.

GIACOMIN, G. et al. Comparative Analysis of Interpolation Methods for Surface Models. **Revista Brasileira de Cartografia**, v. 66, p. 1315–1329, 2014.

GOMES, L. C. et al. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337–350, 2019.

GREGORUTTI, B.; MICHEL, B.; SAINT-PIERRE, P. Correlation and variable importance in random forests. **Statistics and Computing**, v. 27, n. 3, p. 659–678, 2017.

GUO, P. T. et al. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. **Geoderma**, v. 237–238, p. 49–59, 2015.

HEDGE, N. G. et al. Survey paper on Agriculture Yield Prediction Tool using Machine Learning. **International Journal of Advance Research in Computer Science and Management Studies**, v. 5, n. 11, p. 36–39, 2017.

HENGL, T. et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. **PLoS ONE**, v. 10, n. 6, p. 1–26, 2015.

HENGL, T. et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HENGL, T.; HEUVELINK, G. B. M.; STEIN, A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, v. 120, n. 1–2, p. 75–93, 2004.

HEUNG, B. et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62–77, 2016.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214–215, p. 141–154, 2014.

HUO, X. N. et al. Combining geostatistics with moran's i analysis for mapping soil heavy metals in Beijing, China. **International Journal of Environmental Research and Public Health**, v. 9, n. 3, p. 995–1017, 2012.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An Introduction to Applied Geostatistics**. New York: Oxford University Press, 1989.

KESKIN, H.; GRUNWALD, S.; HARRIS, W. G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, v. 339, n. November 2017, p. 40–58, 2019.

KHALEDIAN, Y.; MILLER, B. A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401–418, 2020.

KOTTEK, M. et al. World map of the Köppen-Geiger climate classification updated. **Meteorologische Zeitschrift**, v. 15, n. 3, p. 259–263, 2006.

LEE, S. Developing a bivariate spatial association measure: An integration of Pearson' s r and Moran' s I. **Geographical Systems**, v. 3 p. 369–385, 2001.

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

LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors (Switzerland)**, v. 18, n. 8, p. 1–29, 2018.

LIU, Q.; XIE, W. J.; XIA, J. B. Using Semivariogram and Moran's I Techniques to Evaluate Spatial Distribution of Soil Micronutrients. **Communications in Soil Science and Plant Analysis**, v. 44, n. 7, p. 1182–1192, 2013.

MALLA, R. et al. Soil Fertility Mapping and Assessment of the Spatial Distribution of Sarlahi District , Nepal. **American Journal of Agricultural Science**, v. 7, n. 1, p. 8–16, 2020.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. **Technometrics**, v. 21:2, n. 1, p. 239–245, 1979.

MEIER, M. et al. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. **Revista Brasileira de Ciência do Solo**, v. 42, n. 0, p. 1–22, 2018.

MINASNY, B.; MCBRATNEY, A. B. Latin hipercube sampling as tool for digital soil mapping. **Developments in Soil Science**, v. 31, n. 1997, p. 153–606, 2007.

MOHRI, M.; ROSTAMIZADEH, A. **Foundations of machine learning**. Humana Press, 2018.

MULDER, V. L.; DE BRUIN, S.; SCHAEPMAN, M. E. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. **International Journal of Applied Earth Observation and Geoinformation**, v. 21, n. 1, p. 301–310, 2012.

MUPHY, B.; MULLHER, S.; YURCHARK, R. **GeoStat-Framework/PyKrige v1.5.1(Version v1.5.1)**.

NOGUEIRA MARTINS, R. et al. Site-specific Nutrient Management Zones in Soybean Field Using Multivariate Analysis: An Approach Based on Variable Rate Fertilization. **Communications in Soil Science and Plant Analysis**, v. 51, n. 5, p. 687–700, 2020.

OLIVER, M.; WEBSTER, R. Kriging: a method of interpolation for geographical information systems. **International Journal of Geographical Information System**, v. 4, n. 3, p. 313–332, 1990.

OLSSON, A. M. J.; SANDBERG, G. E. Latin Hypercube Sampling for Stochastic Finite Element Analysis. **Journal of Engineering Mechanics**, v. 128, n. 1, p. 121–125, 2002.

OLSSON, A.; SANDBERG, G.; DAHLBLOM, O. On Latin hypercube sampling for structural reliability analysis. **Structural Safety**, v. 25, n. 1, p. 47–68, 2003.

PARMLEY, K. A. et al. Machine Learning Approach for Prescriptive Plant Breeding. **Scientific reports**, v. 9, n. 1, p. 17132, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 1, p. 2825–2830, 2011.

POULADI, N. et al. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, n. October 2018, p. 85–92, 2019.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System. Open Source Geoespacial Found. Proj.** QGIS Development Team. (2018), 2018.

R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna; Austria. Retrivied form http://www.r-project.org**/R Core Team. (2020), 2020.

REMY, N.; BOUCHER, A.; WU, J. **Applied Geostatistics with SGeMS**. Cambridge: Cambridge University Press, 2009.

REY, S. J.; ANSELIN, L. **PySAL: A Python Library of Spatial Analytical Methods**. **In: Fischer M., Getis A.** (eds). Springer, Berlin, Heidelberg, 2010.

SANTOS, H. G. et al. Sistema Brasileiro de Classificação de Solos. **Embrapa Solos**. 5. ed. Brasilia, DF, 2018.

SEKULIĆ, A. et al. Random forest spatial interpolation. **Remote Sensing**, v. 12, n. 10, p. 1–29, 2020.

SHIELDS, M. D.; ZHANG, J. The generalization of Latin hypercube sampling. **Reliability Engineering and System Safety**, v. 148, p. 96–108, 2016.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Applications of geostatistics to spatial studies of soil properties. **Advances in Agronomy**, v. 38, n. 1, p. 45–94, 1985.

VALENTE, D. S. M. et al. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. **Scientia Agricola**, v. 69, n. 3, p. 173–179, 2012.

VERONESI, F.; SCHILLACI, C. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. **Ecological Indicators**, v. 101, n. December 2018, p. 1032–1044, 2019.

WARNER, J.; SEXAUER, J.; UNNIKRISHNAN, A. **JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2.**

WEBSTER, R.; OLIVER, M.A. **Geostatistics for Environmental Scientists**. 2ª ed. Chichester: John Wiley & Sons, 2007.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of Soil Science**, v. 43, p. 177–192, 1992.

WERE, K. et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, v. 52, p. 394–403, 2015.

WHELAN, B. M.; MCBRATNEY, A. B.; MINASNY, B. VESPER 1.5 - Spatial prediction software for precision agriculture. **6th International Conference on Precision Agriculture**, p. 1–14, 2002.

WONG, D. W. S. Interpolation: Inverse-Distance Weighting. **International Encyclopedia of Geography: People, the Earth, Environment and Technology**, p. 1–7, 2017.

XU, S. et al. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. **Geoderma**, v. 310, p. 29–43, jan. 2018.

3 *SMART-MAP*: UM PLUGIN QGIS OPEN SOURCE PARA MAPEAMENTO DIGITAL UTILIZANDO APRENDIZADO DE MÁQUINA E KRIGAGEM ORDINÁRIA[2]

3.1 Resumo

Algoritmos de Aprendizado de Máquina (ML) têm sido utilizados como alternativa aos métodos convencionais e geoestatísticos no mapeamento digital de atributos do solo. Uma vantagem dos algoritmos de ML é a flexibilidade de se utilizar diferentes camadas de informação como covariáveis. No entanto, os algoritmos de ML apresentam muitas variantes que podem dificultar a sua aplicação por usuários finais. Para preencher essa lacuna, o objetivo desse trabalho foi desenvolver o plugin *Smart-Map*, utilizando modernas ferramentas de inteligência artificial (IA). O plugin foi desenvolvido para ser utilizado como complemento no software de Sistema de Informação Geográfica (SIG) QGIS versão 3. Foram implementados no plugin o método geoestatístico Krigagem Ordinária (OK) e o modelo de ML *Support Vector Machine* (*SVM*), para geração de mapas interpolados com base em aprendizado de máquina e OK. O modelo *SVM* pode utilizar como covariáveis as camadas vetoriais e raster disponíveis no QGIS no momento da interpolação. A seleção de covariáveis no modelo *SVM* foi implementada considerando-se a correlação espacial medida por meio do Índice de Moran (I'Moran). Com objetivo de validar o uso do plugin, foi realizado um estudo de caso com os dados de atributos de solo coletados em uma área de 75 ha. Comparações de performance entre OK e *SVM* foram realizadas para grid amostrais com 38, 75 e 112 pontos amostrados. O R^2 e o RMSE foram utilizados como métricas para avaliar o desempenho dos métodos. O *SVM* mostrou-se superior ao OK para predizer atributos químicos do solo nas três densidades amostrais testadas, sendo, portanto, indicado para predição dos atributos do solo. O plugin desenvolvido neste trabalho representa uma alternativa para geração de mapas interpolados por OK e ML diretamente no software QGIS.

Palavras-chave: Agricultura de precisão, Sistemas de informações geográficas (GIS). Geoprocessamento. Inteligência artificial. Mapeamento de solos.

---

[2] Este capítulo se refere a uma versão em português do artigo "*Smart-Map: an open-source QGIS Plugin for digital mapping using machine learning and ordinary kriging*" submetido no periódico *Computers and Eletronics in Agriculture.* ISSN 0168-1699.

3.2 Abstract

Machine Learning (ML) algorithms have been used as an alternative to conventional and geostatistical methods in the digital mapping of soil attributes. An advantage of ML algorithms is the flexibility to use different layers of information as covariates. However, ML algorithms have many variants that can make their application by end users difficult. In order to fill this gap, the objective of this study was to develop the *Smart-Map* plugin, using modern artificial intelligence (AI) tools. The plugin was developed to be used as a complement in the QGIS Version 3 Geographic Information System (GIS) software. The geostatistical method Ordinary Kriging (OK) and the ML model *Support Vector Machine* (*SVM*) were implemented in the plugin to generate interpolated maps based on ML and OK. The *SVM* model can use as covariates the vector and raster layers available in QGIS at the time of interpolation. The selection of covariates in the *SVM* model was implemented considering the spatial correlation measured using the Moran's Index (I'Moran). In order to validate the use of the plugin, a case study was conducted with data of soil attributes collected in an area of 75 ha. Performance comparisons between OK and *SVM* were performed for sampling grids with 38, 75 and 112 sampled points. R^2 and RMSE were used as metrics to evaluate the performance of the methods. *SVM* was superior to OK in the prediction of soil chemical attributes at the three sample densities tested and was, therefore, indicated for prediction of soil attributes. The plugin developed in this study represents an alternative to generate interpolated maps by OK and ML directly in QGIS software.

Key Words: Precision agriculture. Geographic information systems (GIS). Geoprocessing. Artificial intelligence. Soil mapping.

3.3 Introdução

O mapeamento digital dos atributos do solo e das plantas fornecem informações para aplicação de insumos agrícolas a taxas variadas (MALLA et al., 2020). Entretanto, a eficácia da aplicação depende da qualidade final dos mapas que são, normalmente, obtidos por interpolação com base em amostras georreferenciadas. Quanto maior a densidade amostral, maior será a qualidade final do mapa. No entanto maiores serão os custos com amostragens para a geração dos mapas. Em um sistema de amostragem economicamente viável, uma gama de métodos de interpolação podem ser utilizados, incluindo o método geoestatístico de

Krigagem Ordinária (OK), muito popular no mapeamento digital de solo (VERONESI; SCHILLACI, 2019). O método OK é considerado o melhor estimador, não-viesado, e que minimiza a variância do erro. Entretanto, uma desvantagem da OK é a necessidade de grandes quantidades de pontos amostrais para modelagem da semivariância (POULADI et al., 2019; WEBSTER; OLIVER, 1992).

Recentemente, com o grande volume de informações geradas nos campos de produção, técnicas de Machine Learning (ML) têm sido utilizadas como alternativa à OK para mapeamento digital de atributos do solo (DA MATTA CAMPBELL et al., 2019; GUO et al., 2015; HENGL et al., 2018; HEUNG et al., 2016; SEKULIĆ et al., 2020). Os algoritmos de ML procuram descobrir e quantificar padrões entre os dados disponíveis para fazer predições. Diversos modelos que utilizam algoritmos de ML para predição e mapeamento de atributos de solos têm sido desenvolvidos (HENGL et al., 2018; KHALEDIAN; MILLER, 2020; LIAKOS et al., 2018). Dentre os algoritmos, destacam-se o *Random Forest*, o *Support Vector Machine* (*SVM*), *Cubist, K-Nearest Neighbors*, e *Artificial Neural Networks* (KHALEDIAN; MILLER, 2020; MEIER et al., 2018; PARMLEY et al., 2019). No entanto, para implementar os modelos de ML para mapeamento digital é necessário o domínio de linguagens de programação. Dentre as linguagens de programação, se destacam a linguagem Python (criada por Guido Van Rossum e gerenciado pela Python Software Foundation, Delaware, EUA) e o R (R CORE TEAM., 2020), ambas de código aberto.

Para o desenvolvimento de aplicações utilizando ML, além do conhecimento das linguagens de programação, diversas camadas de dados devem estar disponíveis, como por exemplo: variáveis ambientais e climáticas, dados de sensores de solo e planta, imagens de satélites, mapas de produtividade, modelo digital de elevação, dentre outras. Esses dados podem estar em formato matricial ou vetorial, e em diferentes resoluções espaciais, o que pode tonar a implementação do modelo de interpolação por ML ainda mais complexa. Todos esses dados podem ser utilizados como características (*features)* de entrada nos modelos de ML. Como muitas dessas *features* podem ter uma maior ou menor importância na modelagem, pode ser necessário utilizar técnicas de seleção e eliminação de *features* (GOMES et al., 2019; GREGORUTTI; MICHEL; SAINT-PIERRE, 2017; PARMLEY et al., 2019). Por fim, os modelos de ML são muito flexíveis, dessa forma, deve-se otimizar os hiperparâmetros do algoritmo para obtenção de modelos mais robustos.

Conforme exposto, uma ferramenta computacional que facilite o uso de técnicas de ML no mapeamento digital poderá auxiliar os usuários de softwares de sistemas de informações geográficas (SIG). Dentre estes softwares, o QGIS (QGIS DEVELOPMENT TEAM, 2018) é *open-source*, tem interface amigável e uma comunidade ativa de desenvolvedores e usuários. Programas de computadores gratuitos estão disponíveis para Krigagem Ordinária, como o Vesper (WHELAN; MCBRATNEY; MINASNY, 2002), SGeMS (REMY; BOUCHER; WU, 2009) e KrigMe (VALENTE et al., 2012). Entretanto, nenhum deles estão disponíveis como complemento (plugin) do QGIS. Dada a potencial aplicação de ML e a necessidade de integrar o QGIS à um sistema de mapeamento digital de atributos do solo, esse trabalho teve como objetivo o desenvolvimento de uma ferramenta integrada (plugin) ao software QGIS para mapeamento digital utilizando OK e ML como métodos interpoladores. O plugin para mapeamento digital desenvolvido foi denominado *Smart-Map*.

3.4 Material e Métodos

Implementação do Smart-Map

O fluxograma do software *Smart-Map* é apresentado na Figura 3.1. *Smart-Map* foi registrado no Instituto Nacional de Propriedade Industrial (INPI, Ministério da Economia, Brasil, BR 51 2021 000002-1). A última versão pode ser encontrada no GitHub (https://github.com/gustavowillam/SmartMapPlugin) ou instalada a partir do repositório de plugins do QGIS (https://plugins.qgis.org/plugins/Smart_Map). Um tutorial completo de como instalar e utilizar o software está disponível em: https://github.com/gustavowillam/SmartMapPlugin/wiki. No canal do YouTube: https://www.youtube.com/c/AgriculturaDigital, uma série de vídeos podem ser encontrada para auxiliar na instalação e utilização das principais funcionalidades do plugin. A linguagem Python versão 3.7 foi utilizada para o desenvolvimento do software, sendo compatível com os sistemas operacionais macOS, Linux e Windows. A interface gráfica do usuário (GUI) foi projetada utilizando PyQt5 (Riverbank Computer Limited, Dorchester, United Kingdom). O software é um complemento (plugin) para o software QGIS versão 3.10 ou superior.

Para apresentar a metodologia de OK e ML utilizada pelo *Smart-Map* foi realizado um estudo de caso. No estudo de caso foram comparadas as acurácias da interpolação de atributos de solo utilizando OK e ML para diferentes grids de amostragem com o objetivo de validar o sistema. Para o método de interpolação por OK, foram adotados os protocolos e equações descritos por Isaaks and Srivastava (1989). O plugin desenvolvido permite ao usuário ajustar cinco modelos de semivariogramas teóricos isotrópicos: linear, linear com patamar, exponencial, esférico e gaussiano. Os ajustes dos parâmetros do semivariograma foram calculados com base na minimização da raiz do erro quadrático médio (RMSE). A escolha do melhor modelo foi executada pelo plugin adotando como critério o modelo com o menor RMSE. Para o modelo *SVM*, os hiperparâmetros (parâmetros ajustados pelo usuário) como C e gamma (γ) foram otimizados a partir de um método sistemático de busca em grade (KESKIN; GRUNWALD; HARRIS, 2019; XU et al., 2018), permitindo assim um ajuste automatizado sem a interferência do usuário. A validação cruzada *k-fold* foi utilizada para se obter os valores ótimos destes hiperparâmetros, sendo seu valor default igual a cinco.

Além da geração de mapas por interpolação, o *Smart-Map* é dotado de um algoritmo para executar análise de agrupamento utilizando o método *fuzzy k-means* (BEZDEK; EHRLICH; FULL, 1984). No final do processamento é exibido o mapa de Zonas de Manejo (ZM). Para definir o número ideal de classes pode-se calcular os índices FPI (Índice de Performance Fuzzy) e o NCE (Entropia de Classificação Normalizada). Os índices FPI e NCE são amplamente recomendados na literatura para definir o número apropriado de ZM's (ALBORNOZ et al., 2018; CHEN et al., 2019). Para executar o processo de cluster e definir as ZM's foi utilizado o algoritmo fuzzy k-*means* da biblioteca Python *Scikit-Fuzzy* (WARNER; SEXAUER; UNNIKRISHNAN, 2019). Na Figura 3.2 é apresentada a GUI para interpolação de mapas por OK e por *SVM do Smart-Map*.

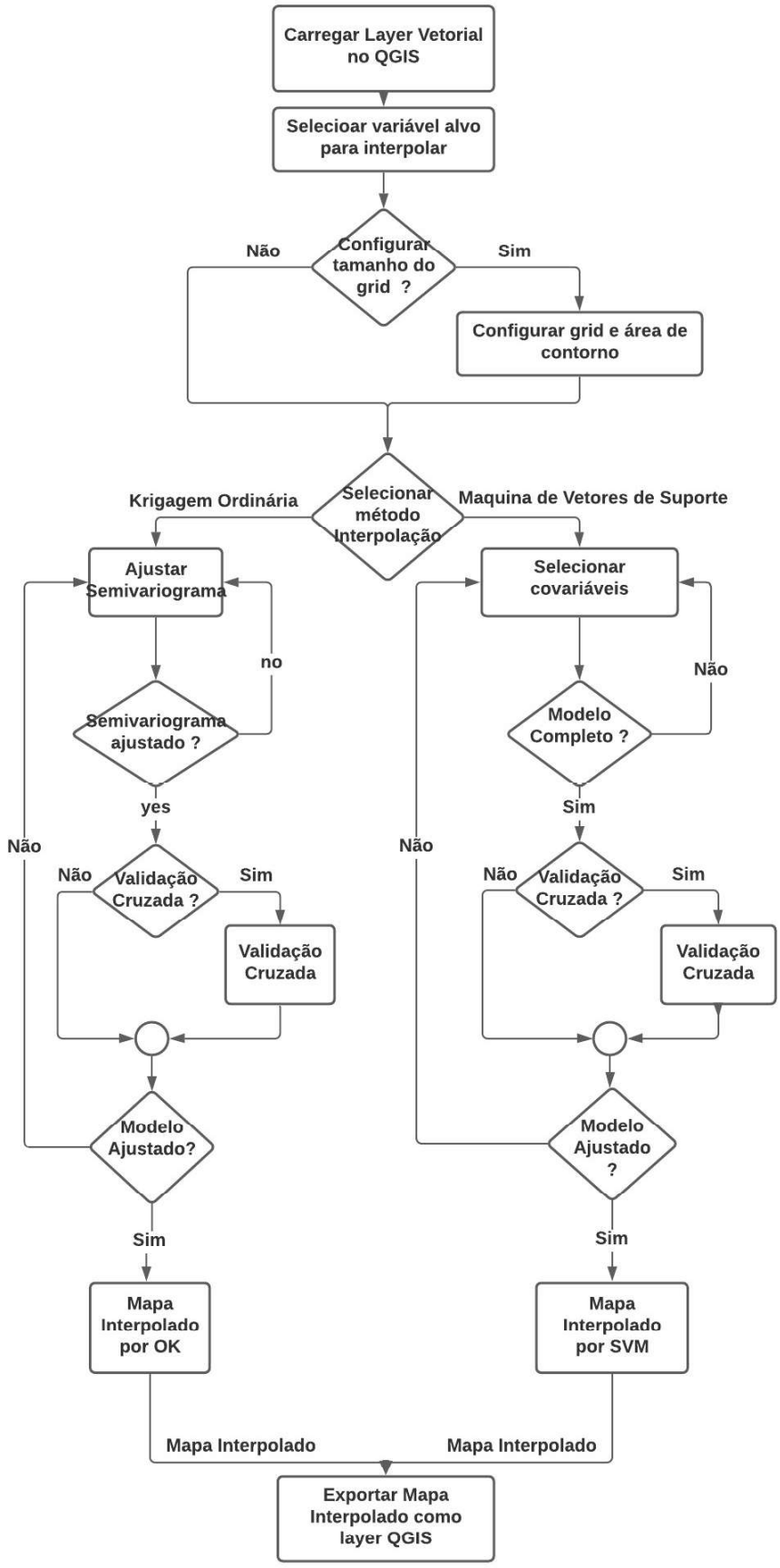
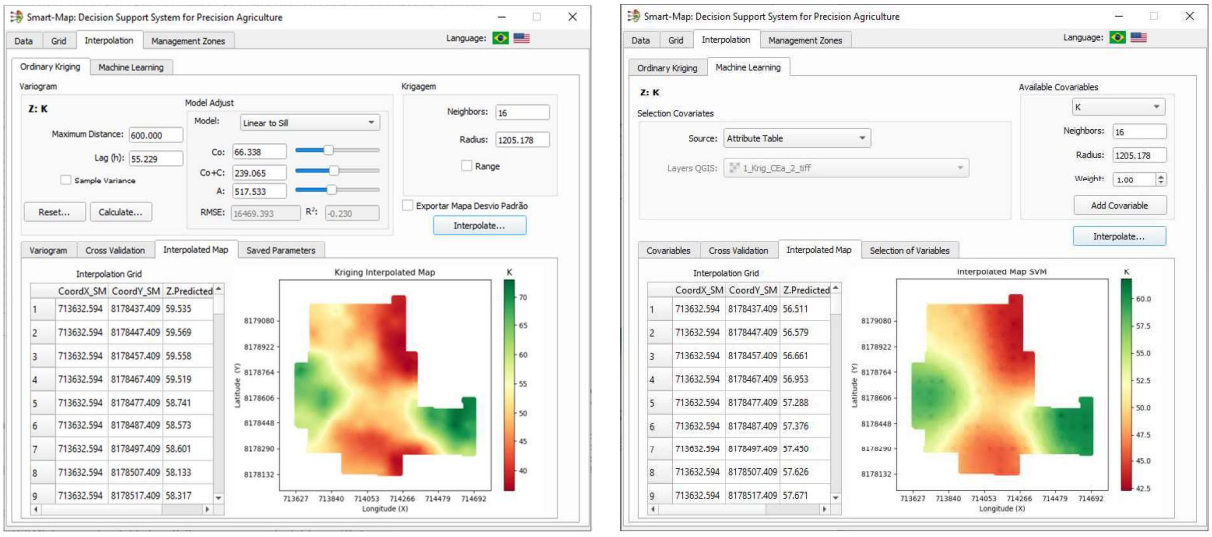


Figura 3.1: Fluxograma das principais etapas de processamento do Smart-Map.



