University of São Paulo
"Luiz de Queiroz" College of Agriculture

Leveraging the application of Earth observation data for mapping and monitoring cropland soils

José Lucas Safanelli

Thesis presented to obtain the degree of Doctor in Science. Area: Soil and Plant Nutrition

Piracicaba
2020

José Lucas Safanelli
Agronomist

Leveraging the application of Earth observation data for mapping and monitoring cropland soils

Advisor:
Prof. Dr. **JOSÉ ALEXANDRE MELO DEMATTÊ**

Thesis presented to obtain the degree of Doctor in Science. Area: Soil and Plant Nutrition

Piracicaba
2020

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

Safanelli, José Lucas

Leveraging the application of Earth observation data for mapping and monitoring cropland soils / José Lucas Safanelli. - - Piracicaba, 2020.

105 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Sensoriamento remoto 2. Mapeamento digital de solo 3. Pedometria 4. Aprendizado de máquina I. Título

To my parents, Zita A. K. Safanelli and José Carlos Safanelli.

AKNOWLEDGMENTS

I thank God for granting me health, blessing me with the strength and determination to pursue the journey of life, and for introducing me to a loving family and good people.

I am grateful to the Soils and Plant Nutrition Graduate Program, the “Luiz de Queiroz” College of Agriculture, and the University of São Paulo, for giving the opportunity to attend a doctorate course, making available the facilities and hosting a supportive staff.

I am so greatefull to the São Paulo Research Foundation (FAPESP, grants number 2016/01597-9 and 2018/21356-1), for providing the scolarship for my doctorate studies in Brazil and funding a international research internship.

To my advisor José A. M. Demattê, I thank you for trusting me, being supportive, and encouraging me with cutting-edge research ideas. I also appreciated the happy hours and relaxation times we had with the GEOCIS members.

I thank the Geo Forschungszentrum (GFZ) Potsdam, Germany, especially Dr. Sabine Chabrillat, for accepting and supervising me during my stay as a visiting student.

To my parents Zita and José Carlos, sisters Carol and Ana, and my girlfriend Nicole, I have no words to describe the love I feel and how essential you are in my life. All my family members (Safanelli and Kochan), especially José Marcos Safanelli (in memorian) and Irene Lenartovicz Kochan (in memorian), I’m so grateful for learning the fundamentals of honesty and benevolence with them.

To roomates and friends that I made during my doctorate: Lucas, André, Rogério, Michel, Felipe, Thiago, Vinícius, Isaias, Geovani, Josimar, and many others… To my colleagues from GEOCIS group… To my friends from Canoinhas, Curitibanos, and Potsdam… I’m greateful for sharing great moments with you.

I also thank all professors, mentors, and professionals in the field of agronomy, which were very supportive and acted as role models during my life, especially Prof. Alexandre ten Caten, Prof. Eduardo Bottega, Eng. Ag. Daniel Uba, Eng. Ag. João Hoffmann, professors from UFSC Curitibanos and professors from ESALQ/USP.

“When a human awakens to a great dream and throws the full force of his soul over it,
all the universe conspires in your favor.”
Johann Wolfgang von Goethe

CONTENTS

RESUMO

Potencializando a aplicação de dados de observação da Terra para o mapeamento e monitoramento de solos agrícolas

O uso e desenvolvimento sustentável de terras agrícolas requer o monitoramento contínuo e a promoção de boas práticas que preservam a qualidade do solo e o proporcionem suas diversas funções. Como a qualidade e o funcionamento do solo pode ser afetado por diversos fatores e intervenções, as quais resultam em mudanças nas escalas temporal e espacial, os sistemas de observação da Terra (OT) tornam-se uma alternativa atrativa de monitoramento devido à capacidade de fornecer dados em tempo hábil, cobrindo grandes áreas geográficas, e revisitando o mesmo lugar na Terra em curtos períodos de tempo. Além disso, como a disponibilidade de informações detalhadas sobre solos de terras agrícolas ainda é um desafio na maioria dos países, e a literatura recente tem apoiado a proposição de que as coleções de dados de OT é uma fonte valiosa para estudos ambientais, este estudo teve como objetivo explorar as coleções de imagens de satélite para o mapeamento e monitoramento de solos agrícolas em grandes extensões geográficas. Para isso, desenvolvemos as rotinas de processamento de grande volume de dados OT em uma plataforma de alto desempenho baseada na nuvem. Com a combinação de recursos extraídos de dados de OT, informações legadas de solo, e algoritmos de aprendizado de máquina, realizamos o mapeamento de média resolução de solos agrícolas sobre as extensões geográficas da Europa e do Brasil. Demonstramos neste estudo que a coleção de imagens do Landsat é uma fonte valiosa para extrair recursos espectrais úteis para o mapeamento e monitoramento solos agrícolas. Imagens de solo exposto, baseados no valor mediano de 37 anos de imagens Landsat, permitiu a predição do teor de argila e carbonatos de cálcio com desempenho moderado no continente Europeu. Além disso, usando o Google Earth Engine, desenvolvemos e disponibilizamos publicamente um pacote para calcular atributos de terreno personalizados para diferentes resoluções espaciais, a qual pode ser explorada em estudos globais. Este pacote também foi particularmente importante para preparar informações adicionais para o mapeamento de solos agrícolas no Brasil. As características extraídas dos dados de OT permitiram a predição de argila, areia, conteúdo de carbono orgânico do solo (COS) e estoque de COS com acurácia satisfatória em solos agrícolas do território brasileiro. Com os mapas resultantes, conseguimos estimar o estoque total de SOC e identificar alguns aspectos relacionados à distribuição dos atributos do solo nas principais regiões agrícolas. Portanto, este estudo apoia a proposição de que dados de OT são uma fonte valiosa para extrair características da paisagem úteis ao mapeamento e monitoraramento de solos agrícolas com resoluções mais precisas, auxiliando na avaliação da distribuição espacial do solo e no entendimento da expansão histórica da agricultura no Brasil e Europa.

Palavras-chave: Sensoriamento remoto, Mapeamento digital do solo, Pedometria, Aprendizado de máquina

ABSTRACT

Leveraging the application of Earth observation data for mapping and monitoring cropland soils

The use and sustainable development of cropland soils requires the continuous monitoring and promotion of good practices that support soil quality and the provision of its several functions. As the soil quality and functioning can be affected by several factors and interventions, resulting in changes at the temporal and spatial scales, Earth observation (EO) systems become an sound alternative for monitoring soils due to ability in providing data in a timely manner, covering large geographical areas, and revisiting the same place in Earth in short periods of time. Furthermore, as the availability of detailed information about cropland soils is still a challenge in most countries, and recent literature has been supporting the proposition that collections of EO data are a valuable source for environmental studies, this study aimed at exploring the collection of satellite images for mapping and monitoring cropland soils over large geographical areas. For this, we developed the routines for processing big EO data within a high-performance cloud-based platform. With the combination of extracted features from EO data, legacy soil datasets, and machine learning algorithms, we performed the medium-resolution mapping of cropland soils over the geographical extents of Europe and Brazil. We demonstrated in this study that the collection of Landsat images is a valuable source for extracting spectral features useful for mapping and monitoring cropland soils. The bare soil composite based on the median of 37 years of Landsat imagery allowed the prediction of clay and calcium carbonates with moderate performance in Europe. In addition to that, using the Google Earth Engine, we developed and made publicly available a package to calculate terrain attributes customized to different spatial resolutions, which can be scaled up to the global extent. This package was particularly important for preparing additional information for mapping the cropland soils in Brazil. The spectral and terrain features extracted from EO data allowed the calibration of prediction models of clay, sand, soil organic carbon (SOC) content, and SOC stock with satisfactory accuracy across the Brazilian cropland soils. With the resulting maps, we were able to estimate the total SOC stock and identify some aspects related to the distribution of soil attributes regarding the main agricultural regions. Therefore, this study supports the proposition that EO data is a valuable source for extracting environmental features for mapping and monitoring cropland soils at finer resolutions, assisting the evaluation of soil spatial distribution and the historical agriculture expansion in Europe and Brazil.

Keywords: Remote sensing, Digital soil mapping, Pedometrics, Machine learning

1. GENERAL INTRODUCTION

Soil is a natural resource that performs multiple functions for the ecosystem. It promotes water and nutrient cycling, gas exchange with the atmosphere, and supports all forms of terrestrial vegetation that impacts the animal and human life. In turn, soil quality and functioning can be affected by several factors and interventions, resulting in changes at the temporal and spatial scales. At the same time that soil and crop management have long supported food security, they also produced significant impacts on the environment with the fragmentation, diminishment, and impoverishment of natural landscapes. The improvement of managed soils for sustainable development, on the other hand, requires the continuous monitoring and promotion of good practices that supports soil quality and the provision of its several functions (Hillel, 2008; McBratney, Field and Koch, 2014).

Nowadays, Earth observation systems have become popular for monitoring the environment due to ability in providing data in a timely manner, covering large geographical areas, and revisiting the same place in Earth in short periods of time (Kuenzer *et al.*, 2014; Chabrillat *et al.*, 2019). In addition, most of these datasets are being distributed to the public with open access (Drusch *et al.*, 2012; Wulder *et al.*, 2016). Machine learning also became popular in spatial and temporal predictions due to the advances in algorithms, high-performance computing processing, and distribution of statistical packages on several programming languages and user interfaces (McBratney, Mendonça Santos and Minasny, 2003; Padarian, Minasny and McBratney, 2020; Wadoux *et al.*, 2020). For soil monitoring, these advanced frameworks are decisive because the traditional ways to analyze soils and understand its spatial and temporal distribution is time-consuming and resource-intensive (Polidoro *et al.*, 2016; Demattê *et al.*, 2019). In this sense, historical collections of Earth observation data have been explored in soil mapping and monitoring using multitemporal satellite imagery (Ben-Dor *et al.*, 2009; Mulder *et al.*, 2011). As example, multispectral reflectance of soils has been explored in the mapping of soil attributes and landscape features (Diek *et al.*, 2017; Demattê *et al.*, 2018; Rogge *et al.*, 2018; Roberts, Wilford and Ghattas, 2019), whereas the frequency of soil exposures has been applied to assess the evolution of conservation agriculture (Demattê *et al.*, 2020).

However, as the availability of detailed information about cropland soils is still a challenge in most countries, and recent literature has been supporting the proposition

that multispectral collections of satellite images are a valuable source for environmental studies, this study aimed at exploring the collection of Landsat images for mapping and monitoring cropland soils over large geographical areas. In addition to that, we also aimed at developing the processing routines of big geospatial data within a high-performance cloud-based platform. Finally, with the combination of extracted features from Earth Observation data, legacy soil datasets, and machine learning algorithms, we also aimed at performing medium-resolution mapping of cropland soils over the geographical extents of Europe and Brazil.

REFERENCES

Ben-Dor, E. *et al.* (2009) “Using Imaging Spectroscopy to study soil properties,” *Remote Sensing of Environment*. Elsevier Inc., 113, pp. S38–S55. doi: 10.1016/j.rse.2008.09.019.

