SMART-MAP: Plugin QGIS para interpolação utilizando Krigagem Ordinária e Machine Learning

Gustavo Willam Pereira[1], Domingos Sárvio Magalhaes Valente[2], Daniel Marçal de Queiroz[2], André Luiz de Freitas Coelho[2]

[1]Instituto Federal de Educação Ciência e Tecnologia Sudeste de Minas Gerais (IFSUDESTEMG) – Muriaé, MG – Brasil

[2]Departamento de Engenharia Agrícola – Universidade Federal de Viçosa (UFV) – Viçosa, MG – Brasil

gustavo.willam@ifsudestemg.edu.br, {valente, queiroz, andre.coelho}@ufv.br

Abstract. *Machine Learning (ML) algorithms have been used as an alternative to conventional and geostatistical methods in the digital mapping of soil attributes. However, ML algorithms have many variants that can make their application difficult for end users. To fill this gap, the objective of this work was to develop the Smart-Map plugin, to be used in QGIS. The Ordinary Kriging (OK) geostatistical method and the ML Support Vector Machine (SVM) model were implemented in the plugin to generate interpolated maps based on machine learning and OK. Performance comparisons between OK and SVM were performed for sample grids with 112 sampled points.*

Resumo. *Algoritmos de Aprendizado de Máquina (ML) têm sido utilizados como alternativa aos métodos convencionais e geoestatísticos no mapeamento digital de atributos do solo. 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, para ser utilizado no QGIS. 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. Comparações de performance entre OK e* SVM *foram realizadas em um grid amostral com 112 pontos amostrados.*

1. 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 and Schillaci 2019].

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 [Campbell et al., 2020; Hengl et al., 2018; Sekulic et al., 2020]. Os algoritmos de ML procuram descobrir e quantificar padrões entre os dados disponíveis para fazer predições. 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.

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 et al., 2002], SGeMS [Remy et al., 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*.

2. Materiais e Métodos

2.1. Estudo de caso para avaliação do plugin Smart-Map

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).

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 com o objetivo de validar o sistema. Além da geração de mapas por interpolação, *Smart-Map* é dotado de um algoritmo para executar análise de agrupamento utilizando o método *fuzzy k-means* [Bezdek et al., 1984]. No final do processamento é exibido o mapa de Zonas de Manejo (ZM).

O estudo de caso foi conduzido em uma área de 119 ha, localizada no município de Tabaporã (10º48’27” S, 56º37’14” W), Mato Grosso, Brasil (Figura 1). Essa área é cultivada com soja, possui altitude média de 325 m, relevo plano e solo predominantemente classificado como Latossolo Vermelho-Amarelo Distrófico (LVAd) 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 um ponto por hectare, totalizando 112 amostras. Foram realizadas as análises laboratoriais para medir a concentração de macronutrientes (P, K^+, MO e ARG). Também foram coletados dados da condutividade elétrica aparente do solo (ECa) utilizando o aparelho Veris U3: plataforma de mapeamento de solo Veris U3® (Veris Technologies Inc., Salina KS USA). A estatística descritiva dos dados é apresentada na Tabela 1.

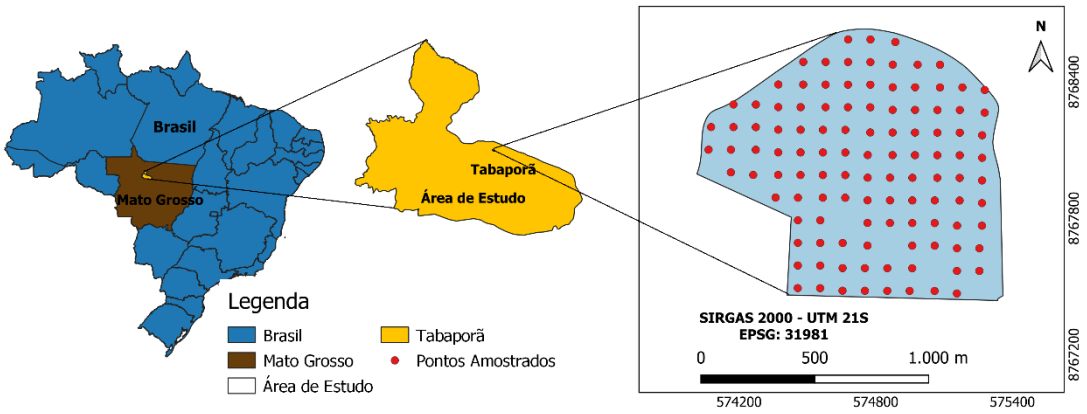

Figura 1. Localização geográfica da área de estudo e distribuição dos pontos amostrais em Tabaporã, Mato Grosso, Brasil.

