University of São Paulo
“Luiz de Queiróz” College of Agriculture

Machine learning for prediction of soil carbon stock changes in sugarcane crop due to straw removal

Ralf Vieira de Araujo

Dissertation presented to obtain the degree of Master in Science. Area: Soil and Plant Nutrition

Piracicaba
2021

Ralf Vieira de Araujo
Biologist

Machine learning for prediction of soil carbon stock changes in sugarcane crop due to straw removal

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **CARLOS EDUARDO PELLEGRINO CERRI**

Dissertation presented to obtain the degree of Master in Science. Area: Soil and Plant Nutrition

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Araujo, Ralf Vieira de

Machine learning for prediction of soil carbon stock changes in sugarcane crop due to straw removal / Ralf Vieira de Araujo. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

50 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Carbono orgânico do solo 2. Machine learning 3. Manejo de resíduos 4. Atributos do solo I. Título

In memorian Alpino Tomaz de Araujo

I DEDICATE

KNOWLEDGEMENTS

I would like to express my gratitude to:

- The Graduate Program in Soils and Plant Nutrition for the support and valuable classes;

- The Center for Nuclear Energy in Agriculture and especially Profs.º Plinio Camargo, José Albertino Bendassolli and Ernani Pinto Jr. for allowing me to start the course and attend the classes;

- My advisor Carlos Eduardo Pellegrino Cerri for the opportunity, guidance, friendship and most of all confidence in an unexpected master's student willing to start in an unexpected study field;

- Prof. André Carlos Ponce de Leon Carvalho for the valuable lessons and kind support on several occasions;

- To Kelly Cristina Ramos da Silva and Saulo Martiello Mastelini for helping me and sharing a little bit of your expertise;

- My parents Aparecida Souza Vieira and Alpino Tomás de Araujo for the unconditional support and faith in me before, now and probably ever;

- My dear wife, best friend and loved Heloíze de Souza Milano for encouraging me or even tell me to keep calm when I've realized that kids and dissertation don't occupy the same space at the same time;

- My kids, Carmen and Angelo for your love and enthusiasm… Dad loves you!

- To my laboratory colleagues Admilson, Dagmar, Lilian, Sandra and Zezinho for the support, friendship and weird meals early in the morning to make our working time a happy time.

CONTENTS

RESUMO

Aprendizado de máquina para predição de alterações nos estoques de carbono do solo em cultivo de cana-de-açúcar devido à remoção da palha

O Brasil estabeleceu políticas energéticas e climáticas que fomentam o uso de biocombustíveis como o etanol da cana-de-açúcar. Uma prática crescente é usar os resíduos da colheita, a palha de cana-de-açúcar, para cogeração de energia elétrica ou para produzir etanol de segunda geração. Neste estudo, objetivou-se prever mudanças de curto prazo nos estoques de carbono orgânico do solo de acordo com a massa de resíduos depositada na colheita utilizando técnicas de aprendizado de máquina (AM). Foram feitas também considerações sobre o atual estado da arte relativo à aplicação de AM nas ciências do solo, com ênfase em solos tropicais e estoques de carbono do solo. Os dados iniciais foram gerados entre 2015 e 2018 em cinco áreas de cultivo comercial de cana-de-açúcar na região centro-sul do Brasil e as variáveis disponíveis relacionam-se ao clima, atributos físicos e químicos do solo, matéria orgânica e variedade cultural. A variável predita (y) foi a taxa de variação do estoque de carbono por área por ano (Mg C ha^{-1} ano^{-1}) em relação à massa seca total da palha. O conjunto de dados inicial foi dividido em treino (80%) e teste (20%) e oito modelos baseados em algoritmos de AM foram desenvolvidos utilizando Random Forest (RF) e Support Vector Machine (SVM) associados a quatro métodos de seleção de atributos. Os resultados foram avaliados pela raiz do erro quadrático médio (RMSE) com validação cruzada no conjunto treino e RMSE da predição no conjunto de teste. Os modelos treinados foram comparados com a adoção de valores médios de y estratificados por massa de palha depositada e camada de solo e entre eles ($p < 0,05$). Todos os modelos AM superaram a generalização de valores médios de y previamente conhecidos. O modelo SVM aplicado ao conjunto de atributos selecionado por RF apresentou melhor desempenho com redução considerável no número de atributos, o que poderia reduzir os custos e esforço de aquisição e processamento de dados em aplicações futuras. Conclui-se que modelos de AM são boas ferramentas para prever mudanças de curto prazo nos estoques de carbono devido à remoção total ou parcial da palha do campo. Os resultados obtidos e metodologia aplicada tem potencial de auxiliar produtores e gestores a identificar ralações de causa-efeito entre as condições locais de cultivo, o manejo da palhada adotado e as variações esperadas no carbono orgânico do solo.

Palavras-chave: Carbono orgânico do solo, Machine learning, Manejo de resíduos, Atributos do solo

ABSTRACT

Machine learning for prediction of soil carbon stock changes in sugarcane crop due to straw removal

Brazil, as other countries, has established energy and climate policies that foster the use of biofuels as sugarcane ethanol, in which a growing practice is to use harvesting residues, the straw, for cogeneration of electricity or to produce second-generation ethanol. In this study, it was aimed to create machine learning (ML) models capable of predict short-term changes in the soil organic carbon stocks according to the mass of sugarcane straw leftover the soil during harvest. Considerations were also made on the current state of the art regarding the application of ML in soil science, with an emphasis on tropical soils and soil carbon stocks. The initial data was generated between 2015 and 2018 in five experimental sites under commercial cultivation of sugarcane in Brazilian south-central region and the available variables were related to climate, soil physical and chemical attributes, organic matter and crop variety. The variable to be predicted (*y*) was the rate of carbon stock change per area per year (Mg C ha^{-1} yr^{-1}) in relation to the total dry mass of straw. The initial dataset was divided into training (80%) and test (20%) and eight ML models were trained using the algorithms Random Forest (RF) and Support Vector Machine (SVM) associated to four feature selection methods. Results were evaluated using 10-fold cross-validation of the root mean squared error (RMSE) in the training set and prediction RMSE in the test set. The trained models were statistically compared among them and to the use of mean y stratified by straw mass deposited and soil layer. All the ML models surpassed the simple generalization of previously known mean values of y. The model SVM associated with RF feature selection performed better with a considerable reduction in the number of attributes, which could reduce the costs and effort of data acquisition and processing in future applications. The achievements indicate that ML models are good tools to predict short-term changes in carbon stocks due to total or partial straw remotion from the field. The obtained results and applied methodology have the potential to help producers and decision-makers interested in identifying cause-effect relationships between in situ crop conditions, straw management and expected soil carbon variations.

Keywords: Soil organic carbon, Machine learning, Waste management, Soil attributes

1. GENERAL INTRODUCTION

Many countries have established energy and climate policies that foster the biofuel use. In 2017, the Brazilian federal government instituted by Law No. 13.576 / 2017 the National Biofuels Policy, called RENOVABIO (GOVERNO FEDERAL, 2017), which has as its main instrument the establishment of annual decarbonization targets for the fuel sector through the increased production and participation of biofuels in the country's energy matrix. RENOVABIO establishes that participating producers will be audited by certifiers and receive grades inversely proportional to their proven GHG emission reductions. They shall receive carbon credits which can then be traded according to the obtained grades. Fuel distributors, in turn, must prove compliance with the decarbonization targets by acquiring carbon credits.

Among biofuels, ethanol receives special attention due to its great potential for emission reduction compared to the use of fossil fuels (LAMPE, 2008). Brazil stands out in the production of sugarcane (*Saccharum officinarum*) ethanol since the 1970s, and is currently the world's largest producer of sugarcane ethanol. For the 2019/20 period, Brazilian cropped area is about 8.481 million hectares with an estimated production of more than 642 million tons of sugarcane and almost 34 billion liters of ethanol (CONAB, 2019). It is expected that by the year 2030, to meet domestic and foreign market demands, Brazilian ethanol production should grow to the range of 40 to 50 billion liters (EPE, 2016; MINISTÉRIO DE MINAS E ENERGIA, 2018).

The major responsible for the expected GHG emissions mitigation due to the increasing participation of ethanol in the Brazilian energy matrix is the soils potential to store carbon. The soil corresponds to the largest terrestrial carbon reservoir, containing about 1500 Pg (Pg=1 Gton) of carbon. This amount, although much smaller than that present in the oceans (40,000 Pg) is about 2 times bigger than that present in the atmosphere (MCCARL; METTING; RICE, 2007). Plants remove carbon from the atmosphere via photosynthesis, and after the decomposition of their tissues, part of their carbon remains partly fixed, or as commonly referred to, "sequestrated" in the soil (LAL, 2004).

Several environmental, soil and coverage factors also strongly affect the potential for soil carbon sequestration, such as the planted crop (WEST; POST, 2002a), temperature (TRUMBORE; CHADWICK; AMUNDSON, 1996; FRØSETH; BLEKEN, 2015), soil granulometry (BAUHUS; PARÉ; CÔTÉ, 1998;

SOUCÉMARIANADIN et al., 2018), pH, access of organic matter to microorganisms (DUNGAIT et al., 2012), among others.

Sugarcane crop management practices also greatly influence the soil capacity to store organic carbon (CERRI et al., 2009; SILVA-OLAYA et al., 2017). One example is the adoption of mechanized harvesting instead of harvesting with pre-burning of straw (GALDOS; CERRI; CERRI, 2009; ZANI et al., 2018). In the first situation, the maintenance of the straw in the field promotes benefits to the soil: it provides nutrients; reduces thermal amplitude; improves aggregation; reduces soil (and consequently carbon) losses by attenuating the direct impact of wind and horizontal speed of water; favors infiltration and storage of water into the soil; decreases the density by better dispersing the tire pressure of the agricultural machinery (CHERUBIN et al., 2018a). On the other hand, the use of the straw for cogeneration of electricity or as a substrate for the production of second-generation ethanol can also be advantageous: It may help to mitigate GHG emissions due to energy expenditure reduction in the mills and ethanol production increment without expanding the cropped area (FRANCO et al., 2013; KARLEN; JOHNSON, 2014).

Thus, a great complexity of interactions dictates the soil carbon dynamics at different spatial and temporal scales (ETTEMA, 2002; AUSTIN et al., 2004). Models that help understand these processes and perform predictions of soil behavior have been formulated and applied for decades in natural and agricultural environments (NIKIFOROFF, 1937; STEVENSON; TANJI, 1982; PALOSUO et al., 2012).

Studies on soil carbon dynamics use a variety of approaches. Some build models based on mathematical equations. Others use simulation models, i.e., a combination of various mathematical models. A comparative review of the assumptions, complexity, and mathematical logic of various approaches identified 250 different soil organic matter models, with a new model emergence rate of approximately 6% $year^{-1}$, many of those being modifications from previous versions (MANZONI; PORPORATO, 2009).

In recent years data mining techniques have also been applied to the study of organic matter (LACOSTE et al., 2012; ABERA et al., 2021; LU et al., 2021). Data mining is especially useful when large databases are available, when interesting non-obvious patterns are intended to be discovered, or when relationships between variables are difficult to understand. Therefore, often the process of mining data is also referred to as knowledge discovery from data, exploratory statistics or statistical

learning (HAN; KAMBER; PEI, 2012; GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Algorithms specially developed to efficiently handle this great complexity of information are known as machine learning algorithms. Nevertheless, the mentioned techniques and tools have not been widely applied in soil science under tropical conditions, especially when disregarding geoprocessing-related research.

In this study it was evaluated the use of machine learning algorithms to predict annual changes in sugarcane soil carbon stocks after partial or total straw removal under tropical conditions. The hypotheses were:

I) Data mining techniques are suitable for identifying the most relevant attributes for understanding the changes in soil carbon stocks in a complex database.

II) Changes in soil carbon stocks as a function of the amount of straw left in the field after harvest in sugarcane crop can be predicted using machine learning tools.

It is expect that the obtained results and suggested approaches may serve as a start point to producers and decision-makers interested in correlate in situ crop conditions and straw management to expected soil carbon variations.