Chabrillat, S. *et al.* (2019) “Imaging Spectroscopy for Soil Mapping and Monitoring,” *Surveys in Geophysics*, 40(3), pp. 361–399. doi: 10.1007/s10712-019-09524-0.

Demattê, J. A. M. *et al.* (2018) “Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images,” *Remote Sensing of Environment*, 212, pp. 161–175. doi: 10.1016/j.rse.2018.04.047.

Demattê, J. A. M. *et al.* (2019) “Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact,” *Geoderma*, 337, pp. 111–121. doi: 10.1016/j.geoderma.2018.09.010.

Demattê, J. A. M. *et al.* (2020) “Bare Earth’s Surface Spectra as a Proxy for Soil Resource Monitoring,” *Scientific Reports*, 10(1), p. 4461. doi: 10.1038/s41598-020-61408-1.

Diek, S. *et al.* (2017) “Barest Pixel Composite for agricultural areas using landsat time series,” *Remote Sensing*. Multidisciplinary Digital Publishing Institute, 9(12), p. 1245. doi: 10.3390/rs9121245.

Drusch, M. *et al.* (2012) “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” *Remote Sensing of Environment*. Elsevier, 120, pp. 25–36. doi: 10.1016/J.RSE.2011.11.026.

Hillel, D. (2008) *Soil in the Environment*. Elsevier. doi: 10.1016/C2009-0-00041-5.

Kuenzer, C. *et al.* (2014) “Earth observation satellite sensors for biodiversity monitoring: potentials and bottlenecks,” *International Journal of Remote Sensing*, 35(18), pp. 6599–6647. doi: 10.1080/01431161.2014.964349.

McBratney, A. ., Mendonça Santos, M. . and Minasny, B. (2003) “On digital soil mapping,” *Geoderma*, 117(1–2), pp. 3–52. doi: 10.1016/S0016-7061(03)00223-4.

McBratney, A., Field, D. J. and Koch, A. (2014) “The dimensions of soil security,” *Geoderma*, 213, pp. 203–213. doi: 10.1016/j.geoderma.2013.08.013.

Mulder, V. L. *et al.* (2011) “The use of remote sensing in soil and terrain mapping — A review,” *Geoderma*. Elsevier, 162(1–2), pp. 1–19. doi: 10.1016/j.geoderma.2010.12.018.

Padarian, J., Minasny, B. and McBratney, A. B. (2020) “Machine learning and soil sciences: a review aided by machine learning tools,” *SOIL*, 6(1), pp. 35–52. doi: 10.5194/soil-6-35-2020.

Polidoro, J. C. *et al.* (2016) *Programa Nacional de Solos do Brasil (PronaSolos)*. 1st ed. Rio de Janeiro, RJ: Embrapa Solos.

Roberts, D., Wilford, J. and Ghattas, O. (2019) “Exposed soil and mineral map of the Australian continent revealing the land at its barest,” *Nature Communications*, 10(1), p. 5297. doi: 10.1038/s41467-019-13276-1.

Rogge, D. *et al.* (2018) “Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014),” *Remote Sensing of Environment*. Elsevier, 205, pp. 1–17. doi: 10.1016/J.RSE.2017.11.004.

Wadoux, A. M. J.-C. *et al.* (2020) “A note on knowledge discovery and machine learning in digital soil mapping,” *European Journal of Soil Science*, 71(2), pp. 133–136. doi: 10.1111/ejss.12909.

Wulder, M. A. *et al.* (2016) “The global Landsat archive: Status, consolidation, and direction,” *Remote Sensing of Environment*. Elsevier, 185, pp. 271–283. doi: 10.1016/j.rse.2015.11.032.

2. MULTISPECTRAL MODELS FROM BARE SOIL COMPOSITES FOR MAPPING TOPSOIL PROPERTIES OVER EUROPE

ABSTRACT

Reflectance of light across the visible, near-infrared and shortwave infrared (VIS-NIR-SWIR, 0.4-2.5 µm) spectral region is very useful for investigating mineralogical, physical and chemical properties of soils, which can reduce the need for traditional wet chemistry analyses. As many collections of multispectral satellite data are available for environmental studies, a large extent with medium resolution mapping could be benefited from the spectral measurements made from remote sensors. In this paper, we explored the use of bare soil composites generated from the large historical collections of Landsat images for mapping cropland topsoil attributes across the European extent. For this task, we used the Geospatial Soil Sensing System (GEOS3) for generating two bare soil composites of 30 m resolution (named synthetic soil images, SYSI), which were employed to represent the median topsoil reflectance of bare fields. The first (framed SYSI) was made with multitemporal images (2006–2012) framed to the survey time of the Land-Use/Land-Cover Area Frame Survey (LUCAS) soil dataset (2009), seeking to be more compatible to the soil condition upon the sampling campaign. The second (full SYSI) was generated from the full collection of Landsat images (1982-2018), which although displaced to the field survey, yields a higher proportion of bare areas for soil mapping. For evaluating the two SYSIs, we used the laboratory spectral data as a reference of topsoil reflectance to calculate the Spearman correlation coefficient. Furthermore, both SYSIs employed machine learning for calibrating prediction models of clay, sand, soil organic carbon (SOC), calcium carbonates ($CaCO_3$), cation exchange capacity (CEC), and pH determined in water, using the gradient boosting regression algorithm. The original LUCAS laboratory spectra and a version of the data resampled to the Landsat multispectral bands were also used as reference of prediction performance using VIS-NIR-SWIR multispectral data. Our results suggest that generating a bare soil composite displaced to the survey time of soil observations did not improve the quality of topsoil reflectance, and consequently, the prediction performance of soil attributes. Despite the lower spectral resolution and the variability of soils in Europe, a SYSI calculated from the full collection of Landsat images can be employed for topsoil prediction of clay and $CaCO_3$ contents with a moderate performance (testing R^2, root mean square error (RMSE) and ratio of performance to interquartile range (RPIQ) of 0.44, 9.59, 1.77, and 0.36, 13.99, 1.54, respectively). Thus, this study shows that although there exist some constraints due to the spatial and temporal variation of soil exposures and among the Landsat sensors, it is possible to use bare soil composites for mapping key soil attributes of croplands across the European extent.

Keywords: remote sensing; digital soil mapping; Google Earth Engine; landsat; LUCAS topsoil data; machine learning

Published as: Safanelli, J.L.; Chabrillat, S.; Ben-Dor, E.; Demattê, J.A.M. Multispectral Models from Bare Soil Composites for Mapping Topsoil Properties over Europe. Remote Sensing 2020, 12(9), 1369; doi.org/10.3390/rs12091369

2.1. INTRODUCTION

Reflectance of light across the visible, near-infrared and shortwave infrared (VIS-NIR-SWIR, 0.4-2.5 µm) spectral region is a valuable property for understanding the nature and composition of many materials. Many studies have shown that this method is very useful for investigating mineralogical, physical and chemical properties of soils, reducing the need of traditional wet chemistry analyses [1–3]. Accordingly, soil spectral libraries (SSL) from the VIS-NIR-SWIR range have become popular around the globe and have been largely studied in combination with chemometrics and machine-learning methods for estimating soil attributes [3–7]. For soil mapping, the large coverage of SSLs from many geographical areas can increase the accuracy of the final maps by covering the high variation of soil types [8]. Nonetheless, some factors regarding the measurement methods (e.g., illumination and geometrical setup, environmental conditions, sample condition, sensor characteristics, etc.) that are governed by the acquisition strategy (e.g., laboratory, field, airborne, and spaceborne) may impact the outcomes.

Protocols of spectral acquisition are distinct in soil research, but many efforts emerged for providing ways of standardizing data acquisition from different setups and sources, mainly for laboratory-level acquisition [3,9]. Protocols for soil reflectance acquisition from other domains (field, air and space) are not yet available, while the laboratory measurements (e.g., the SSL) are considered the most accurate; however, they cannot be directly applied to data collected from air or space [10]. Whereas laboratory conditions allow for better control of reflectance measurements, airborne and satellite sensors are critically influenced by in situ conditions and other factors [11–13]. However, detailed mapping of large spatial extents can be performed using measurements made from large remote-sensing spectral surveys.

Soil reflectance from the remote-sensing domain is limited by the availability of exposed surfaces caused by natural and human-induced factors, such as soil tillage [10,14]. Usually when airborne imagery is taken over bare soil areas, the data acquisition happens under optimal environmental conditions (high sun illumination, clear atmosphere, dried and green-vegetation free) following ground truth measurements under optimal sun radiation, or as recently suggested by [15], using a special assembly that mimics field soil surface spectra in laboratory conditions (SoilPRO®). When Earth Observation satellite sensors are involved, then exposed soil

surface might be limited due to the spatial and temporal dynamics of land use (e.g. crop development). Despite the limited availability of bare surfaces and the unknown field conditions upon image acquisition, archives of satellite images contain historical information about soil exposure that could be explored in automated processing algorithms to generate bare soil representations by combining multitemporal measurements. Recently, several methods for generating bare soil images from historical collections of satellite images from one satellite have been developed, such as the barest pixel composite [16], Soil Composite Mapping Processor (SCMAP) [17], and Geospatial Soil Sensing System (GEOS3) [18], which were applied to develop soil maps at different scales, with more or less success with the modeling of soil properties. The quality of reflectance of the bare soil composites provided depends on whether the adverse conditions that happened upon image acquisition can be minimized (or normalized) and the variability of soil surface can be represented by different shades and colors. Evaluation methods comparing the association with laboratory spectra (used as a reference of soil signal) can provide a way of assessing the quality of bare soil images determined from multitemporal and multispectral remote-sensing means, but a complete correction of effects is still challenging for automated processing systems.

The possibility of transferring prediction models from laboratory to satellite images is another gap that is still in progress for mapping large geographical extents. Transferring prediction models from a SSL to a satellite bare soil composite could be an alternative for making spatially explicit maps using a reflectance image, but the contrast between calibration and application levels, the sampling design and sample support might cause a huge impact on prediction accuracy [11,19]. Some studies have been exploring this approach for mapping small regions using hyperspectral imagery (e.g., [4,20]). In such cases, reference samples measured in different conditions or instruments provided a way to handle the measurements' variability [21,22]. Another effect that might influence the results of the predictions is the temporal discrepancy that can exist between the field surveys and the historical satellite data collection used for generating the bare soil image. This issue can hinder the mapping of dynamic soil properties, such as soil organic carbon, soil moisture, and soil salinity, since the satellite bare soil composites are composed of several years of images.

In this study, we aimed at assessing: a) if bare soil composites generated from the large historical collections of several Landsat satellites can be calculated at the

European (European Union, EU) extent, providing a reliable estimate of topsoil reflectance; b) the effects of generating bare soil composites from a different time frame than the Land-Use/Land-Cover Area Frame Survey (LUCAS) soil data, which was surveyed in 2009; c) if the EU-wide bare soil composite can be and to which degree employed for predicting soil properties using the LUCAS soil database combined with a machine learning approach, despite all the aforementioned issues.