Tabela 1. Estatística descritiva dos atributos do solo na área de estudo

Atributo	Unidade	Média	Desvio Padrão	Valor Mínimo	Valor Máximo	CV (%)
P [1]	(mg dm^{-3})	17,91	10,40	3,40	56,60	58,90
K^+ [2]	(mg dm^{-3})	85,43	35,87	25,00	199,00	42,54
MO [3]	(g dm^{-3})	2,45	0,77	0,80	4,79	0,31
ARG [4]	(g kg^{-1})	0,36	0,12	0,16	0,62	0,33
ECa-020 [5]	(mS m^{-1})	2,57	0,76	0,80	5,94	0,29

[1/] P, Fósforo; [2/] K^+, Potássio; [3/] MO, Matéria Orgânica; [4/] ARG, Argila; [5/] ECa-020, Condutividade Elétrica Aparente do Solo medida a 0,20 m.

2.2. Métodos de Interpolação

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 et al., 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^+, MO e ARG.

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 *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^+, MO e ARG.

Na segunda abordagem (*SVM2*) utilizou-se como *features* as coordenadas (*coordX* e *coordY*), o valor da própria variável interpolada utilizando IDW e a condutividade elétrica aparente do solo (ECa) medida a 0.2m. 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).

Com o objetivo de comparar a performance do método OK e do modelo de ML *SVM* (*SVM1* e *SVM2*) foram interpolados os atributos P, K^+, MO e ARG. 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. O R^2 e RMSE foram calculados conforme Equações 1 e 2, respectivamente, dos dados de teste para P, K^+, MO e ARG.

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

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

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.

Na Figura 2 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_ECa020* representa o valor estimado da ECa 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^+, MO e ARG. Cada linha 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.

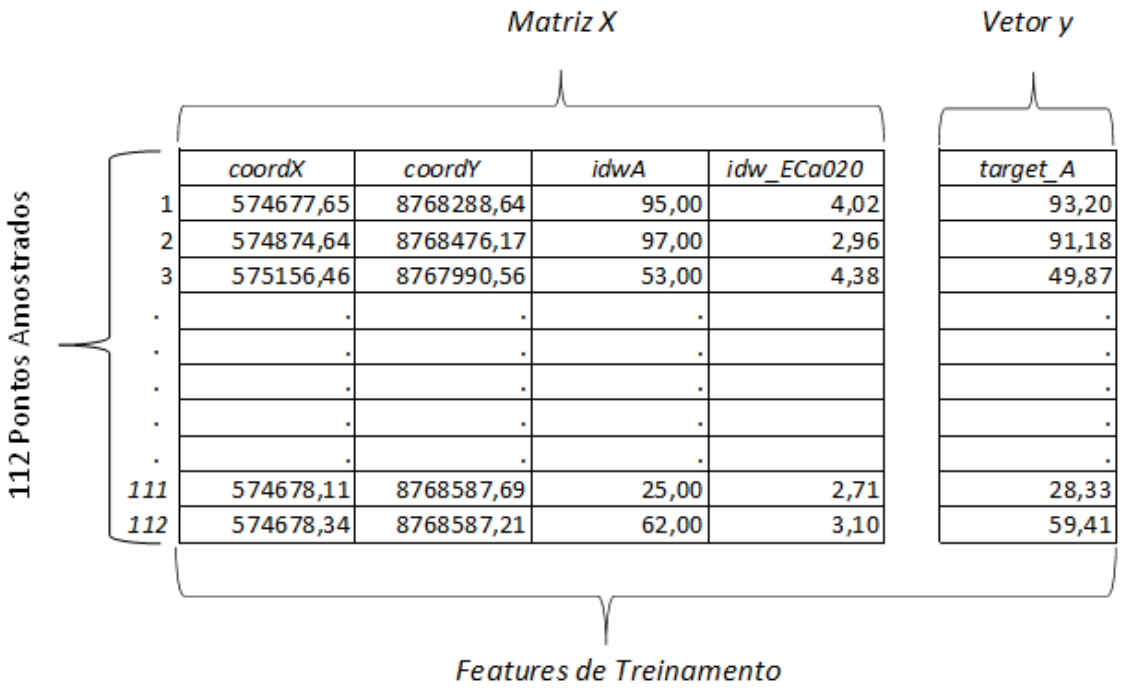

Figura 2. Definição do modelo de ML (*Features* de Treinamento) para os métodos *SVM1* e *SVM2*: *features* (Matriz *X*) e *target* (Vetor *y*).

3. Resultados e discussão

Os mapas interpolados gerados pelos métodos OK, *SVM1* e *SVM2* para os 4 atributos de solos na área de estudo são apresentados na Figura 3. As concentrações mais altas de P podem ser visualizadas na parte central, e de K^+, MO e ARG na região norte do mapa. As transições entre os limites foram mais suaves nos mapas de previsão produzidos pelos métodos OK e *SVM1*. A opção pelos atributos MO e ARG foi por serem componentes com baixa variabilidade temporal.