(a) (b)

Figura 3.2: Interface Gráfica do Smart-Map. (a) Interpolação por OK. (b) Interpolação por SVM.

Estudo de caso para avaliação do plugin *Smart-Map*

Um estudo de caso para avaliação do *Smart-Map* foi conduzido em uma área de 75 ha, localizada entre os municípios de Anápolis e Goianápolis (16º27'48" S e 48º59'38" W), região central do estado de Goiás, Brasil (Figura 3.3). Essa área é cultivada com soja, possui altitude média de 1017 m, relevo plano e solo predominantemente classificado como Latossolos Vermelho-Amarelos (LVA) de acordo com a classificação atualizada da Embrapa Solos (SANTOS et al., 2018). As amostras de solo foram coletadas utilizando um grid regular com densidade amostral de dois pontos por hectare, totalizando 150 amostras. Cada amostra foi composta de 10 amostras simples coletadas em um raio de 3 m. Foram realizadas as análises laboratoriais para medir a concentração de macronutrientes (P, K^+, Ca^{2+} e Mg^{2+}), matéria orgânica, capacidade de troca de cátions a pH 7 e granulometria. Também foram coletados dados da condutividade elétrica aparente do solo (ECa) em cinco diferentes datas (Eca_1 medida em 11/11/2010, Eca_2 medida em 23/11/2010, ECa_3 medida em 04/12/2020, Eca_4 medida em 13/12/2020 e Eca_5 medida em 26/01/2011) utilizando um aparelho portátil fabricado por Landviser, modelo LandMapper® ERM 02. Este aparelho mede a resistividade elétrica do solo por meio de quatro eletrodos que são inseridos no solo. Estes dados, utilizados no presente estudo de caso, foram disponibilizados por Costa et al. (2014). A estatística descritiva dos dados é apresentada na Tabela 3.1.

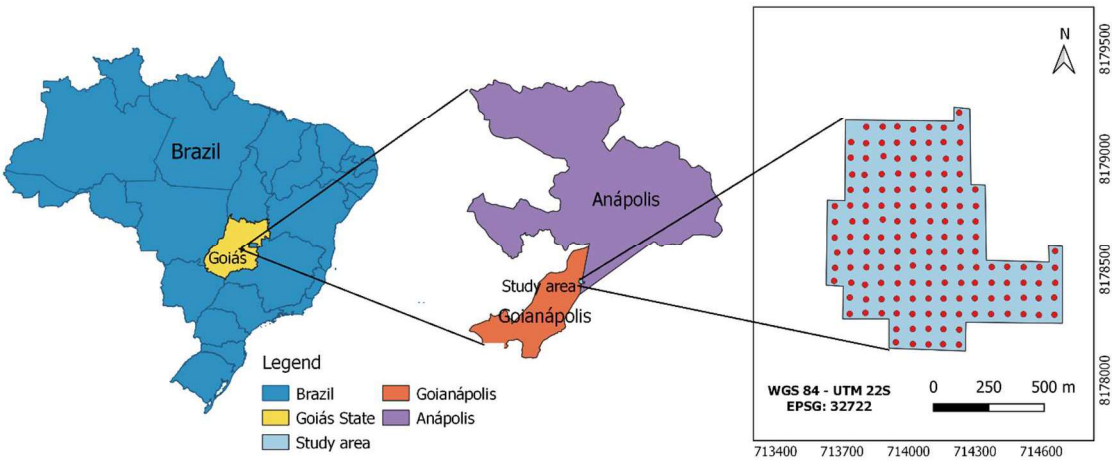


Figura 3.3: Localização geográfica da área de estudo e distribuição dos pontos amostrais em Anápolis/Goianápolis, Goiás, Brasil.

Tabela 3.1: Estatística Descritiva dos atributos de solos na área de estudo, disponibilizado por Costa et al. (2014).

Atributo	Unidade	Valor Mínimo	Valor Máximo	Média	SD(17)	Mediana	CV(%)(18)
P (1)	mg dm^{-3}	1,70	21,60	6,84	3,96	5,85	57,88
K^{+}(2)	mg dm^{-3}	24,00	108,00	52,63	14,20	51,00	26,98
Ca^{2+}(3)	cmolc dm^{-3}	1,90	4,20	3,27	0,46	3,30	14,04
Mg^{2+}(4)	cmolc dm^{-3}	0,60	1,40	0,84	0,14	0,80	16,53
OM(5)	g dm^{-3}	2,50	4,30	3,06	0,30	3,10	9,85
CEC(6)	cmolc dm^{-3}	4,20	9,90	5,95	0,86	5,90	14,41
Altitude(7)	m	987	1025	1011,2	7,63	1012,1	0,75
CLA (8)	g kg^{-1}	26,00	44,00	33,11	3,37	33,00	10,17
SIL(9)	g kg^{-1}	6,00	20,00	10,60	2,94	10,00	27,78
Sand(10)	g kg^{-1}	45,00	65,00	56,28	4,41	56,50	7,84
ECa_1(11)	mS m^{-1}	2,49	8,36	4,92	1,01	4,83	20,62
ECa_2(12)	mS m^{-1}	2,95	10,00	5,95	1,22	5,99	20,56
ECa_3(13)	mS m^{-1}	1,71	9,11	4,54	1,13	4,51	24,86
ECa_4(14)	mS m^{-1}	1,84	7,32	3,98	0,88	3,94	22,09
ECa_5(15)	mS m^{-1}	0,89	5,57	2,65	0,71	2,61	26,67
ECa_Avg(16)	mS m^{-1}	2,17	8,03	4,41	0,84	4,44	19,08

1/ P, Fósforo; 2/ K^{+}, Potássio; 3/ Ca^{2+}, Cálcio; 4/ Mg^{2+}, Magnésio; 5/ OM, Matéria Orgânica; 6/ CEC, Capacidade de Troca de Cátions a pH 7; 7/ Altitude; 8/ CLA, Argila; 9/ SIL, Silte; 10/ Sand, Areia; 11/ ECa_1, Condutividade Elétrica Aparente do Solo medida em 11/11/2010; 12/ ECa_2, Condutividade Elétrica Aparente do Solo medida em 23/11/2010; 13/ ECa_3, Condutividade Elétrica Aparente do Solo medida em 04/12/2010; 14/ ECa_4, Condutividade Elétrica Aparente do Solo medida em 13/12/2010; 15/ ECa_5, Condutividade Elétrica Aparente do Solo medida em 26/01/2011; 16/ ECa_Avg, Valor Médio da Condutividade Elétrica Aparente do Solo (ECa_1, ECa_2, ECa_3, ECa_4, ECa_5); 17/ SD, Desvio Padrão; 18/ CV, Coeficiente de Variação.

Métodos de interpolação e análise de correlação espacial

No estudo de caso apresentado neste trabalho, para realizar a interpolação por OK foi definido um grid de interpolação de 10 x 10 m. Para interpolar cada ponto do grid foram definidos o raio de busca igual ao alcance obtido pelo semivariograma teórico e o número máximo de vizinhos igual a 16. Para a interpolação por OK, o *Smart-Map* utiliza a biblioteca Python *open source PyKrige* (MUPHY; MULLHER; YURCHARK, 2020), com algumas adaptações em seu código para funcionar no ambiente do QGIS.

Para a interpolação por ML foi implementado no *Smart-Map* o modelo de aprendizado supervisionado "*Support Vector Machine*" (*SVM*) disponível na biblioteca Python *open source Scikit-Learn* (PEDREGOSA et al., 2011). No modelo *SVM* foi utilizado o kernel RBF (*Radial Basis Function*). Para a modelagem é necessário construir a matriz X com as *features* e o vetor y com os valores da variável a ser interpolada. Nesse estudo de caso foram interpolados os atributos P, K^+, Ca^{2+} e Mg^{2+}.

Na matriz X, as coordenadas geográficas x e y do ponto a ser interpolado foram adicionadas. Além das coordenadas geográficas, outras *features*, inclusive a *feature* da própria variável foi adicionada na matriz X. Nesse caso, a *feature* é criada com base no cálculo da média ponderada do inverso da distância (IDW) dos vizinhos mais próximos do ponto a ser interpolado. Dessa forma, o valor experimental obtido para o ponto, faz parte do vetor y e não é utilizado para criação da *feature*. Além disso, pode-se utilizar dados de outras *layers* do banco de dados do QGIS (vetorial ou raster) como *features*.

No estudo de caso foi utilizado dois métodos diferentes de modelagem por SVM, que foram denominados como *SVM1* e *SVM2*. Para o método *SVM1* foram utilizados como *features* as coordenadas geográficas (x e y) e o valor da própria variável estimado utilizando o método de interpolação IDW. No *SVM2* foram utilizadas como covariáveis aquelas que eram mais correlacionadas com a variável a ser interpolada, além das coordenadas geográficas (x e y) e o valor da própria variável interpolada utilizando IDW. A seleção das covariáveis foi realizada a partir da correlação espacial do Índice de Moran (I'Moran).

O I'Moran é um dos índices mais populares para avaliação da correlação espacial (HUO et al., 2012) de variáveis regionalizadas. O I'Moran univariado foi

utilizado para comparar o grau de correlação da própria variável em diferentes espaços de distância (autocorrelação espacial). Valor de I'Moran univariado igual a zero significa que a variável em estudo não apresenta correlação espacial. Quando o valor for mais próximo de 1 ou -1, maior será a autocorrelação, ou seja, maior a correlação espacial da variável (GUO et al., 2015; LIU; XIE; XIA, 2013). O I'Moran univariado foi calculado de acordo com a Equação 3.1 (LEGENDRE; FORTIN, 1989). O I'Moran bivariado foi utilizado para medir a correlação espacial entre as covariáveis disponíveis como CEC, OM, Altitude, CLA, SIL, Sand e ECa, com o atributo que foi interpolado. Seu valor foi calculado de acordo com a Equação 3.2 (LEE, 2001).

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} wij} . \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x}).(x_j - \bar{x})}{\sum_{i}^{n}(x_i - \bar{x})^2} \quad (3.1)$$

$$I_{x,y} = \frac{\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{n} w_{ij}(x_j - \bar{x})\right).\left(\sum_{j=1}^{n} w_{ij}(y_j - \bar{y})\right)\right]}{\sqrt{\sum_{i}^{n}(x_i - \bar{x})^2} . \sqrt{\sum_{i}^{n}(y_i - \bar{y})^2}} \quad (3.2)$$

em que: n é o número de observações na área em estudo; x_i, x_j representam os valores observados dos atributos de solo a serem interpolados nos pontos i, j; y_i, y_j representam os valores observados da covariável selecionada nos pontos i, j; $\bar{x}$ a média de *x*; $\bar{y}$ a média de y; w_{ij} são os elementos da matriz de pesos espaciais com valor 0 na diagonal ($w_{ii} = 0$).

A seleção do subconjunto ótimo de covariáveis para o método *SVM2* foi realizada considerando-se o I'Moran bivariado. Covariáveis que apresentaram maior correlação espacial com a variável a ser interpolada foram adicionadas gradativamente ao método *SVM2*. Com o objetivo de verificar a significância do I'Moran, o pseudo p-valor foi obtido a partir de 999 permutações entre os pontos do grid amostral aos níveis de 1% e 5% de probabilidade. Para o cálculo do I'Moran o *Smart-Map* utiliza a biblioteca Python *open-source PySAL* (REY; ANSELIN, 2010).

Geração de cenários e critérios de desempenho para comparação entre os métodos de interpolação

Com o objetivo de comparar a performance do método OK e do modelo de ML *SVM* (*SVM1* e *SVM2*) em diferentes densidades de amostragem, realizou-se a redução do grid regular de 150 pontos em grids com menor densidade (25%, 50% e 75%) de pontos. Obteve-se três grids com 38, 75 e 112 pontos, respectivamente. Esses pontos foram utilizados para modelagem do semivariograma no método OK e definição do conjunto de treinamento no modelo *SVM* e os pontos restantes (que não participaram da modelagem e do treinamento) foram utilizados para teste (verificação da acurácia final da predição). Na Figura 3.4 é apresentado o grid original de 150 pontos e os grids reduzidos compostos por dados de modelagem e teste. No grid de 38 pontos (Figura 3.4b), 38 pontos foram utilizados para modelagem e 112 pontos para teste. No grid de 75 pontos (Figura 3.4c), 75 pontos foram utilizados para modelagem e 75 pontos foram utilizados para teste. No grid de 112 pontos (Figura 3.4d), 112 pontos foram utilizados para modelagem e 38 pontos foram utilizados para teste.

A partir da redução do grid amostral foram gerados os mapas interpolados utilizando os conjuntos de pontos de treinamento para o método OK e o modelo de ML *SVM* nas três densidades de grids amostrais. Nesse estudo de caso, foram interpolados os atributos P, K^+, Ca^{2+} e Mg^{2+}. Para modelagem foi utilizada a validação cruzada *leave-one-out* (*LOOCV*). Foram calculados o Coeficiente de Determinação (R^2) e a Raiz Quadrada do Erro Quadrático Médio (RMSE) da validação cruzada para cada modelo e para cada atributo interpolado. Para medir a acurácia final de cada mapa obtido pela interpolação de P, K^+, Ca^{2+} e Mg^{2+}, após a modelagem, foram utilizados os conjuntos de teste. Para isso nos mesmos locais onde se encontravam os pontos de teste, foram extraídos os valores interpolados de P, K^+, Ca^{2+} e Mg^{2+}. O R^2 e RMSE foram calculados conforme Equações 3.3 e 3.4, respectivamente, dos dados de teste para P, K^+, Ca^{2+} e Mg^{2+} para os diferentes grids de amostragem.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x_i})^2} \tag{3.4}$$

em que: $\hat{x}_i$ representa o valor estimado do atributo de solo no ponto i; $\bar{x}$ a média dos *n* pontos amostrados do atributo de solo; x_i o valor observado do atributo de solo no ponto i; e n, o número de pontos amostrados.

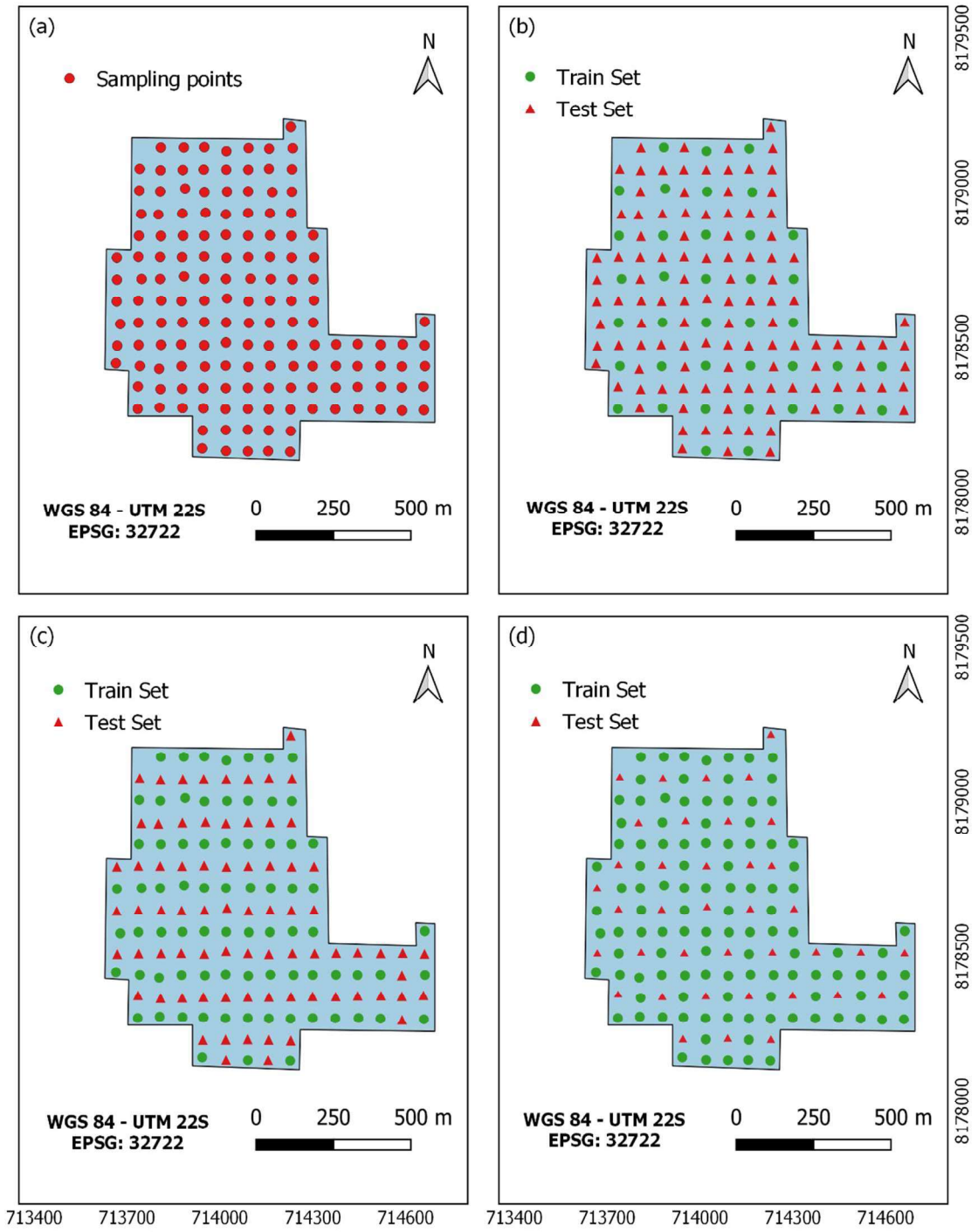


Figura 3.4: Número de pontos amostrais em: (a) grid original 150 pontos; (b) 38 pontos para treinamento e 112 pontos para teste; (c) 75 pontos para treinamento e teste; (d) 112 pontos para treinamento e 38 pontos para teste.

Definição e seleção de features para o modelo SVM

Para definir as *features* a serem inseridas no modelo *SVM*, o usuário define os parâmetros do IDW como o valor da ponderação (p), o raio de busca e o número de vizinhos (n) a serem considerados para cálculo da covariável. No estudo de caso foram utilizados como padrão um raio de busca igual a máxima distância entre os pontos amostrados, número de vizinhos máximos igual a 16 e o peso da ponderação igual a 1. A Figura 3.5 mostra como diferentes layers com número de observações diferentes é interpolada por meio do método IDW para ajustar o número de observações à tabela de atributo carregada pelo plugin.

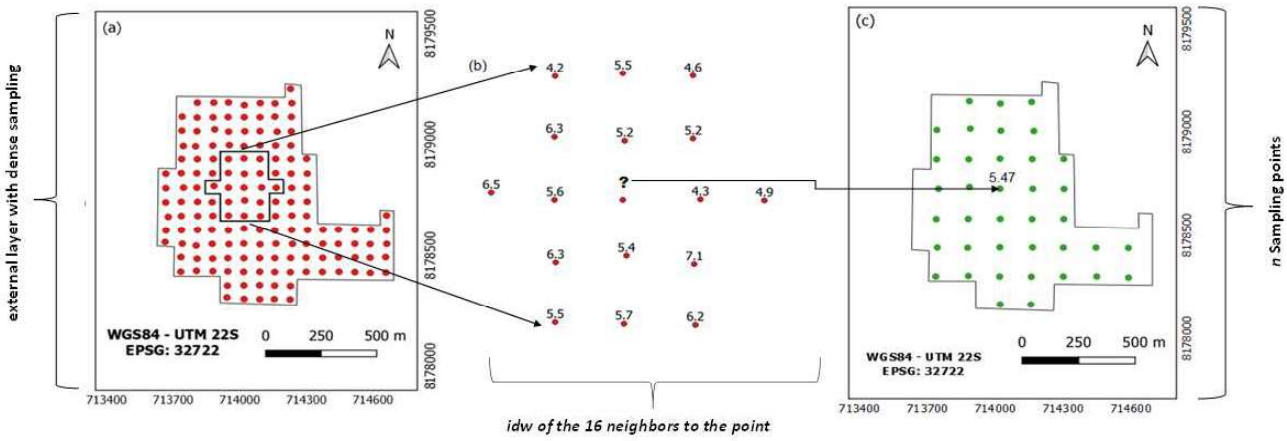


Figura 3.5: Construção do modelo de ML: (a) Layer com malha densa de pontos amostrados. (b) Seleção dos 16 vizinhos mais próximos ao ponto que se deseja estimar o valor do atributo utilizando o IDW (idwA). (c) Valor interpolado pelo método IDW do atributo selecionado (Atribute_A) da layer QGIS.