References

ABERA, W., TAMENE, L., ABEGAZ, A., HAILU, H., PIIKKI, K., SODERSTROM, M., GIRVETZ, E., SOMMER, R., 2021. ESTIMATING SPATIALLY DISTRIBUTED SOC SEQUESTRATION POTENTIALS OF SUSTAINABLE LAND MANAGEMENT PRACTICES IN ETHIOPIA. J. ENVIRON. MANAGE. 286. HTTPS://DOI.ORG/10.1016/J.JENVMAN.2021.112191

AUSTIN, A.T., YAHDJIAN, L., STARK, J.M., BELNAP, J., PORPORATO, A., NORTON, U., RAVETTA, D.A., SCHAEFFER, S.M., 2004. WATER PULSES AND BIOGEOCHEMICAL CYCLES IN ARID AND SEMIARID ECOSYSTEMS. OECOLOGIA 141, 221–235. HTTPS://DOI.ORG/10.1007/S00442-004-1519-1

BAUHUS, J., PARÉ, D., CÔTÉ, L., 1998. EFFECTS OF TREE SPECIES, STAND AGE AND SOIL TYPE ON SOIL MICROBIAL BIOMASS AND ITS ACTIVITY IN A SOUTHERN BOREAL FOREST. SOIL BIOL. BIOCHEM. HTTPS://DOI.ORG/10.1016/S0038-0717(97)00213-7

CERRI, C.C., MAIA, S.M.F., GALDOS, M.V., PELLEGRINO CERRI, C.E., FEIGL, B.J., BERNOUX, M., 2009. BRAZILIAN GREENHOUSE GAS EMISSIONS: THE IMPORTANCE OF AGRICULTURE AND LIVESTOCK. SCI. AGRIC. 66, 831–843. HTTPS://DOI.ORG/10.1590/S0103-90162009000600017

Cherubin, M.R., Oliveira, D.M.S., Feigl, B.J., Pimentel, L.G., Lisboa, I.P., Gmach, M.R., Varanda, L.L., Morais, M.C., Satiro, L.S., Popin, G.V., De Paiva, S.R., Dos Santos, A.K.B., De Vasconcelos, A.L.S., De Melo, P.L.A., Cerri, C.E.P., Cerri, C.C., 2018. Crop residue harvest for bioenergy production and its implications on soil functioning and plant growth: A review. Sci. Agric. 75, 255–272. https://doi.org/10.1590/1678-992x-2016-0459

Conab, 2019. ACOMPANHAMENTO DA SAFRA BRASILEIRA Cana-de- açúcar. Cia. Nac. Abastecimento. https://doi.org/10.1371/journal.pone.0175940

Dungait, J.A.J., Hopkins, D.W., Gregory, A.S., Whitmore, A.P., 2012. Soil organic matter turnover is governed by accessibility not recalcitrance. Glob. Chang. Biol. 18, 1781–1796. https://doi.org/10.1111/j.1365-2486.2012.02665.x

EPE, 2016. NOTA TÉCNICA DEA 13/15 Demanda de Energia 2050.

Ettema, C., 2002. Spatial soil ecology. Trends Ecol. Evol. 17, 177–183. https://doi.org/10.1016/S0169-5347(02)02496-5

Franco, H.C.J., Pimenta, M.T.B., Carvalho, J.L.N., Magalhães, P.S.G., Rossell, C.E.V., Braunbeck, O.A., Vitti, A.C., Kölln, O.T., Rossi Neto, J., 2013. Assessment of sugarcane trash for agronomic and energy purposes in Brazil. Sci. Agric. 70, 305–312. https://doi.org/10.1590/S0103-90162013000500004

Frøseth, R.B., Bleken, M.A., 2015. Effect of low temperature and soil type on the decomposition rate of soil organic carbon and clover leaves, and related priming effect. Soil Biol. Biochem. 80, 156–166. https://doi.org/10.1016/J.SOILBIO.2014.10.004

Galdos, M.V., Cerri, C.C., Cerri, C.E.P., 2009. Soil carbon stocks under burned and unburned sugarcane in Brazil. Geoderma 153, 347–352. https://doi.org/10.1016/J.GEODERMA.2009.08.025

Goldschmidt, R., Passos, E., Bezerra, E., 2015. Data Mining : Conceitos, técnicas, algoritmos, orientações e aplicações, Elsevier.

Governo Federal, 2017. Política Nacional de Biocombustíveis (RenovaBio) [WWW Document]. URL http://legislacao.anp.gov.br/?path=legislacao-federal/leis/2017&item=lei-13.576--2017 (accessed 9.11.18).

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

KARLEN, D.L., JOHNSON, J.M.F., 2014. CROP RESIDUE CONSIDERATIONS FOR SUSTAINABLE BIOENERGY FEEDSTOCK SUPPLIES. BIOENERGY RES. 7, 465–467. HTTPS://DOI.ORG/10.1007/S12155-014-9407-Y

LACOSTE, M., MICHOT, D., VIAUD, V., WALTER, C., MINASNY, B., MCBRATNEY, A.B., 2012. HIGH RESOLUTION 3D MAPPING FOR SOIL ORGANIC CARBON ASSESSMENT IN A RURAL LANDSCAPE, IN: MINASNY, B AND MALONE, BP AND MCBRATNEY, AB (ED.), DIGITAL SOIL ASSESSMENTS AND BEYOND. CRC PRESS-TAYLOR \& FRANCIS GROUP, 6000 BROKEN SOUND PARKWAY NW, STE 300, BOCA RATON, FL 33487-2742 USA, PP. 341–345.

LAL, R., 2004. SOIL CARBON SEQUESTRATION TO MITIGATE CLIMATE CHANGE. GEODERMA. HTTPS://DOI.ORG/10.1016/J.GEODERMA.2004.01.032

LAMPE, M. VON., 2008. BIOFUEL SUPPORT POLICIES : AN ECONOMIC ASSESSMENT. OECD.

LU, H., LI, S., MA, M., BASTRIKOV, V., CHEN, X., CIAIS, P., DAI, Y., ITO, A., JU, W., LIENERT, S., LOMBARDOZZI, D., LU, X., MAIGNAN, F., NAKHAVALI, M., QUINE, T., SCHINDLBACHER, A., WANG, J., WANG, Y., WARLIND, D., ZHANG, S., YUAN, W., 2021. COMPARING MACHINE LEARNING-DERIVED GLOBAL ESTIMATES OF SOIL RESPIRATION AND ITS COMPONENTS WITH THOSE FROM TERRESTRIAL ECOSYSTEM MODELS. ENVIRON. RES. LETT. 16. HTTPS://DOI.ORG/10.1088/1748-9326/ABF526

MANZONI, S., PORPORATO, A., 2009. SOIL CARBON AND NITROGEN MINERALIZATION: THEORY AND MODELS ACROSS SCALES. SOIL BIOL. BIOCHEM. 41, 1355–1379. HTTPS://DOI.ORG/10.1016/J.SOILBIO.2009.02.031

MCCARL, B.A., METTING, F.B., RICE, C., 2007. SOIL CARBON SEQUESTRATION. CLIM. CHANGE 80, 1–3. HTTPS://DOI.ORG/10.1007/S10584-006-9174-7

MINISTÉRIO DE MINAS E ENERGIA, 2018. CENÁRIOS DE OFERTA DE ETANO E DEMANDA DE CICLO OTTO 2018-2030.

NIKIFOROFF, C.C., 1937. SOME GENERAL ASPECTS OF THE CHERNOZEM FORMATION. SOIL SCI. SOC. AM. J. 1, 333. HTTPS://DOI.ORG/10.2136/SSSAJ1937.03615995000100000060X

PALOSUO, T., FOEREID, B., SVENSSON, M., SHURPALI, N., LEHTONEN, A., HERBST, M., LINKOSALO, T., ORTIZ, C., RAMPAZZO TODOROVIC, G., MARCINKONIS, S., LI, C., JANDL, R., 2012. A MULTI-MODEL COMPARISON OF SOIL CARBON ASSESSMENT OF A CONIFEROUS FOREST STAND. ENVIRON. MODEL. SOFTW. 35, 38–49. HTTPS://DOI.ORG/10.1016/J.ENVSOFT.2012.02.004

Silva-Olaya, A.M., Cerri, C.E.P., Williams, S., Cerri, C.C., Davies, C.A., Paustian, K., 2017. Modelling SOC response to land use change and management practices in sugarcane cultivation in South-Central Brazil. Plant Soil 410, 483–498. https://doi.org/10.1007/s11104-016-3030-y

Soucémarianadin, L.N., Cécillon, L., Guenet, B., Chenu, C., Baudin, F., Nicolas, M., Girardin, C., Barré, P., 2018. Environmental factors controlling soil organic carbon stability in French forest soils. Plant Soil 426, 267–286. https://doi.org/10.1007/s11104-018-3613-x

Stevenson, F.J., Tanji, K.K., 1982. Modeling of the Soil Nitrogen Cycle, in: Nitrogen in Agricultural Soils. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, pp. 721–772. https://doi.org/10.2134/agronmonogr22.c19

Trumbore, S.E., Chadwick, O.A., Amundson, R., 1996. Rapid Exchange Between Soil Carbon and Atmospheric Carbon Dioxide Driven by Temperature Change. Science (80-.). 272, 393–396. https://doi.org/10.1126/science.272.5260.393

West, T.O., Post, W.M., 2002. Soil Organic Carbon Sequestration Rates by Tillage and Crop Rotation. Soil Sci. Soc. Am. J. 66, 1930. https://doi.org/10.2136/sssaj2002.1930Zani, C.F., Barneze, A.S., Robertson, A.D., Keith, A.M., Cerri, C.E.P., McNamara, N.P., Cerri, C.C., 2018. Vinasse application and cessation of burning in sugarcane management can have positive impact on soil carbon stocks. PeerJ 6. https://doi.org/10.7717/peerj.5398

2. MACHINE LEARNING APPLIED TO SOIL SCIENCES AND SOIL ORGANIC CARBON: A REVIEW WITH EMPHASIS ON TROPICAL SOILS

ABSTRACT

The improvement and adoption of artificial intelligence (AI) techniques and Machine learning (ML) has increased dramatically over the last 20 years. A bibliographical search was conducted at the Web of Science database to identify the number of studies and some trends in ML usage for soil science, with emphasis on soil organic carbon (SOC) in tropical soils. It was revealed that in soil sciences, the adoption of ML techniques is in accelerated growth, but in countries with tropical soils only 40 papers were related to SOC research; most of them were published in the past 2 years and originated from Australia. The main task to which ML is addressed is regression, and the most frequent algorithms are Random Forest, Cubist and Boosted Decision Trees, with an apparent growing trend in the application of Random Forest and Extreme Gradient Boosting. Cross-validation is the most frequent validation strategy. At the end, some gaps and opportunities in the area are discussed.

Keywords: Data mining; Machine learning; Soil organic carbon; Soil sciences; Tropical soils; Weathered soils

2.1. Introduction

Clustering, modelling and predicting outcomes from data are common goals in soil science. Statistical approaches addressing to solve and interpret relations among soil and environmental attributes are adopted since the development of this research field.

The improvement and adoption of artificial intelligence (AI) techniques have increased dramatically over the last 20 years driven by the reduction of costs associated with the capture, transmission and storage of data. This has made it easier and faster to program computers, to create models and take actions from incoming data in instead of predefined commands. Among AI variety of possibilities, the machine learning (ML) algorithms (also referred to as data mining algorithms) are by far the most widely adopted in all sorts of research and industry applications (JORDAN; MITCHELL, 2015). They associate statistical approaches and computational solutions to deal with a large amount of data and investigate patterns among variables that are not readily recognized by human eyes and computationally costly through standard

statistical techniques. This is the reason why the process of mining data is often referred to as knowledge discovery from data or exploratory statistics (HAN; KAMBER; PEI, 2012).

The mentioned tools were previously restricted to the computer sciences but more recently, it is noticeable the increasing use of ML in soil science publications. Some reviews were devoted to examining the current state of the art in the implementation of ML tools in soil sciences (XIAO et al., 2019; PADARIAN; MINASNY; MCBRATNEY, 2020; SRIVASTAVA; SHUKLA; BANSAL, 2021; WADOUX; MCBRATNEY, 2021) and soil organic carbon (SOC) (ANGELOPOULOU et al., 2019; AHIRWAL et al., 2021), however little or no emphasis was given to countries with tropical soils.

The present work aims to clarify some of the concepts and tasks related to AI and ML, demonstrate its acceptance in SOC studies in tropical soils and to observe the most used techniques. It is expected that this literature review could help to demystify this exciting research field and encouraged more soil scientists to try new AI and ML approaches in their areas.

2.2. Development

2.2.1. Main terms and procedures in machine learning

A ML or data mining algorithm is the part of an AI system responsible for creating a rule or predict a value given the provided data and requires careful data preparation to obtains reliable insights from complex patterns.

Roughly speaking, ML tasks can be classified as supervised and unsupervised. Supervised tasks refer to situations where correctly labeled examples are presented to the algorithm for training, as in classification (discrete target variable) or regression (continuous target). Unsupervised tasks are clustering, creation of association rules and summarization. In these applications, there are not correctly labeled examples and heuristic solutions should be used to best achieve the goal.

The main steps that precede a ML technique usage are, in the sequence: data cleaning (removal of incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data), data integration (combination of multiple data sources), data selection (to keep only the data relevant to the task) and data transformation (changed

to a better organized or task compatible format) (HAN; KAMBER; PEI, 2012), and are commonly referred to as the most time-consuming part of any ML application (PYLE; EDITOR; CERRA, 1999). This is due to the fact that data, if not treated properly, can negatively influence the accuracy and precision of data mining tasks. Occurrence of unrealistic outliers, ambiguities in the information encoding, attributes monitored at different scales of time or space, unbalanced classes (that is, some may occur at relative frequencies so different that one or more are ignored by the algorithm) are examples of frequent situations that deserve special attention.

The following steps include the application of methods to extract patterns (ML algorithm), model evaluation using statistics metrics and validation strategies and knowledge presentation.