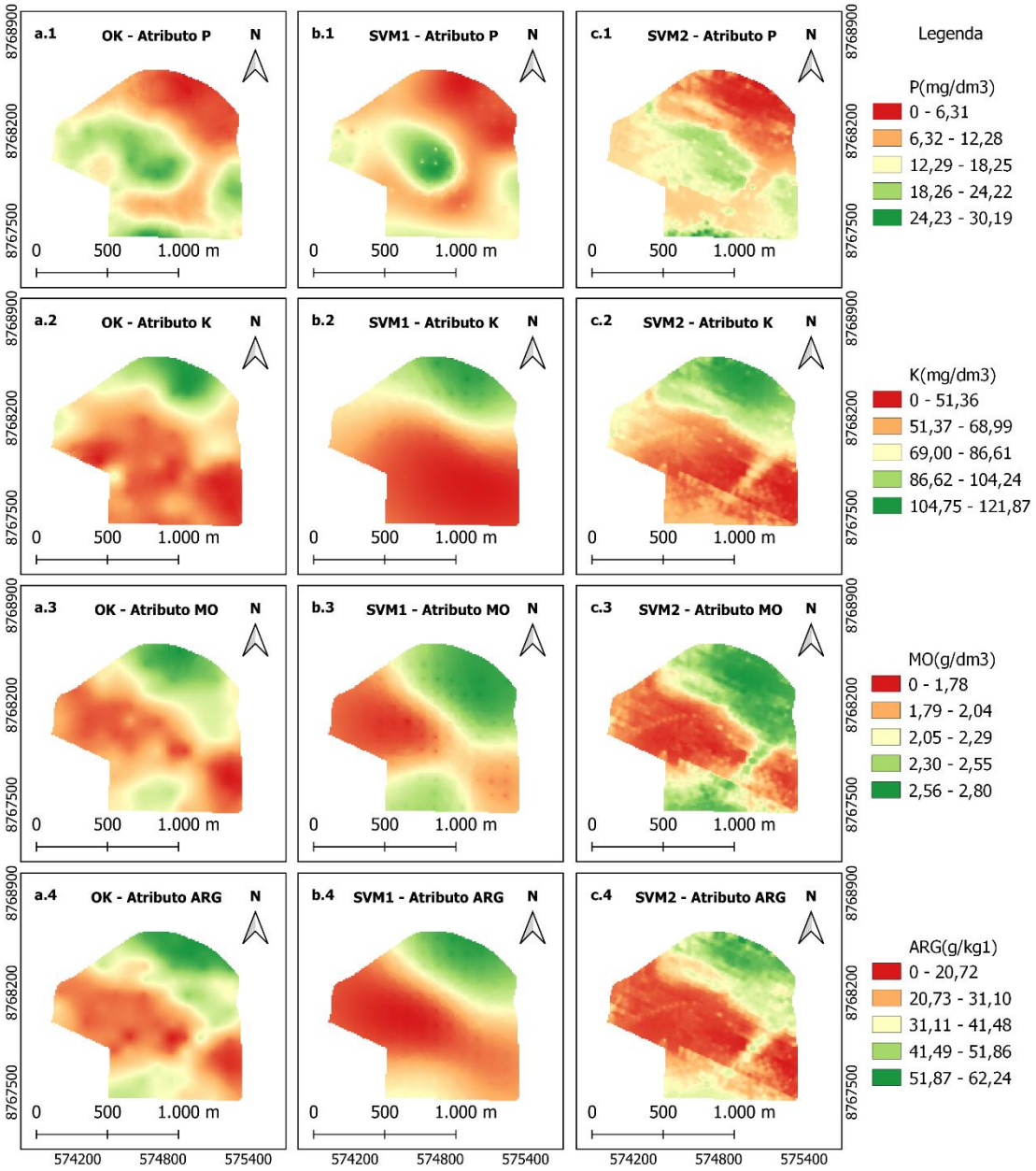

Figura 3. Mapas obtidos por interpolação de Fósforo (P): (a.1-c.1); Potássio (K^+): (a.2-c.2); Matéria Orgânica (MO): (a.3-c.3) e Argila (ARG): (a.4-c-4). Métodos de Interpolação: OK (a.1-a-4), *SVM1* (b.1-b.4), *SVM2* (c.1-c.4).

Os valores de R^2 e RMSE para os atributos de solo P, K^+, MO e ARG, para os métodos OK, *SVM1 e SVM2* são apresentados na Figura 4.