Para o *SVM1* utilizou-se como *features* as coordenadas (*coordX* e *coordY*) do ponto e o valor do IDW da variável (y) utilizando os 16 vizinhos mais próximo do ponto amostrado, dentro do raio de busca definido do atributo a ser estimado. A variável a ser interpolada (*y*) representa o atributo de solo observado, para o qual se quer predizer seus valores em locais não amostrados. No caso específico deste estudo de caso, as variáveis são P, K^{+}, Ca^{2+} e Mg^{2+}.

Na segunda abordagem (*SVM2*) utilizou-se como *features* as coordenadas (*coordX* e *coordY*), e o IDW de 12 covariáveis disponíveis na área de estudo: OM, CEC, Altitude, CLA, SIL, Sand, ECa_1, ECa_2, ECa_3, ECa_4, ECa_5, ECa_Avg. Nesse caso, as *features* utilizadas foram do grid original com 150 pontos. Isso foi feito uma vez que o objetivo de usar o *SVM* é aproveitar informações que foram

densamente amostradas na área. Essas informações podem ser facilmente obtidas por sensores ou são informações que não se modificam ou modificam muito lentamente ao longo dos anos (apresentam baixa variabilidade temporal).

A métrica de acurácia R^2 da validação cruzada *LOOCV* foi aplicada para cada subconjunto de covariável adicionada. O subconjunto de covariáveis que apresentou o melhor valor de R^2 foi escolhido para definir o modelo *SVM* a ser utilizado para a variável a ser interpolada. Esta seleção foi executada considerando todas as *features* para os grids com 38, 75 e 112 pontos amostrais.

Na Figura 3.6 é apresentado o modelo de ML para os métodos *SVM1* e *SVM2* foi construído dividido em *features* (Matriz *X*) e variável a ser interpolada (Vetor *y*). Na Matriz X, *coordX* e *coordY* são as coordenadas x e y do ponto amostrado, respectivamente; *idwA* representa o valor estimado para a variável com base no IDW utilizando os 16 vizinhos mais próximos do ponto amostrado do atributo a ser interpolado; *idw_At1*, *idw_At2*, *idw_Atn* representa o valor estimado com base no IDW utilizando os 16 vizinhos mais próximos do ponto amostrado das *features* selecionadas. No vetor y, *target_A* representa os valores amostrados do atributo a ser interpolado, sendo no presente estudo P, K^+, Ca^{2+} e Mg^{2+}. Cada linha dos dados de treinamento representa uma amostra do grid. A Matriz X formada pelas colunas (*coordX*, *coordY* e *idwA*) e o vetor y foram as entradas do conjunto de treinamento do método *SVM1*. Para o método *SVM2* foram utilizadas todas as colunas da Matriz X e o Vetor y como entradas. Os dados de entrada foram padronizados em média zero e desvio padrão um.

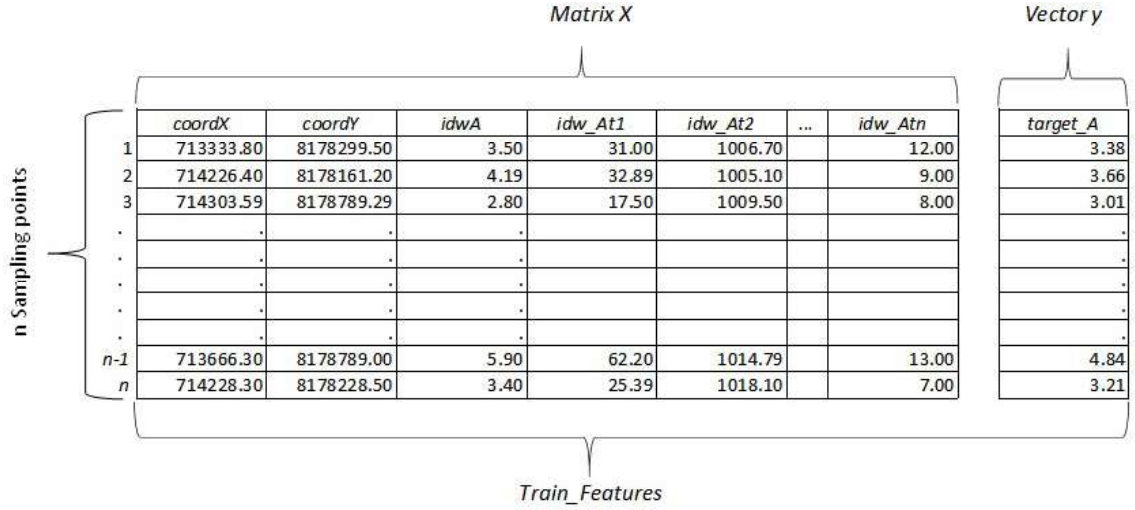


Figura 3.6: Definição do modelo de ML (Train_Features) para os métodos SVM1 e SVM2: features (Matriz X) e target (Vetor y).

3.5 Resultados e discussão

Correlação espacial e seleção de covariáveis para o modelo SVM

Para a análise de correlação espacial nas três diferentes densidades de grids amostrais, foi utilizado o I'Moran bivariado para medir a correlação entre os teores dos macronutrientes P, K^+, Ca^{2+} e Mg^{2+}, e as covariáveis com maior estabilidade temporal (CEC, OM, Altitude, CLA, SIL, Sand, ECa_1, ECa_2, ECa_3, ECa_4, ECa_5, ECa_Avg). Na Figura 3.7 são apresentados os valores de I'Moran univariado para as variáveis a serem interpoladas (P, K^+, Ca^{2+} e Mg^{2+}) e bivariado entre as variáveis a serem interpoladas e as covariáveis com maior estabilidade temporal para as densidades amostrais de 38, 75 e 112 pontos.

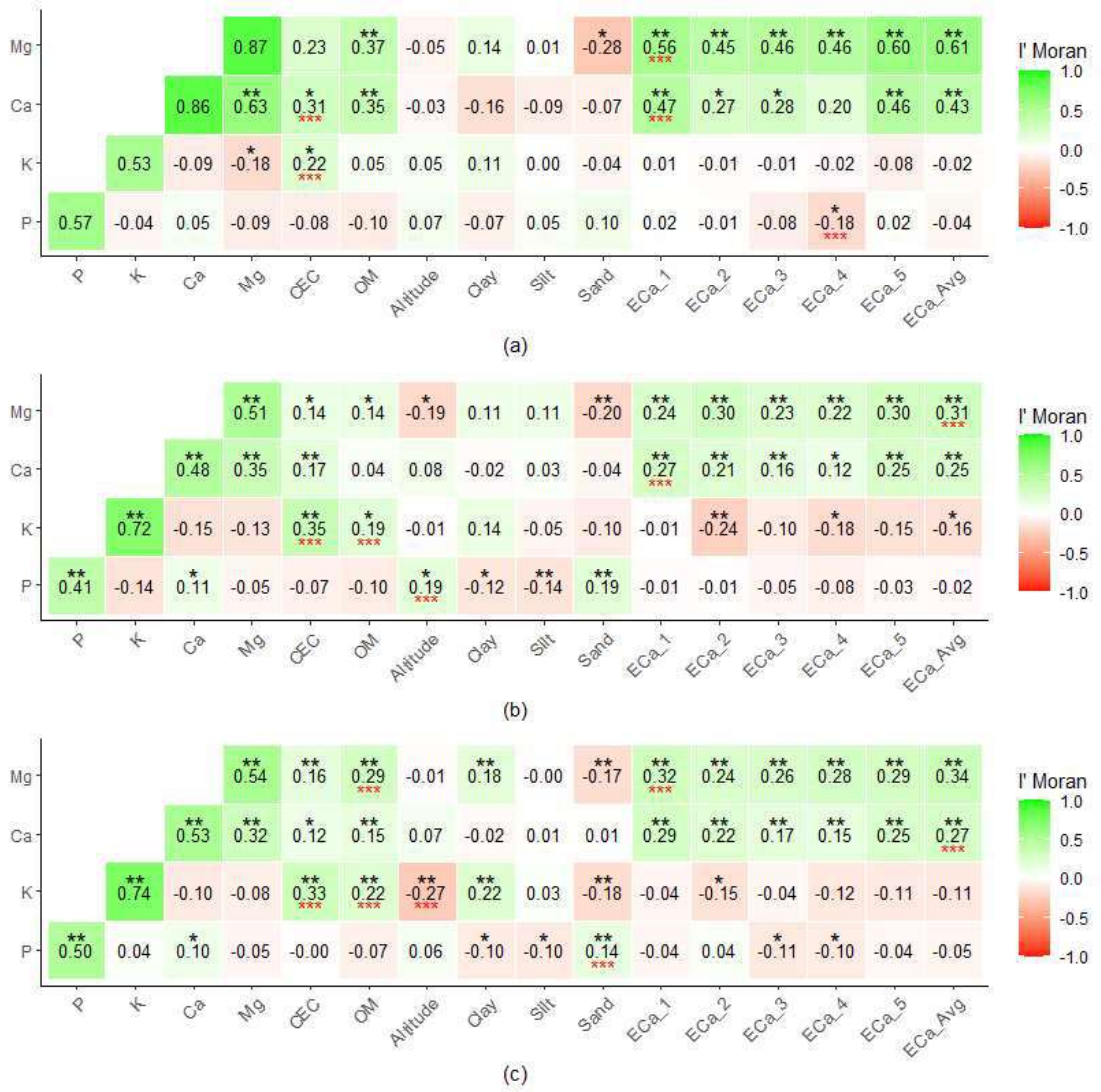


*Figura 3.7: Índice de Moran Global univariado para os atributos de solo P, K^+, Ca^{2+} e Mg^{2+} e bivariado entre os atributos de solo P, K^+, Ca^{2+} e Mg^{2+} e as covariáveis na área de estudo para os grids amostrais do conjunto de treinamento com: (a) 38 pontos; (b) 75 pontos; (c) 112 pontos. *, ** indicando nível de significância à 0,05 e 0,01 respectivamente. *** covariáveis utilizadas pelo método SVM2 para interpolar os atributos de solo P, K^+, Ca^{2+} e Mg^{2+}.*

Na Figura 3.7 é possível observar que o método de interpolação *SVM2* apresentou potencial para ser aplicado para algumas covariáveis, pois resultou em correlações significativas. A condutividade elétrica aparente do solo (ECa) medida em cinco diferentes datas apresentou correlação positiva significativa com os atributos Mg^{2+}*e* Ca^{2+}, com valores variando de 0,12 (entre Ca^{2+} e ECa_4, grid de 75 pontos) a 0,61 (entre Mg^{2+} e ECa_Avg, grid de 38 pontos). Para a interpolação desses dois atributos do solo foi utilizada a Eca como covariável no método *SVM2* nas três densidades de grids amostrais (Figura 3.7). No grid de 38 pontos amostrais (Figura 3.7a) foi utilizado a covariável Eca_1 para os atributos Mg^{2+} e Ca^{2+} e CEC para Ca^{2+}.

No grid de 75 pontos (Figura 3.7b) foi utilizada a Eca_Avg para o atributo Mg^{2+} e a Eca_1 para o atributo CA^{2+}. Por fim, no grid com 112 pontos amostrais (Figura 3.7c) foram utilizadas como covariáveis de interpolação os atributos OM e Eca_1 para MG^{2+} e Eca_Avg para CA^{2+}.

Por outro lado, a ECa apresentou baixa correlação com os atributos P e K^+, tendo, portanto, menor potencial de uso como covariáveis para interpolar P e K^+. A Eca_4 foi utilizada para interpolar somente o atributo P no grid de 38 pontos, uma vez que a correlação foi significativa com I'Moran de -0,18 (Figura 3.7a). Para o mesmo grid foi utilizado como covariável a CEC para o atributo K^+. Para o grid de 75 pontos (Figura 3.7b) foram utilizadas as covariáveis CEC e OM para o atributo K^+ e a covariável Altitude para o atributo P. Para o grid com 112 pontos (Figura 3.7c), o atributo K^+ utilizou as covariáveis CEC, OM e Altitude e o atributo P utilizou Sand como covariáveis para interpolação.

Comparação entre os métodos OK e SVM

Para o conjunto de treinamento, em três diferentes densidades de grids amostrais, os valores de R^2 (Figura 3.8) mostram que o método *SVM2* foi superior para os quatro atributos de solo analisados (P, K^+, Ca^{2+} e Mg^{2+}), exceto para K^+ no grid de 75 pontos. O I'Moran univariado para o atributo K^+ foi de 0,72 e significativo ao nível de 1% de probabilidade no grid de 75 pontos, como mostrado na Figura 3.7b. Valores de R^2 para o método *SVM2* no conjunto de treinamento variaram entre 0,16 e 0,38. Em comparação com o método *SVM1*, o método *SMV2* obteve R^2 superior para todos os atributos analisados em todas as densidades de pontos dos grids amostrais.

Em relação aos menores valores de R^2, OK se destacou por apresentar os menores coeficientes de determinação para os atributos P, K^+ e Ca^{2+} no grid de 38 pontos (Figura 3.8). O I'Moran univariado para P e K^+ foram baixos e não significativos para os dois atributos de solo analisados (Figura 3.7a). Conforme mencionado por Pouladi et al. (2019) e Webster and Oliver (1992), a Krigagem Ordinária necessita de uma quantidade mínima de pontos amostrais para uma boa modelagem do semivariograma. Para o grid de 38 pontos, o método *SVM2* obteve melhor desempenho em relação ao método OK com valores de R^2 no intervalo de 0,19 a 0,38. Para o atributo P, os três métodos apresentaram os menores valores de R^2. Estes dados corroboram com a Figura 3.7, em que os valores de I'Moran univariado e

bivariado foram baixos para o atributo P. Em geral, OK, *SVM1*, e *SVM2* apresentaram menores valores de R² para o grid de 38 pontos amostrais, comparado aos grids com maior densidade amostral.

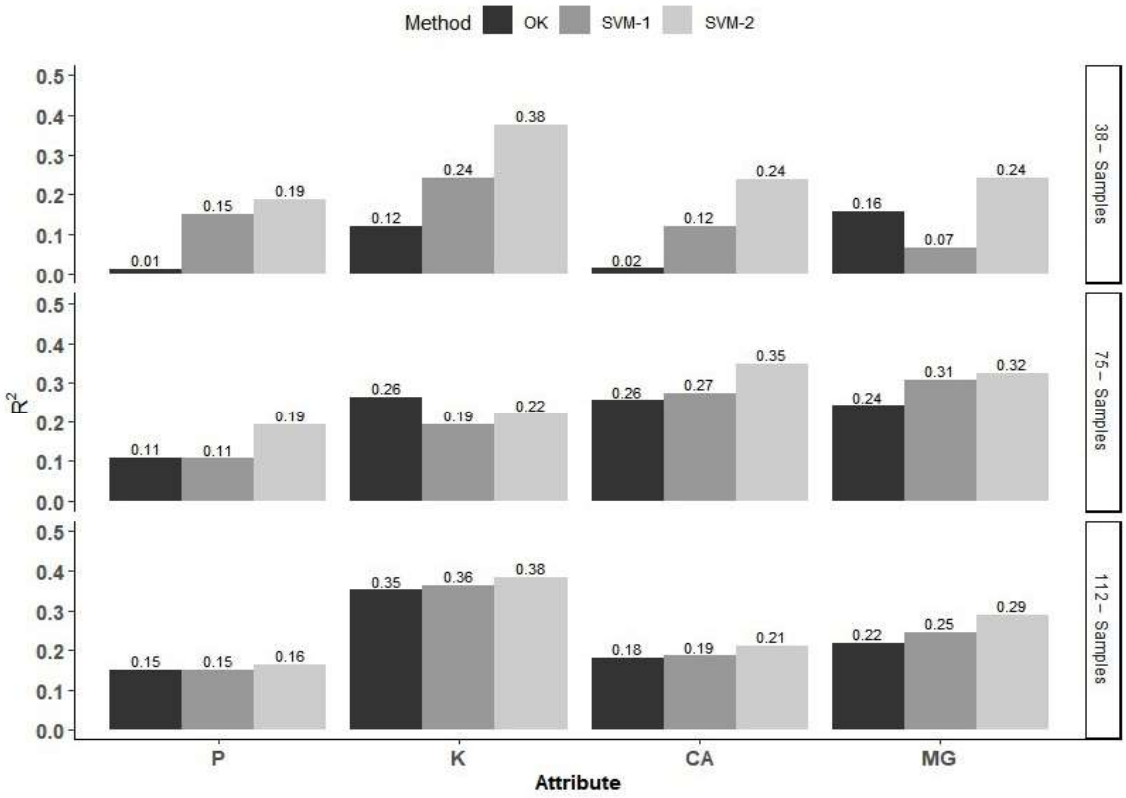


Figura 3.8: Coeficiente de Determinação (R^2) calculado para os atributos P, K^+, Ca^{2+} e Mg^{2+} em três diferentes grids amostrais para o conjunto de treinamento.

Assim como no conjunto de treinamento, os valores de R^2 também foram superiores para o método *SVM2* no conjunto de testes (Figura 3.9). O menor coeficiente de correlação para o método *SVM2* foi para o atributo P no grid amostral com 38 pontos no conjunto de teste (R^2 = 0,15). A baixa performance do *SVM2* para a predição do atributo P está relacionado a covariável adicionada no grid de 112 pontos do conjunto de treinamento. A covariável Sand utilizada pelo método *SVM2* apresentou I'Moran bivariado de 0,14 com o atributo P (Figura 3.7c). Este valor foi o menor utilizado por uma covariável adicionada ao método *SVM2*. Covariáveis que apresentam baixo valor de I'Moran bivariado com o atributo a ser interpolado podem não contribuir ou contribuir de forma não significativa para uma melhor performance do método *SVM2*. Para Gregorutti et al. (2017), a baixa correlação entre as variáveis

preditoras com a variável dependente (y) impacta diretamente na performance do modelo de Machine Learning.

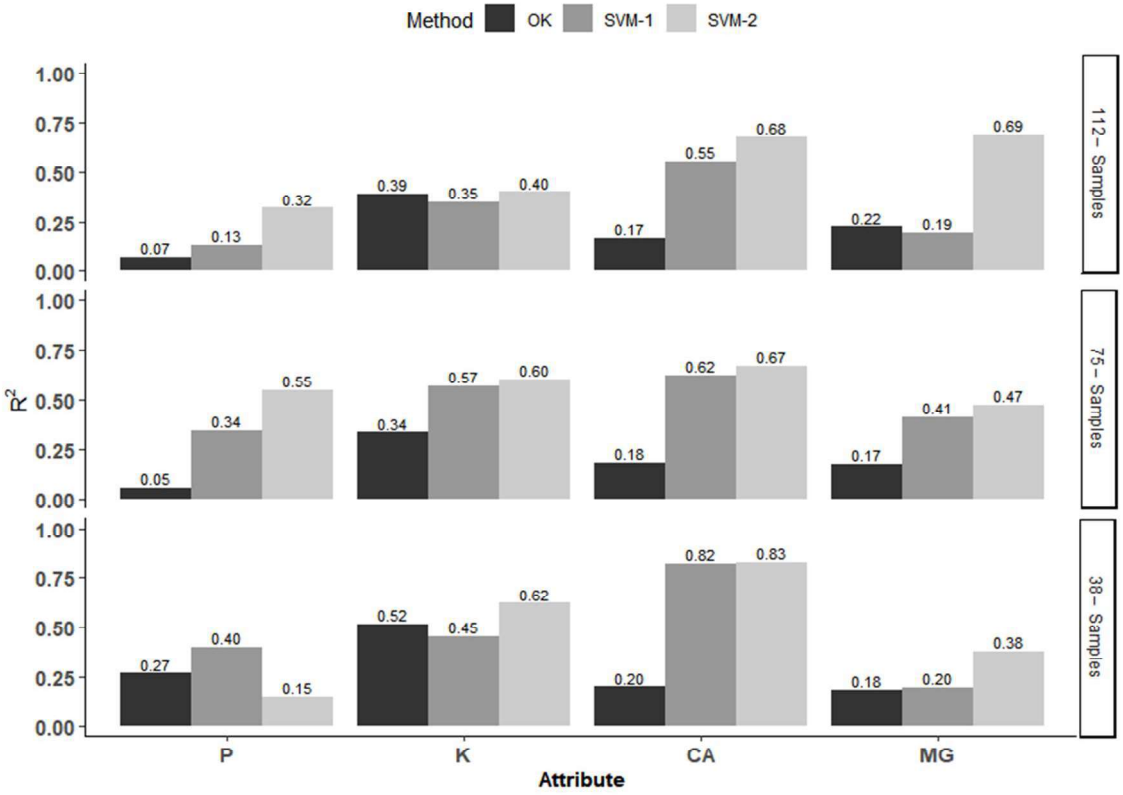


Figura 3.9: Coeficiente de Determinação (R^2) calculado para os atributos P, K^+, Ca^{2+} e Mg^{2+} em três diferentes grids amostrais para o conjunto de teste.