2.2.2. Adoption of Machine Learning Techniques in Soil Science and Soil Organic Carbon

The first publications related to AI may be tracked to the 1950´s (CAMPAIGNE; HOWARD, 1959; MCCARTHY et al., 2006), however it was on the late 1990´s that the subject started to arouse greater academic interest, inclusive at the soil.

In June of 2021, a series of searches were performed at the Web of Science database (WoS), to evaluate the progress in acceptance of ML techniques in soil sciences. Considering *“data mining” OR “machine learning”* as TOPIC among articles and reviews, 145,528 documents could be recovered at WoS. However, when the term “*soil*” was added, i.e. *(“data mining” OR “ML”) AND “soil”*, the number of recovered documents reduced to 2,504. It´s interesting to point out that around 60% of the papers focused directly on ML applications at soil science were published in the last two and a half years period (2019 - June 2021) and in the same period, the mean annual increase in publications was 44% (Figure 1-A). The same query was conducted using the SCOPUS database to verify if similar patterns can be observed. Maintaining the combination of the terms as TOPIC, a total of 2,528 documents were retrieved, thus confirming the low number of publications focusing directly on soil and ML.

As a next step, only the publications containing the term “soil” in the title were considered among them. A total of 947 could be identified and almost half of them (48%) originate from China and USA (Figure 1-B).

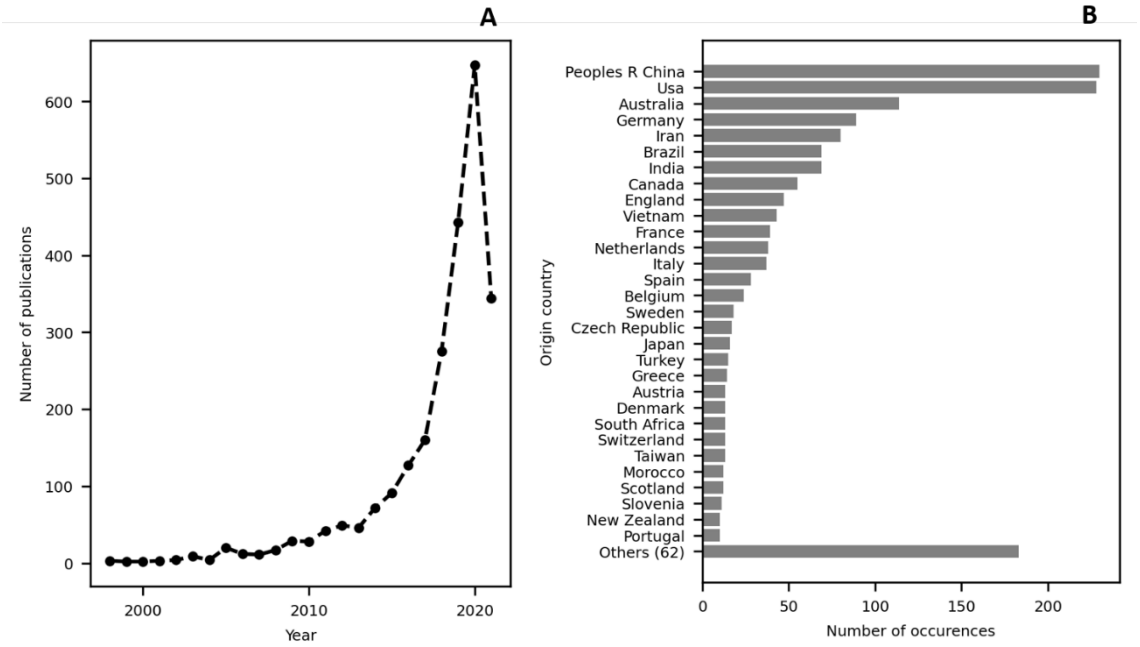

Figure 1. A) Number of publications per year where terms “data mining” or “machine learning” were related to soil studies. **B)** Top origin countries from publications containing the term “soil” in the title (10 or more published papers). The labeled bar “others (62)” refers to the group: Ireland, Malaysia, Norway, Russia, Mexico, Pakistan, Algeria, Iraq, Israel, Romania, Colombia, Hungary, Poland, Saudi Arabia, Argentina, Finland, Nigeria, Philippines, Serbia, Zambia, Cameroon, Croatia, Ecuador, Egypt, Georgia, Kenya, Slovakia, Burkina Faso, Chile, Democratic Republique of the Congo, Estonia, Indonesia, Jordan, Nepal, New Caledonia, Sri Lanka, Tanzania, Thailand, United Arab Emirates, Armenia, Bangladesh, Botswana, Costa Rica, Cote d’Ivoire, Ethiopia, Ghana, Haiti, Latvia, Lebanon, Madagascar, Mali, Montenegro, Mozambique, Peru, Qatar, Senegal, Tunisia, Uganda, Uzbekistan, Wales, Yemen, Zimbabwe.

Studies are even more scarce when considering only publications from countries with tropical soils (389) and limiting to the ones also containing the term “carbon” in the title, 40 papers could be recovered from 10 countries, being 62% of them from Australia (Figure 2). This, along with previously presented information, demonstrates how small is the article production in the area among countries with highly weathered soils (except Australia). This subset of papers (40) was selected for further reviews.

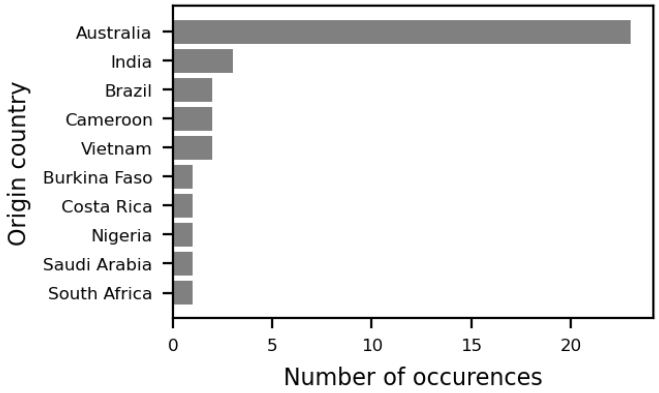

Figure 2. Tropical countries where terms “data mining” or “machine learning” were related to soil studies and words “soil” and “carbon were mentioned in the article title.

2.2.3. Main techniques in use and their applications

To identify major trends in ML usage in SOC from weathered soils, the selected subset of papers was evaluated accordingly to the main ML task applied, algorithm tested, applied metrics and validation strategies.

In 39 out of the 40 papers, regression algorithms were used and frequently more than one. Three of them solved classification tasks (but two of them also involved regression tasks) and only one also made use of a clustering technique (Table 1).

Tabela 1. Identification of the type of task for which machine learning algorithms were used in SOC studies in weathered soils.

Machine learning task	Number of occurrences	Article
Regression	38	(SPENCER et al., 2005; KRISHNAN et al., 2007; BUI; HENDERSON; VIERGEVER, 2009; CHAKRABORTY et al., 2013; LACOSTE et al., 2014; SREENIVAS et al., 2014, 2016; HOBLEY et al., 2015; HOBLEY; BALDOCK; WILSON, 2016; HOBLEY; WILSON, 2016; RUDIYANTO et al., 2016, 2018; SOMARATHNA; MALONE; MINASNY, 2016; MARTINEZ-ESPANA et al., 2017; ROUDIER et al., 2017; HOUNKPATIN et al., 2018; SANDERMAN et al., 2018; WANG et al., 2018a, 2018b; DING et al., 2018; HAMZEHPOUR; SHAFIZADEH-MOGHADAM; VALAVI, 2019; MUKHERJEE et al., 2019; ROSSEL et al., 2019; GOMES et al., 2019; MOURA-BUENO et al., 2020; RODRIGUEZ-VEIGA et al., 2020; SILATSA et al., 2020; BENKE et al., 2020; TAGHIZADEH-MEHRJARDI et al., 2020; GHOLIZADEH et al., 2020; ABERA et al., 2021; ABRAMOFF et al., 2021; GUEVARA-ESCOBAR et al., 2021; AHIRWAL et al., 2021; PHAM et al., 2021; VASUDEVA et al., 2021; VENTER et al., 2021; GOYDARAGH et al., 2021)
Classification	3	(SPENCER et al., 2005; RODRIGUEZ-VEIGA et al., 2020; COSTA et al., 2021)
Clustering	1	(MOURA-BUENO et al., 2020)

Considering all the applications, 17 algorithms or combinations of them were used. Random Forest, Cubist, Boosted Decision Trees, Artificial Neural Networks, Support Vector Machine and Extreme Gradient Boosting are, in the mentioned order, the most frequent adopted techniques in supervised tasks (regression and classification) and a tendency of increase in Random Forest usage appears to be ongoing in the recent years (Figure 3). In the single paper where a ML clustering task was identified the chosen algorithm was the K-Means.

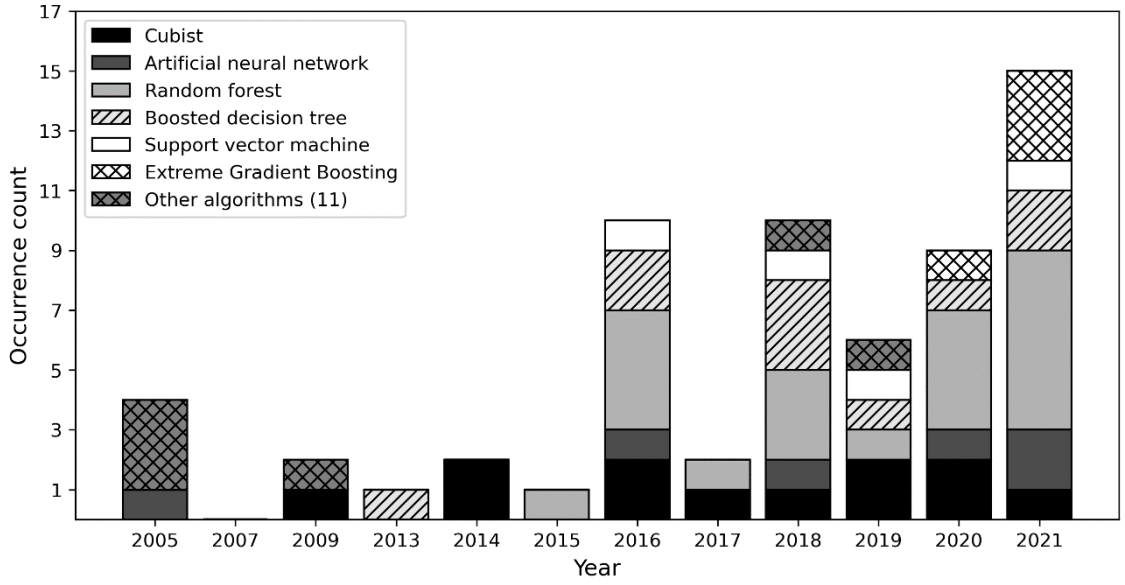

Figure 3. Progression in the usage of the most common algorithms in SOC studies in weathered soils. The label “other algorithms” refers to the group: quantile regression forest (2 publications), bagging with decision trees (2), conditional inference trees (2), Lasso regression (1), multiplicative adaptative regression splines (1), generalised linear mixed model (1), K-means (1), stacking with various diferent models (1) and light gradient boosting machine regression (1).

Among the metrics for regression models the preferred ones are the coefficient of determination (R^2) and the root mean squared error (RMSE), present in 72% of and 65% of the publications, respectively. Other frequent statistics applied are the mean absolute error, mean squared error (both present in 17% of studies) and Linn´s concordance correlation coefficient (15%). Other 11 metrics were used in smaller frequencies, being used in only one or two publications: standard deviation error, median absolute relative error, Pearson´s correlation, model efficiency factor, ratio of performance to interquartile distance, mean squared deviation, ratio of performance deviation, normalized root mean squared error, mean total deviance, mean residual deviance and a customized measure.

In the classification ML tasks the chosen metrics were the accuracy, probability of agreement and area under the receiver operating curve. In the clustering task the applied metric was not specified.

A decisive step in a ML pipeline is to property validate the trained models, especially in supervised tasks. It´s not uncommon to develop models with great performance in the training set but with a much different prediction capacity when new unlabeled examples are presented. To overcome this difficulty, different strategies can be adopted as cross-validation, hold out and leave one out.

In the cross-validation, the provided training set is split n times, with a smaller fraction of it kept out of the training process. That fraction is later used as a test set for performance comparison. Each split part of the dataset is chosen to test a model trained in the remaining examples. In the end, a mean result of the n training processes can be evaluated. Cross-validation was the most applied validation strategy in the reviewed publications, being used in 25 articles (62%).

Leave one out is a variation of cross-validation where the number of splits and training processes is equal to the number of examples in the provided training set, i.e, at any split only one example is left as the test set. Leave one out can be computationally expensive and time-consuming, being identified in only one of the selected publications. The hold-out validation strategy was present in 12 publications (30%) and consists of separating a percentage of the dataset to be used as a test set.