2.2. MATERIAL AND METHODS

2.2.1. Bare Soil Composites

Bare soil composites were produced for an extent covering most of the European countries, situated between longitude –12 and 34 degrees, and latitude of 33 and 73 degrees. The algorithm Geospatial Soil Sensing System (GEOS3 [18]), developed within the Google Earth Engine [23], was used to generate the multispectral bare soil composites and the soil exposure frequencies from the collections of 30-m-resolution Landsat images from 1982 to 2018. The GEOS3 was slightly modified to adapt to several Landsat satellites with variable spectral characteristics but considering that the multispectral bands are positioned in equivalent spectral regions and their small differences do not affect the bare soil composites (Table 1 from Appendix A). GEOS3 is a data-mining algorithm that extracts soil features from the collection of historical images and aggregates the spatially bare soil fragments into a synthetic soil image (SYSI). The SYSI is the reflectance image of the bare soil composite, while the frequency of soil exposure is denominated as soil frequency (SF). SF is determined by the proportion of a given pixel location was identified as bare soil to the total number of pixel occurrences from the time interval of the collection. To identify bare soil pixels from single satellite images, a set of identification rules were used. They were based on spectral indices coupled with quality assessment bands, which removed cloud, cloud shadow, inland water, snow, photosynthetic vegetation and non-photosynthetic vegetation (crop residues). In this study, a pixel was flagged soil when it had the Normalized Difference Vegetation Index (NDVI, Equation (1)) values falling between the range of –0.05 and 0.30 (masking out green vegetation), and Normalized Burn Ration 2 index (NBR2, Equation (2)) values between the range of –0.15 and 0.15 (masking out crop residues). These thresholds were defined based on histogram and

density plot analysis calculated from the LUCAS spectral measurements (Figure 1 from Appendix A). The flagged soil pixels were used to select each reflectance band on each acquisition time. Then the bare soil composite was composed by aggregating the multitemporal bare soil pixels by their median value. More detailed information about GEOS3, the spectral indices and sensitivity analysis of spectral indices thresholds are described in [18] or elsewhere [13,24–26].

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

$$NBR2 = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2} \quad (2)$$

where red (~630–690 nm), near infrared (NIR: ~760-900 nm), shortwave infrared 1 (SWIR1: ~1550–1750 nm) and shortwave infrared 2 (SWIR2: ~2080–2350 nm) are the harmonized spectral bands from Landsat 4 Thematic Mapper (L4 TM), Landsat 5 Thematic Mapper (L5 TM), Landsat 7 Enhanced Thematic Mapper Plus (L7 ETM+), and Landsat 8 Operational Land Imager (L8 OLI), respectively.

The bare soil composites (SYSIs) were produced using a harmonized and merged collection of surface reflectance images from the L4 TM (available images from 1982 to 1993), L5 TM (available images from 1984 to 2012), L7 ETM+ (available images from 1999 to present), and L8 OLI (available images from 2013 to present) available in Google Earth Engine datasets catalog [27–29]. Surface reflectance bands of each sensor were harmonized to a common band name that represented the same spectral range between the sensors (Table 1 from Appendix A). Although the relative spectral responses of instruments are slightly different and may cause effects on a time-series analysis [30], no significant effects were identified after merging the Landsat collections for aggregating SYSI over time. Some studies have also used this merging approach for increasing the availability of surface reflectance images, e.g., in [25,31], and consequently, improving the bare soil representation.

2.2.2. Reflectance Evaluation and Soil Dataset

To test the influence of temporal time frame considered for the calculation of the bare soil composite, two SYSIs were produced considering two different

collections. The first SYSI was produced using a temporal subset from the full Landsat archive defined by 3 years before and after the LUCAS field survey of 2009 (2006–2012), which was called framed SYSI. The second SYSI was generated considering the full-time interval (1982–2018), being called full SYSI. We have generated the framed SYSI to check if significant changes would become evident when comparing its performance to the full SYSI. The two SYSI were compared using the correlation analysis with the reference topsoil spectra determined in laboratory, as well as by the performance after prediction. For evaluating the consistency of both SYSIs by the correlation analysis, we resampled the LUCAS spectra to the mean multispectral response of the four Landsat sensors and used the Spearman rank correlation analysis [32]. We used rank correlation analysis because we compared two domains (satellite and laboratory) that have different sensor and measurement characteristics, which interfere with the spectral response of soils. Therefore, this method was used to minimize the domain discrepancies and check if the rank of a sample from the first domain (laboratory) correspond to the rank of the same sample in the second domain (satellite). Despite the differences on the signal to noise ratio, the soil conditions and the temporal variability of satellite images when comparing laboratory and remote sensing data, the correlation coefficient gives an opportunity to understand if the patterns displayed in SYSI are linked to the soil reflectance, which considers in this case the laboratory measurements as the reference of the soil spectral response. For correlation analysis and prediction models, image data was sampled by intersecting the LUCAS coordinates with the SYSI bands.

Absorbance spectra and attributes data from the topsoil samples (0–20 cm) of the EU LUCAS database surveyed in 2009 were used in this study [33]. Approximately 20,000 topsoil samples were collected in 25 EU member states (EU-27 except Bulgaria and Romania). The soil sampling was undertaken within the frame of the Land-Use/Land-Cover Area Frame Survey, which represents one million points distributed in a grid of 2 × 2 km. The sample collected at each location followed a composite sampling strategy which comprised five topsoil (0–20 cm) subsamples that were mixed to form a single composite sample. The first subsample was taken at the coordinate point of the pre-established LUCAS point, whereas the remaining four are taken 2 m from the central one following the cardinal directions (North, East, South and West). Vegetation residues, grass, and litter, if present, were removed from the surface before sampling and from the composite sample. Soil samples have been analyzed for basic

soil properties, including particle size distribution (soil texture), pH, organic carbon, carbonates, nitrogen, phosphorus, potassium, cation exchange capacity (CEC) and absorbance spectra, which was determined in the full continuous spectrum from 400 to 2500 nm and spectral resolution of 2 nm [33].

In this study, only samples from the main land cover type of croplands (category B) were selected from the 20,000 samples because most of the bare soils come from croplands, thus they have a more homogenized soil layer on the surface due to tillage operations. In the end, only 7142 samples were used from the original LUCAS dataset (Figure 2 in Appendix A). As the LUCAS spectral data are delivered in absorbance (A), we transformed the spectra to reflectance (R) by R = 1/10^A in order to correspond to the reflectance data of SYSIs. Besides the correlation analysis, the reflectance of LUCAS was also used to calibrate two reference models for comparing the prediction performance. For the first reference model, the LUCAS reflectance spectra (from 400 to 2500 nm) without any preprocessing method were transformed by principal component analysis (PCA) to reduce the spectral dimensionality, where the first five components that had a cumulative variance of more than 99% were submitted to model calibration. For the second reference model, the LUCAS reflectance spectra were resampled to the Landsat multispectral bands using the relative spectral responses of the four Landsat sensors [34], which were averaged to a single multispectral dataset for model calibration and correlation analysis.

We also assessed the median reflectance generated from the full SYSI by inspecting its reflectance dispersion when the topsoil was identified as bare by GEOS3. Three random sites from France, Germany and Spain that were available in the EU LUCAS soil dataset (LUCAS IDS 9219, 1392, and 4364, respectively) were used for comparing subset images of the original true color composition and their respective bare soil masks. In this visualization step, three random scenes identified as bare in the L5 TM, L7 ETM + and L8 OLI collections were buffered by 1 km around the LUCAS geographical coordinate in order to visually compare the bare soil masked by GEOS3. At the same sites, the minimum, median, maximum and 0.25 and 0.75 percentiles of the reflectance were collected, making possible the evaluation of the soil reflectance dispersion. Site characteristics were also provided together with the spectral response to complement the SYSI reflectance evaluation.

2.2.3. Prediction Models of Soil Properties

Clay, sand, soil organic carbon (SOC), calcium carbonate (CaCO3), pH determined in water (pH H2O), and cation exchange capacity (CEC) data were selected to make prediction models. For this, we randomly split the dataset into training (80%) and test sets (20%). The prediction models were calibrated based either on the reflectance from the framed or the full SYSI, using the reflectance bands as predictors: blue (~450-520 nm), green (~520-600 nm), red (~630-690 nm), NIR (~760-900 nm), SWIR1 (~1550-1750 nm) and SWIR2 (~2080-2350 nm). Quantile regression from the gradient boosting trees (GBT) of scikit-learn Python library was used as the machine learning algorithm for calibrating prediction models [35]. Model tuning was performed using 10-fold cross-validation for the 0.50 percentile (median) of the training set trying different combinations of hyperparameters submitted to a grid search [36,37]. Learning rate was optimized from the set of 0.10, 0.15 and 0.20 units in order to control the magnitude of learning with the increasing number of trees. The number of estimators tested were 100, 250 and 500 trees to control the maximum number of trees for learning. These hyperparameters impact the forest used to ensemble the estimates. The maximum depth was tested in the set of respectively 5, 8 and 10 layers. The minimum samples at each split were optimized in the set of respectively 50, 100 and 200 samples. The minimum samples of each leaf were tested in the set of 5, 10, 20 samples. The maximum features used in each individual tree split was optimized in the set of 2, 4 and 6 predictors (6 is equals to the maximum number of predictors, i.e., the reflectance bands). These latter hyperparameters control individual trees in the GBT algorithm and are used to prevent specific learning of samples and reduce overfitting [38], making more robust and generalization models for spatially predicting the full geographical area. The best estimator was defined by the minimum root mean square error (RMSE, Equation 3) from the 10-fold cross-validation of the training set, and the final hyperparameters are presented in Table 2 of Appendix A.

Before the model calibration, the soil attributes values were logit-transformed to constrain the predicted range to a defined maximum and minimum limit [39], with the predictions back transformed for generating the maps and assessing the performance. To assess the model performance, the following parameters were calculated from the test set: the RMSE (Equation 3) was measured to evaluate the model’s inaccuracy; the coefficient of determination (R2, Equation 4) was calculated to

evaluate the explained variance of models; and the ratio of performance to interquartile range (RPIQ, Equation 5) was estimated to assess the consistency between the predicted values and the testing dataset variability [40].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

$$RPIQ = \frac{IQR_y}{RMSE_{\hat{y}}} \quad (5)$$

where y is the vector of measured values, $\hat{y}$ is the vector of predicted values, $\bar{y}$ is the mean of vector y, and IQR is the interquartile range defined by the differences of 75th and 25th percentiles.