Os valores de RMSE para os atributos de solo P, K^+, Ca^{2+} e Mg^{2+}, para os métodos OK, *SVM1 e SVM2* são apresentados nas Tabelas 3.2 e 3.3 para os conjuntos de treinamento e teste, respectivamente. Os valores de RMSE estão relacionados ao R^2. Como esperado, os valores de RMSE tendem a ser menores para valores maiores de R^2 (Figura 3.8 e Tabela 3.2) para o conjunto de treinamento, (Figura 3.9 e Tabela 3.3) e para o conjunto de teste. Resultados semelhantes foram observados em outros estudos (DA MATTA CAMPBELL et al., 2019; GOMES et al., 2019). Com valores de RMSE menores (Tabela 3.2), pode-se inferir que o OK foi superior ao *SVM1* na predição de P, em que o R^2 foi semelhante (R^2 = 0,11 e 0,15 no grid de 75 e 112 pontos respectivamente) como mostrado na Figura 3.8.

Tabela 3.2: Valores de RMSE encontrados para P, K^+, Ca^{2+} e Mg^{2+} para os grids amostrais com grid de 38, 75 e 112 pontos amostrais para o conjunto de treinamento.

Densidade	38 amostras			75 amostras			112 amostras		
Variável*	OK	SVM1	SVM2	OK	SVM1	SVM2	OK	SVM1	SVM2
P	3,24	2,92	2,85	2,91	3,19	2,80	3,36	3,47	3,32
K^+	11,57	10,87	8,94	8,73	9,21	9,03	10,33	10,27	10,09
Ca^{2+}	0,46	0,42	0,40	0,40	0,40	0,38	0,40	0,40	0,39
Mg^{2+}	0,12	0,12	0,11	0,10	0,10	0,10	0,11	0,10	0,10

*P, K^+ em (mg/dm^3), e Ca^{2+}, Mg^{2+} em ($cmolc/dm^3$).

Tabela 3.3: Valores de RMSE encontrados para P, K^+, Ca^{2+} e Mg^{2+} para os grids amostrais com densidade de 38, 75 e 112 pontos amostrais para o conjunto de teste.

Densidade	112 amostras			75 amostras			38 amostras		
Variável *	OK	SVM1	SVM2	OK	SVM1	SVM2	OK	SVM1	SVM2
P	3,40	3,36	3,22	3,59	3,04	2,74	2,75	1,94	2,79
K^+	9,74	10,05	9,70	12,01	11,77	11,41	9,04	9,46	8,14
Ca^{2+}	0,41	0,29	0,28	0,41	0,26	0,25	0,41	0,24	0,23
Mg^{2+}	0,11	0,11	0,07	0,12	0,10	0,10	0,15	0,14	0,10

*P, K^+ em (mg/dm^3), e Ca^{2+}, Mg^{2+} em ($cmolc/dm^3$).

Mapas dos atributos de solo

Para gerar os mapas de atributos de solo utilizou-se as amostras selecionadas a partir dos conjuntos de treinamento com 38, 75, 112 pontos amostrais, como mostrado na Figura 3.4. O conjunto de 150 pontos também foi utilizado para realizar a interpolação e obter os mapas interpolados. Os atributos P, K^+, Ca^{2+} e Mg^{2+} foram interpolados utilizando os métodos OK, *SVM1 e SVM2,* obtendo-se, os mapas em quatro densidades de pontos. Para a obtenção dos mapas utilizou-se um grid com células de tamanho 10 x 10 m, totalizando 7.388 pontos interpolados. Cada atributo interpolado apresentou um padrão diferente de variabilidade espacial (Figura 3.10, 3.11, 3.12 e 3.13). Isso pode estar associado às características de mobilidade do atributo no solo, forma do relevo, formação dos solos e manejo do solo ao longo dos anos.

O RMSE apresentado na Tabela 3.3 pode ser interpretado como o erro da interpolação para cada mapa obtido por interpolação em cada densidade do grid amostral e para cada atributo de solo. Esse erro foi calculado com base no conjunto de teste, pois nos mesmos locais onde se encontravam os pontos de teste, foram

extraídos os valores dos mapas obtidos por interpolação de P, K^+, Ca^{2+} e Mg^{2+}, calculando dessa forma o RMSE. Para os mapas obtidos por interpolação do atributo P no grid com 38 pontos amostrais do conjunto de treinamento (Figura 3.10 a.1, b.1 e c.1), o método *SVM2* apresentou menor erro (RMSE = 3,22 mg/dm^3), considerando o conjunto de 112 pontos de teste, de acordo com a Tabela 3.3. Para o grid com densidade de 75 pontos amostrais no conjunto de treinamento e teste (Figura 3.10 a.2, b.2 e c.2), também o *SVM2* apresentou o menor RMSE (2,74 mg/dm^3), seguido do *SVM1* e OK (Tabela 3.3). Para o grid com densidade de 112 pontos amostrais no conjunto de treinamento (Figura 3.10 a.3, b.3 e c.3) e 38 pontos de teste, o mapa obtido por interpolação pelo método *SVM1* apresentou o menor RMSE (1,94 mg/dm^3), seguindo pelo OK. O *SVM2* apresentou o maior erro (RMSE = 2,79 mg/dm^3). Para o mapa obtido por interpolação no grid com 150 pontos amostrais do conjunto de treinamento não foi possível calcular o erro, uma vez que não foi separado nenhum ponto observado para o conjunto de teste. Para essa densidade, os maiores teores de P concentram-se distribuídos na parte central do mapa (Figura 3.10 a.4, b.4 e c.4).

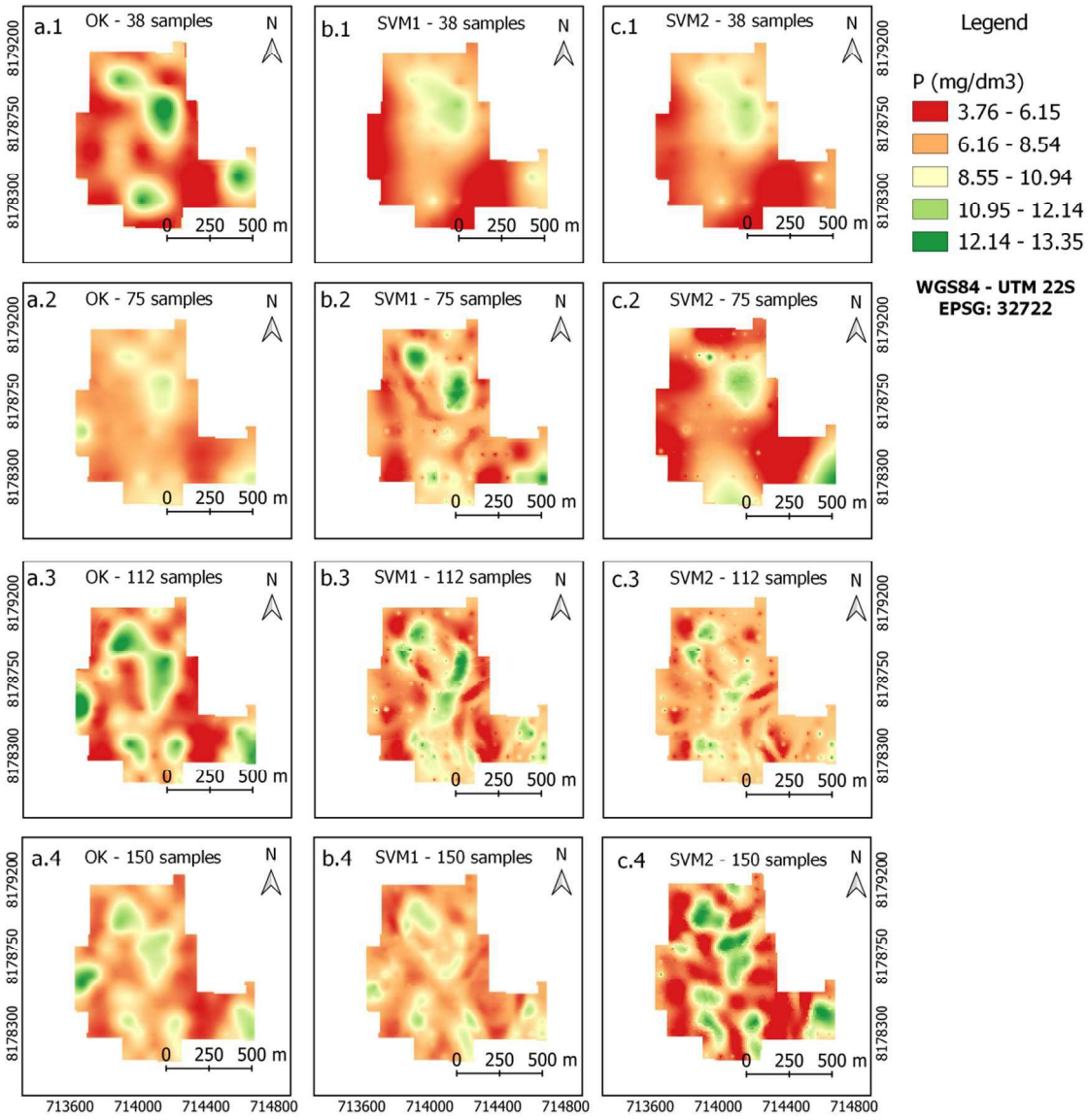


Figura 3.10: Mapas obtidos por interpolação de Fósforo (P): (a) OK, (b) SVM1, (c) SVM2; Conjunto de pontos (treinamento): 38 (a.1-c.1), 75 (a.2-c.2), 112 (a.3-c.3) e 150 pontos (a.4-c.4).

Para os mapas obtidos por interpolação do atributo K^+ nos grids com 38 (Figura 3.11 a.1, b.1 e c.1), 75 (Figura 3.11 a.2, b.2 e c.2) e 112 (Figura 3.11 a.3, b.3 e c.3) amostras do conjunto de treinamento, o método *SVM2* apresentou menor erro, seguido por OK nos conjuntos com 38 e 112 e *SVM1* no conjunto de 75 pontos (Tabela 3.3). Para mapa obtido por interpolação no grid de 150 pontos as maiores concentrações de K^+ estão localizadas na parte leste e na parte oeste do mapa (Figura 3.11 a.4, b.4 e c.4).

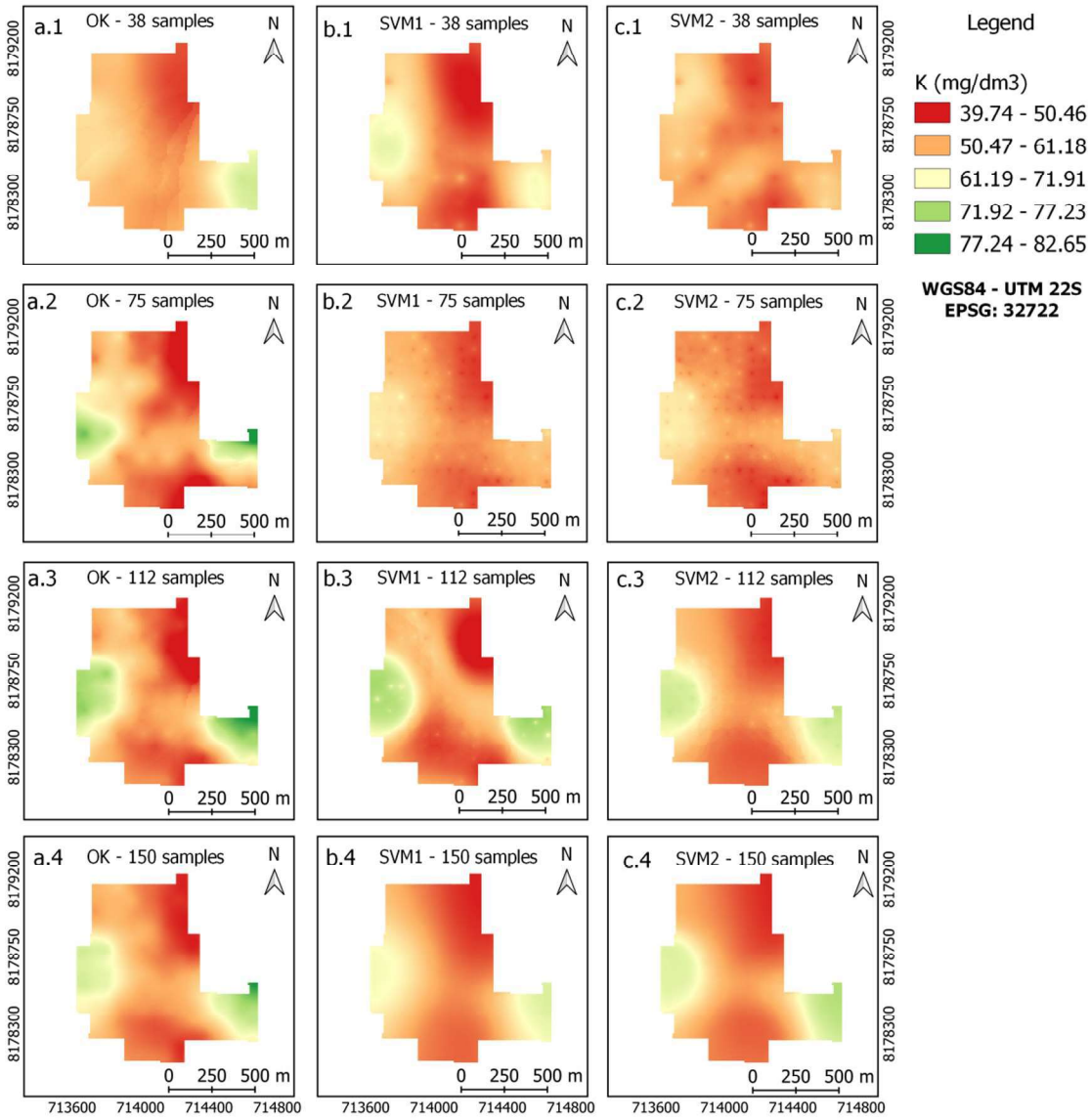


Figura 3.11: Mapas obtidos por interpolação de Potássio (K^+): (a) OK, (b) SVM1, (c) SVM2; Conjunto de pontos (treinamento): 38 (a.1-c.1), 75 (a.2-c.2), 112 pontos (a.3-c.3), 150 pontos (a.4-c.4).

O método *SVM2* obteve o menor erro de interpolação nas 3 densidades de grids amostrais, seguido pelo *SVM1* e OK para o atributo Ca^{2+} como pode ser observado na Tabela 3.3. Para o grid com densidade de 150 pontos amostrais Ca^{2+} apresentou valores mais elevados nas partes norte e centro da área de estudo (Figura 3.12 a.4, b.4 e c.4) para os três métodos interpoladores.

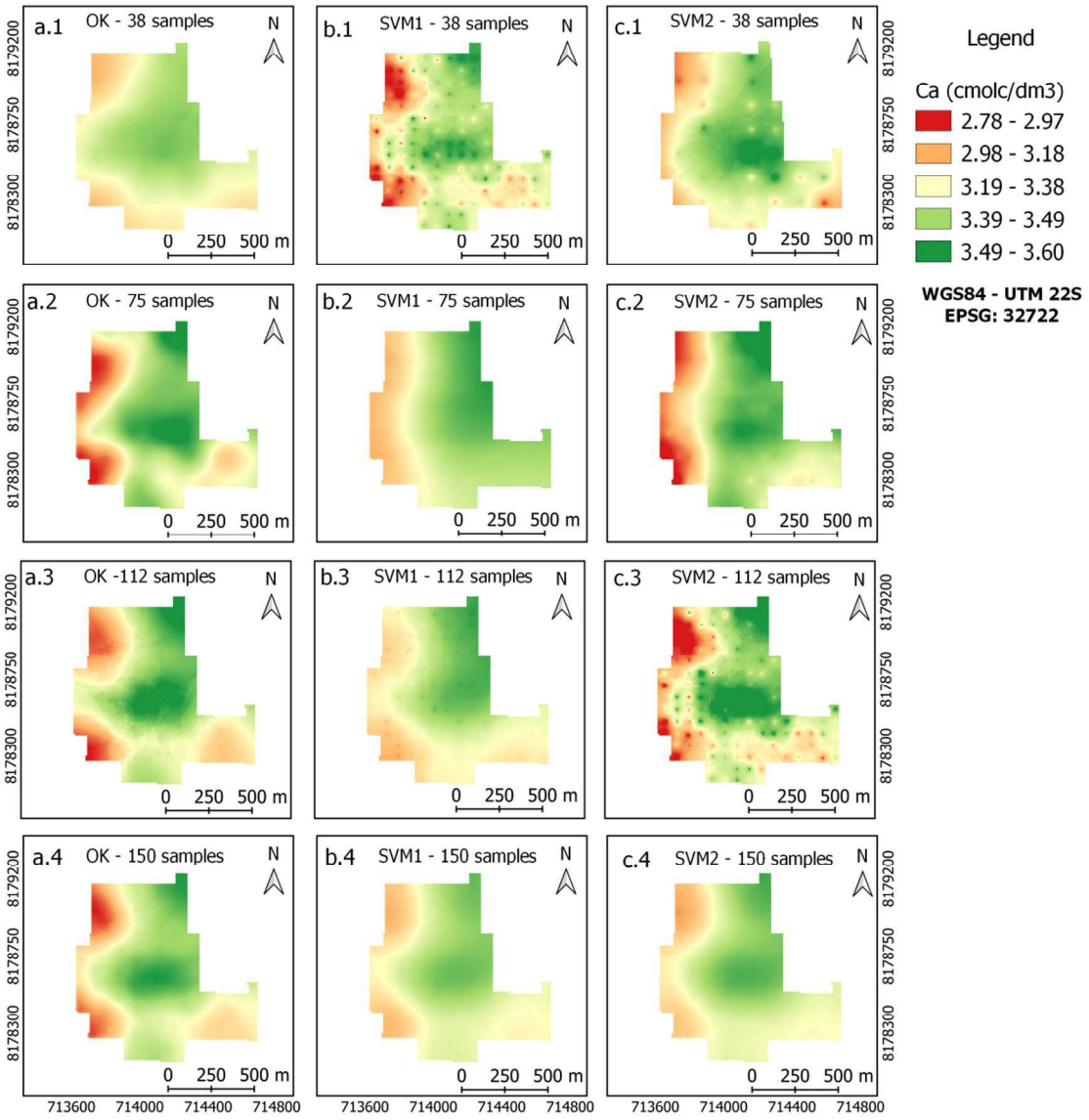


Figura 3.12: Mapas obtidos por interpolação de Cálcio (Ca^{2+}): (a) OK, (b) *SVM1*, (c) *SVM2*; Conjunto de pontos (treinamento): 38 (a.1-c.1), 75 (a.2-c.2), 112 pontos (a.3-c.3), 150 pontos (a.4-c.4).

O atributo Mg^{2+} assim como o Ca^{2+} e o K^{+} apresentou o menor erro para os mapas obtidos por interpolação pelo método *SVM2*. No grid de 75 pontos amostrais para o conjunto de treino e teste (Figura 3.13 a.2, b.2 e c.2), os métodos *SVM1* e *SVM2* obtiveram o mesmo erro (RMSE = 0,10 cmolc/dm^3). Como o R^2 do *SVM2* (0,47) foi superior ao R^2 do *SVM1* (0,41) de acordo com a Figura 3.9, podemos inferir que a performance do *SVM2* foi superior ao do *SVM1*. O mesmo ocorreu para o grid de 38 pontos do conjunto de treinamento (Figura 3.13 a.1, b.1 e c.1), e o grid de 112 amostras no conjunto de teste, o erro dos métodos OK e *SVM1* foram de 0,11

cmolc/dm^3. A OK foi superior pois apresentou R^2 maior (Figura 3.9). Para o grid com 150 pontos amostrais, o mapa obtido por interpolação apresentou comportamento espacial com os maiores valores concentrados na parte norte da área (Figura 3.13 a.4, b.4 e c.4).

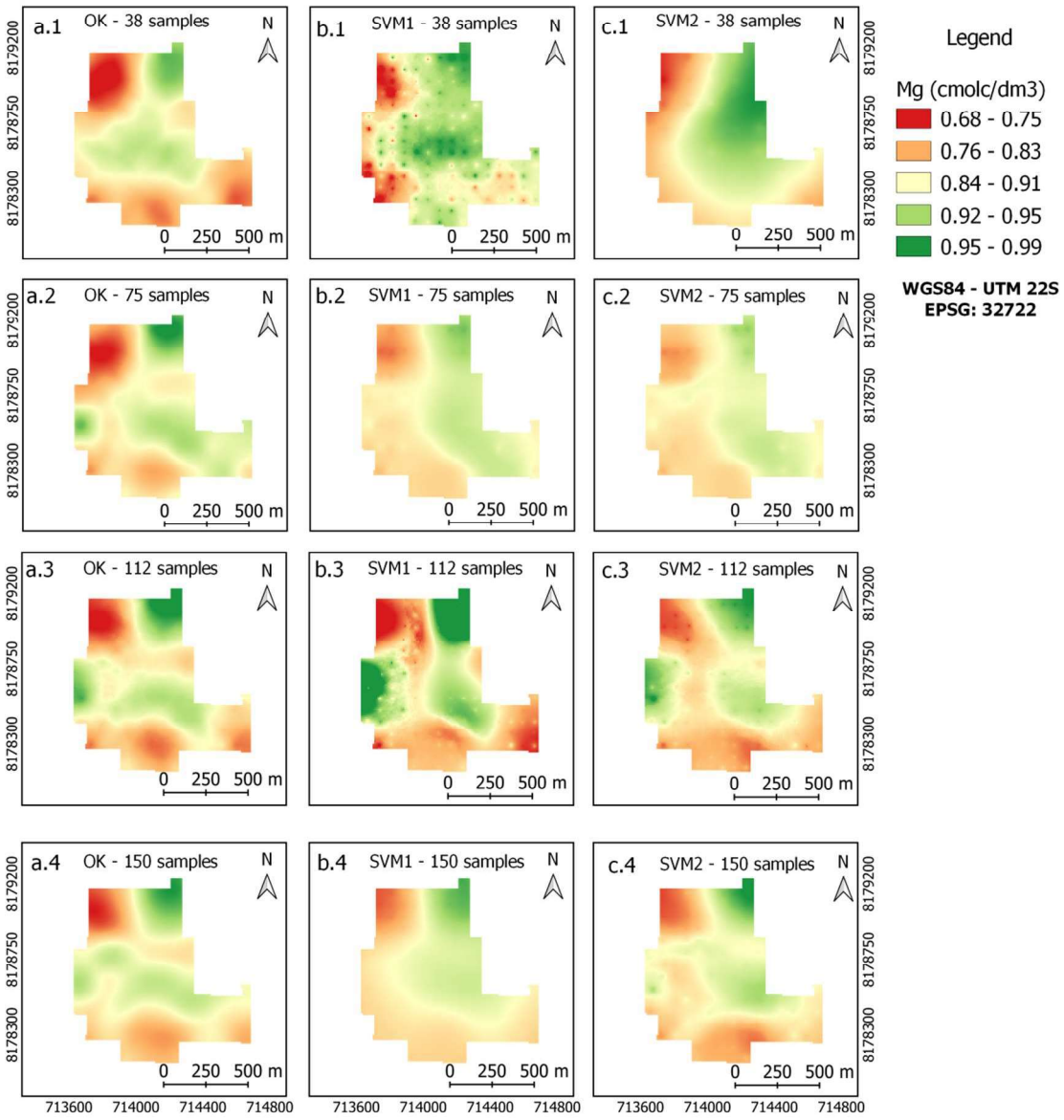


Figura 3.13: Mapas obtidos por interpolação de Magnésio (Mg^{2+}): (a) OK, (b) SVM1, (c) SVM2; Conjunto de pontos (treinamento): 38 (a.1-c.1), 75 (a.2-c.2), 112 pontos (a.3-c.3), 150 pontos (a.4-c.4).