2.3. Final remarks and future perspectives

Several environmental, soil and cover factors strongly affect the soil carbon sequestration potential, such as crop (WEST et al., 2002), temperature (TRUMBORE; CHADWICK; AMUNDSON, 1996; FRØSETH; BLEKEN, 2015), texture (BAUHUS; PARÉ; CÔTÉ, 1998; SOUCÉMARIANADIN et al., 2018), pH (CURTIN; CAMPBELL; JALIL, 1998), access of organic matter to microorganisms (DUNGAIT et al., 2012) among others (FRANCAVIGLIA et al., 2017). The ML solutions represent a powerful tool to help soil researchers unveil these relations and make an assumption on the future behavior of SOC. Perhaps, as important as predicting SOC carbon fluctuation is to select the most important attributes to monitor it on small and large scales.

Several algorithms applied in the area were written many years ago and now, with the advances in computer hardware and software, are being taken to new usage frontiers. Others are quite recent, and its application is still evolving following the ML field steps. Random Forest and several other ensemble algorithms inspired in Decision Trees as Cubist and Extreme Gradient Boosting are well accepted and shall grow in usage along with the popularization of ML among soil scientists. The simpler ones as decision trees itself will probably remain more restricted due to the inferior prediction and generalization power compared to the ensemble algorithms.

The present literature review has made clear that SOC research field in tropical and highly weathered soils has a lot more to offer and benefit from ML. Except for

Australia, which is quite advanced in the area, studies are still incipient or null in most developing countries. Yet, it was identified that some ML tasks, especially unsupervised ones as search for association rules and summarization, were not contemplated in any of the articles found and therefore represented an opportunity to create new applications.

References

Abera, W., Tamene, L., Abegaz, A., Hailu, H., Piikki, K., Soderstrom, M., Girvetz, E., Sommer, R., 2021. Estimating spatially distributed SOC sequestration potentials of sustainable land management practices in Ethiopia. J. Environ. Manage. 286. https://doi.org/10.1016/j.jenvman.2021.112191

Abramoff, R.Z., Georgiou, K., Guenet, B., Torn, M.S., Huang, Y., Zhang, H., Feng, W., Jagadamma, S., Kaiser, K., Kothawala, D., Mayes, M.A., Ciais, P., 2021. How much carbon can be added to soil by sorption? Biogeochemistry 152, 127–142. https://doi.org/10.1007/s10533-021-00759-x

Ahirwal, J., Nath, A., Brahma, B., Deb, S., Sahoo, U.K., Nath, A.J., 2021. Patterns and driving factors of biomass carbon and soil organic carbon stock in the Indian Himalayan region. Sci. Total Environ. 770. https://doi.org/10.1016/j.scitotenv.2021.145292

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. https://doi.org/10.3390/rs11060676

Bauhus, J., Paré, D., Côté, L., 1998. Effects of tree species, stand age and soil type on soil microbial biomass and its activity in a southern boreal forest. Soil Biol. Biochem. https://doi.org/10.1016/S0038-0717(97)00213-7

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B., Hopley, J., 2020. Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma 366. https://doi.org/10.1016/j.geoderma.2020.114210

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global Biogeochem. Cycles 23. https://doi.org/10.1029/2009GB003506

Campaigne, H., Howard, 1959. Some experiments in machine learning, in: Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference on XX - IRE-AIEE-ACM '59 (Western). ACM Press, New York, New York, USA, pp. 173–175. https://doi.org/10.1145/1457838.1457868

Chakraborty, S., Weindorf, D.C., Ali, M.N., Li, B., Ge, Y., Darilek, J.L., 2013. Spectral data mining for rapid measurement of organic matter in unsieved moist compost. Appl. Opt. 52, B82–B92. https://doi.org/10.1364/AO.52.000B82

Costa, M.D. de P., Lovelock, C.E., Waltham, N.J., Young, M., Adame, M.F., Bryant V, C., Butler, D., Green, D., Rasheed, M.A., Salinas, C., Serrano, O., York, P.H., Whitt, A.A., Macreadie I, P., 2021. Current and future carbon stocks in coastal wetlands within the Great Barrier Reef catchments. Glob. Chang. Biol. 27, 3257–3271. https://doi.org/10.1111/gcb.15642

Curtin, D., Campbell, C.A., Jalil, A., 1998. Effects of acidity on mineralization: pH-dependence of organic matter mineralization in weakly acidic soils. Soil Biol. Biochem. 30, 57–64. https://doi.org/10.1016/S0038-0717(97)00094-1

Ding, F., Van Zwieten, L., Zhang, W., Weng, Z. (Han), Shi, S., Wang, J., Meng, J., 2018. A meta-analysis and critical evaluation of influencing factors on soil carbon priming following biochar amendment. J. SOILS SEDIMENTS 18, 1507–1517. https://doi.org/10.1007/s11368-017-1899-6

Dungait, J.A.J., Hopkins, D.W., Gregory, A.S., Whitmore, A.P., 2012. Soil organic matter turnover is governed by accessibility not recalcitrance. Glob. Chang. Biol. 18, 1781–1796. https://doi.org/10.1111/j.1365-2486.2012.02665.x

Francaviglia, R., Di Bene, C., Farina, R., Salvati, L., 2017. Soil organic carbon sequestration and tillage systems in the Mediterranean Basin: a data mining approach. Nutr. Cycl. Agroecosystems 107, 125–137. https://doi.org/10.1007/s10705-016-9820-z

Frøseth, R.B., Bleken, M.A., 2015. Effect of low temperature and soil type on the decomposition rate of soil organic carbon and clover leaves, and related priming effect. Soil Biol. Biochem. 80, 156–166. https://doi.org/10.1016/J.SOILBIO.2014.10.004

Gholizadeh, A., Saberioon, M., Rossel, R.A.V., Boruvka, L., Klement, A., 2020. Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. Geoderma 357. https://doi.org/10.1016/j.geoderma.2019.113972

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Fernandes Filho, E.I., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

Goydaragh, M.G., Taghizadeh-Mehrjardi, R., Jafarzadeh, A.A., Triantafilis, J., Lado, M., 2021. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. CATENA 202. https://doi.org/10.1016/j.catena.2021.105280

Guevara-Escobar, A., Gonzalez-Sosa, E., Cervantes-Jimenez, M., Suzan-Azpiri, H., Elisa Queijeiro-Bolanos, M., Carrillo-Angeles, I., Hugo Cambron-Sandoval, V., 2021. Machine learning estimates of eddy covariance carbon flux in a scrub in the Mexican highland. BIOGEOSCIENCES 18, 367–392. https://doi.org/10.5194/bg-18-367-2021

Hamzehpour, N., Shafizadeh-Moghadam, H., Valavi, R., 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. CATENA 182. https://doi.org/10.1016/j.catena.2019.104141

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hobley, E., Wilson, B., Wilkie, A., Gray, J., Koen, T., 2015. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil 390, 111–127. https://doi.org/10.1007/s11104-015-2380-1

Hobley, E.U., Baldock, J., Wilson, B., 2016. Environmental and human influences on organic carbon fractions down the soil profile. Agric. Ecosyst. \& Environ. 223, 152–166. https://doi.org/10.1016/j.agee.2016.03.004

Hobley, E.U., Wilson, B., 2016. The depth distribution of organic carbon in the soils of eastern Australia. ECOSPHERE 7. https://doi.org/10.1002/ecs2.1214

Hounkpatin, O.K.L., de Hipt, F.O., Bossa, A.Y., Welp, G., Amelung, W., 2018. Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). CATENA 166, 298–309. https://doi.org/10.1016/j.catena.2018.04.013

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 255–60. https://doi.org/10.1126/science.aaa8415

Krishnan, P., Bourgeon, G., Lo Seen, D., Nair, K.M., Prasanna, R., Srinivas, S., Muthusankar, G., Dufy, L., Ramesh, B.R., 2007. Organic carbon stock map for soils of southern India: A multifactorial approach. Curr. Sci. 93, 706–710.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311. https://doi.org/10.1016/j.geoderma.2013.07.002

Martinez-Espana, R., Bueno-Crespo, A., Soto, J., Janik, L.J., Soriano-Disla, J.M., 2017. Influence of Multivariate Modeling in the Prediction of Soil Carbon by a Portable Infrared Sensor, in: Analide, C and Kim, P (Ed.), INTELLIGENT ENVIRONMENTS 2017, Ambient Intelligence and Smart Environments. IOS PRESS, NIEUWE HEMWEG 6B, 1013 BG AMSTERDAM, NETHERLANDS, pp. 88–97. https://doi.org/10.3233/978-1-61499-796-2-88

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., 2006. A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag. 27.

Moura-Bueno, J.M., Diniz Dalmolin, R.S., Horst-Heinen, T.Z., ten Caten, A., Vasques, G.M., Dotto, A.C., Grunwald, S., 2020. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Sci. Total Environ. 737. https://doi.org/10.1016/j.scitotenv.2020.139895

Mukherjee, J., Bhowmick, A.R., Ghosh, P.B., Ray, S., 2019. Impact of environmental factors on the dependency of litter biomass in carbon cycling of Hooghly estuary, India. Ecol. Inform. 51, 193–200. https://doi.org/10.1016/j.ecoinf.2019.03.007

Padarian, J., Minasny, B., Mcbratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. SOIL 6, 35–52. https://doi.org/10.5194/soil-6-35-2020

Pham, Tien Dat, Yokoya, N., Nguyen, T.T.T., Le, N.N., Ha, N.T., Xia, J., Takeuchi, W., Pham, Tien Duc, 2021. Improvement of Mangrove Soil Carbon Stocks Estimation in North Vietnam Using Sentinel-2 Data and Machine Learning Approach. GISCIENCE \& Remote Sens. 58, 68–87. https://doi.org/10.1080/15481603.2020.1857623

Pyle, D., Editor, S., Cerra, D.D., 1999. Data Preparation for Data Mining, The Morgan Kaufmann Series in Data Management Systems. https://doi.org/10.1080/713827180

Rodriguez-Veiga, P., Carreiras, J., Smallman, T.L., Exbrayat, J.-F., Ndambiri, J., Mutwiri, F., Nyasaka, D., Quegan, S., Williams, M., Balzter, H., 2020. Carbon Stocks and Fluxes in Kenyan Forests and Wooded Grasslands Derived from Earth Observation and Model-Data Fusion. Remote Sens. 12. https://doi.org/10.3390/rs12152380

Rossel, R.A.V., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12, 547+. https://doi.org/10.1038/s41561-019-0373-z

Roudier, P., Malone, B.P., Hedley, C.B., Minasny, B., McBratney, A.B., 2017. Comparison of regression methods for spatial downscaling of soil organic carbon stocks maps. Comput. Electron. Agric. 142, 91–100. https://doi.org/10.1016/j.compag.2017.08.021

Rudiyanto, Minasny, B., Setiawan, B.I., Arif, C., Saptomo, S.K., Chadirin, Y., 2016. Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands. Geoderma 272, 20–31. https://doi.org/10.1016/j.geoderma.2016.02.026

Rudiyanto, Minasny, B., Setiawan, B.I., Saptomo, S.K., McBratney, A.B., 2018. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313, 25–40. https://doi.org/10.1016/j.geoderma.2017.10.018

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M.F., Benson, L., Bukoski, J.J., Carnell, P., Cifuentes-Jara, M., Donato, D., Duncan, C., Eid, E.M., zu Ermgassen, P., Lewis, C.J.E., Macreadie, P.I., Glass, L., Gress, S., Jardine, S.L., Jones, T.G., Nsombo, E.N., Rahman, M.M., Sanders, C.J., Spalding, M., Landis, E., 2018. A global map of mangrove forest soil carbon at 30 m spatial resolution. Environ. Res. Lett. 13. https://doi.org/10.1088/1748-9326/aabe1c

Silatsa, F.B.T., Yemefack, M., Tabi, F.O., Heuvelink, G.B.M., Leenaars, J.G.B., 2020. Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in Cameroon. Geoderma 367. https://doi.org/10.1016/j.geoderma.2020.114260

Somarathna, P.D.S.N., Malone, B.P., Minasny, B., 2016. Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. GEODERMA Reg. 7, 38–48. https://doi.org/10.1016/j.geodrs.2015.12.002

Soucémarianadin, L.N., Cécillon, L., Guenet, B., Chenu, C., Baudin, F., Nicolas, M., Girardin, C., Barré, P., 2018. Environmental factors controlling soil organic carbon stability in French forest soils. Plant Soil 426, 267–286. https://doi.org/10.1007/s11104-018-3613-x

Spencer, M., McCullagh, J., Whitfort, T., Reynard, K., 2005. An Application into Using Artificial Intelligence for Estimating Organic Carbon, in: Zerger, A and Argent, RM (Ed.), MODSIM 2005: INTERNATIONAL CONGRESS ON MODELLING AND SIMULATION: ADVANCES AND APPLICATIONS FOR MANAGEMENT AND DECISION MAKING: ADVANCES AND APPLICATIONS FOR MANAGEMENT AND DECISION MAKING. pp. 84–90.