2.2.4. Spatial Prediction and Uncertainty

For the development of maps of soil properties, the best spectral model derived from the framed or full SYSI was selected from the evaluation metrics and employed for predicting cropland soils across the European extent. In this step, we used the CORINE land-cover map of 2012 (version 20) to restrict our predictions, considering in this case, the grouping of ‘non-irrigated arable land’, ‘permanently irrigated land’, ‘rice fields’, and ‘annual crops associated with permanent crops’ classes [41]. Soil attribute maps of the median estimate (0.50 percentile) and 90% prediction interval defined by 0.05 and 0.95 percentiles were produced. Uncertainty at the pixel level was defined as the ratio between the 90% prediction interval to the median estimate, which was then converted to the percent scale. Thus, as the uncertainty is standardized to the median estimate, we can compare both the spatial patterns and the differences between the uncertainty maps. Additionally, to visually assess the regional variability of the predicted clay map, four different locations were selected based on the spatial variability of soils and previous works developed at the same sites [13,42–45].

The predictions were made for separated tiles of 1 × 1 degree using multi-core processing. Overviews (levels 2 [pixel resolution of 60 m], 4 [pixel resolution of 120 m],

and 8 [pixel resolution of 240 m]) and virtual raster were used for mosaicking the tiles and displaying the original 30-m-resolution maps over the European extent. The statistical analysis, machine learning and map visualizations were performed using free and open source software: Python 3.6 [46], GDAL [47], R 3.4 [48], and Quantum GIS 3.4 [49].

2.3. RESULTS

2.3.1. Bare Soil Composites

The full SYSI produced across the croplands within the European extent reveals different patterns represented by the true color composition (Figure 1a). Soil surfaces in Southern Europe are brighter than the other regions, probably because of the semi-arid conditions relatively dominated by iron oxides, sand and clays [44]. Conversely, the Eastern part of the map have a darker shade, which could be linked to soil types with dark surface horizons, such Chernozems and Phaeozems [50]. The frequency of soil exposure (Figure 1b), which represents how many times a site was identified as bare during the time interval of the multitemporal collection, also gives an opportunity to understand the degree of soil disturbance, which can be linked to natural factors, such as the type of land cover, or even human-induced factors, such as the disturbances caused by tillage operations. The remaining white areas in both panels are due to the absence of bare surface information that was masked by a cropland reference map of 2012. Despite the fact that framed SYSI had, in general, similar patterns as the full SYSI, framed SYSI yielded a lower proportion of bare soils and is not presented in this section. However, the statistical evaluation of both SYSIs is described below.

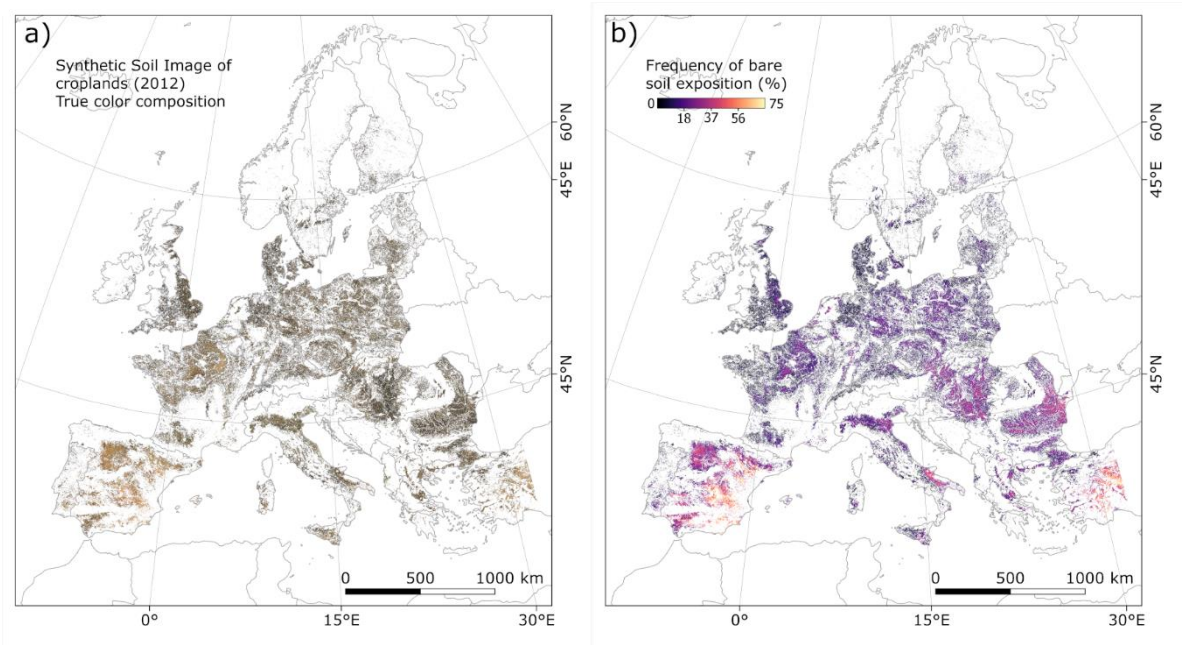

Figure 1. a) Full synthetic soil image (SYSI) of croplands, over the European extent; b) the equivalent bare soil frequency (SF).

2.3.2. Soil Dataset and Reflectance Evaluation

The summary statistics of the soil attributes used in this study are demonstrated in Table 1. Clay values ranged from 1% to 79%, with a mean value of 22% and standard deviation (SD) of almost 13%, indicating a higher predominance of medium textured soils in the dataset (mean and SD sand values: 36% and 25%, respectively). SOC values were variable and ranged from 0 to 43.84%, with a low mean value of 1.68% and SD of 1.56%, probably because of the absence of higher SOC content samples that are common in other land cover classes, such as grasslands and forests. Calcium carbonate content ($CaCO_3$) varied between 0% and 88%, with a mean value of 9%, while pH determined in water had a mean value close to 7 (SD of 1.01). The mean value of cation exchange capacity (CEC) was estimated in 15.30 $cmol_c\ kg^{-1}$, with a SD of 9.40 $cmol_c\ kg^{-1}$.

Table 1 also shows the summary statistics of the different reflectance sources and highlights the discrepancies between LUCAS resampled reflectance with the median reflectance of framed and full SYSI. Laboratory spectral measurements resampled to Landsat multispectral range had a higher reflectance intensity for all the bands, with the mean value being in general twice the value of the reflectance of bare soil composites. The contrasting intensities can be linked to different measurement

and soil conditions of the two acquisition levels, where in the laboratory the illumination and geometrical factors, and sample conditions, are controlled. Also, there could be a mixing of patterns in the field-of-view at the satellite level that reduces the overall reflectance. The field-of-view and the signal-to-noise ratio of sensors is another factor that is very contrasting between laboratory and spaceborne sensors, even for field reflectance measurements.

Simple correlations between the laboratory and bare soil composite bands were moderate with coefficients varying from 0.49 to 0.66 (Table 2). Although the reflectance values have different amplitude, as demonstrated in Table 1, we can see that the correlation between both the sources still exists. Framed SYSI had a slightly lower correlation with the laboratory reflectance than the full SYSI. For the framed SYSI, the correlation coefficients ranged from 0.49 to 0.62, while for the full SYSI they ranged from 0.53 to 0.66. This result gives a first idea that framing the generation of a bare soil composite to the soil sample survey time might not improve the reflectance accuracy when using the GEOS3 methodology, which is based on the median reflectance of the bare soil pixels. Here, an assumption is that the longer time frame of the full SYSI provides a more stable median reflectance that is less affected by dynamic effects of the bare soils. Additionally, there could be a limitation of generating a bare soil composite using a shorter historical collection, which would retrieve only a few bare soil exposures along the time series that would not be sufficient for estimating a robust median value closer to ideal bare soil conditions (e.g., dried, vegetation-free). Therefore, we selected the full SYSI as the best representative of the bare topsoil reflectance for the further analyses.

Table 1. Descriptive statistics of soil samples subset from the Land-Use/Land-Cover Area Frame Survey (LUCAS) topsoil database (n = 7142).

Variable[1]	Min.[2]	Mean	SD[3]	Median	IQR[4]	Max.[5]
Soil Attributes						
Clay (%)	1.00	21.94	12.58	21.00	16.00	79.00
Sand (%)	1.00	36.02	25.12	31.00	41.00	97.00
SOC (%)	0.00	1.68	1.56	1.38	0.94	43.84
$CaCO_3$ (%)	0.00	9.01	15.91	0.30	11.30	88.20
pH H_2O	3.55	7.05	1.01	7.33	1.58	8.93
CEC ($cmol_c\ kg^{-1}$)	0.00	15.30	9.40	13.80	11.30	188.10
Resampled reflectance from laboratory						
Blue	0.03	0.15	0.05	0.14	0.05	0.50
Green	0.04	0.21	0.06	0.20	0.08	0.61
Red	0.05	0.27	0.07	0.27	0.10	0.67
NIR	0.10	0.36	0.08	0.36	0.10	0.75
SWIR1	0.17	0.48	0.08	0.48	0.10	0.81
SWIR2	0.15	0.45	0.07	0.45	0.09	0.74
Reflectance from framed SYSI[6]						
Blue	0.03	0.08	0.02	0.08	0.02	0.15
Green	0.04	0.12	0.03	0.12	0.03	0.23
Red	0.03	0.15	0.04	0.15	0.05	0.33
NIR	0.05	0.23	0.05	0.23	0.07	0.43
SWIR1	0.02	0.30	0.06	0.29	0.08	0.54
SWIR2	0.02	0.24	0.05	0.24	0.07	0.43
Reflectance from full SYSI						
Blue	0.04	0.08	0.02	0.08	0.02	0.14
Green	0.04	0.12	0.03	0.12	0.03	0.22
Red	0.04	0.15	0.04	0.15	0.05	0.33
NIR	0.05	0.23	0.05	0.23	0.06	0.43
SWIR1	0.03	0.29	0.06	0.29	0.08	0.53
SWIR2	0.03	0.24	0.05	0.24	0.06	0.42

[1]Soil attributes: soil organic carbon (SOC), calcium carbonate ($CaCO_3$), pH determined in water solution (pH H_2O), cation exchange capacity (CEC). Reflectance bands: blue (~450–520 nm), green (~520–600 nm), red (~630–690 nm), near infrared (NIR: ~760–900 nm), shortwave infrared 1 (SWIR1: ~1550–1750 nm) and shortwave infrared 2 (SWIR2: ~2080–2350 nm). [2]Minimum value. [3]Standard deviation (SD). [4]Interquartile range (IQR). [5]Maximum value. [6]Synthetic soil image (SYSI).

Table 2. Spearman correlation between laboratory resampled reflectance and synthetic soil images.

Correlation[1]	Blue	Green	Red	NIR	SWIR1	SWIR2
Resampled~Framed SYSI	0.60	0.62	0.62	0.60	0.59	0.49
Resampled~Full SYSI	0.63	0.66	0.66	0.65	0.63	0.53
Framed SYSI~Full SYSI	0.90	0.93	0.94	0.93	0.94	0.93

[1]All correlations are significant at $p < 0.05$. Reflectance bands: blue (~450–520 nm), green (~520–600 nm), red (~630–690 nm), near infrared (NIR: ~760–900 nm), shortwave infrared 1 (SWIR1: ~1550–1750 nm) and shortwave infrared 2 (SWIR2: ~2080–2350 nm).