3.6 Conclusões

O plugin *Smart-Map* desenvolvido está disponível para download no site GitHub (https://github.com/gustavowillam/SmartMapPlugin) e no repositório de plugins do QGIS (https://plugins.qgis.org/plugins/Smart_Map). Foram implementadas técnicas para mapeamento digital de atributos do solo utilizando Krigagem Ordinária e Machine Learning. A interpolação por Machine Learning permite importar dados das layers QGIS de banco de dados tipo raster e vetorial para serem utilizados como covariáveis na interpolação. Os mapas gerados pelo plugin podem ser exportados para o QGIS em formato shapefile e/ou raster.

No estudo de caso, foram comparados a interpolação utilizando três métodos. Krigagem ordinária (OK), o método de Machine Learning que utiliza como covariável o próprio atributo interpolado por IDW *(SVM1*) e com utilização de covariáveis (*SVM2*). Dessa forma, pode-se concluir:

1) O método *SVM2* foi superior aos demais modelos na predição de atributos químicos do solo nas três densidades de pontos dos grids amostrais. Os valores de R^2 foram superiores em 11 das 12 combinações entre os 4 atributos de solos interpolados em três densidades de pontos dos grids amostrais, considerando-se o conjunto de treinamento.

2) Considerando o RMSE do conjunto de teste, *SVM2* apresentou o menor erro para a predição dos mapas obtidos por interpolação para os quatro atributos de solos nas três densidades amostrais, exceto para o atributo P no método *SVM1* com grid de 38 pontos no conjunto de teste.

3) Uma dificuldade encontrada pelos algoritmos de ML para problemas de mapeamento e predição de atributos do solo é lidar com o número excessivo de covariáveis a serem utilizadas no modelo. Correlação espacial do I'Moran mostrou-se eficiente para seleção de covariáveis de maior importância para o modelo.

4) Em áreas com baixa correlação espacial dos atributos de solos e poucos pontos amostrados, técnicas de ML é uma alternativa ao método OK. Principalmente quando estão disponíveis covariáveis com maior número de pontos e que apresentam um nível de correlação significativa com as variáveis a serem interpoladas. Os resultados neste trabalho confirmaram a viabilidade e aplicabilidade de técnicas de

ML, em especial o método "*Support Vector Machine*", para predição e mapeamento de atributos químicos do solo em escala regional.

3.7 Referências

ALBORNOZ, E. M. et al. Development and evaluation of an automatic software for management zone delineation. **Precision Agriculture**, v. 19, n. 3, p. 463–476, 2018.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: the fuzzy c-means clustering algorithm. **Comput. Geosci.**, v. 10, p. 191–203, 1984.

CHEN, S. et al. Delineation of management zones and optimization of irrigation scheduling to improve irrigation water productivity and revenue in a farmland of Northwest China. **Precision Agriculture**, v. 21, n. 3, p. 655–677, 2019.

COSTA, M. M. et al. Moisture content effect in the relationship between apparent electrical conductivity and soil attributes. **Acta Scientiarum - Agronomy**, v. 36, n. 4, p. 395–401, 2014.

DA MATTA CAMPBELL, P. M. et al. Digital mapping of soil attributes using machine learning. **Revista Ciencia Agronomica**, v. 50, n. 4, p. 519–528, 2019.

GOMES, L. C. et al. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, n. January, p. 337–350, 2019.

GREGORUTTI, B.; MICHEL, B.; SAINT-PIERRE, P. Correlation and variable importance in random forests. **Statistics and Computing**, v. 27, n. 3, p. 659–678, 2017.

GUO, P. T. et al. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. **Geoderma**, v. 237–238, p. 49–59, 2015.

HENGL, T. et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HEUNG, B. et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62–77, 2016.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214–215, p. 141–154, 2014.

HUO, X. N. et al. Combining geostatistics with moran's i analysis for mapping soil heavy metals in Beijing, China. **International Journal of Environmental Research and Public Health**, v. 9, n. 3, p. 995–1017, 2012.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An Introduction to Applied Geostatistics**. New York: Oxford University Press, 1989.

KESKIN, H.; GRUNWALD, S.; HARRIS, W. G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, v. 339, n. November 2017, p. 40–58, 2019.

KHALEDIAN, Y.; MILLER, B. A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401–418, 2020.

LEE, S. Developing a bivariate spatial association measure: An integration of Pearson' s r and Moran' s I. **Geographical Systems**, v. 3 p. 369–385, 2001.

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

LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors (Switzerland)**, v. 18, n. 8, p. 1–29, 2018.

LIU, Q.; XIE, W. J.; XIA, J. B. Using Semivariogram and Moran's I Techniques to Evaluate Spatial Distribution of Soil Micronutrients. **Communications in Soil Science and Plant Analysis**, v. 44, n. 7, p. 1182–1192, 2013.

MALLA, R. et al. Soil Fertility Mapping and Assessment of the Spatial Distribution of Sarlahi District, Nepal. **American Journal of Agricultural Science**, v. 7, n. 1, p. 8–16, 2020.

MEIER, M. et al. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. **Revista Brasileira de Ciência do Solo**, v. 42, n. 0, p. 1–22, 2018.

MUPHY, B.; MULLHER, S.; YURCHARK, R. **GeoStat-Framework/PyKrige v1.5.1(Version v1.5.1)**.

PARMLEY, K. A. et al. Machine Learning Approach for Prescriptive Plant Breeding. **Scientific reports**, v. 9, n. 1, p. 17132, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 1, p. 2825–2830, 2011.

POULADI, N. et al. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, n. October 2018, p. 85–92, 2019.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System. Open Source Geoespacial Found. Proj.**QGIS Development Team. (2018), 2018.

R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna; Austria**. Retrivied form http://www.r-project.org/R Core Team. (2020), 2020.

REMY, N.; BOUCHER, A.; WU, J. **Applied Geostatistics with SGeMS**. Cambridge: Cambridge University Press, 2009.

REY, S. J.; ANSELIN, L. **PySAL: A Python Library of Spatial Analytical Methods. In: Fischer M., Getis A. (eds).** Springer, Berlin, Heidelberg, 2010.

SANTOS, H. G. et al. Sistema Brasileiro de Classificação de Solos. **Embrapa Solos**. 5. ed. Brasilia, DF, 2018.

SEKULIĆ, A. et al. Random forest spatial interpolation. **Remote Sensing**, v. 12, n. 10, p. 1–29, 2020.

VALENTE, D. S. M. et al. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. **Scientia Agricola**, v. 69, n. 3, p. 173–179, 2012.

VERONESI, F.; SCHILLACI, C. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. **Ecological Indicators**, v. 101, 2018, p. 1032–1044, 2019.

WARNER, J.; SEXAUER, J.; UNNIKRISHNAN, A. **JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2.**

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of Soil Science**, v. 43, p. 177–192, 1992.

WHELAN, B. M.; MCBRATNEY, A. B.; MINASNY, B. VESPER 1.5 - Spatial prediction software for precision agriculture. **6th International Conference on Precision Agriculture**, p. 1–14, 2002.

XU, S. et al. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. **Geoderma**, v. 310, p. 29–43, 2018.

4 TÉCNICAS DE AMOSTRAGEM DE SOLO PARA AGRICULTURA DE PRECISÃO: ANÁLISE DE ESTUDOS DE CASO NO BRASIL

4.1 Resumo

A agricultura de precisão (AP) tem como objetivo maximizar a lucratividade e reduzir os impactos ambientais dos sistemas de produção agrícola. Para isso, é necessário conhecer a variabilidade espacial de atributos do solo nas áreas exploradas. O número de amostras necessário para isso é superior ao adotado nas práticas de agricultura convencional. Entretanto, o custo de uma amostragem densa poderá, muitas vezes, tornar a caracterização da variabilidade espacial inviável. Uma forma de otimizar a amostragem seria com utilização de sensores. Logo, a condutividade elétrica aparente (ECa) é uma ferramenta alternativa para caracterizar a variabilidade espacial no campo, devido sua facilidade de aquisição e baixo custo. Por meio de zonas de manejo (ZM) definidas com base na ECa, pode-se reduzir o número de amostras de solo necessárias para se caracterizar a variabilidade espacial dos atributos do solo. O objetivo deste estudo foi analisar as diferentes técnicas de amostragem para caracterização da variabilidade espacial de atributos do solo. Foram implementados os métodos de amostragem por célula, por zonas de manejo, pelo método convencional e por grid, gerado em duas diferentes densidades amostrais. A raiz quadrada do erro quadrático médio (RMSE) foi utilizada para avaliar o desempenho dos métodos de amostragem medida através da validação cruzada *leave-one-out cross-validation* (*LOOCV*). Para melhor interpretação dos resultados gerados, foi determinado a autocorrelação espacial dos atributos de solo pelo Índice de Moran univariado. Os trabalhos foram conduzidos em seis diferentes áreas, localizadas em seis diferentes estados brasileiros. Os resultados revelam que a amostragem por célula apresentou o melhor desempenho, seguido da amostragem por grid de maior densidade. A amostragem por ZM definidas pela ECa apresentou uma performance superior em relação aos métodos de amostragem com grid de menor densidade e amostragem convencional. Portanto, o método de amostragem em células e por zonas de manejo com base na ECa são uma alternativa aos métodos de amostragem em grid.

Palavras-chave: Mapeamento digital de atributos do solo. Zonas de manejo. Krigagem ordinária. Geoprocessamento.

4.2 Abstract:

Precision agriculture (PA) aims to maximize profit and reduce the environmental impacts of agricultural production systems. For this, it is necessary to know the spatial variability of the soil in the explored areas. The number of samples required for this is greater than that adopted in conventional agricultural practices. However, the cost of dense sampling can often make the characterization of spatial variability unfeasible. One way to optimize sampling would be with the use of sensors. Therefore, apparent electrical conductivity (ECa) is an alternative tool to characterize the spatial variability in the field, due to its ease of acquisition and low cost. Through management zones (ZM) defined based on ECa, the number of soil samples needed to characterize the spatial variability of soil attributes can be reduced. Thus, this work aims to analyze the different sampling techniques to characterize the spatial variability of soil attributes. Sampling methods by cell, by management zones, by the conventional method and by grid, generated in two different sampling densities, were implemented. The square root mean square error (RMSE) was used to evaluate the performance of the sampling methods measured through the leave-one-out cross-validation (LOOCV) cross-validation. For better interpretation of the generated results, the spatial autocorrelation of soil attributes was determined using the univariate Moran Index. The work was carried out in six different areas, located in six different Brazilian states. The results reveal that cell sampling showed the best performance, followed by higher density grid sampling. The ZM sampling defined by ECa presented a superior performance in relation to sampling methods with lower density grid and conventional sampling. Therefore, the method of sampling in cells and by management zones based on ECa are an alternative to the methods of sampling in grid.

Key words: Digital soil mapping. Management zones, Ordinary kriging. Geoprocessing.

4.3 Introdução

A agricultura de precisão (AP) pode ser definida como o gerenciamento dos insumos agrícola levando em consideração a variabilidade espacial e temporal dos fatores de produção (COELHO et al., 2020; COSTA et al., 2014). Essa variabilidade espacial é geralmente descrita por meio de mapas dos atributos do sistema solo-planta. Normalmente, para se gerar os mapas de atributos do solo visando à

recomendação de corretivos e fertilizantes à taxas variadas, é necessário realizar uma amostragem densa do solo (ADAMCHUK et al., 2007; FERGUSON; HERGERT, 2009). Após a amostragem, uma interpolação dos atributos é feita para obtenção dos mapas de variabilidade espacial. A exatidão dos mapas está diretamente relacionada à densidade amostral, à correlação entre estas amostras, e ao número total de amostras (GUO-SHUN et al., 2010; LAWRENCE et al., 2020). No entanto, reduzir pela metade a distância entre amostras significa aproximadamente quadruplicar o número de amostras, consequentemente aumentando os custos de amostragem.

O método de amostragem em grid gera maiores custos, principalmente quando realizado com elevada densidade amostral. Para reduzir os custos, muitos agricultores adotam um sistema de amostragem com menor densidade amostral, normalmente com uma amostra a cada três a cinco hectares. Após a obtenção dessas amostras, os mapas são interpolados por métodos de Krigagem Ordinária (OK) ou Inverso da Distância Ponderada (IDW) (BORGES et al., 2020; COELHO et al., 2018; OUABO; SANGODOYIN; OGUNDIRAN, 2020; PARREIRAS et al., 2020). No entanto, essa prática poderá gerar erros maiores que assumir a média dos atributos do solo para toda a área.

Alternativamente ao método de amostragem em grid, pode-se realizar uma amostragem em células (DHAKAL; AMADA; ANIYA, 2000; WOLLENHAUPT; WOLKOWSKI, 1994) ou em zonas de manejo (ALBORNOZ et al., 2018; BOTTEGA et al., 2017; MALLARINO, 2001; VALENTE et al., 2012). Na amostragem em células, divide-se a área em pequenos quadrados ou retângulos, normalmente de um a cinco hectares cada. Dentro de cada célula, coleta-se várias amostras simples que irá formar uma amostra composta. A amostra composta representa a média dos atributos em toda a célula. No caso da amostragem realizada por zonas de manejo (ZM), o processo é similar à amostragem em células, a diferença está no processo de definição das ZM. Enquanto as células são definidas sem nenhum critério a priori, as ZM são definidas com base em mapas de produtividade, de condutividade elétrica aparente do solo, de matéria orgânica, de teor de argila, dentre outros atributos (ALBORNOZ et al., 2018; MOHARANA et al., 2020; PACCIORETTI; CÓRDOBA; BALZARINI, 2020; VALLENTIN et al., 2020).

O delineamento de ZM é uma forma técnica e economicamente viável para aplicação de insumos agrícolas à taxas variadas, principalmente, fertilizantes e

sementes (CHEN et al., 2019; SCHENATTO et al., 2017). Por meio da definição de ZM, pode-se reduzir o número de amostras de solo necessárias para obter uma boa representação dos atributos do solo (GAVIOLI et al., 2016). Dentre as variáveis utilizadas para definir as zonas de manejo, a ECa é uma alternativa aos métodos de coleta de dados convencionais, sendo frequentemente utilizada no estabelecimento de ZM (FARID et al., 2016; GUASTAFERRO et al., 2010; LI et al., 2007; MORAL et al., 2019; MORAL; REBOLLO; SERRANO, 2020; SCHENATTO et al., 2017). As medições da ECa são adequadas para caracterizar a distribuição espacial dos atributos do solo por serem confiáveis, rápidas e fáceis de realizar, especialmente quando se usa o princípio da resistividade elétrica (CORWIN; LESCH, 2003; CORWIN; SCUDIERO, 2020; QUEIROZ et al., 2020; SHADDAD et al., 2016; ÜNAL; KABAŞ; SÖZER, 2020).

Alguns estudos foram realizados para comparação entre os métodos de amostragem em grid em diferentes densidades amostrais (HOFMAN; BRUS, 2021; KHALEDIAN; MILLER, 2020). Estabeleceu-se que uma boa densidade amostral seria de duas amostras compostas a cada hectare (FERGUSON; HERGERT, 2009). Amostragens em grids com densidade amostral inferior a essa poderia levar a erros mais elevados que considerar a média do atributo na área. Além disso, amostragens realizadas em células ou zonas de manejo, definidas utilizando a ECa do solo, poderia gerar erros compatíveis aos grids de amostragem utilizados atualmente. Portanto o objetivo desse trabalho foi comparar os diferentes métodos de amostragem em diferentes regiões do Brasil. Especificamente pretende-se avaliar o desempenho dos métodos de amostragem em células, amostragem em grid com maior e menor densidade, amostragem por ZM definidas a partir da ECa, e a amostragem convencional utilizando a média do atributo na área.

4.4 Material e métodos

Áreas de Estudo

Nesse estudo foram utilizadas seis áreas localizadas nos estados da Bahia (BA), Minas Gerais (MG), Mato Grosso (MT), Goiás (GO), Mato Grosso do Sul (MS) e São Paulo (SP), conforme apresentado na Figura 4.1. Nessas áreas foram coletadas amostras de solo em grid e dados de condutividade elétrica aparente do solo. O objetivo foi utilizar áreas que estão localizadas em diferentes regiões brasileiras e sob diferentes tipos de cultivo para testar a aplicação de zonas de manejo em diferentes

condições. Na Tabela 4.1 são apresentadas algumas informações das áreas de estudo.

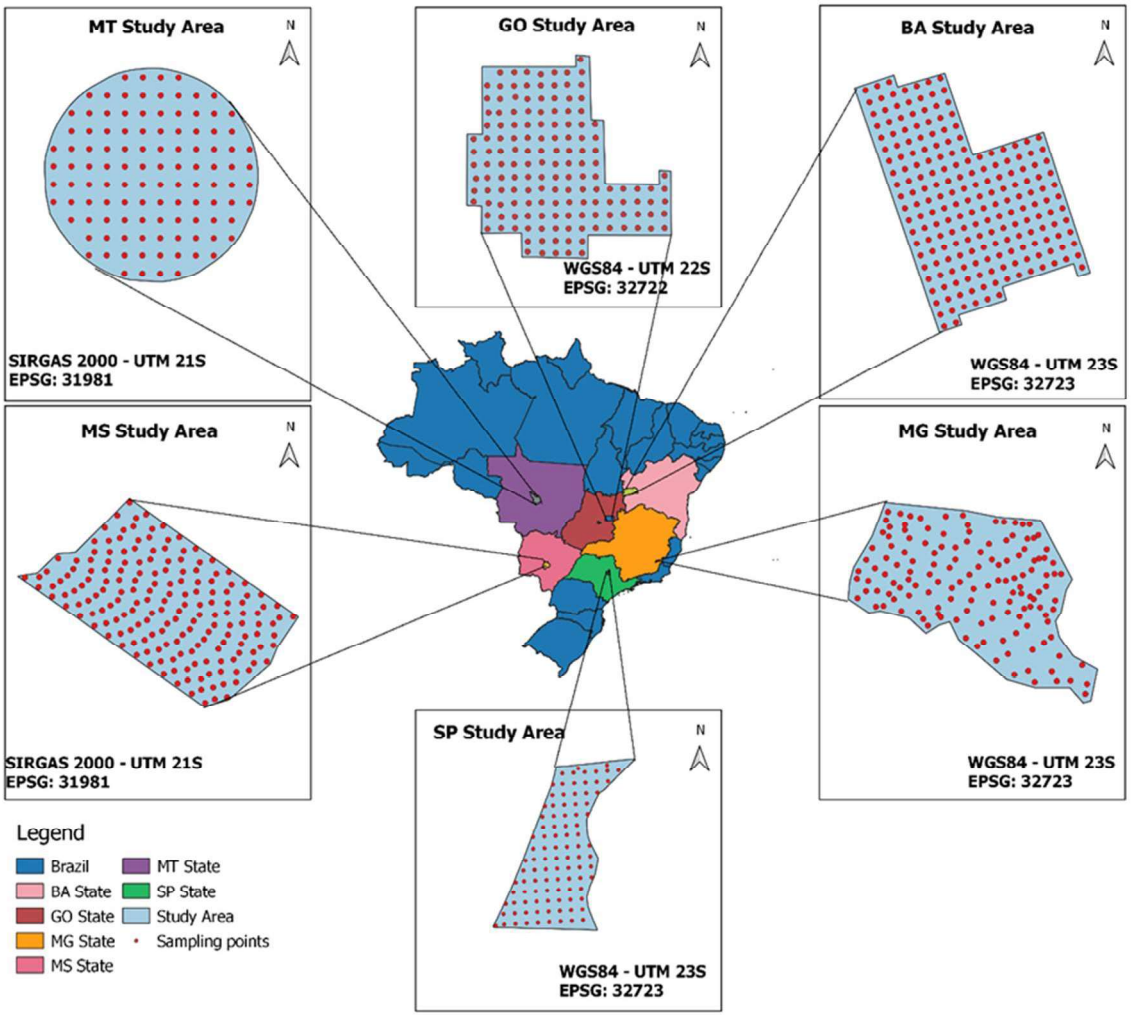


Figura 4.1: Localização geográfica das áreas de estudo e distribuição dos pontos amostrais.

Tabela 4.1: Informações sobre as áreas de estudo.