Sreenivas, K., Dadhwal, V.K., Kumar, S., Harsha, G.S., Mitran, T., Sujatha, G., Suresh, G.J.R., Fyzee, M.A., Ravisankar, T., 2016. Digital mapping of soil organic and inorganic carbon status in India. Geoderma 269, 160–173. https://doi.org/10.1016/j.geoderma.2016.02.002

Sreenivas, K., Sujatha, G., Sudhir, K., Kiran, D.V., Fyzee, M.A., Ravisankar, T., Dadhwal, V.K., 2014. Spatial Assessment of Soil Organic Carbon Density Through Random Forests Based Imputation. J. Indian Soc. Remote Sens. 42, 577–587. https://doi.org/10.1007/s12524-013-0332-x

Srivastava, P., Shukla, A., Bansal, A., 2021. A comprehensive review on soil classification using deep learning and computer vision techniques. Multimed. Tools Appl. 80, 14887–14914. https://doi.org/10.1007/s11042-021-10544-5

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. Remote Sens. 12. https://doi.org/10.3390/rs12071095

Trumbore, S.E., Chadwick, O.A., Amundson, R., 1996. Rapid Exchange Between Soil Carbon and Atmospheric Carbon Dioxide Driven by Temperature Change. Science (80-.). 272, 393–396. https://doi.org/10.1126/science.272.5260.393

Vasudeva, V., Nandy, S., Padalia, H., Srinet, R., Chauhan, P., 2021. Mapping spatial variability of foliar nitrogen and carbon in Indian tropical moist deciduous sal (Shorea robusta) forest using machine learning algorithms and Sentinel-2 data. Int. J. Remote Sens. 42, 1139–1159. https://doi.org/10.1080/01431161.2020.1823043

Venter, Z.S., Hawkins, H.-J., Cramer, M.D., Mills, A.J., 2021. Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa. Sci. Total Environ. 771. https://doi.org/10.1016/j.scitotenv.2021.145384

Wadoux, A.M.J.-C., McBratney, A.B., 2021. Hypotheses, machine learning and soil mapping. Geoderma 383. https://doi.org/10.1016/j.geoderma.2020.114725

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Liu, D.L., Simpson, M., McGowen, I., Sides, T., 2018a. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. Ecol. Indic. 88, 425–438. https://doi.org/10.1016/j.ecolind.2018.01.049

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., Liu, D.L., 2018b. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Sci. Total Environ. 630, 367–378. https://doi.org/10.1016/j.scitotenv.2018.02.204

West, Tristram O, Post, Wilfred M, West, T O, Post, W M, 2002. Soil Organic Carbon Sequestration Rates by Tillage and Crop Rotation: A Global Data Analysis, Published in Soil Sci. Soc. Am. J.

Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J.A., Huete, A.R., Ichii, K., Ni, W., Pang, Y., Rahman, A.F., Sun, G., Yuan, W., Zhang, L., Zhang, X., 2019. Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. Remote Sens. Environ. 233. https://doi.org/10.1016/j.rse.2019.111383

3. MACHINE LEARNING FOR PREDICTION OF SOIL CARBON STOCK CHANGES IN SUGARCANE CROP DUE TO STRAW REMOVAL

ABSTRACT

Brazil, as other countries, has established energy and climate policies that foster the use of biofuels as sugarcane ethanol, in which a growing practice is to use harvesting residues, the straw, for cogeneration of electricity or to produce second-generation ethanol. In this study, it was aimed to create machine learning (ML) models capable of predict short-term soil carbon stock changes according to the mass of sugarcane straw leftover the soil during harvest. The initial data was generated between 2015 and 2018 in five experimental sites under commercial cultivation of sugarcane in Brazilian south-central region and the available variables were related to climate, soil physical and chemical attributes, organic matter and crop variety. The variable to be predicted (*y*) was the rate of carbon stock change per area per year in relation to the total dry mass of straw (Mg C ha^{-1} yr^{-1}). The initial dataset was divided into training (80%) and test (20%) and eight ML models were trained using algorithms Random Forest (RF) and Support Vector Machine (SVM) associated to four feature selection methods. Results were evaluated using 10-fold cross-validation of the root mean squared error (RMSE) in the training set and prediction RMSE in the test set. The trained models were statistically compared among them and to the use of mean y stratified by straw mass deposited and soil layer. All the ML models surpassed the simple generalization of previously known mean values of y. The model SVM associated with RF feature selection performed slightly better with a considerable reduction in the number of attributes, which could reduce the costs and effort of data acquisition and processing in future applications. The achievements indicate that ML models are good tools to predict short-term changes in carbon stocks due to total or partial straw remotion from the field. The obtained results and applied methodology have the potential to help producers and decision-makers interested in correlate in situ crop conditions and straw management to expected soil carbon variations.

Keywords: Soil organic carbon; Machine learning; Straw management; Feature selection; Random forest; Support vector machines

3.1. Introduction

Governments aiming to mitigate greenhouse gas (GHG) emissions should take advantage of soil capacity to store carbon and implement public politics which stimulates the preservation or even increase of soil carbon stocks in cropped lands. Brazil, as other countries, has established energy and climate policies that foster the

use of biofuels, and sugarcane ethanol receives special attention due to the great potential for emission mitigation in comparison to fossil fuels. This should encourage more sustainable management practices which positively affect soil carbon sequestration(POST; KWON, 2000; LAL, 2004; MCCARL; METTING; RICE, 2007).

In sugarcane cropping, a growing practice uses harvesting residues, the straw, for cogeneration of electricity or as a substrate to produce second-generation ethanol. Although it may help to reduce industry´s energy expenditure and allow an ethanol production increment without area expansion (FRANCO et al., 2013; KARLEN; JOHNSON, 2014), the organic matter kept in the field promotes several additional benefits to the soil (CHERUBIN et al., 2018b), including desirable inputs of carbon, nutrients and physical protection.

It becomes clear that, in the short future, sugarcane management strategies shall take soil carbon fluctuation into account, despite the complexity of interactions that dictate it´s dynamics at different spatial and temporal scales (ETTEMA, 2002; AUSTIN et al., 2004).

Data mining techniques and machine learning (ML) algorithms comprise a set of tools specially developed to handle complex and large volumes of information and generate optimized models based on presented data. Efficient ML models were developed to address a range of challenges related to the organic matter (TAGHIZADEH-MEHRJARDI; NABIOLLAHI; KERRY, 2016; FRANCAVIGLIA et al., 2017; FARHATE et al., 2018; ANGELOPOULOU et al., 2019) and may help producers to take data-driven decisions on straw management and predict soil carbon stocks fluctuation at the same time.

In this study a series of information from field experiments in commercial sugarcane crops in the south-central region of Brazil were compiled. The main goals were: I) create ML models capable of predicting short term soil carbon stocks change according to the mass of sugarcane straw leftover the soil during harvest and II) select the most important attributes to feed the models, which reduces the cost and effort of data acquisition.

3.2. Material and Methods

3.2.1. Dataset description

The initial data was generated between 2015 and 2018 in five experimental sites under commercial cultivation of sugarcane in Brazil (Figure 4-A). The site selection was made intending to represent contrasting and representative sugarcane cropped situations in terms of soil and climate in the Brazilian south-central region (Figure 4-B). The treatments consisted of masses of sugarcane straw left in the field during the harvest. For two locations — Capivarí (22º59' S, 47º30' W) and Valparaíso (21°13' S, 50°52' W) at São Paulo state — the straw masses represented percentages of 0 (i.e. total removal), 25, 50, 75 and 100% from the total residues generated per area. For the three other locations — Quatá (22º14' S, 50º41' W) in São Paulo state, Chapadão do Céu (18º23' S, 52º39' W) and Quirinópolis (18° 26' S, 50° 27' W) in Goiás state — the straw masses represented percentages of 0, 50 and 100% of the mean mass of straw produced per area yearly in sugarcane in Brazil (MENANDRO et al., 2017). The experiments started in the first harvest after soil preparation. From the second harvest ahead, the straw left in the field added to the remaining and not completely decomposed from the previous year.

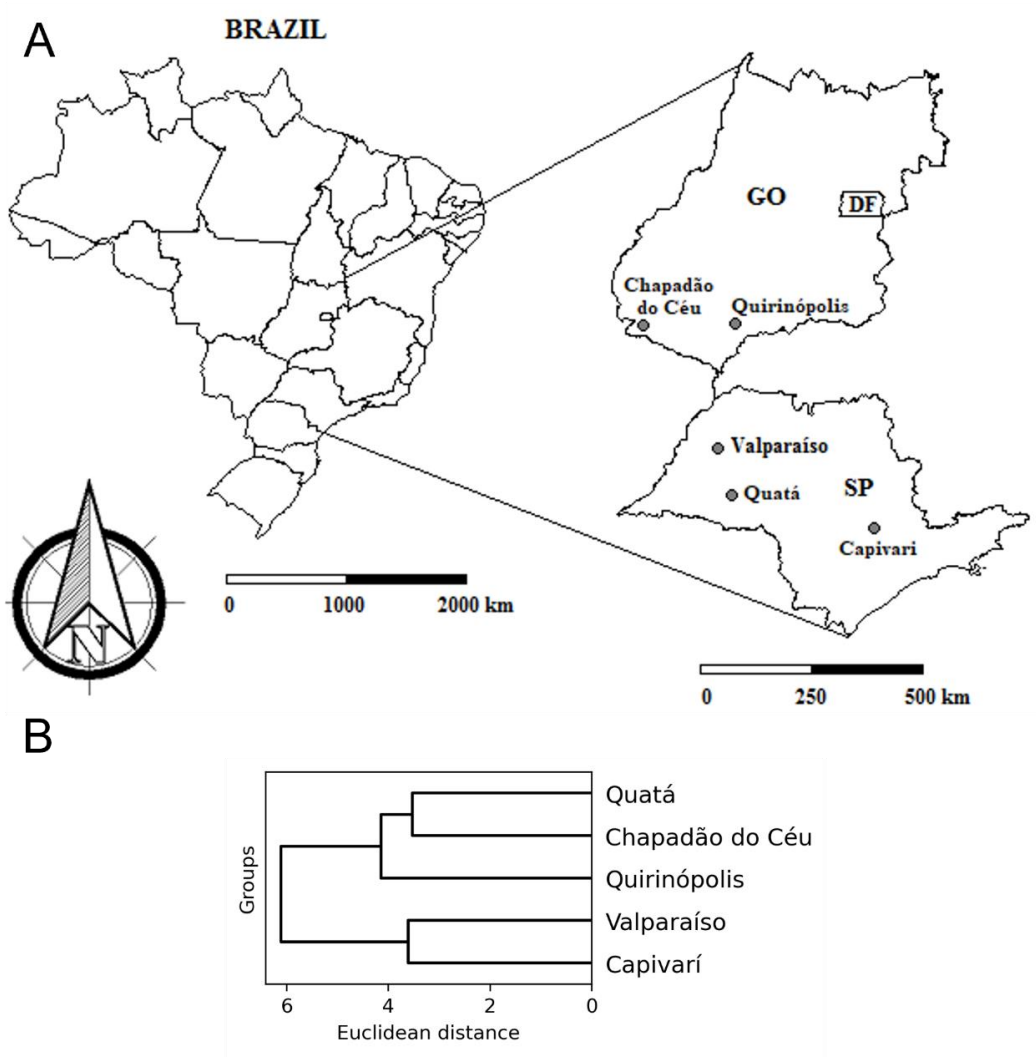

Figure 4. A) Geographic locations of the sampling sites. Data sampled between the 2015 - 2018 period. **B**) Hierarchical clustering dendrogram from data similarity considering climatic data and soil granulometry.

The available variables were related to climate, soil physical and chemical attributes, organic matter and crop variety (Table 2 and Figure 5). Information corresponding to the annual temperature, precipitation and climatological classification patterns were obtained from (ALVARES et al., 2013) and the database provided by the authors at http://www.ipef.br/geodatabase/.