Another way of assessing the median reflectance generated from the full SYSI was by inspecting the dispersion of the reflectance when the surface was identified as bare by GEOS3 (Figure 2). Different sampling sites from France, Germany and Spain from the LUCAS dataset were used in this additional evaluation. The land use and the geographical characteristics of the sites affects the amount of bare soil that can be extracted by GEOS3 (Figure 2a), regardless the Landsat sensor. For example, the selected site in Spain (LUCAS ID 4364) provided a higher proportion of bare soils than the sites from Germany and France, which can be probably linked to the dryer climate and also to a higher intensive land use. Despite the occurrence of extreme values as demonstrated by the minimum and maximum values in the spectral plots (Figure 2b), we can observe that the median statistics provide a reasonable estimate of bare soil reflectance, with a lower dispersion defined by the interquartile range (0.75 and 0.25 percentiles, represented by the gray shadow). The site in France has the highest median intensity, with a higher decrease of reflectance between 2000 and 2500 nm, which can be linked to its higher clay and $CaCO_3$ contents. Overall, full SYSI provided a good estimate for the bare soil reflectance and was used in the further steps of the study for predicting cropland soil attributes across the European extent.

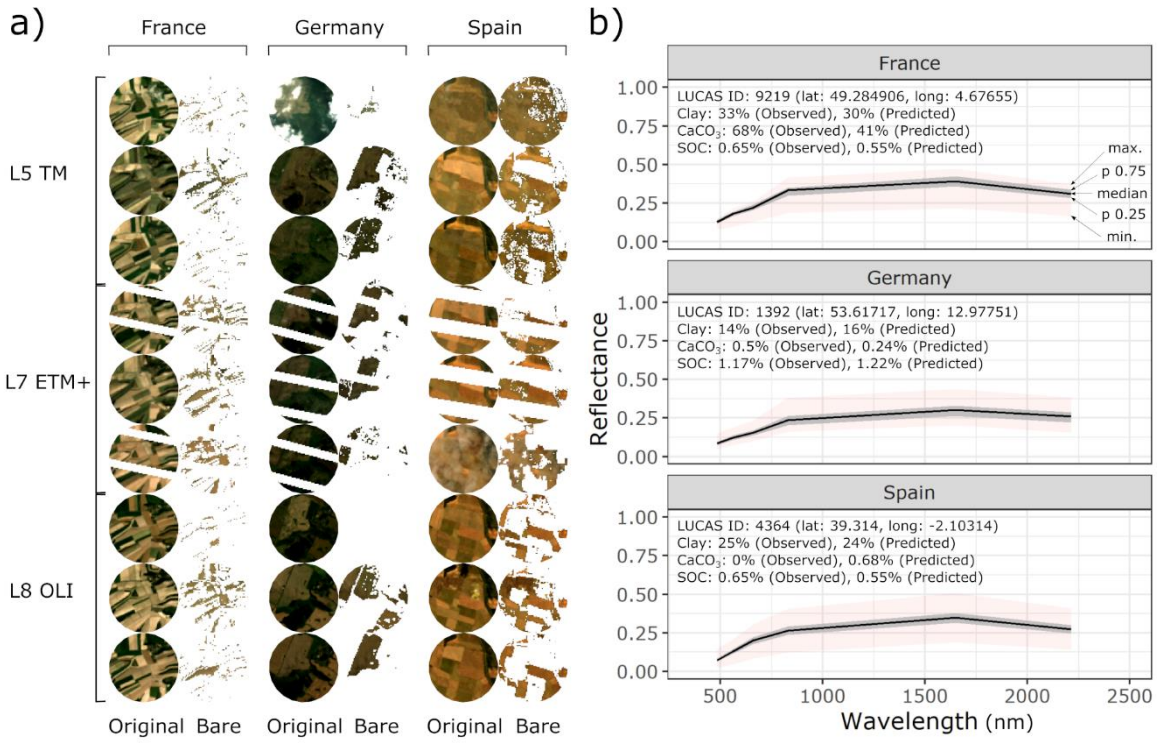

Figure 2. a) Example of original scenes (true color composition) and bare soil masks from the Landsat collection (left panels), in this case considering the Landsat 5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and Landsat 8 Operational Land Imager (OLI), centered at the three LUCAS sampling points and with a circle buffer of 1000 m; **b)** Dispersion of the full SYSI reflectance for the same sites, where the minimum, 0.25 percentile, median, 0.75 percentile and maximum values are provided. Full SYSI reflectance is defined by the median estimate, attenuating the influence of extreme values. Soil site characteristics are provided together with the spectral patterns.

2.3.3. Prediction Models

Table 3 shows the performance of the gradient boosting tree regressions for each of the six soil attributes using four different reflectance data. The LUCAS original and resampled reflectance datasets deliver the best performances for almost all the attributes in both training and testing sets, confirming its superior composition for calibrating spectral prediction models. However, the soil organic carbon was unable to be predicted (R2 < 0.35 and RPIQ < 1.5 in training and testing sets) even considering the LUCAS laboratory data. In particular, texture attributes and calcium carbonate content had the best calibration and testing performance for all datasets (R2 > 0.35 and RPIQ > 1.5), except for framed SYSI. Additionally, we can identify from the testing results that there is an inferior performance for both SYSI models compared to the reference models of laboratory data. This result can be associated to measurement

characteristics and soil surface conditions of satellite acquisition level, and possibly to the effect of integrating pixel values to point coordinates (support change). Nonetheless, the results demonstrate that the full SYSI can still be used as a covariate for building prediction models based on its median reflectance values. Another point that is worth mentioning is that framing the SYSI to soil survey period does not improve the prediction performance, suggesting that the full SYSI have the most reliable estimate of the median topsoil reflectance after using a denser historical collection (37 years).

Table 3. Performance of prediction models of soil properties (n = 7142) using reflectance data.

Attribute[1]	Reflectance data[2]	[3]R^2	RMSE[4]	RPIQ[5]	R^2	RMSE	RPIQ
		Training set (80%)			Testing set (20%)		
Clay (%)	Original	0.80	5.57	2.87	0.58	8.35	2.04
	Resampled	0.78	5.80	2.76	0.49	9.18	1.85
	Framed SYSI	0.53	8.58	1.87	0.36	10.28	1.65
	Full SYSI	0.67	7.20	2.22	0.44	9.59	1.77
Sand (%)	Original	0.66	14.57	2.81	0.42	19.26	2.23
	Resampled	0.72	13.27	3.01	0.37	20.07	2.14
	Framed SYSI	0.56	16.54	2.42	0.22	22.39	1.92
	Full SYSI	0.68	14.10	2.84	0.25	21.93	1.96
SOC (%)	Original	0.35	1.09	0.86	0.24	1.52	0.58
	Resampled	0.25	1.31	0.72	0.13	1.62	0.54
	Framed SYSI	0.10	1.44	0.66	0.04	1.69	0.52
	Full SYSI	0.16	1.39	0.68	0.06	1.68	0.52
$CaCO_3$ (%)	Original	0.59	10.97	1.70	0.54	11.89	1.82
	Resampled	0.76	8.48	2.30	0.47	12.70	1.70
	Framed SYSI	0.47	12.56	1.55	0.31	14.50	1.49
	Full SYSI	0.51	12.18	1.60	0.36	13.99	1.54
pH H_2O	Original	0.62	0.63	2.48	0.39	0.80	2.05
	Resampled	0.62	0.62	2.52	0.31	0.85	1.93
	Framed SYSI	0.46	0.74	2.10	0.14	0.94	1.73
	Full SYSI	0.45	0.75	2.08	0.21	0.90	1.80
CEC ($cmol_c\ kg^{-1}$)	Original	0.70	4.38	2.32	0.38	7.66	1.46
	Resampled	0.66	5.35	2.11	0.32	8.02	1.39
	Framed SYSI	0.39	7.16	1.58	0.22	8.60	1.30
	Full SYSI	0.54	6.24	1.81	0.28	8.25	1.35

[1]Soil attributes: Soil organic carbon (SOC), calcium carbonate ($CaCO_3$), pH determined in water solution (pH H_2O), and cation exchange capacity (CEC). [2]Synthetic Soil Image (SYSI); Original and Resampled terms refer to the LUCAS absorbance spectra (450 to 2500 nm) used as reference prediction models, where the original was converted to reflectance and reduced by principal component analysis, and the resampled was converted to reflectance and resampled to the Landsat multispectral bands. [3]Coefficient of determination (R^2). [4]Root mean squared error (RMSE). [5]Ratio of performance to interquartile range (RPIQ).

2.3.4. Spatial Predictions and Uncertainty

The predicted clay content of croplands across the European geographical extent (using the full SYSI) showed the predominance of soils with low to medium clay content (Figure 3a). Soils with low clay estimates prevail from the central of Western

to Eastern Europe. Conversely, soils rich in clay are more evident in some regions of the United Kingdom, Spain, Italy and the Southeastern of Europe. The uncertainty of predictions (Figure 3b), estimated as the 90% prediction interval standardized to the median estimate, reveals that croplands with higher estimates of clay can have in some cases, a high uncertainty, i.e., there is a high variation around the predicted median value. This observation coincides with the croplands in the North of Italy and in the United Kingdom.

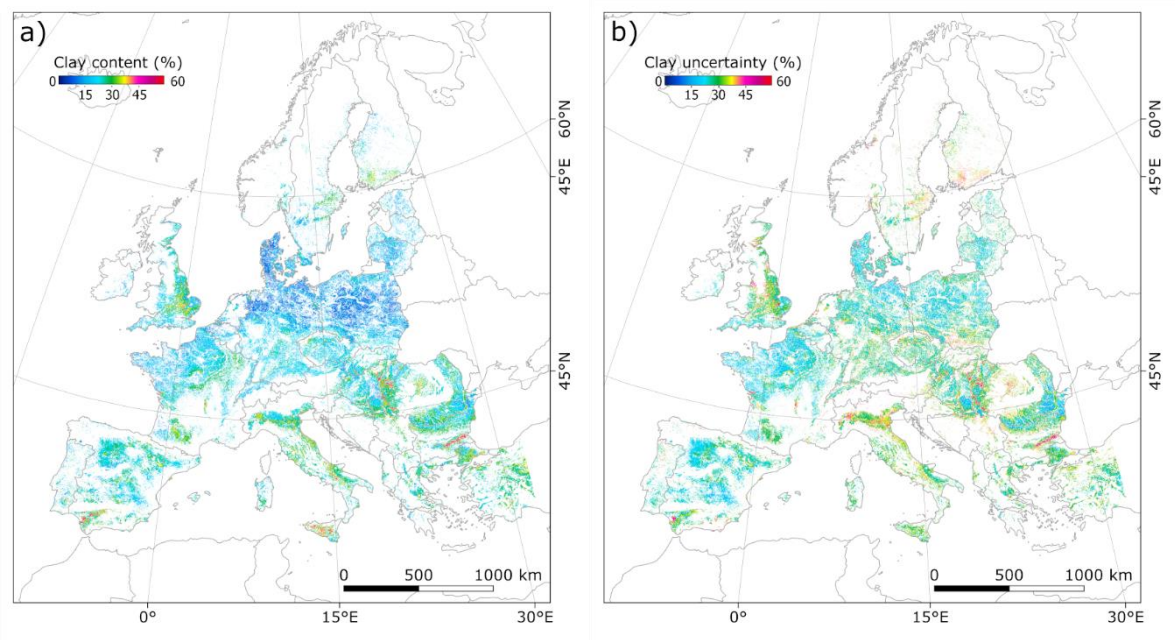

Figure 3. Soil European clay map using the full synthetic soil image (SYSI) as model predictor; b) the clay uncertainty map.