Áreas	BA	MG	MT	GO	MS	SP
Area(1)	204	20,2	100,2	75	90	90,04
Nr. Amostras	204	141	111	150	181	119
Município	São Desidério	Araponga	Lucas do Rio Verde	Anápolis	Sidrolândia	Descalvado
Latitude (Sul)	12°25'12"	20º42'33"	13°24'32"	16°28'20"	21°1'54"	21°49'04"
Longitude (Oeste)	45°29'46"	42º34'17"	56°6'56"	49°00'32"	55°2'51"	47°44'11"
Altitude(2)	493m	904m	424m	1011m	490m	757m
Cultura(3)	Soja	Café	Soja	Soja	Soja	Cana-de-Açucar
Ano de Coleta	2017	2009	2020	2010	2010	2017
Classficação do Solo(4)	LAd	LVAd	LVAd	LVA	LVd	LVd
Método Obtenção ECa(5)	RE	RE	Veris U3	RE	RE	EMI

[1/] Area, Área em hectare (ha); [2/] Altitude média (em metros) para cada área de estudo; [3/] Cultura existente na área na época da obtenção das amostras de solo; [4/] Classificação do Solo: Classificação atualizada de acordo com Embrapa Solos (SANTOS et al., 2018): LAd - Latossolo Amarelo Distrófico; LVAd - Latossolo Vermelho Amarelo Distrófico; LVA – Latossolo Vermelho Amarelo, LVd - Latossolo Vermelho Distrófico; [5/] Método utilizado para obtenção da ECa: RE: Resistividade Elétrica: medida com o medidor portátil fabricado pela Landviser®, modelo Landmapper® ERM-02 (League City, Texas, United States); Veris U3: plataforma de mapeamento de solo Veris U3® (Veris Technologies Inc., Salina KS USA), EMI: Indução Eletromagnética medida com o aparelho EM38-MK2® (Geonics, Ontário, Canadá).

Fonte: (BA) Nogueira Martins et al., 2020; (MG) Valente et al., 2012; (GO) Costa et al., 2014; (MS) Bottega et al., 2013; (SP) Sanches, 2018.

Nesse estudo, foram considerados 10 atributos do solo, determinados em laboratório a partir das amostras coletadas em campo. Foram selecionados aqueles atributos que eram comuns à todas as áreas nas coletas de dados. Os valores da média e desvio padrão para esses dez atributos de solo (pH, P, K^+, CA^{2+}, MG^{2+}, OM, CEC, V, Clay, Silt) são apresentados na Tabela 4.2.

Tabela 4.2: Estatística descritiva dos atributos de solo nas áreas de estudo.

Area	Atributo*	pH (1)	P (2)	K^+ (3)	Ca^{2+} (4)	Mg^{2+} (5)	OM (6)	CEC (7)	V(%) (8)	Clay (9)	Silt (10)
BA	Média	5,53	13,67	66,78	1,66	0,65	0,77	5,07	49,16	20,25	5,17
	SD (11)	0,29	12,64	20,06	0,35	0,18	0,45	0,73	8,92	2,03	1,62
MG	Média	5,76	4,36	89,21	2,70	1,21	5,42	11,57	36,75	53,57	7,49
	SD	0,50	17,20	32,06	1,20	0,48	1,03	1,91	15,59	4,90	2,04
MT	Média	5,76	28,20	75,30	3,27	1,00	3,43	8,00	56,54	64,47	16,47
	SD	0,50	7,48	16,43	0,51	0,23	0,21	0,59	7,22	2,37	2,65
GO	Média	6,74	6,84	52,63	3,27	0,84	3,06	5,95	72,26	33,11	10,60
	SD	0,29	3,96	14,20	0,46	0,14	0,30	0,86	11,05	3,37	2,94
MS	Média	5,52	24,81	201,18	3,69	0,92	3,76	11,20	45,92	63,43	23,90
	SD	0,30	15,95	55,98	0,75	0,21	0,25	0,81	0,81	2,90	3,96
SP	Média	5,60	25,08	70,50	4,08	1,44	2,95	7,69	81,32	362,4	99,3
	SD	0,12	10,64	25,30	1,19	0,43	0,77	1,74	4,64	154,9	57,7

[1/] pH, Acidez Ativa em Água; [2/] P, Fósforo; [3/] K^+, Potássio; [4/] Ca^{2+}, Cálcio; [5/] Mg^{2+}, Magnésio; [6/] OM, Matéria Orgânica; [7/] CEC, Capacidade de Troca de Cátions a pH 7; [8/] V, Saturação por Bases; [9/] Clay, Argila; [10/] Silt, Silte; [11/] SD, Desvio Padrão.

*P, K^+ em mg/dm^3, e Ca^{2+}, Mg^{2+}, CEC em $cmolc/dm^3$, e OM em g/dm^3, e Clay, Silt em g/kg.
Fonte: (BA) Nogueira Martins et al., 2020; (MG) Valente et al., 2012; (GO) Costa et al., 2014; (MS) Bottega et al., 2013; (SP) Sanches, 2018.

Métodos de amostragem implementados

No presente trabalho foram simulados seis métodos de amostragem, sendo assim definidos como: método convencional (CONV); por células retangulares (CEL);

por zonas de manejo (ZM); por células geradas aleatoriamente (CEL-RND); e por grid em duas densidades de pontos amostrais (GRID-1 e GRID-2).

No método de amostral convencional (CONV), a estimativa do atributo foi considerada como a média dos pontos da área de estudo. Na amostragem em células retangulares (CEL), Figura 4.2, cada área de estudo foi dividida em células retangulares, com aproximadamente o mesmo tamanho. O tamanho de cada célula foi proporcional ao tamanho da área. Na área BA os tamanhos das células variaram de 2,3 à 11,2 ha, visto que esta é a maior área de estudo, enquanto que na área MG os tamanhos das células ficaram entre 0,4 à 1,7 ha, pois foi a menor área de estudo. Nas demais áreas as células foram geradas com tamanhos de um a seis hectares aproximadamente, dispostas como apresentado na Figura 4.2. Os valores dos atributos de solo para cada célula foi o valor médio dos pontos amostrais que ficaram dentro da célula.

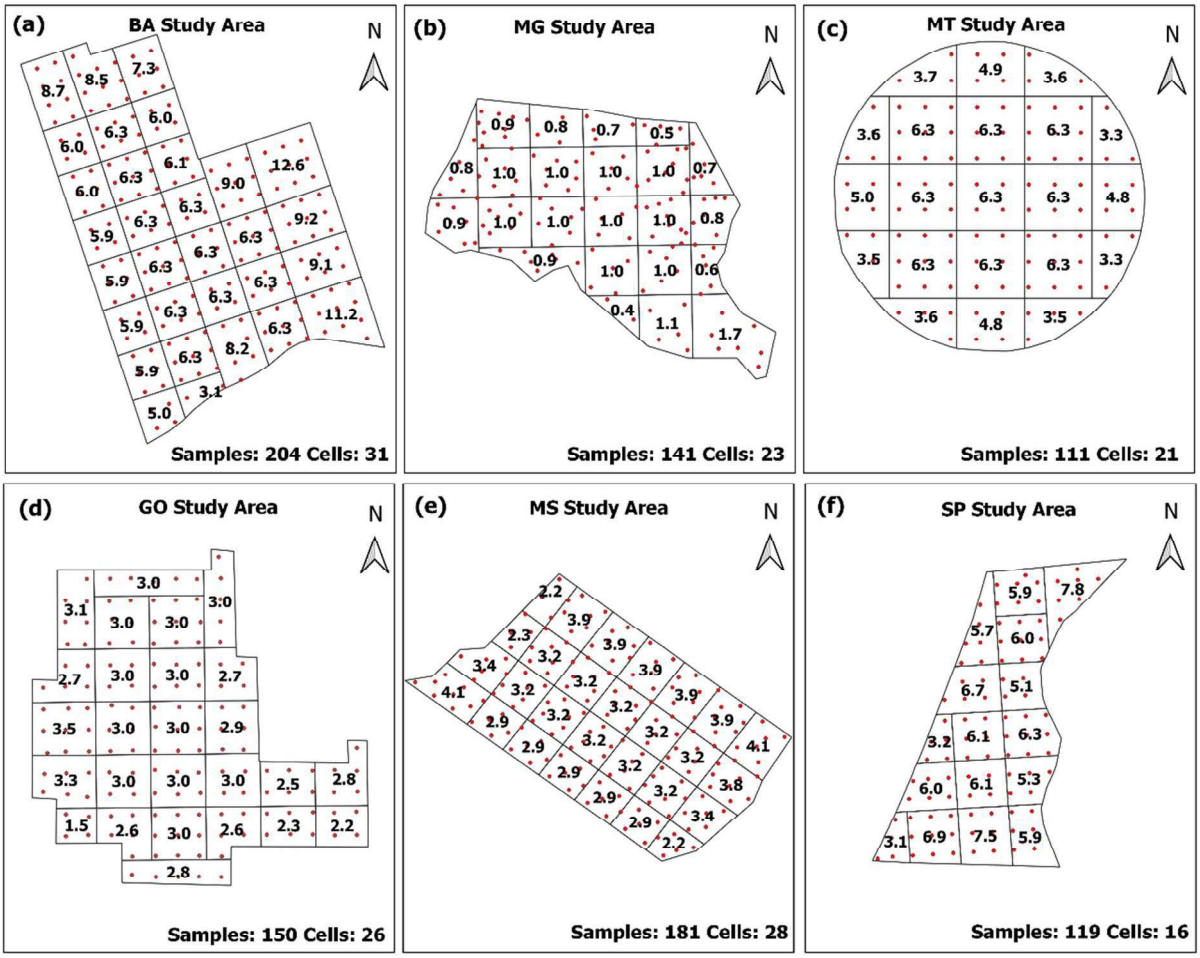


Figura 4.2. Método de amostragem por célula com suas respectivas áreas em hectares (ha), para as áreas de estudo: a) BA; b) MG; c) MT; d) GO; e) MS; f) SP.

Para o método de amostragem por ZM (Figura 4.3), foram geradas zonas de manejo com base no mapa de ECa do solo. Para a definição dos agrupamentos foi utilizado o algoritmo de agrupamento *fuzzy k-means* (BEZDEK; EHRLICH; FULL, 1984) com três classes. Os agrupamentos gerados foram processados e analisados para geração das ZM. As ZM's foram geradas utilizando o plugin *Smart-Map*. O plugin é um complemento para o software QGIS (QGIS DEVELOPMENT TEAM, 2018) versão 3.10 ou superior desenvolvido em Python (criada por Guido Van Rossum e gerenciado pela Python Software Foundation, Delaware, EUA). *Smart-Map* está disponível no GitHub (https://github.com/gustavowillam/SmartMapPlugin) ou pode ser instalado a partir do repositório de plugins do QGIS (https://plugins.qgis.org/plugins/Smart_Map). Após a definição das ZM's, foram utilizadas ferramentas de geoprocessamento do QGIS para ajustar as classes geradas. Classes pequenas foram unidas às áreas maiores, áreas muitas longas e estreitas foram divididas em áreas menores, sendo estas divisões realizadas preferencialmente nas partes estreitas da ZM, e classes iguais separadas espacialmente foram consideradas ZM's distintas. Esse processo foi executado manualmente dentro do QGIS para cada área de estudo. Na Figura 4.3 é apresentado o número de zonas de manejo geradas em cada área.

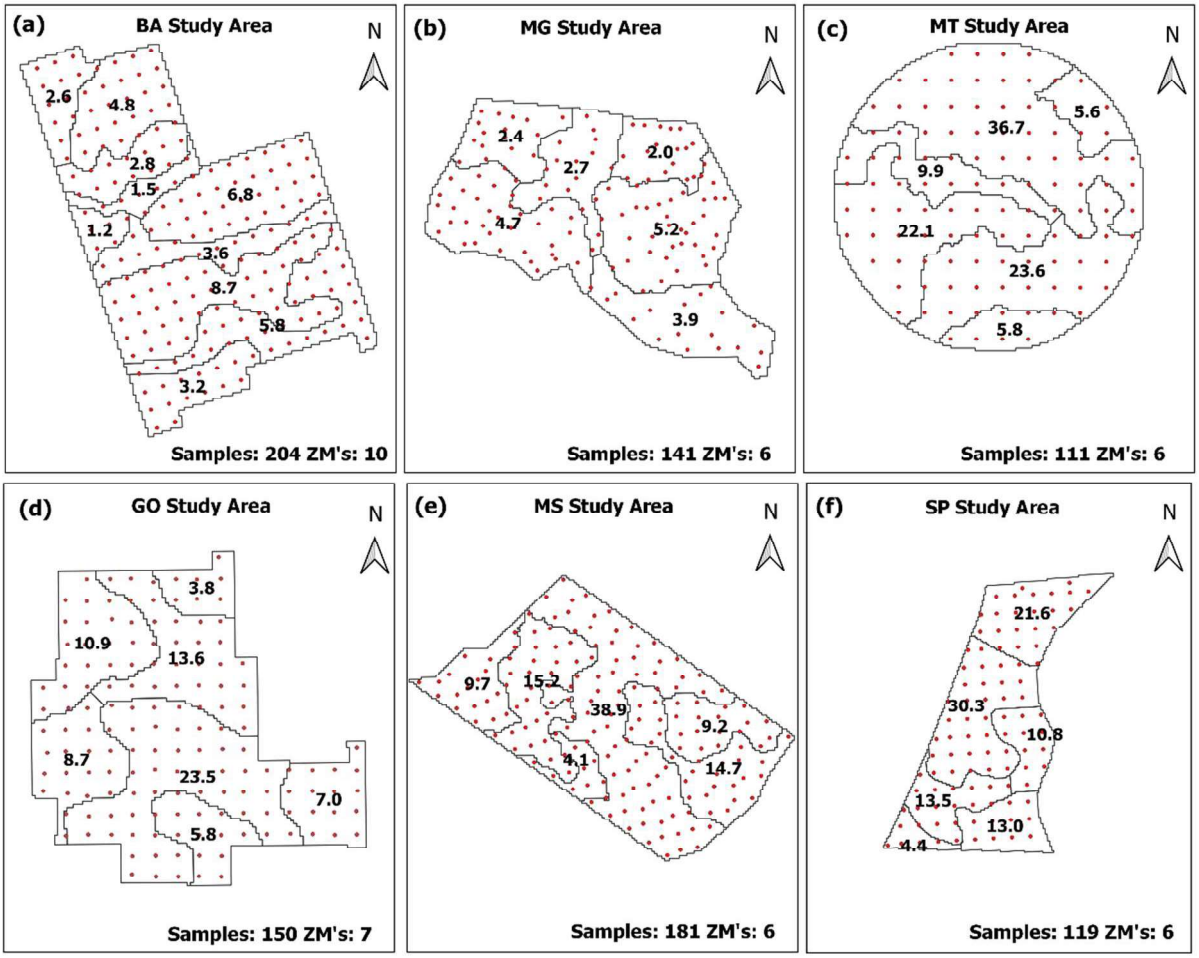


Figura 4.3. Método de amostragem definida por zonas de manejo geradas a partir da ECa do solo e suas respectivas áreas em hectares (ha), para as áreas de estudo: a) BA; b) MG; c) MT; d) GO; e) MS; f) SP.

Para verificar se a ECa realmente contribuiu como critério para definição do esquema de amostragem definido pelas ZMs delimitadas para cada área de estudo, foram geradas células aleatórias com o mesmo número de zonas de manejo, ou seja, o mesmo número de amostras compostas. Esse método foi definido como CEL-RND. Para cada processo de geração de células aleatórias, pode-se determinar os erros para cada atributo. Foram geradas 1000 células aleatórias em cada área de estudo, através de um script Python executado dentro do QGIS. Dessa forma, foi possível verificar a distribuição dos erros para as células aleatórias para cada área e atributo. Na Figura 4.4 é apresentado um exemplo de células geradas de forma aleatória nas áreas de estudo. As células aleatórias foram definidas a partir de pontos aleatórios distribuídos na área de estudo. Após a definição aleatória dos pontos utilizou-se o diagrama de *Voronoi* (FORTUNE, 1987), disponível no QGIS, para definição das células.

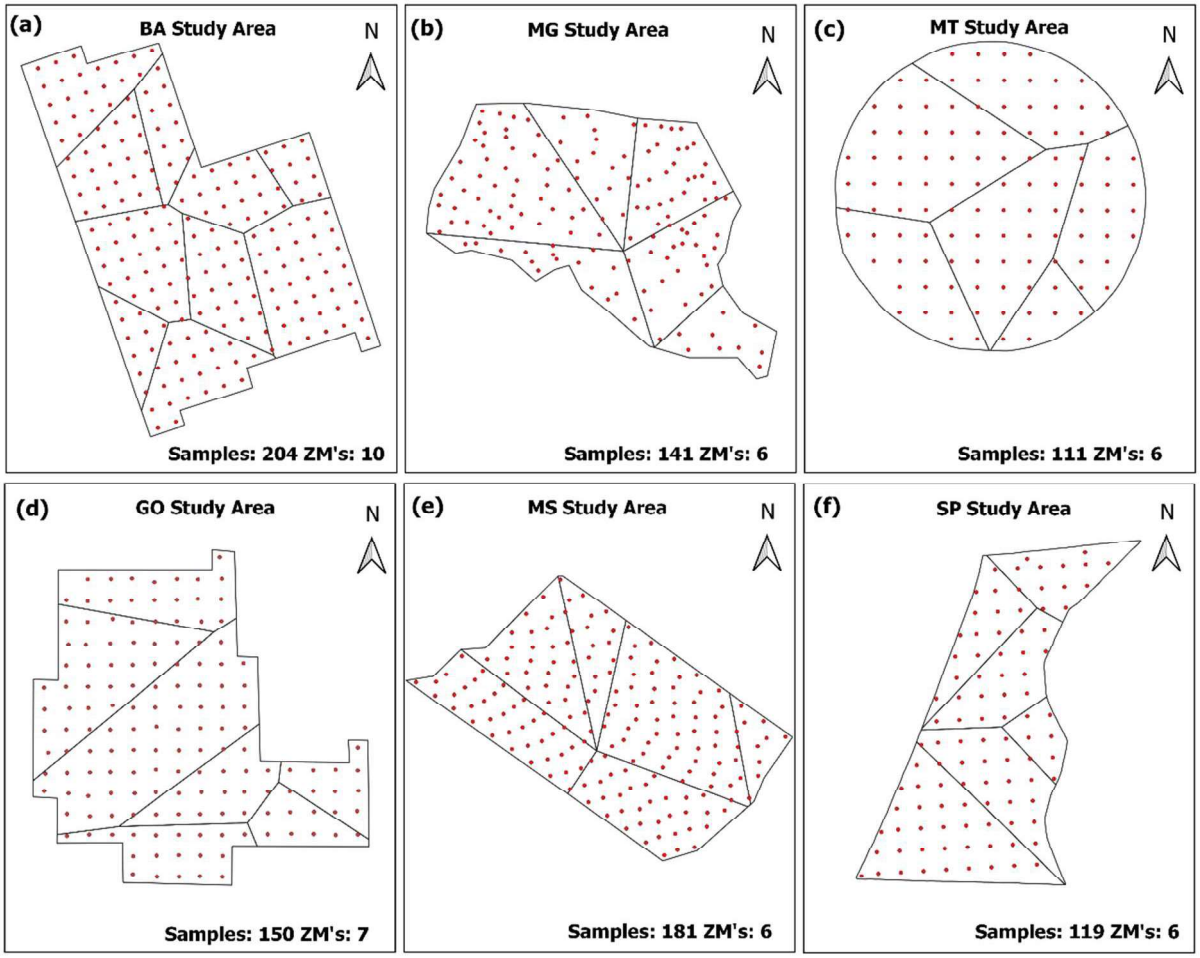


Figura 4.4. Método de amostragem definida por células aleatórias para as áreas de estudo: a) BA; b) MG; c) MT; d) GO; e) MS; f) SP.

Para a amostragem por grid utilizou-se o grid de amostragem original de cada área de estudo apresentados na Figura 4.1. Assim o grid original foi denominado método GRID-1. A partir do GRID-1 (original), foi realizada uma redução da densidade amostral para formar o GRID-2 (Figura 4.5). Em ambos os grids, realizou-se a interpolação por OK utilizando-se o plugin *Smart-Map.*

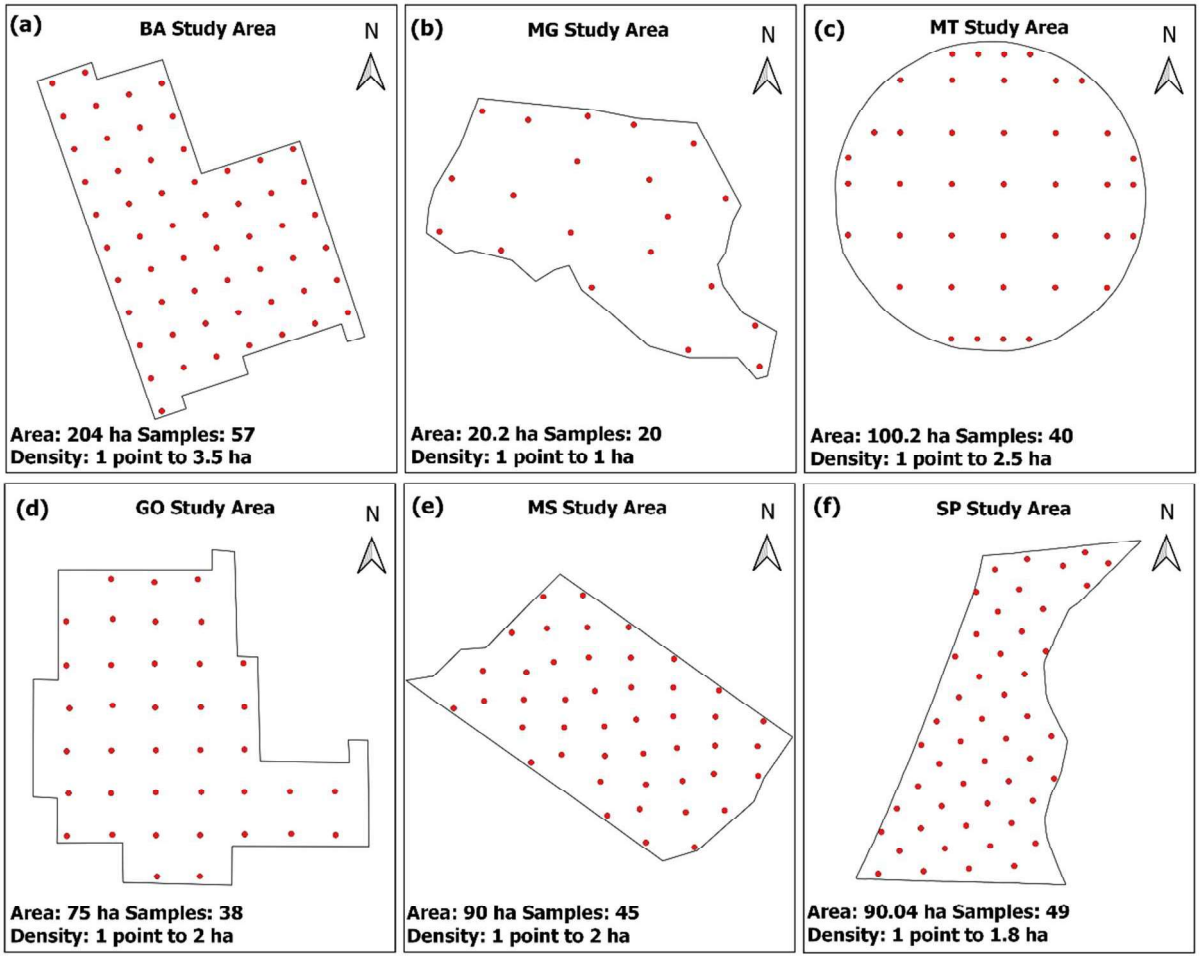


Figura 4.5: Distribuição dos pontos amostrais para o método GRID-2 nas áreas de estudo: a) BA; b) MG; c) MT; d) GO; e) MS; f) SP.