Tabela 2. Description of initially available attributes from the study sites

Description	Unit or Labels	Attribute encoding
Altitude above sea level	m	altitude
Average annual rainfall	mm	yearly precipitation
Average annual maximum temperature	ºC	max yearly temperature
Average annual minimum temperature	ºC	min yearly temperature
Average annual temperature	ºC	mean yearly temperature
Straw mass left in the field in the first year of the experiment	Mg ha^{-1}	straw mass
Distance from surface (depth) of the assessed soil layer	cm	soil layer upper limit
Thickness of the assessed soil layer	cm	soil layer thickness
Nitrogen content in cane straw	g kg^{-1}	straw N
Phosphorus content in cane straw	mg dm^{-3}	straw P
Potassium content in cane straw	$mmol_c$ dm^{-3}	straw k
Calcium content in cane straw	$mmol_c$ dm^{-3}	straw Ca
Magnesium content in cane straw	$mmol_c$ dm^{-3}	straw Mg
Sulfur content in cane straw	$mmol_c$ dm^{-3}	straw S
Initial soil density in layer 0-10 cm	Mg m^{-3} (= g cm^{-3})	bulk density 0-10 cm
Initial density of soil in the layer 10-20 cm	Mg m^{-3} (= g cm^{-3})	bulk density 10-20 cm
Initial pH in the soil layer 0-10 cm	Adimensional	soil ph 0-10 cm
Initial pH in the soil layer 10-20 cm	Adimensional	soil ph 10a20_i
Initial phosphorus content of the soil in the 0-10 cm layer	mg dm^{-3}	soil P 0-10 cm
Initial phosphorus content of the soil in the 10-20 cm layer	mg dm^{-3}	soil P 10-20 cm
Initial potassium content of the soil in the 0-10 cm layer	$mmol_c$ dm^{-3}	soil K 0-10 cm
Initial potassium content of the soil in the 10-20 cm layer	$mmol_c$ dm^{-3}	soil K 10-20 cm
Initial calcium content of soil in the 0-10 cm layer	$mmol_c$ dm^{-3}	soil Ca 0-10 cm i
Initial calcium content of the soil in the 10-20 cm layer	$mmol_c$ dm^{-3}	soil Ca 10-20 cm
Initial magnesium content of the soil in the 0-10 cm layer	$mmol_c$ dm^{-3}	soil Mg 0-10 cm
Initial magnesium content of the soil in the 10-20 cm layer	$mmol_c$ dm^{-3}	soil Mg 10-20 cm
Soil sand content	g kg^{-1}	sand content
Soil clay content	g kg^{-1}	clay content
Cultivated sugarcane variety (Dummy variables were created)	yes (1), no (0)	CTC-14, RB86-7515 or RB96-6928
Köppen climate classification	Aw (1), Cwa (0)	Köppen climate
Target variable. It describes the rate of change of the soil carbon stock in relation to the treatment without straw removal. The negative values indicate straw tons per hectare loss per year.	Mg C ha^{-1} $year^{-1}$	Rate of carbon stock change

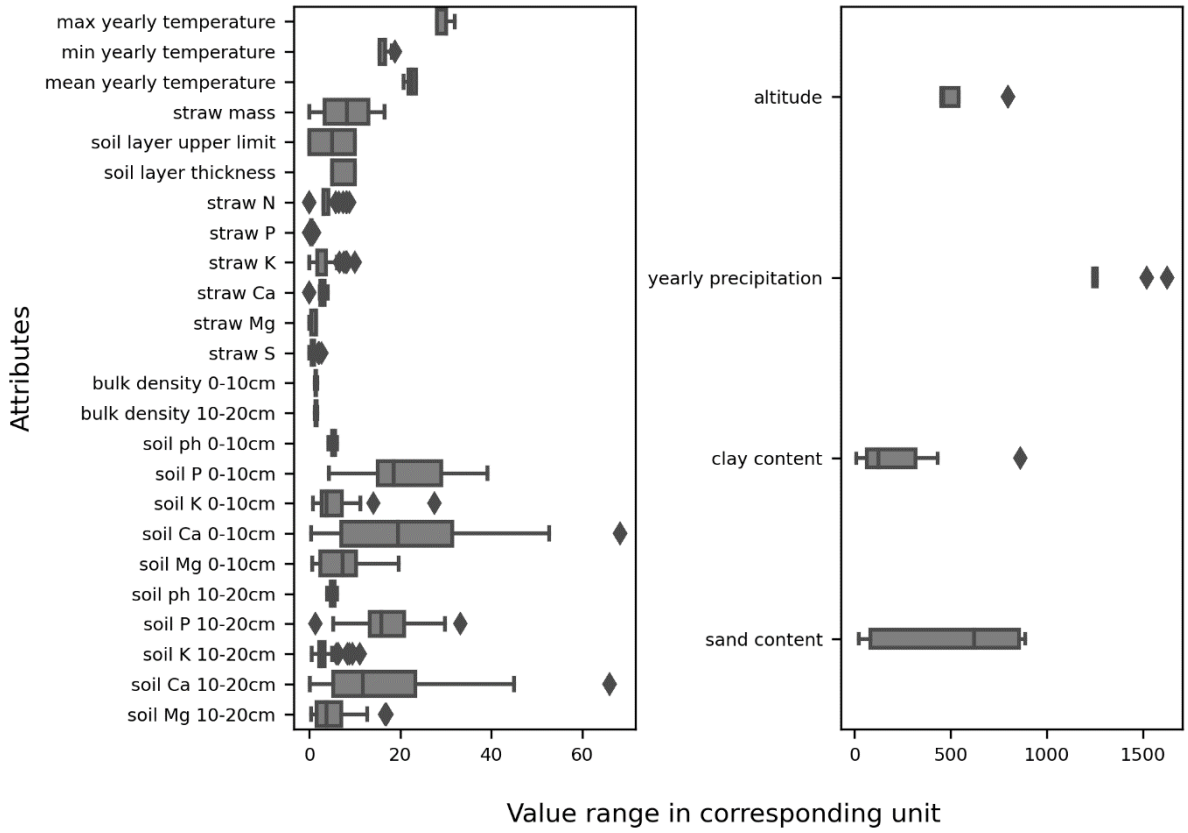

Figure 5. Value distribution of numerical attributes used in the models. Individual measurement units were described in table 1.

3.2.2. Data integration and preprocessing

Since data were obtained from different field experiments, they were adjusted for standardizing the evaluation periodicity and thickness of evaluated soil layers. All numerical variables were summarized and critically evaluated for their permanence in the dataset given their procurement cost and effort and relationship with others. As an example, only sand and clay content were chosen as granulometry metrics because silt content can be deducted from them. Also, for future application purposes, it was decided to include only attributes sampled before the installation of the treatments. Missing data were not frequent and were filled with median values of the corresponding location, treatment and soil layer.

The variable to be predicted (*y*) was the rate of carbon stock change per area per year in relation to the total dry mass of straw (given in megagrams of carbon per hectare per year, Mg C ha^{-1} yr^{-1}). It expresses how much carbon should be lost or gained from a soil layer stock per area in one year if the amount of straw left in the field

is modified. It was calculated for every experimental site, treatment and soil layer as $y = (C\ stock_{100\%} - C\ stock_{treatment})\ time^{-1}$, where $C\ stock_{100\%}$ is the median of the final carbon stocks for the control treatment (100% of sugarcane straw left in the field) and $C\ stock_{treatment}$ is the median of final carbon stocks for the evaluated treatment (Figure 6).

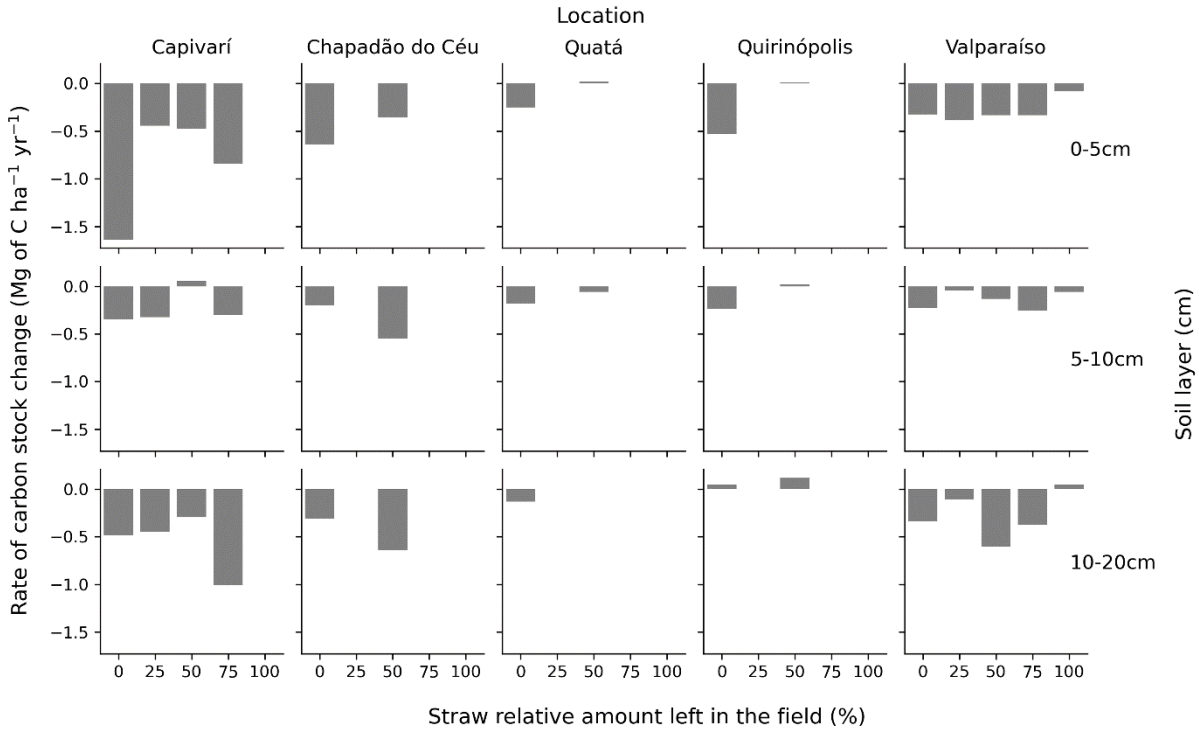

Figure 6. Calculated rate of carbon change in evaluated locations and soil layer depths.

Most algorithm implementations require only numerical data for the training process, so the qualitative variables were transformed. Sugarcane varieties were derived into *n* new ones (*n* equal to the number of labels at the given attribute), indicating the presence of the label (1) or absence (0). Binary encoding was used for the Köppen climate classification indicating Aw (0) or Cwa (1). The description of the sampled soil layer was derived into two continuous attributes: initial depth and thickness, using the unit in cm.

As a final prepossessing step, to prevent attributes with greater values from standing out over others during the learning process, data were transformed to z-scores according to the formula: $z = (x - u)\ s^{-1}$, where "*u*" is the mean of attribute values and "*s*" is the standard deviation.

The final dataset after data integration and preprocessing consisted of 305 instances and 32 attributes.

3.2.3. Attribute selection and regression algorithms implementation

Attribute selection and ML algorithm implementation were done using the Scikit-learn library implementation (PEDREGOSA et al., 2011) in Python language.

Initially, the dataset was randomly partitioned in two: 80% of the data for the learning process (training set) and 20% for performance check (test set or validation set). Two ML algorithms were trained to predict y:

Random Forest (RF) (BREIMAN, 2001): The RF is an ensemble predictor which uses bootstrap resampling to create n new datasets (usually $500 < n < 3000$). Statistically, around ⅓ of examples in the original dataset is not included for each new bootstrap dataset, the out-of-bag samples. Decision trees are grown in each or those resampled datasets using a function that searches for the attribute and its value threshold that minimizes a data impurity measure (for example, Gini or Entropy index) at each node split. However, only a subset of the original attributes is sampled for the selection of the best node split. To evaluate the model in regression tasks, each tree uses its out-of-bag examples as a test set and the predicted value is the mean of all the voting trees which did not use the example in the learning process. The error is calculated as a mean difference between predicted and real values for the examples.

Support Vector Machine (SVM) (BOSER; GUYON; VAPNIK, 1992; CORTES; VAPNIK, 1995): When used for classification tasks SVM creates a hyperplane that separate the classes of the examples. To choose the best fit for the hyperplane, boundary limits are created using some of the examples located next to the decision frontier, the so-called support vectors. If it is not possible to do the classification using a hyperplane directly, the kernel function is applied to map the examples in a higher dimension. The same main idea applies for regression tasks, however the algorithm goal becomes to find a hyperplane capable of include most of the samples inside the boundary limits and thus minimize an error function.

Previous papers have aimed to measure the differences in carbon stocks in the evaluated study sites according to the amount of straw left in the field (TENELLI et al., 2021). Once we intended to develop ML models with the potential to perform predictions also in different locations, for a comparison purpose the means of y in the

training set were calculated stratified by treatment and soil layer (location not specified) and values were used as predictors for y in the test set, a more straightforward but easily generalized approach.

The three prediction strategies (two ML methods and use of stratified means) were implemented with 10-fold cross-validation. The trained algorithms were then used to predict y in the test set for a performance check and to prevent overfitting.

To select the most critical attributes for prediction of *y* and to develop simpler models, we have tested three feature selection methods:

Recursive Feature Elimination (RFE): To select a subset of features from a rank of feature importances as some filter methods do have the disadvantage of ignoring possible interactions between attributes. The RFE (GUYON; WESTON; BARNHILL, 2002) is an exhaustive search method which overcomes this drawback by iteratively: I) training the chosen ML model, II) rank the feature importance according to the selected metric and III) exclude the feature with the smallest ranking criterion. These three steps are repeated continuously until only one feature lasts. The performance metric can be used alone or with cross-validation to choose the best subset.

PCA: Differently from PCA dimensionality reduction due to data transformation, in the applied technique (JOLLIFFE, 1972)**,** the main idea is that in a PCA analysis, every eigenvector whose eigenvalue ratio value is less than 0.70 has a small contribution to represent the data variance. Therefore, the most dominant variable in each one of these eigenvalues (i.e., with the highest absolute coefficient value) is also of little relevance and can be ruled out.

RF embedded feature importance: The same heuristics previously described for the RF applies. In sequence, to estimate the feature importance for prediction, the values of the attributes are randomly permuted across all trees, one attribute at a time. The bigger the increase in error, the bigger the importance of the attribute and the analyst can decide graphically or through a threshold limit (desired number of attributes or importance value) for the permanence of the features in the dataset.