Calcium carbonate content was also estimated across the croplands within the European extent using the full SYSI (Figure 4a). In this map, lower carbonate contents predominate across the European extent, which is possibly related to low abundance (mean and median values) already expressed in the samples from the LUCAS dataset (Table 1). The Champagne region in France and some soils of Spain had the highest estimates of $CaCO_3$, although the uncertainty was also considered moderately high (Figure 4b). Other sites in Italy, Greece and the Mediterranean region also had significant estimates of $CaCO_3$ as displayed by light blue shades in Figure 4a. These mapped areas coincide with calcium-rich lithology that forms $CaCO_3$ rich soils, suggesting that multispectral reflectance from bare soil composites can be relevant for mapping this soil attribute over croplands.

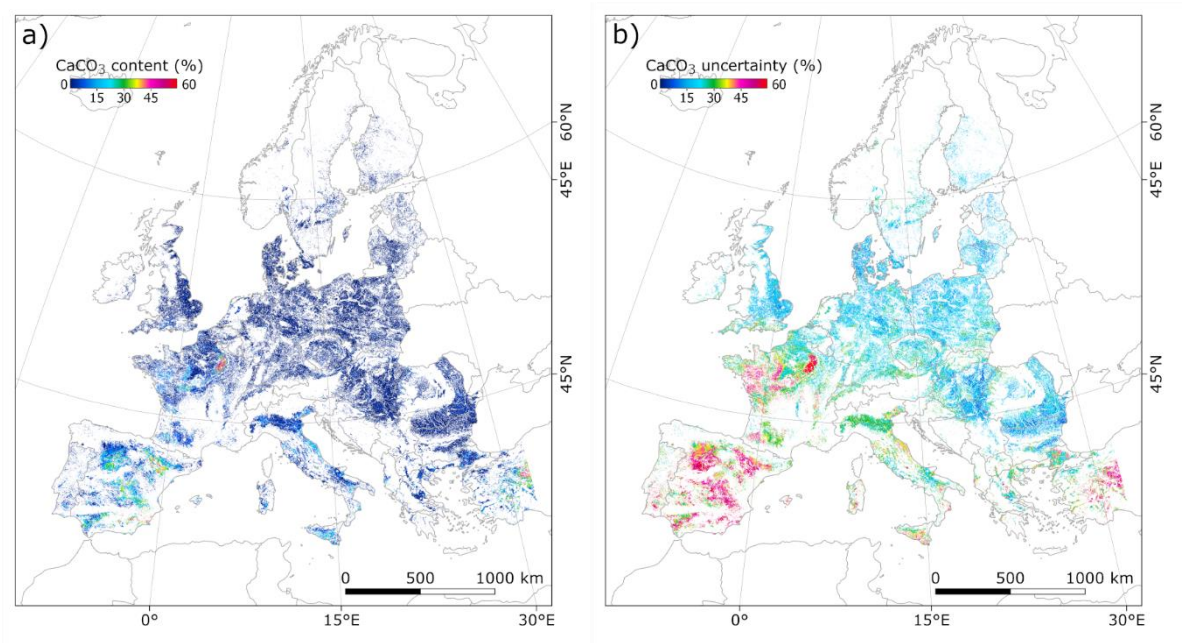

Figure 4. Soil European CaCO3 map using the full synthetic soil image (SYSI) as model predictor; b) the CaCO3 uncertainty map.

The clay maps were also checked at the regional level by inspecting the predictions produced from the full SYSI (Figure 5). Four different regions across the European extent were selected to check the spatial variability of clay, which included sites from Germany (Figure 5a), Spain (Figure 5b), France (Figure 5c), and Italy (Figure 5d). The regional maps seem to aid the recognition of the local variability of clay content, which can be important for a more efficient management of croplands, for example. The spatial patterns in these images can be linked to lithological variability, which affects the soil texture. In general, the comparison of these regional mapping fits well with previous published results for these regions [13,42–45]. The true color composites (full SYSI) for the same sites of Figure 5 are provided in Figure 3 in Appendix A.

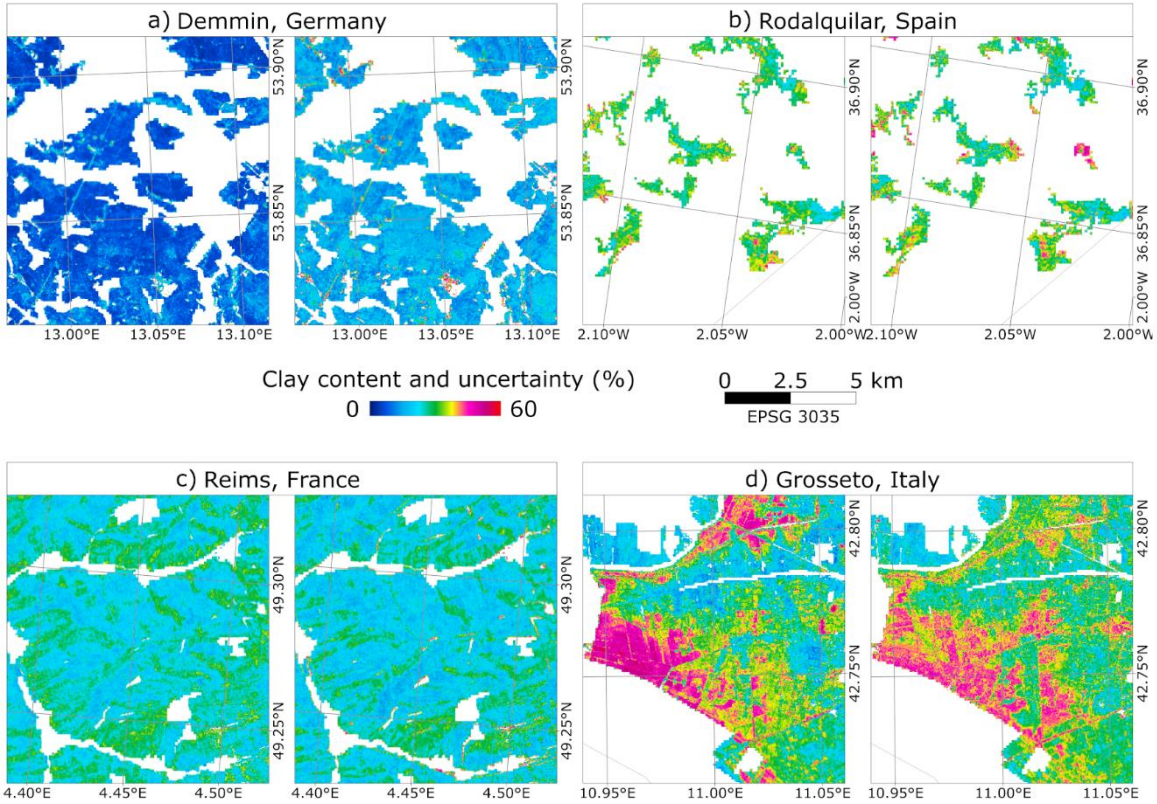

Figure 5. a) Regional clay map (left) and uncertainty (right) near Demmin, Germany; b) regional clay map (left) and uncertainty (right) near Rodalquilar, Spain; c) regional clay map (left) and uncertainty (right) near Reims, France; d) regional clay map (left) and uncertainty (right) near Grosseto, Italy. Note: the regional maps were masked by croplands of the CORINE 2012 map.

2.4. DISCUSSION

Framing the SYSI generation to the soil observation survey was a more reasonable approach for generating a bare soil composite, especially for predicting dynamic soil properties, such as SOC and chemical soil properties. However, the median reflectance estimated by the full SYSI had a higher correlation with the reference spectra collected in the laboratory, which resulted in more accurate predictions by machine learning. Some studies have already pointed out that denser collections of historical images increase the probability of retrieving more bare soils over a given area, which increases the mapped area and the quality of topsoil reflectance [16,18,31].

After employing the reflectance from bare soil composites to machine learning, the evaluation metrics revealed that $CaCO_3$ and clay attributes yielded the best prediction performances (Table 2). Many studies have demonstrated that clay and calcium carbonates have distinctive absorption characteristics in VIS-NIR-SWIR region

[11,51,52] including albedo, shape, and absorption features, which could explain why these two soil attributes had the highest accuracies. In the bare soil composites, although specific absorption features are not depicted due to the lower resolution of multispectral data, the intensity and shape were influenced by soil constituents, in accordance with the findings of [53]. Furthermore, studies that compared different acquisition levels for predicting either clay or carbonate topsoil contents confirm that laboratory-based spectra give the best estimates, although field and aerial data can also be employed for depicting the spatial variability of clay and $CaCO_3$ [11,52].

In the work performed by [11], the effectiveness of continuum-removed absorption features for clay and $CaCO_3$ prediction from different hyperspectral acquisition levels was tested. The spectral consistency changes from laboratory to aerial sensors were also assessed in that study. Their results confirmed that simple models based on absorption features are efficient in predicting clay and $CaCO_3$ estimates regardless of the source [34]. Our results obtained from a multispectral approach confirmed that it is possible to map these two soil attributes using the bare soil reflectance and a large SSL as ground data for calibration. However, the prediction of soil attribute using multispectral reflectance relies on the total apparent reflectance value (albedo) rather than specific absorption features [53], which are usually employed in a specific group of soils that are measured by the same protocol (sample preparation, spectroradiometer, and lightening configuration).