Critérios de desempenho para comparação entre os métodos de amostragem implementados

Para calcular o desempenho de cada método de amostragem foi utilizado o método de validação cruzada *leave-one-out (LOOCV)*. O método consistiu em remover um ponto do conjunto de dados, em seguida, esse ponto era estimado pelos métodos de amostragem descritos anteriormente (CONV, CEL, ZM, GRID-1, GRID-2, CEL-RND). Em seguida o ponto removido era retornado para o conjunto de dados, e repetia-se o processo para o próximo ponto até finalizar todos os pontos de conjunto de dados. De posse dos valores observados e estimados para cada método de amostragem, determinou-se o RMSE, para os atributos pH, P, K^{+}, CA^{2+}, MG^{2+}, OM, CEC, V, Clay, Silt, nas seis áreas de estudo e para cada método de amostragem descrito, conforme Equação 4.1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \quad (4.1)$$

em que: $\hat{x}_i$ representa o valor estimado do atributo de solo no ponto i; x_i o valor observado do atributo de solo no ponto i; e n, o número de pontos amostrados.

Para os métodos de amostragem CEL, ZM, e CEL-RND a estimativa do ponto removido consistiu na determinação do valor médio dos pontos que estão dentro de cada célula ou ZM. Para o método GRID-1 e GRID-2, a estimativa do ponto removido foi determinada pelo método OK após ajuste do semivariograma de cada atributo. Para o método CONV, a estimativa do ponto removido foi determinada pela média dos pontos restantes.

Para auxiliar na interpretação dos resultados encontrados, foi determinada a autocorrelação espacial de cada atributo em cada densidade amostral, calculada por meio do Índice de Moran univariado. Seu valor foi calculado de acordo com a Equação 4.2 (LEGENDRE; FORTIN, 1989), considerando-se o grid com todos os pontos (Figura 4.1) e o grid com o número de pontos reduzidos (Figura 4.5). O pseudo p-valor foi obtido a partir de 999 permutações entre os pontos do grid amostral com o objetivo de verificar sua significância ao nível de 5% de probabilidade.

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} wij} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x}).\left(x_j - \bar{x}\right)}{\sum_{i}^{n}(x_i - \bar{x})^2} \quad (4.2)$$

em que: n é o número de observações na área em estudo; x_i, x_j representam os valores observados dos atributos de solo a serem interpolados nos pontos i, j; $\bar{x}$ a média de x; w_{ij} são os elementos da matriz de pesos espaciais com valor 0 na diagonal ($w_{ii} = 0$).

4.5 Resultados e discussão

Na Figura 4.6 o RMSE é apresentado para cada atributo de solo nas seis áreas de estudo. Cada eixo do gráfico (vértice do polígono) representa uma área de estudo. Para cada polígono tem-se o valor do RMSE encontrado para cada área de estudo. Como são cinco polígonos para cada atributo, é possível visualizar o valor do RMSE do atributo nas seis áreas de estudo. Quanto mais próximo da borda do gráfico se encontra o vértice do polígono, maior é o RMSE, consequentemente pior é o método para o atributo na área em estudo.

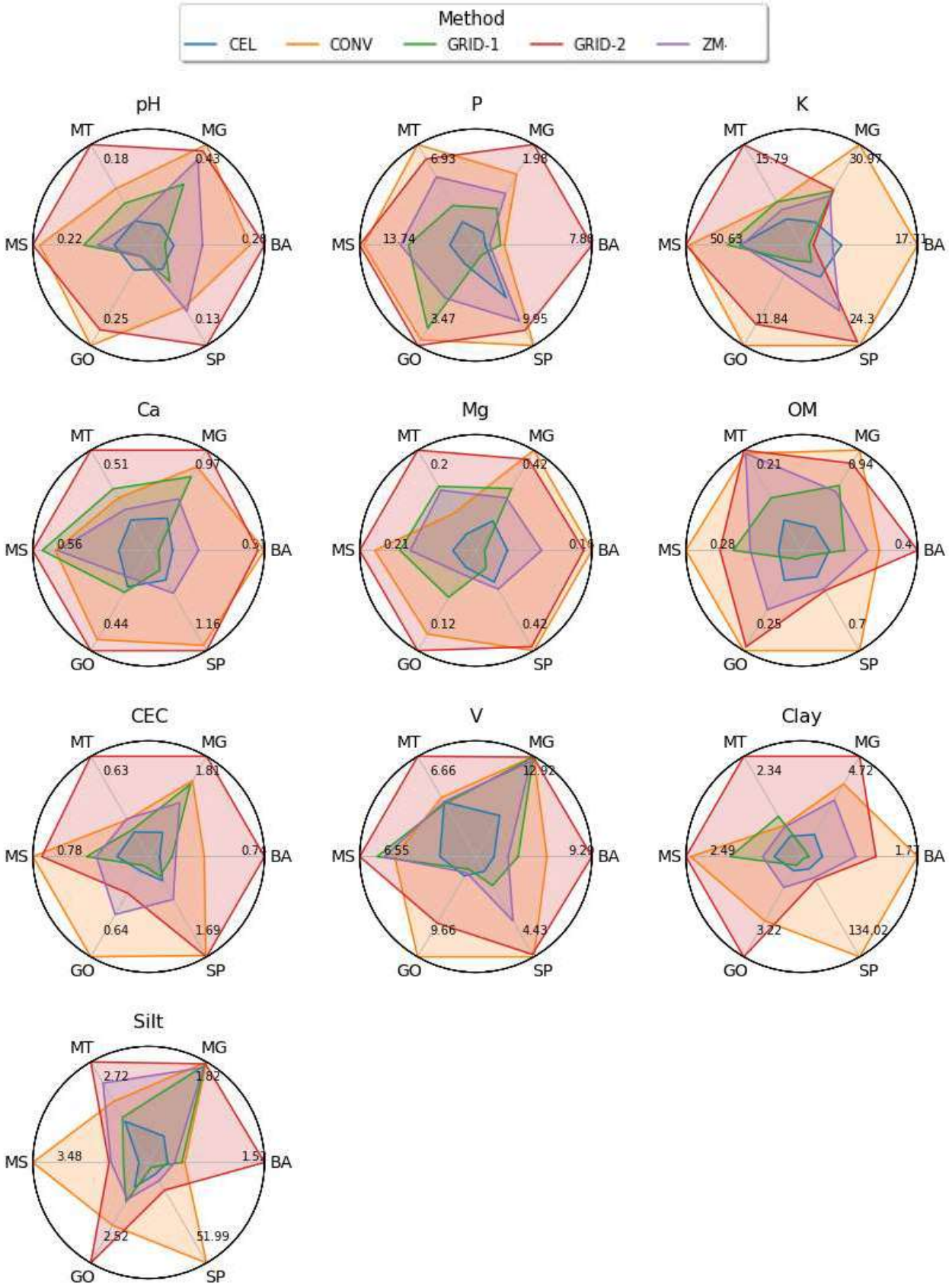


Figura 4.6: RMSE dos métodos de amostragem CEL, CONV, GRID-1, GRID-2, ZM, para os atributos de solo pH, P, K^{+}, Ca^{2+} MG^{2+}, OM, CEC, V, Clay e Silt nas seis áreas em estudo.

Nas Tabelas 4.3 e 4.4 são apresentados a correlação espacial medida por meio do Índice de Moran para as seis áreas de estudo em duas densidades amostrais. Analisando a Tabela 4.3, os atributos que possuem Índice de Moran superior a 0,70 e significativos, o método de amostragem GRID-1 tende a ser o melhor método de amostragem. A Tabela 4.4 está relacionada ao método GRID-2. Como o método GRID-2 utilizou uma baixa densidade de pontos, como pode ser visualizado na Figura 4.5, poucos atributos apresentaram uma boa performance, mesmo com valores altos e significativos para o Índice de Moran. Apenas os atributos K^+ na BA, CEC em GO e Clay em SP, apresentaram desempenho similar com as demais técnicas de amostragem. Para esses atributos o Índice de Moran foi superior a 0,90 e significativos ao nível de 1% ($p \leq 0,001$).

Tabela 4.3: Índice de Moran e p-value para os dez atributos de solos analisados nas seis áreas de estudo para o grid de maior densidade de pontos amostrados (Figura 4.1).

Area	Atributo**	pH (1)	P (2)	K^+ (3)	Ca^{2+} (4)	Mg^{2+} (5)	OM (6)	CEC (7)	V(%) (8)	Clay (9)	Silt (10)
BA	Moran	0,71	0,58	0,84	0,80	0,77	0,57	0,64	0,67	0,84	0,53
	p-vaue	0,001*	0,002*	0,001*	0,001*	0,001*	0,003*	0,001*	0,001*	0,001*	0,108
MG	Moran	0,49	0,56	0,61	0,48	0,52	0,62	0,58	0,45	0,77	0,57
	p-vaue	0,009*	0,001*	0,001*	0,015*	0,001*	0,001*	0,001*	0,066	0,001*	0,001*
MT	Moran	0,54	0,59	0,54	0,51	0,48	0,59	0,57	0,52	0,52	0,56
	p-vaue	0,078	0,003*	0,132	0,334	0,419	0,002*	0,016*	0,206	0,227	0,031*
GO	Moran	0,74	0,63	0,72	0,62	0,64	0,77	0,80	0,73	0,70	0,56
	p-vaue	0,001*	0,001*	0,001*	0,001*	0,001*	0,001*	0,001*	0,001*	0,001*	0,004*
MS	Moran	0,39	0,52	0,53	0,50	0,55	0,47	0,54	0,46	0,56	0,69
	p-vaue	0,152	0,004*	0,003*	0,009*	0,001*	0,048*	0,001*	0,131	0,001*	0,001*
SP	Moran	0,64	0,75	0,76	0,72	0,83	0,87	0,70	0,65	0,95	0,89
	p-vaue	0,005*	0,001*	0,001*	0,001*	0,001*	0,001*	0,001*	0,002*	0,001*	0,001*

1/ pH, Acidez Ativa em Água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CEC, Capacidade de Troca de Cátions a pH 7; 8/ V, Saturação por Bases; 9/ Clay, Argila; 10/ Silt, Silte.

*Significância ao nível de 5% de probabilidade.

**P, K^+ em mg/dm^3, e Ca^{2+}, Mg^{2+}, CEC em $cmolc/dm^3$, e OM em g/dm^3, e Clay, Silt em g/kg.

Tabela 4.4: Índice de Moran e p-value para os dez atributos de solos analisados nas seis áreas de estudo para o grid de menor densidade de pontos amostrados (Figura 4.5).

Area	Atributo**	pH (1)	P (2)	K^+ (3)	Ca^{2+} (4)	Mg^{2+} (5)	OM (6)	CEC (7)	V(%) (8)	Clay (9)	Silt (10)
BA	Moran	0,57	0,60	0,92	0,47	0,48	0,50	0,58	0,58	0,78	0,51
	p-vaue	0,080	0,057	0,001*	0,001*	0,003*	0,476	0,076	0,073	0,001*	0,487
MG	Moran	0,74	0,51	0,76	0,73	0,74	0,68	0,69	0,75	0,76	0,74
	p-vaue	0,042*	0,103	0,041*	0,076	0,039*	0,201	0,165	0,032*	0,028*	0,065
MT	Moran	0,69	0,73	0,67	0,68	0,70	0,61	0,67	0,68	0,49	0,72
	p-vaue	0,141	0,040*	0,259	0,187	0,095	0,465	0,221	0,206	0,033*	0,031
GO	Moran	0,89	0,57	0,53	0,86	0,87	0,89	0,93	0,89	0,88	0,89
	p-vaue	0,030*	0,346	0,194	0,312	0,179	0,039*	0,002*	0,007*	0,147	0,020*
MS	Moran	0,58	0,58	0,42	0,63	0,71	0,49	0,68	0,66	0,57	0,89
	p-vaue	0,269	0,255	0,277	0,059	0,002*	0,188	0,008*	0,026*	0,001*	0,001*
SP	Moran	0,50	0,71	0,62	0,62	0,66	0,85	0,60	0,54	0,91	0,83
	p-vaue	0,28	0,011*	0,036*	0,033*	0,006*	0,001*	0,061	0,401	0,001*	0,001*

1/ pH, Acidez Ativa em Água; 2/ P, Fósforo; 3/ K^+, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CEC, Capacidade de Troca de Cátions a pH 7; 8/ V, Saturação por Bases; 9/ Clay, Argila; 10/ Silt, Silte.

*Significância ao nível de 5% de probabilidade.

**P, K^+ em mg/dm^3, e Ca^{2+}, Mg^{2+}, CEC em $cmolc/dm^3$, e OM em g/dm^3, e Clay, Silt em g/kg.

Para melhor sintetizar os resultados obtidos, na Tabela 4.5, é apresentado a área dos polígonos para cada método de amostragem e atributo nas seis áreas de estudo. Na Tabela 4.5, quanto menor a área, menor é o RMSE, melhor o desempenho do método. O método células (CEL) apresentou o menor erro para os atributos que tiveram Índice de Moran inferior a 0,70. Para atributos que apresentaram Índice de Moran superior a 0,70, e significativo ao nível de 5% de probabilidade, o método de amostragem GRID-1 apresentou menores erros nas estimativas. Pelos resultados apresentados, verifica-se que a amostragem em grid, seguida pela interpolação, é o método que deve ser aplicado quando existir uma forte dependência espacial. Em situações, no qual o método de amostragem é realizado com baixa densidade amostral, o método de amostragem em células apresenta melhor desempenho.

Tabela 4.5: Área calculada para os cinco métodos de amostragem para os dez atributos de solos analisados nas seis áreas de estudo.

Atributo	Unidade	CEL	CONV	GRID-1	GRID-2	ZM
pH (1)		0,118	0,161	0,125	0,168	0,134
P (2)	mg dm^{-3}	89,069	121,934	95,360	133,314	105,247
K^{+} (3)	mg dm^{-3}	1040,468	1436,965	1046,725	1240,781	1110,594
Ca^{2+} (4)	cmolc dm^{-3}	0,757	0,980	0,798	1,026	0,820
Mg^{2+} (5)	cmolc dm^{-3}	0,107	0,152	0,115	0,160	0,125
OM (6)	g dm^{-3}	0,363	0,542	0,369	0,482	0,429
CEC (7)	cmolc dm^{-3}	7,557	11,577	7,694	11,655	9,007
V (8) (%)		137,816	173,540	150,401	179,812	151,650
Clay(9)	g kg^{-1}	102,221	348,146	81,122	176,855	133,598
Silt(10)	g kg^{-1}	50,785	106,389	54,962	75,962	61,379

1/ pH, Acidez Ativa em Água; 2/ P, Fósforo; 3/ K^{+}, Potássio; 4/ Ca^{2+}, Cálcio; 5/ Mg^{2+}, Magnésio; 6/ OM, Matéria Orgânica; 7/ CEC, Capacidade de Troca de Cátions a pH 7; 8/ V, Saturação por Bases; 9/ Clay, Argila; 10/ Silt, Silte.

De acordo com os resultados apresentados na Tabela 4.5, o método GRID-1 apresentou-se como o segundo melhor método, seu desempenho foi inferior ao método de amostragem por célula (CEL). Entretanto, o método GRID-1 necessita de uma quantidade suficiente de amostras para permitir um bom ajuste de um modelo teórico de semivariância (GIACOMIN et al., 2014; POULADI et al., 2019; WEBSTER; OLIVER, 1992), e estas amostras devem ter uma boa correlação espacial para produzir bons resultados.

Pela análise da Tabela 4.5, pode-se verificar que o método de amostragem por zonas de manejo (ZM) apresentou o terceiro melhor resultado, mesmo utilizando o menor número de amostras compostas. O método de zonas de manejo utilizou entre 6 a 10 amostras compostas, número bem menor que o utilizado pelo método de amostragem por células (que necessitou de 16 a 31 amostras compostas), como pode ser visualizado na Figura 4.3. No entanto, para definir as zonas de manejo foi utilizado uma amostragem densa de condutividade elétrica aparente do solo. A amostragem por zonas de manejo é similar à amostragem realizada por células, no entanto, as zonas de manejo são definidas com base em algum critério. Nesse trabalho foi utilizada apenas a condutividade elétrica aparente do solo, no entanto, recomenda-se utilizar também outros atributos para definir as zonas de manejo (CHEN et al., 2019; MORAL et al., 2019; SCHENATTO et al., 2017; VALLENTIN et al., 2020).

O método de amostragem GRID-2, que utilizou uma amostragem com baixa densidade e interpolação por OK, gerou os piores resultados (Tabela 4.5). Conforme apresentado na Figura 4.6, pode-se verificar que os resultados foram piores que o método convencional (CONV) para os atributos pH, P, Ca^{2+}, Mg^{2+}, CEC, V, ou seja, considerar a média dos atributos na área. Percebe-se que a área ocupada (Figura 4.6), em boa parte dos atributos, a área do método GRID-2 foi maior. Isso provavelmente ocorreu devido a redução da dependência espacial entre as amostras (Tabela 4.4). Além disso, a reduzida quantidade de amostras pode prejudicar a modelagem do semivariograma para realizar a Krigagem (GIACOMIN et al., 2014; POULADI et al., 2019). Apesar da redução na quantidade de amostras, o método GRID-2 utilizou maior número de pontos que o método de zonas de manejo e praticamente o mesmo número de pontos que o método de células. De acordo com Tabela 4.4, é possível perceber que somente para atributos com dependência espacial superior a 0,90 e significativa, apresentaram bons resultados nas estimativas.

A Tabela 4.6 apresenta os valores médios, o desvio padrão e a autocorrelação espacial do Índice de Moran medida para a ECa do solo em cada área de estudo, para as diferentes profundidades onde a ECa foi obtida. Na Tabela 4.7 é apresentado a comparação do erro (RMSE) do método por zonas de manejo em relação ao erro gerado para as 1000 células aleatórias. O valor corresponde ao percentual em que o método de zonas de manejo apresentou erro superior ao método de células aleatórias. Dessa forma, quanto menor o valor percentual na Tabela 4.7, melhor o resultado obtido para o método da ZM. Veja que para a área de Goiás (GO), o erro para a maioria dos atributos foi inferior ao erro gerado pelas células aleatórias. Isso indica que as zonas de manejo definidas pela condutividade elétrica aparente do solo apresentou boa eficiência na segmentação da área. Por outro lado, na área do Mato Grosso do Sul (MS) o desempenho do método foi pior que o método de células aleatórias.

Pela análise da Tabela 4.6, a área MS obteve baixa correlação medida através do Índice de Moran e não significativo ($p > 0,05$). Outro fator que pode influenciar de forma negativa é a alta variabilidade dos atributos dentro das ZM's definidas. Nas demais áreas, o resultado foi favorável à zonas de manejo para alguns atributos. Por isso, é importante destacar que a utilização somente da condutividade elétrica aparente do solo poderá não ser uma boa estratégia para geração de zonas de

manejo. Nesses casos, será melhor definir zonas de manejo acrescentando outras informações, como, histórico de uso da área, mapas de atributos que apresentem baixa variabilidade temporal e mapas de produtividade (MINASNY; MCBRATNEY, 2007b; PANTAZI et al., 2015; VALLENTIN et al., 2020).

Tabela 4.6: Dados da condutividade elétrica aparente do solo obtida em diferentes profundidades nas seis áreas de estudo.

Area	Profundidade[1]	Nr. Amostras	Média[2]	SD[3]	Moran	p-value
BA	0,50	489.722	1,61	0,54	0,823	0,001*
	100	489.722	0,80	0,31	0,727	0,001*
	200	489.722	0,61	0,20	0,830	0,001*
MG	0,20	783	1,71	0,85	0,416	0,001*
	0,40	783	1,15	0,49	0,514	0,001*
MT	0,30	18.844	5,08	0,72	0,737	0,001*
GO	0,20	30	5,95	1,22	0,663	0,001*
MS	0,20	181	6,64	2,34	0,478	0,051
	0,40	181	5,55	1,73	0,510	0,060
SP	0,375	22.807	93,72	64,44	0,981	0,001*
	0,750	22.807	1,73	16,72	0,984	0,001*

[1/] Profundidade (em centímetros) onde foram realizadas as medições da ECa; [2/] Valor Médio da ECa obtida em diferentes profundidades para cada área de estudo, medida em mS m^{-1}; [3/] SD, Desvio Padrão.