For each chosen subset of attributes and model in use, the model hyperparameters were adjusted by 50 random search iterations. To RF the adjusted hyperparameters were the number of grown trees (500, 1000, 1500, 2000, 2500 or 3000), maximum tree depth (5, 6 or 7) and a maximum number of attributes used by

tree (3, 5 or the squared root of the total number of attributes). For SVM the adjusted hyperparameters were the kernel function (radial basis or polynomial), C error regularization parameter (10^{-3}, 10^{-2}, 0.1, 0.5 or 1), the kernel coefficient gamma (10^{-4}, 10^{-3}, 10^{-2}, 0.1 or 1) and degree which is only used by the polynomial kernel function (2, 3, 4 or 5).

The models were evaluated in the training and test set using Pearson correlation coefficient, coefficient of determination (R^2) and the root mean square error (RMSE) between measured and predicted values, the last calculated according to the equation 1 :

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}} \quad (1)$$

where,

$\hat{Y}$= predicted value for the class attribute; Y = Observed value and n = number of examples

3.2.4. Statistical comparison of the machine learning models

The results were compared using confidence intervals ($p < 0.05$) of the empirical frequency distribution. To do so, the 10 RMSE measurements obtained from cross-validation in the training sets were resampled with reposition (bootstrap resample) 5 10^3 times and mean value of each new resampled RMSE set was used to create the empirical frequency distribution of errors. The confidence intervals of those simulated distributions were used for comparison.

3.3. Results and Discussion

Due to the subset selection and random search tuning different hyperparameters were adopted in each trained ML model (Table 3). In general, RF has differed in number of grown trees and maximum number of attributes randomly selected for each tree. SVM only differed in gamma parameter in one of the models.

The eight models yielded similar prediction errors among them considering $p < 0.05$, with the exception of RF using the wrapper RFE which was surpassed by the

SVM associated to RF embedded feature importance selection, this last one identified as the best performance model (Figure 7).

Tabela 3. Random search hyperparameter tunning for each Random Forest and Support Vector Machine trained model.

		Adjusted hyperparameters		
Algorithm	Selection method	Number o trees	Tree depth (maximum)	Attributes per tree (maximum)
Random Forest	None (all attributes)	1500	7	5
	Recursive feature elimination	2500	7	3
	Random forest feature importance	500	7	5
	PCA	500	7	5

		Adjusted hyperparameters			
	Selection method	Kernel function	Regularization parameter	Gamma	Function degree
Support Vector Machine	None (all attributes)	Radial basis	1	5	-
	Recursive feature elimination	Radial basis	1	1	-
	Random forest feature importance	Radial basis	1	1	-
	PCA	Radial basis	1	1	-

When using the mean values for rate of carbon change according to the approximate percentages of straw left in the field and soil layer, the obtained RMSE was 0.35 Mg C ha^{-1} yr^{-1} in the training set and 0.34 C ha^{-1} yr^{-1} in the test set, with low correlation between measured and predicted y. All the ML-trained models performed better than simply assuming the mean values for the training and test set with reductions in prediction error from 55 up to 74% and 40 up to 55% for SVM and RF, respectively (Table 4). Correlations and R^2 between measured and predicted y were also consistently higher for all the implemented ML models.

Tabela 4. Prediction errors and correlations between observed and predicted values for the rate of carbon stock change in training and test set for the evaluated models.

Prediction strategy	Selection method	Training set			Test set		
		Mean RMSE ± sd	R^2	Y x Ŷ correlation	RMSE	R^2	Y x Ŷ correlation
		Mg C ha^{-1} yr^{-1}	---	---	Mg C ha^{-1} yr^{-1}	---	---
Calculated medians by the straw relative amount left (%) and soil layer	None	0.35 ± 0.07	0.16	0.32	0.34	0.09	0.32
Random Forest	None (all attributes)	0.21 ± 0.06	0.78	0.95	0.17	0.76	0.76
	Recursive feature elimination	0.21 ± 0.05	0.87	0.94	0.17	0.77	0.90
	Random forest feature importance	0.19 ± 0.06	0.91	0.91	0.15	0.81	0.81
	PCA	0.19 ± 0.06	0.92	0.97	0.15	0.83	0.93
Support Vector Machine	None (all attributes)	0.16 ± 0.05	0.94	0.94	0.09	0.93	0.98
	Recursive feature elimination	0.16 ± 0.05	0.94	0.94	0.09	0.94	0.94
	Random forest feature importance	0.14 ± 0.05	0.94	0.99	0.09	0.94	0.98
	PCA	0.19 ± 0.06	0.86	0.94	0.21	0.64	0.81

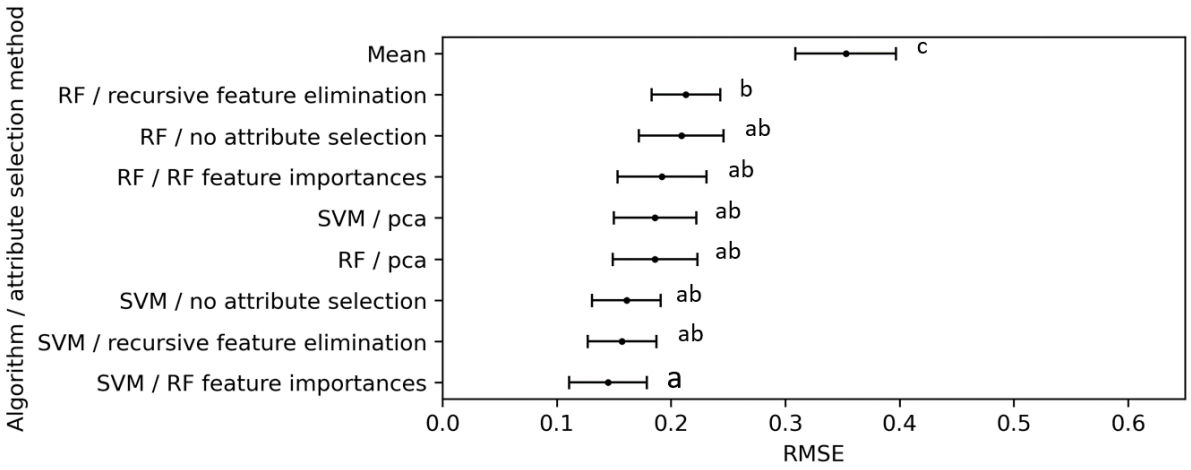

Figure 7. Statistical comparison from bootstrapped 10-fold cross validation RMSE results. Confidence interval bars followed by the same letter indicate no statistical difference ($p < 0.05$).

From a mathematical perspective, more attributes can many times be beneficial for prediction models but this is not always true. They add complexity and usually extra costs to obtain and process the data, which are not always compensated in the same proportion by the prediction improvement (FACELI et al., 2011). In this study, the obtained results corroborate this affirmation. The selected number of attributes for each model has varied largely and was just weakly correlated to performance improvement (given in RMSE reduction) considering the cross-validation prediction errors (Pearson correlation of - 0.2).

The feature selection methods have permitted reductions in the number of predictors from 70 (RFE for RF) up to 83% (PCA) in comparison to the full list of attributes (figure 8-A). A hierarchical clustering (Ward algorithm) identified two major groups of subsets: one formed by the RFE for RF and PCA (same attribute subset for both ML models) subsets and other containing RFE for SVM and RF embedded feature importance selection (same attribute subset for both ML models). Besides the number of attributes, both groups differ due to the selection of the attribute “straw mass” wich defined the experimental treatments.

It could be considered curious that “straw mass” was not used as input for all the models once previous studies conducted in the same experimental sites support that straw removal rate is directly related to soil organic carbon depletion (BLANCO-CANQUI, 2013; CHERUBIN et al., 2018b; MORAIS et al., 2020) and many others demonstrate direct relationship between crop residues intake and increase in soil organic carbon stocks (WEST; POST, 2002b). A possible explanation is that in a short-term straw removal effects in the carbon balance are difficult to be related if single thin soil layers are considered, once the mentioned previous works were mostly based on 0-10 or 0-30 cm soil layers or longer evaluation periods. Even so, the mass of crop residue was used by the the best performance model (SVM with the RF embedded feature importance selection method). Possibly other optimization based ML algorithms with adjustable input weights as Artificial Neural Networks (MCCULLOCH; PITTS, 1943) may also be a good option to capture its subtle influence.

The most frequently selected attributes were: soil layer upper limit (derived from soil layer) and K content in the 0-10 cm soil layer (6 times), P content in 0-10 cm soil layer (5 times) and soil clay content (4 times). This is suggestive that these attributes shall be considered prior to others in case of limited data collection resources for future applications similitar to the one here presented.

The top soil layers represent the interface where most of the straw decomposition takes place, so it was expected that at least one of the attributes containing its derivated information would be chosen. Crop residue decomposition also plays an important role in providing K and P to the soil. Potassium is the most abundant macronutrient in vegetal tissue and P, in turn, is also inputted to the soil in the form of soluble compounds during the decomposition process. In addition, crop residues are beneficial to preserve soil P by reducing losses by runoff, erosion and a number of other physical and chemical mechanisms (MENANDRO et al., 2017; LISBOA et al., 2018; SOLTANGHEISI et al., 2021). In the available data, the soil clay content was probably considered the best granulometry indicator, which directly affect soil carbon depletion due to straw management (CHERUBIN et al., 2021).

For the evaluated algorithms Altitude, Köppen climate classification, mean yearly temperature and straw S content were not used as inputs in any of the selected subsets (Figure 8-B).

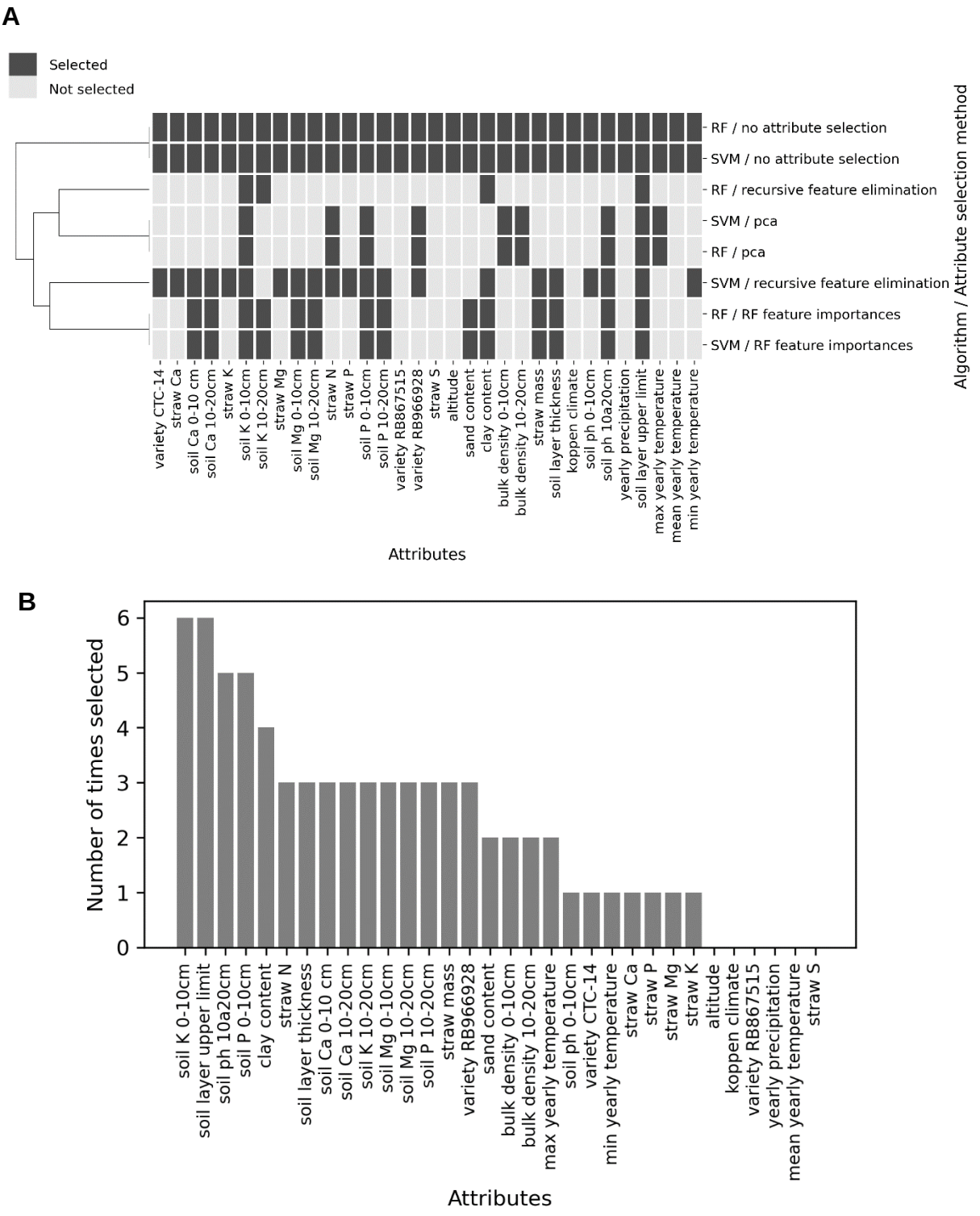

Figure 8. A) Predictive attributes in each trained model associated to a feature selection method and similarity comparizon using Ward algorithm for agglomerative hierarchical clustering. **B)** Number of times that each attribute was selected.