Soil spectral libraries (employed in laboratory) have been extensively explored in soil spectroscopy and digital soil mapping and have become an alternative for traditional wet analysis for some specific attributes. Transferring multispectral or hyperspectral prediction models from laboratory to bare soil images seems to be challenging. The uncertainty is high because not only do the acquisition means hamper the predictions, but also the temporal and surface conditions degrade the signal of soils in satellite images [11,20,54]. Differences in reflectance intensities make difficulty for the transfer of a prediction model without standardizing the data, as the model's coefficients of one level might yield biased estimates on the other. In addition, there are other aspects that pose limitations when integrating soil spectral libraries to multispectral bare soil composites. Since the soil spectral reflectance is determined by several soil properties that usually vary in space and time, the sampling design must consider the variability of soils in adequate scales, i.e., usually at the regional or local level. It is also important to take into consideration the temporal changes that can

happen for some soil attributes, such as SOC, pH and soil elements. These characteristics can affect model calibration, resulting in poor validation performance and the generalization of predictions, similarly to what was found for SOC, pH and CEC in this study (Table 3). Furthermore, the change of support from pixels to coordinate points may also be assessed when integrating SSL to bare soil images. The optimal condition for integrating these data takes into account the pixel variability within a certain extent around the point coordinate, followed by the fitting of a spatial model to predict a value exactly at the point coordinate [1,55]. This approach could correct the differences between the point and pixel support but are still subjected to some additional definitions, such as the extent size around the point coordinate (number of pixels) and other spatial model parameters. Similarly, the subsampling design of soil samples at the coordinate points may also affect the integration, i.e., whether the point was comprised of a single instance (point support) or many composite samples (areal support). Thus, as these additional factors can impact the results derived from bare soil composites, they may be addressed in future works.

Nonetheless, there is still room for exploring the transferring approach using bare soil composites, mainly considering forthcoming hyperspectral images and more advanced machine learning frameworks [56,57]. Bare soil composites derived from hyperspectral imagers, either aerial or orbital, can be exploited in future investigations because they provide images with more spectral bands and higher spectral resolution, which can be associated to specific absorption features. This is the case of the current hyperspectral in orbit PRISMA (PRecursore IperSpettrale della Missione Applicativa [58]) and of the other upcoming hyperspectral imagers, such as the German EnMAP (Environmental Mapping and Analysis Program [59]). Other machine-learning frameworks and standardization methods can also be investigated on transferring prediction models from laboratory to satellite levels. This could be the case of convolution neural networks using the model transfer approach by fine tuning the models to the acquisition and/or spatial domains [20,54]. In the work performed by [20], the researches successfully transferred a clay prediction model calibrated from LUCAS laboratory spectral data to a hyperspectral aerial image covering a small region in Spain, reaching an R^2 of 0.60 and RMSE of 8.62% using convolutional neural networks.

2.5. CONCLUSIONS

This study supports the proposition that bare soil composites can be generated over the European extent for developing topsoil prediction models of clay and calcium carbonates of cropland soils. Our approach used the median reflectance of 37 years of Landsat imagery for reducing extreme estimates along the multitemporal survey. Further, prediction models were established using gradient boosting tree regressions coupled with a subset of the EU LUCAS soil dataset. We propose that this approach can be added to digital soil mapping seeking to improve the topsoil prediction of croplands at a regional or local level, since the topsoil of this land use is more homogeneous due to tillage practices. We also found that generating a bare soil composite displaced to the survey time of soil survey did not affect the prediction accuracy of relatively stable soil attributes, i.e., clay and calcium carbonates. In fact, a denser historical collection increases the chances of retrieving more exposed surfaces, which improves the representation of diverse soils.

ACKNOWLEDGMENTS

J.L.S. is grateful to the German Research Center for Geosciences (GFZ-Potsdam, section 1.4) for accepting and providing facilities as a visiting Ph.D. student. The authors are also grateful to the Geotechnologies in Soil Science group, and to the European Soil Data Centre for making available the LUCAS topsoil data. This research was funded by São Paulo Research Foundation, grants number 2014/2262-2, 2016/01597-9 and 2018/21356-1.

REFERENCES

1. Ben-Dor, E.; Chabrillat, S.; Demattê, J.A.M.; Taylor, G.R.; Hill, J.; Whiting, M.L.; Sommer, S. Using Imaging Spectroscopy to study soil properties. *Remote Sens. Environ.* **2009**, *113*, S38–S55.
2. Summers, D.; Lewis, M.; Ostendorf, B.; Chittleborough, D. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. *Ecol. Indic.* **2011**, *11*, 123–131.

3. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Brown, D.J.; Demattê, J.A.M.; Shepherd, K.D.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; et al. A global spectral library to characterize the world's soil. *Earth-Science Rev.* **2016**, *155*, 198–230.
4. Nocita, M.; Stevens, A.; Toth, G.; Panagos, P.; van Wesemael, B.; Montanarella, L. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biol. Biochem.* **2014**, *68*, 337–347.
5. Gholizadeh, A.; Saberioon, M.; Carmon, N.; Boruvka, L.; Ben-Dor, E. Examining the Performance of PARACUDA-II Data-Mining Engine versus Selected Techniques to Model Soil Carbon from Reflectance Spectra. *Remote Sens.* **2018**, *10*, 1172.
6. Demattê, J.A.M.; Dotto, A.C.; Paiva, A.F.S.; Sato, M. V.; Dalmolin, R.S.D.; de Araújo, M. do S.B.; da Silva, E.B.; Nanni, M.R.; ten Caten, A.; Noronha, N.C.; et al. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* **2019**, *354*, 113793.
7. Ng, W.; Minasny, B.; Montazerolghaem, M.; Padarian, J.; Ferguson, R.; Bailey, S.; McBratney, A.B. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* **2019**, *352*, 251–267.
8. Ramirez-Lopez, L.; Wadoux, A.M.J.-C.; Franceschini, M.H.D.; Terra, F.S.; Marques, K.P.P.; Sayão, V.M.; Demattê, J.A.M. Robust soil mapping at the farm scale with vis–NIR spectroscopy. *Eur. J. Soil Sci.* **2019**, *70*, 378–393.
9. Ben-Dor, E.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245–246*, 112–124.
10. Chabrillat, S.; Ben-Dor, E.; Cierniewski, J.; Gomez, C.; Schmid, T.; van Wesemael, B. Imaging Spectroscopy for Soil Mapping and Monitoring. *Surv. Geophys.* **2019**, *40*, 361–399.
11. Lagacherie, P.; Baret, F.; Feret, J.-B.; Madeira Netto, J.; Robbez-Masson, J.M. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sens. Environ.* **2008**, *112*, 825–835.
12. Diek, S.; Chabrillat, S.; Nocita, M.; Schaepman, M.E.; de Jong, R. Minimizing soil moisture variations in multi-temporal airborne imaging spectrometer data for digital soil mapping. *Geoderma* **2019**, *337*, 607–621.

13. Castaldi, F.; Chabrillat, S.; Don, A.; van Wesemael, B. Soil Organic Carbon Mapping Using LUCAS Topsoil Database and Sentinel-2 Data: An Approach to Reduce Soil Moisture and Crop Residue Effects. *Remote Sens.* **2019**, *11*, 2121.
14. Ustin, S.L.; Roberts, D.A.; Gamon, J.A.; Asner, G.P.; Green, R.O. Using Imaging Spectroscopy to Study Ecosystem Processes and Properties. *Bioscience* **2004**, *54*, 523–534.
15. Ben-Dor, E.; Granot, A.; Notesco, G. A simple apparatus to measure soil spectral information in the field under stable conditions. *Geoderma* **2017**, *306*, 73–80.
16. Diek, S.; Fornallaz, F.; Schaepman, M.E.; de Jong, R. Barest Pixel Composite for agricultural areas using landsat time series. *Remote Sens.* **2017**, *9*, 1245.
17. Rogge, D.; Bauer, A.; Zeidler, J.; Mueller, A.; Esch, T.; Heiden, U. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). *Remote Sens. Environ.* **2018**, *205*, 1–17.
18. Demattê, J.A.M.; Fongaro, C.T.; Rizzo, R.; Safanelli, J.L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sens. Environ.* **2018**, *212*, 161–175.
19. Tao, C.; Wang, Y.; Cui, W.; Zou, B.; Zou, Z.; Tu, Y. A transferable spectroscopic diagnosis model for predicting arsenic contamination in soil. *Sci. Total Environ.* **2019**, *669*, 964–972.
20. Liu, L.; Ji, M.; Buchroithner, M. Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery. *Sensors* **2018**, *18*, 3169.
21. Andrew, A.; Fearn, T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 51–56.
22. Feudale, R.N.; Woody, N.A.; Tan, H.; Myles, A.J.; Brown, S.D.; Ferré, J. Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 181–192.
23. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27.

24. Gallo, B.; Demattê, J.; Rizzo, R.; Safanelli, J.; Mendes, W.; Lepsch, I.; Sato, M.; Romero, D.; Lacerda, M. Multi-Temporal Satellite Images on Topsoil Attribute Quantification and the Relationship with Soil Classes and Geology. *Remote Sens.* **2018**, *10*, 1571.
25. Poppiel, R.R.; Lacerda, M.P.C.; Safanelli, J.L.; Rizzo, R.; Oliveira, M.P.; Novais, J.J.; Demattê, J.A.M. Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil. *Remote Sens.* **2019**, *11*, 2905.
26. Fongaro, C.; Demattê, J.; Rizzo, R.; Lucas Safanelli, J.; Mendes, W.; Dotto, A.; Vicente, L.; Franceschini, M.; Ustin, S. Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images. *Remote Sens.* **2018**, *10*, 1555.
27. USGS Landsat 8 Surface Reflectance Code LaSRC Product Guide. Available online: https://www.usgs.gov/media/files/landsat-8-surface-reflectance-code-lasrc-product-guide (accessed on 11 March 2019).
28. USGS Landsat 4-7 Surface Reflectance Code LEDAPS Product Guide. Available online: https://www.usgs.gov/media/files/landsat-4-7-surface-reflectance-code-ledaps-product-guide (accessed on 11 March 2019).
29. Wulder, M.A.; White, J.C.; Loveland, T.R.; Woodcock, C.E.; Belward, A.S.; Cohen, W.B.; Fosnight, E.A.; Shaw, J.; Masek, J.G.; Roy, D.P. The global Landsat archive: Status, consolidation, and direction. *Remote Sens. Environ.* **2016**, *185*, 271–283.
30. Chastain, R.; Housman, I.; Goldstein, J.; Finco, M.; Tenneson, K. Empirical cross sensor comparison of Sentinel-2A and 2B MSI, Landsat-8 OLI, and Landsat-7 ETM+ top of atmosphere spectral characteristics over the conterminous United States. *Remote Sens. Environ.* **2019**, *221*, 274–285.
31. Roberts, D.; Wilford, J.; Ghattas, O. Exposed soil and mineral map of the Australian continent revealing the land at its barest. *Nat. Commun.* **2019**, *10*, 5297.
32. Hollander, M.; Wolfe, D.A.; Chicken, E. *Nonparametric Statistical Methods*; Volume 751.; John Wiley & Sons, 2013; ISBN 1118553292.
33. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the largest expandable soil dataset for Europe: a review. *Eur. J. Soil Sci.* **2018**, *69*, 140–153.
34. Ben-Dor, E.; Banin, A. Evaluation of several soil properties using convolved TM spectra. In *Monitoring Soils in the Environment with Remote Sensing and GIS*; ORSTOM: Paris, France, 1996; pp. 135–149.