*Significância ao nível de 5% de probabilidade.

Tabela 4.7: Valor percentual em que o RMSE do método zonas de manejo foi maior que o RMSE gerado pelas 1000 células aleatórias.

Atributo	Unidade	BA	MG	MT	GO	MS	SP
pH [(1)]		25,0%	63,9%	12,4%	6,0%	98,8%	82,7%
P [(2)]	mg dm^{-3}	7,4%	62,8%	31,6%	6,7%	92,4%	51,0%
K^+ [(3)]	mg dm^{-3}	10,5%	53,2%	8,2%	4,8%	6,0%	48,3%
Ca^{2+} [(4)]	cmolc dm^{-3}	17,5%	37,2%	3,1%	0%	86,8%	5,1%
Mg^{2+} [(5)]	cmolc dm^{-3}	79,5%	27,4%	13,1%	0,2%	90,8%	3,9%
OM [(6)]	g dm^{-3}	100%	15,2%	68,6%	5,2%	99,7%	11,9%
CEC [(7)]	cmolc dm^{-3}	17,5%	34,3%	65,0%	37,4%	36,0%	7,0%
V [(8)] (%)		23,7%	79,0%	11,5%	0,1%	98,7%	31,5%
Clay[(9)]	g kg^{-1}	99,1%	55,1%	24,0%	13,0%	33,3%	0%
Silt[(10)]	g kg^{-1}	45,5%	78,8%	91,9%	68%	38,6%	0,5%

[1/] pH, Active Acidity in water; [2/] P, Phosphorus; [3/] K^+, Potassium; [4/] Ca^{2+}, Calcium; [5/] Mg^{2+}, Magnesium; [6/] OM, Organic Matter; [7/] CEC, Cation Exchangeable Capacity at PH 7; [8/] V, Basis Saturations; [9/] Clay, Argila; [10/] Silt, Silte.

Na Figura 4.7, é apresentado os histogramas para os métodos de amostragem utilizados. Os erros (RMSE) do método de células aleatórias (CEL-RND) para cada

área e cada um dos atributos foram padronizados em média zero e desvio padrão um. Os erros dos demais métodos foram padronizados com base no método de células aleatórias. Dessa forma, pode-se visualizar a distribuição dos erros de cada método em relação ao método de células aleatórias.

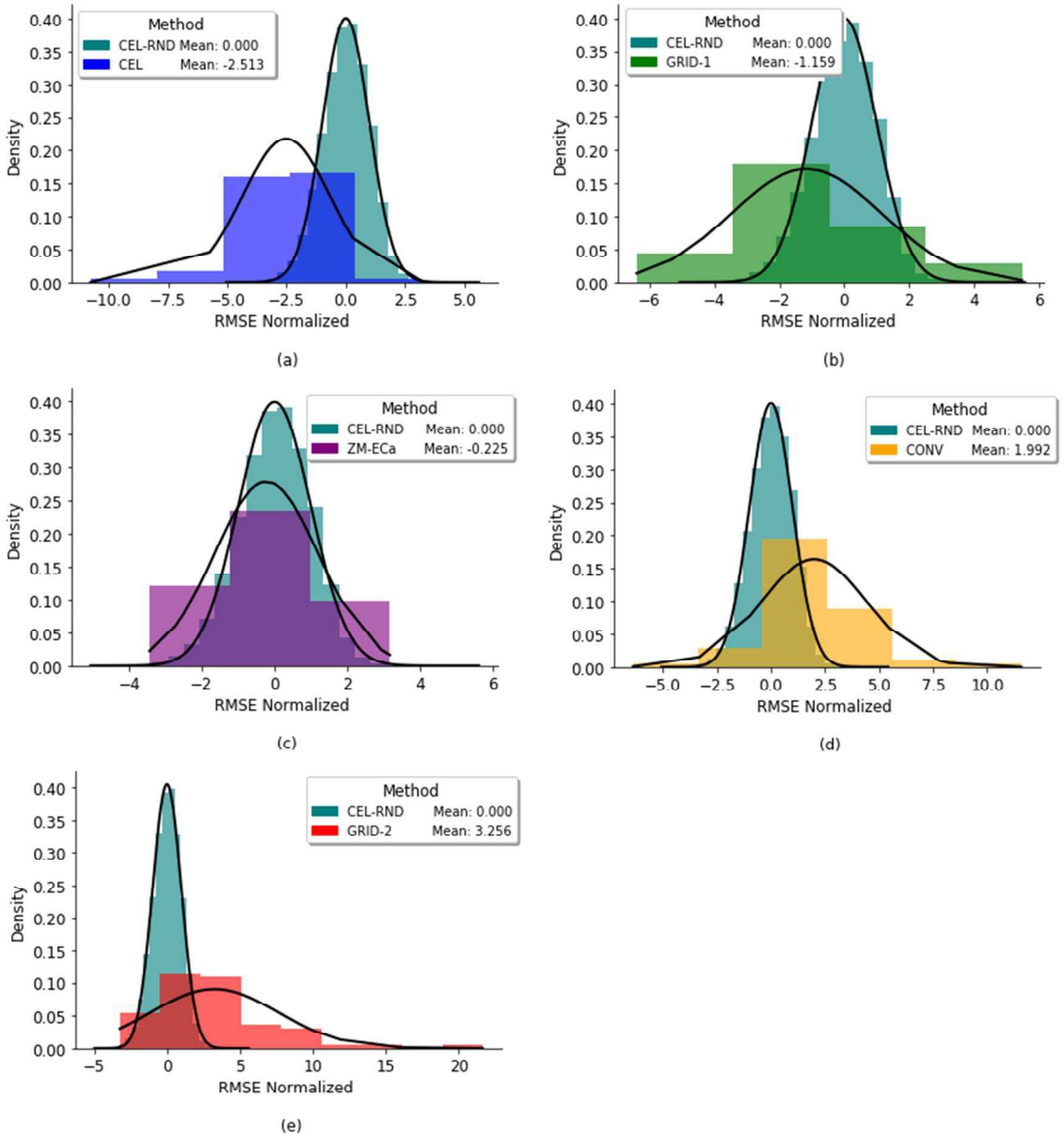


Figura 4.7: Histograma gerado para as 1000 combinações do método CEL-RND. (a) Comparação entre os métodos de amostragem CEL-RND e CEL; (b) Comparação entre os métodos de amostragem CEL-RND e GRID-1; (c) Comparação entre os métodos de amostragem CEL-RND e ZM; (d) Comparação entre os métodos de amostragem CEL-RND e CONV; (e) Comparação entre os métodos de amostragem CEL-RND e GRID-2.

Observa-se na Figura 4.7(a) que o método de células (CEL) apresenta erro médio inferior ao erro médio obtido pelo método de células aleatórias (CEL-RND). Para os métodos GRID-1 e ZM, os valores de erro médio também tenderam a ser

menor que o erro médio obtido quando se utilizou as células aleatórias. Situação diferente foi observado para os métodos CONV e GRID-2, em que esses métodos se mostraram inferiores ao método das células aleatórias.

No Brasil, boa parte das amostragens de solo realizada para prescrição das dosagens de corretivos e fertilizantes serem aplicadas a taxas variadas são realizados com amostragem em grid com 3 a 5 amostra por hectare. Dessa forma, os resultados apresentados na Figura 4.7 indicam que para essa densidade amostral, recomenda-se que se realize uma amostra composta por célula ou dirigida por zonas de manejo.

4.6 Conclusões

Neste estudo de caso foram comparadas seis técnicas de amostragem de solo. A amostragem por célula (CEL) mostrou-se como a técnica mais eficiente, uma vez que os pontos georreferenciados dentro de cada célula tendem a apresentar uma baixa variação entre seus valores. Para a amostragem em grid utilizou-se duas densidades de pontos, uma de maior densidade, denominada de GRID-1 e outra de menor densidade denominada de GRID-2. O método GRID-1 obteve um melhor desempenho quando comparado com os demais métodos, exceto com o método CEL. Entretanto o bom desempenho do GRID-1 está diretamente relacionado ao número de pontos de coleta de amostras simples em cada célula e a correlação espacial desses pontos determinada por meio do Índice de Moran univariado. Com valores de Índices de Moran superior a 0,70 e significativos ($p \leq 0,05$), o método GRID-1 tende a ser a melhor técnica de amostragem.

As técnicas de amostragem convencional (CONV) e amostragem em grid com baixa densidade de pontos (GRID-2) apresentaram os maiores erros (RMSE). O baixo desempenho da técnica GRID-2 pode estar relacionada à baixa densidade de pontos amostrados nas áreas de estudo e consequentemente à baixa correlação espacial entre estes pontos, ou altas correlações espaciais, porém não significativas. Embora tenha apresentado um desempenho inferior à amostragem por CEL e GRID-1, a técnica de amostragem por zonas de manejo definidas por meio da condutividade elétrica aparente do solo (ZM) apresentou-se como uma alternativa, uma vez que dados podem ser coletados de forma mais adensada na área, por meio do uso de sensores, sendo, portanto, um processo de aquisição de dados de baixo custo.

4.7 Referências

ADAMCHUK, V. I. et al. Evaluation of an on-the-go technology for soil pH mapping. **Precision Agriculture**, v. 8, n. 3, p. 139–149, 2007.

ALBORNOZ, E. M. et al. Development and evaluation of an automatic software for management zone delineation. **Precision Agriculture**, v. 19, n. 3, p. 463–476, 2018.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: the fuzzy c-means clustering algorithm. **Comput. Geosci.**, v. 10, p. 191–203, 1984.

BORGES, L. F. et al. Web software to create thematic maps for precision agriculture. Pesquisa Agropecuaria Brasileira, v. 55, 2020.

BOTTEGA, E. L. et al. Spatial variability of soil attributes in no a no-tillage system with crop rotation in the brazilian savannah. **Revista Ciência Agronômica**, v. 44, n. 1, p. 1–9, 2013.

BOTTEGA, E. L. et al. Precision agriculture applied to soybean: Part I - Delineation of management zones. **Australian Journal of Crop Science**, v. 11, n. 5, p. 573–579, 2017.

CHEN, S. et al. Delineation of management zones and optimization of irrigation scheduling to improve irrigation water productivity and revenue in a farmland of Northwest China. **Precision Agriculture**, v. 21, n. 3, p. 655–677, 2019.

COELHO, A. L. F. et al. An open-source spatial analysis system for embedded systems. **Computers and Electronics in Agriculture**, v. 154, n. September, p. 289–295, 2018.

COELHO, A. L. F. et al. Development of a variable-rate controller for a low-cost precision planter. **Applied Engineering in Agriculture**, v. 36, n. 2, p. 233–243, 2020.

CORWIN, D. L.; LESCH, S. M. Application of Soil Electrical Conductivity to Precision Agriculture. **Agronomy Journal**, v. 95, n. 3, p. 455, 2003.

CORWIN, D. L.; SCUDIERO, E. Field-scale apparent soil electrical conductivity. **Soil Science Society of America Journal**, v. 84, n. 5, p. 1405–1441, 2020.

COSTA, M. M. et al. Moisture content effect in the relationship between apparent electrical conductivity and soil attributes. **Acta Scientiarum - Agronomy**, v. 36, n. 4, p. 395–401, 2014.

DHAKAL, A. S.; AMADA, T.; ANIYA, M. Landslide hazard mapping and its evaluation using GIS: An investigation of sampling schemes for a grid-cell based quantitative method. **Photogrammetric Engineering and Remote Sensing**, v. 66, n. 8, p. 981–989, 2000.

FARID, H. U. et al. Delineating site-specific management zones for precision agriculture. **Journal of Agricultural Science**, v. 154, n. 2, p. 273–286, 2016.

FERGUSON, R. B.; HERGERT, G. W. Soil sampling for precision agriculture. **Precision Agriculture**, p. 1–4, 2009.

FORTUNE, S. A sweepline algorithm for Voronoi diagrams. **Algorithmica**, v. 2, n. 1, p. 153–174, 1987.

GAVIOLI, A. et al. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture**, v. 127, p. 302–310, 2016.

GIACOMIN, G. et al. Comparative Analysis of Interpolation Methods for Surface Models. **Revista Brasileira de Cartografia**, v. 66, n. January 2014, p. 1315–1329, 2014.

GUASTAFERRO, F. et al. A comparison of different algorithms for the delineation of management zones. **Precision Agriculture**, v. 11, n. 6, p. 600–620, 2010.

GUO-SHUN, L. et al. Comparison of kriging interpolation precision with different soil sampling intervals for precision agriculture. **Soil Science**, v. 175, n. 8, p. 405–415, 2010.

HOFMAN, S. C. K.; BRUS, D. J. How many sampling points are needed to estimate the mean nitrate-N content of agricultural fields? A geostatistical simulation approach with uncertain variograms. **Geoderma**, v. 385, 2021.

KHALEDIAN, Y.; MILLER, B. A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401–418, 2020.

LAWRENCE, P. G. et al. Guiding soil sampling strategies using classical and spatial statistics: A review. **Agronomy Journal**, v. 112, n. 1, p. 493–510, 2020.

LEGENDRE, P.; FORTIN, M.-J. Spatial pattern and ecological analysis. **Vegetatio**, v. 80, n. December 2009, p. 107–138, 1989.

LI, Y. et al. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. **Computers and Electronics in Agriculture**, v. 56, n. 2, p. 174–186, 2007.

MALLARINO, A. Management Zones Soil Sampling: A Better Alternative to Grid and Soil Type Sampling? **Proceedings of the Integrated Crop Management Conference**, 2001.

MINASNY, B.; MCBRATNEY, A. B. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. **Geoderma**, v. 140, n. 4, p. 324–336, 2007b.

MOHARANA, P. C. 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**, v. 21, n. 2, p. 426–448, 2020.

MORAL, F. J. et al. Mapping Soil Properties And Delineating Management Zones Based On Electrical Conductivity In A Hedgerow Olive Grove. **American Society of Agricultural and Biological Engineers**, v. 62, n. 3, p. 749–760, 2019.

MORAL, F. J.; REBOLLO, F. J.; SERRANO, J. M. Delineating site-specific management zones on pasture soil using a probabilistic and objective model and geostatistical techniques. **Precision Agriculture**, v. 21, n. 3, p. 620–636, 2020.

NOGUEIRA MARTINS, R. et al. Site-specific Nutrient Management Zones in Soybean Field Using Multivariate Analysis: An Approach Based on Variable Rate Fertilization. **Communications in Soil Science and Plant Analysis**, v. 51, n. 5, p. 687–700, 2020.

OUABO, R. E.; SANGODOYIN, A. Y.; OGUNDIRAN, M. B. Assessment of ordinary kriging and inverse distance weighting methods for modeling chromium and

cadmium soil pollution in e-waste sites in Douala, Cameroon. **Journal of Health and Pollution**, v. 10, n. 26, 2020.

PACCIORETTI, P.; CÓRDOBA, M.; BALZARINI, M. FastMapping: Software to create field maps and identify management zones in precision agriculture. **Computers and Electronics in Agriculture**, v. 175, n. November 2019, p. 105556, 2020.

PANTAZI, X. E. et al. Data fusion of proximal soil sensing and remote crop sensing for the delineation of management zones in arable crop precision farming. **CEUR Workshop Proceedings**, v. 1498, p. 765–776, 2015.

PARREIRAS, T. C. et al. Using unmanned aerial vehicle and machine learning algorithm to monitor leaf nitrogen in coffee. **Coffe Science**, 2020.

POULADI, N. et al. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, October 2018, p. 85–92, 2019.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System. Open Source Geoespacial Found. Proj.**QGIS Development Team. 2018, 2018.

QUEIROZ, D. M. et al. Development and testing of a low-cost portable apparent soil electrical conductivity sensor using a beaglebone black. **Applied Engineering in Agriculture**, v. 36, n. 3, p. 341–355, 2020.

SANCHES, G. M. **Spatial and temporal variability of soil attributes and their relationship with crop yield, topographic parameters and apparent electrical conductivity (eca) in sugarcane fields**. 2018. 96p. Tese (Doutorado) - Universidade Estadual de Campinas. Campinas, 2018.

SANTOS, H. G. et al. Sistema Brasileiro de Classificação de Solos. **Embrapa Solos**. 5. ed. Brasilia, DF, 2018.

SCHENATTO, K. et al. Normalization of data for delineating management zones. **Computers and Electronics in Agriculture**, v. 143, n. February, p. 238–248, 2017.

SHADDAD, S. M. et al. Data fusion techniques for delineation of site-specific management zones in a field in UK. **Precision Agriculture**, v. 17, n. 2, p. 200–217, 2016.

ÜNAL, İ.; KABAŞ, Ö.; SÖZER, S. Real-time electrical resistivity measurement and mapping platform of the soils with an autonomous robot for precision farming applications. **Sensors (Switzerland)**, v. 20, n. 1, 2020.

VALENTE, D. S. M. et al. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. **Scientia Agricola**, v. 69, n. 3, p. 173–179, 2012.

VALLENTIN, C. et al. Delineation of management zones with spatial data fusion and belief theory. **Precision Agriculture**, v. 21, n. 4, p. 802–830, 2020.

WEBSTER, R.; OLIVER, M. A. **Geostatistics for Environmental Scientists**. John Wiley & Sons, 2 ed. Chichester, 2007.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of Soil Science**, v. 43, p. 177–192, 1992.

WOLLENHAUPT, N. C.; WOLKOWSKI, R. P. Grid Soil Sampling. **Better Crops**, v. 78, n. 4, p. 6–8, 1994.

5 CONCLUSÕES GERAIS

O presente trabalho resultou em um sistema de código aberto integrado ao software QGIS para auxiliar na interpolação de atributos físicos e químicos do solo através de técnicas como Krigagem Ordinária e Aprendizado de Máquina. Além da interpolação de atributos do solo o sistema permite a geração de Zonas de Manejo e integração com as layers do QGIS. O sistema está disponível para download no site GitHub (https://github.com/gustavowillam/SmartMapPlugin) e no repositório de plugins do QGIS (https://plugins.qgis.org/plugins/Smart_Map). Assim, o presente trabalho, através desse sistema torna a agricultura de precisão mais acessível aos usuários como produtores rurais, empresas de consultoria agrícola e instituições de ensino, que podem fazer o uso do sistema em suas disciplinas de agricultura de precisão.

No segundo capítulo, foram implementadas as técnicas de interpolação Krigagem Ordinária (*OK*) e Aprendizado de Máquina "*Support Vector Machine*" (*SVM*). No estudo de caso, foram interpolados dez atributos de solos em cinco diferentes densidades de grids amostrais em duas áreas de estudo. Para o método *SVM* foi utilizada como *features* para o modelo de ML apenas a própria variável a ser interpolada, obtida pelo interpolador Inverso da Distância Ponderada (IDW) dos pontos amostrados da vizinhança. Os resultados apontaram que o método OK foi superior ao método *SVM* nos grids de maior densidade de pontos amostrais e correlação espacial medida pelo Índice de Moran superior a 0,70 significativo ao nível de 5% de probabilidade. Logo os resultados confirmaram que a técnica de *Machine Learning* se apresenta como uma alternativa para a interpolação de atributos do solo tanto em áreas com alta e baixa densidade amostral e pouca correlação espacial entre os pontos amostrados.

No terceiro capítulo foi apresentado o *Smart-Map*. Um plugin QGIS "*open source*" desenvolvido em Python capaz de realizar a predição de atributos de solo através de técnicas de *Machine Learning* e Krigagem Ordinária. Lançado em janeiro/2021, *Smart-Map* contabiliza até a presente data mais de 5.000 downloads (https://plugins.qgis.org/plugins/Smart_Map). No estudo de caso implementado para demonstrar a viabilidade do plugin foram interpolados os mapas para os atributos P, K^+, Ca^{2+}, Mg^{2+} em três diferentes densidades de grids amostrais em uma área do cerrado brasileiro com área de 90 ha e 181 pontos coletados. Com a utilização de

covariáveis como *features* para o modelo de ML, o método *SVM* foi superior a *OK* para os quatro atributos de solo nas três densidades de grids amostrais, considerando o conjunto de treinamento. Para o conjunto de teste o *SVM* só não apresentou desempenho superior na interpolação do atributo P no grid de 38 pontos amostrais. Tal fato se deu devido a baixa correlação espacial entre o atributo P e a covariável utilizada pelo método *SVM*. Novas funcionalidades estão previstas como trabalhos futuros: implementação de uma tabela de recomendação de insumos a taxa variada; configuração pelo usuário para eliminação de outliers; reamostragem de pontos para uma malha densa; determinação de linhas de plantio de forma mais eficiente, dentre outras funcionalidades.

Por fim, no quarto capítulo foi implementa uma estratégia de amostragem de atributos do solo utilizando a ECa para definição de zonas de manejo através dos mapas obtidos por interpolação por ECa (ZM). Essa estratégia foi comparada com outras estratégias de amostragem: células (CEL), convencional (CONV), células geradas aleatoriamente (CEL-RND) e com o método de amostragem em grid medido em duas densidades diferentes (GRID-1 e GRID-2). Um estudo de caso foi implementado utilizando seis áreas de estudo para determinar a performance das estratégias de amostragem. O RMSE foi utilizado como métrica para medir a acuraria das estratégias implementadas. A estratégia de amostragem por célula mostrou-se mais eficiente obtendo o menor RMSE, seguido da amostragem em grid em uma malha mais densa (GRID-1), utilizando como método interpolador a Krigagem Ordinária. Entretanto para o método OK com amostragem mais densa deve-se considerar o custo para amostragem dos atributos de solo analisados. A ECa pode ser obtida através do uso de sensores, de forma mais adensada e baixo custo. Os erros produzidos pela amostragem ZM foram inferiores aos erros produzidos pela amostragem em grid em uma malha com menor densidade de pontos (GRID-2), e estratégia de amostragem convencional (CONV). Portanto a estratégia de amostragem por ZM apresentou-se como uma alternativa para definição de ZM's, reduzindo curtos e fornecendo uma variabilidade espacial da área.