To better observe the advantage of using a ML method instead of adopting and generalize mean values obtained previously, the test set predictions of y from the

best model were compared to the means by soil layer and straw relative amount left in the field (Figure 9).

Assuming the application of SVM with the RF embedded feature importance selection method for prediction of the rate of carbon change (y) an approximate RMSE of 0.27 Mg C ha^{-1} yr^{-1} would be obtained for the soil first 20 cm (RMSE of 0.09 Mg C ha^{-1} yr^{-1} described in Table 3 times three for the evaluated layers 0-5, 5-10 and 10-20 cm). The same logic applied to the use of mean values from previous experiments, even considering the same straw removal proportion and soil layers would result in an error of 1.02 Mg C ha^{-1} yr^{-1} in the soil first 20 cm or, in other words, an increase in prediction error of 0.76 Mg C ha^{-1} yr^{-1}.

Lead by the observed differences in prediction accuracy, some extrapolations were imagined. The Brazilian sugarcane cropped area in 2020 was estimated at 8.6 million hectares (CONAB, 2020). Assuming an overstated possibility of partial or total straw removal in the full cropped area would cause a national prediction error of 6.5 million Mg C yr-1. Such a deviation would certainly cause severe impacts to sugarcane carbon footprint calculations and to sustainable management strategies taking soil carbon fluctuation into account.

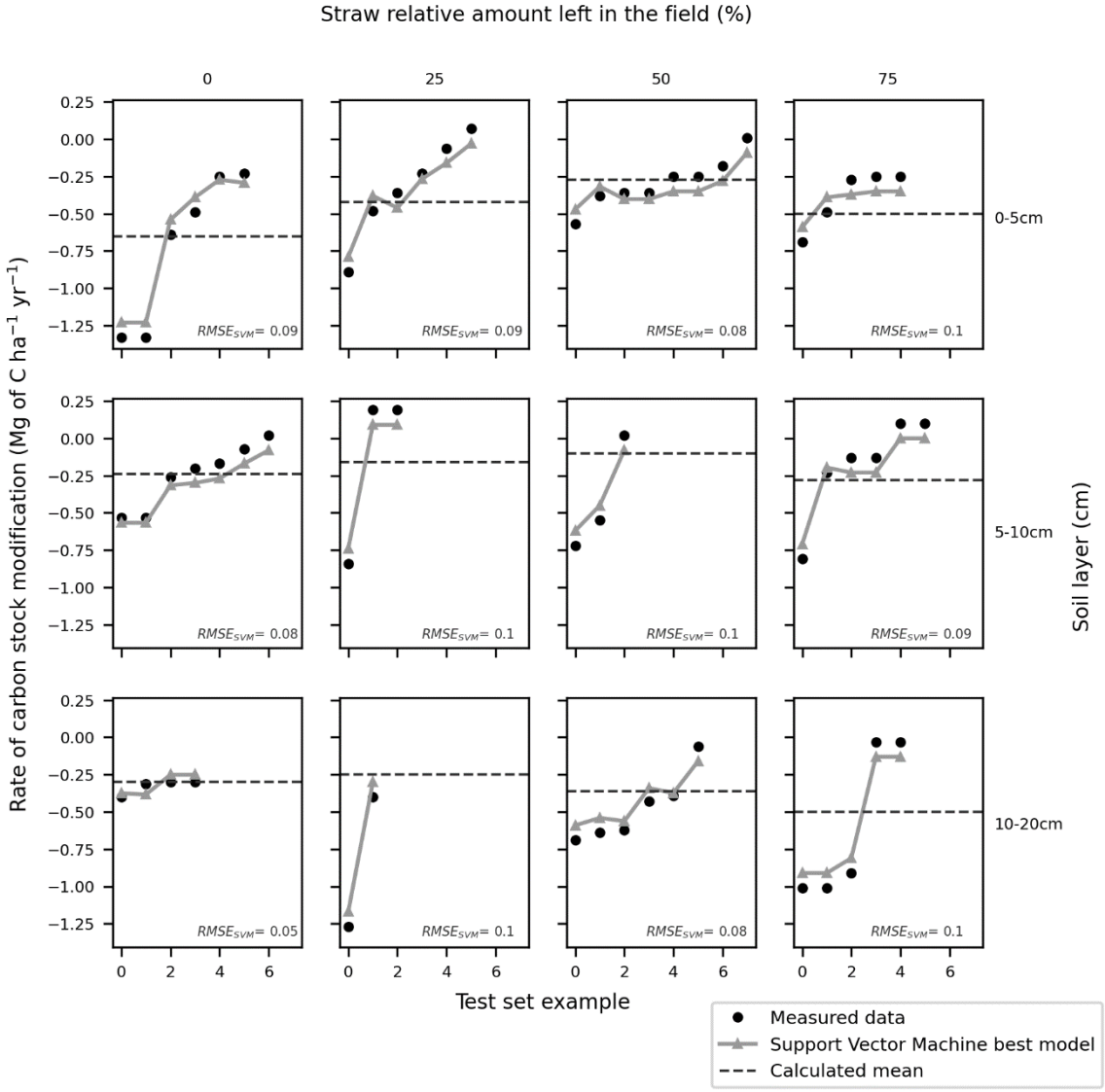

Figure 9. Calculated and predicted values of rate of carbon change in the validation set stratified by relative amount of straw left in the field and soil layer depth. Only the best models of each evaluated algorithm are exibited.

3.4. Conclusion

The adoption of ML techniques can considerably improve the capacity to predict short-term changes in soil carbon stocks due to straw total or partial remotion in relation to simple generalization of previously obtained measurements. Yet, the selection of smaller attribute subsets using ML tools did not greatly affect the prediction errors, or in some circumstances even improved the prediction power as observed when the algorithm Support Vector Machine was trained in a subset selected by Random Forest embedded feature importance.

The obtained results have the potential to impact sugarcane management practices intended to improve soil organic carbon preservation if the required attributes to serve as inputs for the models can be measured.

References

Alvares, C.A., Stape, J.L., Sentelhas, P.C., De Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorol. Zeitschrift. https://doi.org/10.1127/0941-2948/2013/0507

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D., 2019. Remote sensing techniques for soil organic carbon estimation: A review. Remote Sens. https://doi.org/10.3390/rs11060676

Austin, A.T., Yahdjian, L., Stark, J.M., Belnap, J., Porporato, A., Norton, U., Ravetta, D.A., Schaeffer, S.M., 2004. Water pulses and biogeochemical cycles in arid and semiarid ecosystems. Oecologia 141, 221–235. https://doi.org/10.1007/s00442-004-1519-1

Blanco-Canqui, H., 2013. Crop Residue Removal for Bioenergy Reduces Soil Carbon Pools: How Can We Offset Carbon Losses? Bioenergy Res. https://doi.org/10.1007/s12155-012-9221-3

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92. https://doi.org/10.1145/130385.130401

Breiman, L., 2001. Random forests. Mach. Learn. https://doi.org/10.1023/A:1010933404324

Cherubin, M.R., Bordonal, R.O., Castioni, G.A., Guimarães, E.M., Lisboa, I.P., Moraes, L.A.A., Menandro, L.M.S., Tenelli, S., Cerri, C.E.P., Karlen, D.L., Carvalho, J.L.N., 2021. Soil health response to sugarcane straw removal in Brazil. Ind. Crops Prod. 163. https://doi.org/10.1016/j.indcrop.2021.113315

Cherubin, M.R., Oliveira, D.M. da S., Feigl, B.J., Pimentel, L.G., Lisboa, I.P., Gmach, M.R., Varanda, L.L., Morais, M.C., Satiro, L.S., Popin, G.V., Paiva, S.R. de, Santos, A.K.B. dos, Vasconcelos, A.L.S. de, Melo, P.L.A. de, Cerri, C.E.P., Cerri, C.C., 2018. Crop residue harvest for bioenergy production and its implications on soil functioning and plant growth: A review. Sci. Agric. https://doi.org/10.1590/1678-992x-2016-0459

CONAB, 2020. Acompanhamento da Safra Brasileira de Cana-de-açúcar: safra 2019/2020. Obs. Agrícola 6.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. https://doi.org/10.1023/A:1022627411411

Ettema, C., 2002. Spatial soil ecology. Trends Ecol. Evol. 17, 177–183. https://doi.org/10.1016/S0169-5347(02)02496-5

Faceli, K., Lorena, A.C., Gama, J., Carvalho, A.C.P.L.F., 2011. Inteligência artificial : uma abordagem de aprendizado de máquina, Livros Técnicos e Científicos.

Farhate, C.V.V., Souza, Z.M. de, Oliveira, S.R. de M., Tavares, R.L.M., Carvalho, J.L.N., 2018. Use of data mining techniques to classify soil CO2 emission induced by crop management in sugarcane field. PLoS One 13, e0193537. https://doi.org/10.1371/journal.pone.0193537

Francaviglia, R., Di Bene, C., Farina, R., Salvati, L., 2017. Soil organic carbon sequestration and tillage systems in the Mediterranean Basin: a data mining approach. Nutr. Cycl. Agroecosystems 107, 125–137. https://doi.org/10.1007/s10705-016-9820-z

Franco, H.C.J., Pimenta, M.T.B., Carvalho, J.L.N., Magalhães, P.S.G., Rossell, C.E.V., Braunbeck, O.A., Vitti, A.C., Kölln, O.T., Rossi Neto, J., 2013. Assessment of sugarcane trash for agronomic and energy purposes in Brazil. Sci. Agric. 70, 305–312. https://doi.org/10.1590/S0103-90162013000500004

Guyon, I., Weston, J., Barnhill, S., 2002. Gene Selection for Cancer Classification using Support Vector Machines.

Jolliffe, I.T., 1972. Discarding Variables in a Principal Component Analysis. I: Artificial Data. Appl. Stat. https://doi.org/10.2307/2346488

Karlen, D.L., Johnson, J.M.F., 2014. Crop Residue Considerations for Sustainable Bioenergy Feedstock Supplies. BioEnergy Res. 7, 465–467. https://doi.org/10.1007/s12155-014-9407-y

Lal, R., 2004. Soil carbon sequestration to mitigate climate change. Geoderma. https://doi.org/10.1016/j.geoderma.2004.01.032

Lisboa, I.P., Cherubin, M.R., Lima, R.P., Cerri, C.C., Satiro, L.S., Wienhold, B.J., Schmer, M.R., Jin, V.L., Cerri, C.E.P., 2018. Sugarcane straw removal effects on plant growth and stalk yield. Ind. Crops Prod. https://doi.org/10.1016/j.indcrop.2017.11.049

Mccarl, B.A., Metting, F.B., Rice, C., 2007. Soil carbon sequestration. Clim. Change 80, 1–3. https://doi.org/10.1007/s10584-006-9174-7

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5. https://doi.org/10.1007/BF02478259

Menandro, L.M.S., Cantarella, H., Franco, H.C.J., Kölln, O.T., Pimenta, M.T.B., Sanches, G.M., Rabelo, S.C., Carvalho, J.L.N., 2017. Comprehensive assessment of sugarcane straw: implications for biomass and bioenergy production. Biofuels, Bioprod. Biorefining 11. https://doi.org/10.1002/bbb.1760

Morais, M.C., Siqueira-Neto, M., Guerra, H.P., Satiro, L.S., Soltangheisi, A., Cerri, C.E.P., Feigl, B.J., Cherubin, M.R., 2020. Trade-offs between sugarcane straw removal and soil organic matter in Brazil. Sustain. 12. https://doi.org/10.3390/su12229363

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12.

Post, W.M., Kwon, K.C., 2000. Soil carbon sequestration and land-use change: processes and potential. Glob. Chang. Biol. 6, 317–327. https://doi.org/10.1046/j.1365-2486.2000.00308.x

Soltangheisi, A., Haygarth, P.M., Pavinato, P.S., Cherubin, M.R., Teles, A.P.B., Bordonal, R. de O., Carvalho, J.L.N., Withers, P.J.A., Martinelli, L.A., 2021. Long term sugarcane straw removal affects soil phosphorus dynamics. Soil Tillage Res. 208. https://doi.org/10.1016/j.still.2020.104898

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266, 98–110. https://doi.org/10.1016/J.GEODERMA.2015.12.003

Tenelli, S., Bordonal, R.O., Cherubin, M.R., Cerri, C.E.P., Carvalho, J.L.N., 2021. Multilocation changes in soil carbon stocks from sugarcane straw removal for bioenergy production in Brazil. GCB Bioenergy. https://doi.org/10.1111/gcbb.12832

West, T.O., Post, W.M., 2002. Soil Organic Carbon Sequestration Rates by Tillage and Crop Rotation. Soil Sci. Soc. Am. J. 66, 1930. https://doi.org/10.2136/sssaj2002.1930