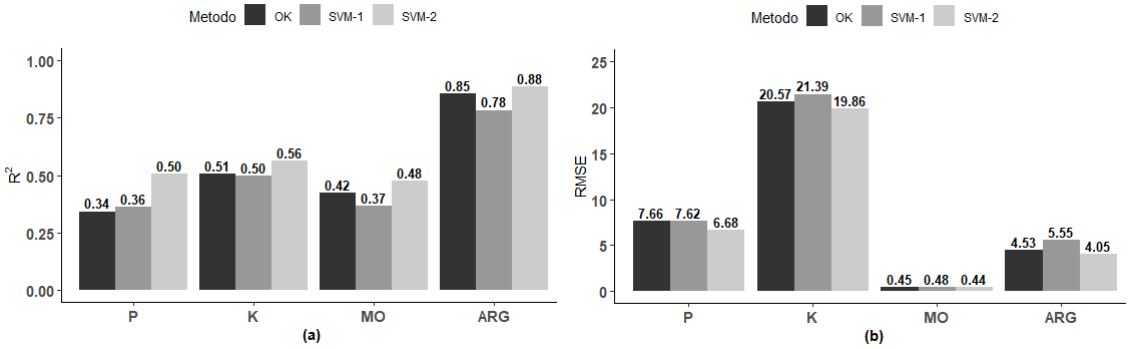

Figura 4. (a) Coeficiente de Determinação (R^2). (b) RMSE, calculados para os atributos P, K^+, MO e ARG.

O método *SVM1* foi superior ao método OK apenas para o atributo P. O método *SVM2* foi superior para os quatro atributos de solo analisados (P, K^+, MO e ARG), obtendo os maiores valores de R^2 e os menores valores de RMSE, quando comparado aos métodos OK e *SVM1*. A adição da ECa como covariável no método *SVM2* foi determinante para o bom desempenho do método em relação a *OK* e *SVM1*. Outras covariáveis poderiam ser adicionadas ao modelo de ML, como índices de vegetação obtidas a partir de análise de imagens, mapas de produtividade, dentre outras de fácil aquisição e de forma mais adensada, medidas através de sensores, produzindo custos menores para aquisição.

4. 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 que o método *SVM2* foi superior aos demais métodos na predição de atributos do solo. Em áreas quando estão disponíveis covariáveis com maior número de pontos e que apresentam um nível de correlação significativa com as variáveis a serem interpoladas técnicas de ML é uma alternativa ao método OK. 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 do solo em escala regional.

Como trabalhos futuros pretende-se implementar no plugin métodos de interpolação como Inverso da Distância Ponderada (IDW), Co-Krigagem e outros métodos que utilizam Machine Learning como Random Forest, métodos Ensemble, XGBoost, CatBoost, AdaBoost.

5. Referências Bibliográficas

Bezdek, J. C., Ehrlich, R. and Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. In Comput. Geosci., v. 10, pages 191–203.

Campbell, P. M. M., Francelino, M. R., Filho, E. I. F., Rocha, P.A. and Azevedo, B. C. (2019) Digital mapping of soil attributes using machine learning. Revista Ciência Agronômica, v. 50, n. 4, p. 519–528.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M. and Graler, B. (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, v. 6, p. e5518.

Malla, R., Shrestha, S., Khadka, D. and Bam, C. R. (2020). 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.

Muphy, B., Mullher, S., Yurchark, R. GeoStat-Framework/PyKrige v1.5.1 (Version v1.5.1).

Pedregosa, F., Varoquaux, G., Granfort, A., Michel, V. and Thirion, B. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, n. 1, p. 2825–2830.

QGIS Development Team. (2018). QGIS Geographic Information System. Open Source Geoespacial Found. Proj.QGIS Development Team.

Remy, N., Boucher, A. and Wu, J. (2009). Applied Geostatistics with SGeMS. Cambridge: Cambridge University Press, 2009.

Santos, H. G., Carvalho Júnior, W., Dart, R. O., Áglio, M. L. D., Sousa, J. S., Pares, J. G., Fontana, A., Martins, A. L. S. and Oliveira, A. P. O. (2018) Sistema Brasileiro de Classificação de Solos. 5. ed. Brasilia, DF.

Sekulic, A., Kilibarba, M., Heuvelink, G. B. M., Nikolic, M. and Bajat, B. (2020) Random forest spatial interpolation. Remote Sensing, v. 12, n. 10, p. 1–29.

Valente, D. S. M., Queiroz, D. M., Pinto, F. A. C., Santos, N. T. and Santos, F. L. (2012). Definition of management zones in coffee production fields based on apparent soil electrical conductivity. Scientia Agricola, v. 69, n. 3, p. 173–179.

Veronesi, F. and Schillaci, C. (2019). Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecological Indicators, v. 101, p. 1032–1044.

Whelan, B. M., Mcbratney, A. B., Minasny, B. (2002). VESPER 1.5 - Spatial prediction software for precision agriculture. 6th International Conference on Precision Agriculture, p. 1–14.