35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
36. Folberth, C.; Baklanov, A.; Balkovič, J.; Skalský, R.; Khabarov, N.; Obersteiner, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. For. Meteorol.* **2019**, *264*, 1–15.
37. Schratz, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol. Modell.* **2019**, *406*, 109–120.
38. Dev, V.A.; Eden, M.R. Formation lithology classification using scalable gradient boosted decision trees. *Comput. Chem. Eng.* **2019**, *128*, 392–404.
39. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93.
40. Malone, B.P.; Minasny, B.; McBratney, A.B. *Using R for Digital Soil Mapping*; Progress in Soil Science; Springer International Publishing: Cham, Switzerland 2017; ISBN 978-3-319-44325-6.
41. *European Landscape Dynamics*; Feranec, J., Soukup, T., Hazeu, G., Jaffrain, G., Eds.; CRC Press: Boca Raton, FL, USA, 2016; ISBN 9781315372860.
42. Escribano, P.; Schmid, T.; Chabrillat, S.; Rodríguez-Caballero, E.; García, M. Optical Remote Sensing for Soil Mapping and Monitoring. In *Soil Mapping and Process Modeling for Sustainable Land Use Management*; Elsevier: Chennai, India, 2017; pp. 87–125.
43. Bianchini, S.; Solari, L.; Soldato, M.D.; Raspini, F.; Montalti, R.; Ciampalini, A.; Casagli, N. Ground Subsidence Susceptibility (GSS) Mapping in Grosseto Plain (Tuscany, Italy) Based on Satellite InSAR Data Using Frequency Ratio and Fuzzy Logic. *Remote Sens.* **2019**, *11*, 2015.
44. Steinberg, A.; Chabrillat, S.; Stevens, A.; Segl, K.; Foerster, S. Prediction of Common Surface Soil Properties Based on Vis-NIR Airborne and Simulated EnMAP Imaging Spectroscopy Data: Prediction Accuracy and Influence of Spatial Resolution. *Remote Sens.* **2016**, *8*, 613.
45. Meersmans, J.; Martin, M.P.; Lacarce, E.; De Baets, S.; Jolivet, C.; Boulonne, L.; Lehmann, S.; Saby, N.P.A.; Bispo, A.; Arrouays, D. A high resolution map of French soil organic carbon. *Agron. Sustain. Dev.* **2012**, *32*, 841–851.

46. Python Software Foundation. Python Language Reference, version 3.6. 2016. Available online: https://www.python.org/ (accessed on 11 March 2019).
47. GDAL/OGR contributors. GDAL/OGR Geospatial Data Abstraction Software Library. 2019. Available online: https://gdal.org (accessed on 11 March 2019).
48. R Core Team. R: A language and environment for statistical computing. 2018. Available online: https://www.r-project.org/ (accessed on 11 March 2019).
49. QGIS Development Team. QGIS Geographic Information System. 2019. Available online: http://qgis.osgeo.org (accessed on 11 March 2019).
50. Jones, A.; Panagos, P.; Barcelo, S.; Bouraoui, F.; Bosco, C.; Dewitte, O.; Gardi, C.; Erhard, M.; Hervás, J.; Hiederer, R. *The state of soil in Europe*; A Contribution of the JRC to the European Environment Agency's Environment State and Outlook Report. European Commission: Luxembourg, 2012.
51. Ben-Dor, E.; Banin, A. Near-Infrared Reflectance Analysis of Carbonate Concentration in Soils. *Appl. Spectrosc.* **1990**, *44*, 1064–1069.
52. Gomez, C.; Lagacherie, P.; Coulouma, G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* **2008**, *148*, 141–148.
53. Ben-Dor, E.; Banin, A. Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm). *Int. J. Remote Sens.* **1995**, *16*, 3509–3528.
54. Padarian, J.; Minasny, B.; McBratney, A.B. Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma* **2019**, *340*, 279–288.
55. Lobell, D.B.; Lesch, S.M.; Corwin, D.L.; Ulmer, M.G.; Anderson, K.A.; Potts, D.J.; Doolittle, J.A.; Matos, M.R.; Baltes, M.J. Regional-scale Assessment of Soil Salinity in the Red River Valley Using Multi-year MODIS EVI and NDVI. *J. Environ. Qual.* **2010**, *39*, 35–41.
56. Castaldi, F.; Chabrillat, S.; Jones, A.; Vreys, K.; Bomans, B.; van Wesemael, B. Soil Organic Carbon Estimation in Croplands by Hyperspectral Remote APEX Data Using the LUCAS Topsoil Database. *Remote Sens.* **2018**, *10*, 153.
57. Ben-Dor, E.; Patkin, K.; Banin, A.; Karnieli, A. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data - a case study over clayey soils in Israel. *Int. J. Remote Sens.* **2002**, *23*, 1043–1062.

58. Loizzo, R.; Guarini, R.; Longo, F.; Scopa, T.; Formaro, R.; Facchinetti, C.; Varacalli, G. Prisma: The Italian Hyperspectral Mission. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018; pp. 175–178.
59. Guanter, L.; Kaufmann, H.; Segl, K.; Foerster, S.; Rogass, C.; Chabrillat, S.; Kuester, T.; Hollstein, A.; Rossner, G.; Chlebek, C.; et al. The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation. *Remote Sens.* **2015**, *7*, 8830–8857.

3. TERRAIN ANALYSIS IN GOOGLE EARTH ENGINE: A METHOD ADAPTED FOR HIGH-PERFORMANCE GLOBAL-SCALE ANALYSIS

ABSTRACT

Terrain analysis is an important tool for modeling environmental systems. Seeking to use the cloud-based computing capabilities of Google Earth Engine (GEE), we customized an algorithm for calculating terrain attributes, such as slope, aspect, and curvatures, for different resolution and geographical extents. The calculation method is based on geometry and elevation values estimated within a 3x3 spheroidal window and it does not rely on projected elevation data. Thus, partial derivatives of terrain are calculated considering the great circle distances of reference nodes of the topographic surface. The algorithm was developed using the JavaScript programming interface of the online code editor of GEE and can be loaded as a custom package. The algorithm also provides an additional feature for making the visualization of terrain maps with a dynamic legend scale, which is useful for mapping different extents: from local to global. We compared the consistency of the proposed method with an available but limited terrain analysis tool of GEE, which resulted in a correlation of 0.89 and 0.96 for aspect and slope over a near-global scale, respectively. In addition to this, we compared the slope, aspect, horizontal and vertical curvature of a reference site (Mount Ararat) to their equivalent attributes estimated on the System for Automated Geospatial Analysis (SAGA), which achieved a correlation between 0.96 and 0.98. The visual correspondence of TAGEE and SAGA confirms its potential for terrain analysis. The proposed algorithm can be useful for making terrain analysis scalable and adapted to customized needs, benefiting from the high-performance interface of GEE. The package code and a minimal reproducible example are available in <https://github.com/zecojls/tagee>.

Keywords: topographic surface; terrain modeling; global terrain dataset.

Published as: Safanelli, J.L.; Poppiel, R.R.; Ruiz, L.F.C.; Bonfatti, B.R.; Mello, F.A.O.; Rizzo, R.; Demattê, J.A.M. Terrain Analysis in Google Earth Engine: A Method Adapted for High-Performance Global-Scale Analysis. ISPRS International Journal of Geo-Information. 2020, 9(6), 400; doi.org/10.3390/ijgi9060400

3.1. INTRODUCTION

Terrain analysis is essential for modeling environmental systems [1–3]. The variability of landforms is frequently used to understand, map or model geomorphological, hydrological, and biological processes [4–7]. Elevation has a strong relationship with terrestrial temperature, vegetation type, and with the potential energy accumulated on a slope. The aspect and derived products, such as Northernness and Easternness attributes, can be linked to the potential solar irradiation on terrain. The

Slope gradient, for example, controls the overland and subsurface flow velocity and runoff rate. Similarly, curvatures are associated with acceleration and dispersion of water and sediment flows, which impacts the erosion and soil water content [8].

The public availability of elevation data with global coverage, such as the digital elevation model (DEM) derived from NASA's Shuttle Radar Topography Mission (SRTM DEM, [9]) and the digital surface model from the Advanced Land Observing Satellite (AW3D30 DSM, [10]), has promoted the exploration of topographic features in different contexts using processing tools available in several geographic information systems (GIS) [4,11,12]. However, despite the popularization of many global elevation datasets, it is important to pay attention to their quality when used for modelling purposes, as the acquisition mean and other production aspects can significantly impact the outputs [13,14]. In addition, analyzing big geospatial datasets can pose some limitations to traditional GIS. This becomes more critical with the availability of new digital datasets, which are providing better temporal and spatial resolutions due to advances in sensor technologies [15].

The Global Multi-resolution Terrain Elevation Data 2010 [16] and the global suit of terrain attributes [2] are examples of datasets that were produced using large computational tasks for mapping the global extent and in different spatial resolutions, which demanded optimized processing architectures. In general, high performance architectures are based on splitting the data in smaller subsets (tiles) to take the advantage of distributed computing operations. Recently, with the advent and popularization of cloud-based interfaces for processing big geospatial data, e.g., Google Earth Engine [17], the Pangeo software packages [18], and Actinia REST service [19], computational tasks applied to terrain analysis could be scaled and customized directly by the user.

Earth Engine (GEE) is a cloud-based platform developed by Google that supports the global-scale analysis of big catalogs of Earth Observation data [17]. It has been used to map global forest change in the 21st century [20], Earth's surface water change [21], global urban areas [11], wildfire progression [22], global bare surface change [23], and others. In this sense, GEE becomes compelling not because the distributed processing tasks are executed on the server-side of Google, but also due to the increasing availability of many global geospatial datasets that could be explored in topographic mapping. There exist several available topographical data within GEE, such as the global SRTM DEM, AW3D30 DSM, Global 30 Arc-Second Elevation data

(GTOPO30 DEM, [24]), and others. Thus, GEE characteristics could permit the customization of high-performance terrain analysis with minimal user input and any computational processing on the user side. In fact, GEE provides three algorithms for calculating slope, illumination, and aspect of terrain, but lacks in providing calculation methods of other terrain information, such as the curvatures and landscape characterization.

In addition, a common obstacle of global terrain analysis in common GIS is the need for projecting DEMs onto projected coordinate systems, which ensures the elevation data is equally spaced on a plane square grid [25]. This step is complicated because it is difficult to define a projected system that minimizes terrain distortions over a global extent [26]. Moreover, as many available global DEMs are referenced by geographical coordinate systems and some researchers continue to apply square-grid algorithms to them, the algorithms should consider the geometry and specificity of global spheroidal DEMs [25]. This aspect is important because the application of square-grid methods to spheroidal equal angular DEMs leads to substantial computational errors in models of morphometric variables [25].

In this paper, we aimed at describing and making available an user-friendly processing algorithm for performing terrain analysis in GEE. This algorithm takes advantage of GEE's high-performance architecture for making the computational analysis scalable, adapted to customized needs, and requiring minimal user input. For this, the proposed package takes advantage of a calculation method adapted for spheroidal elevation grids, which favors the global-scale analysis of different DEM resolutions without projecting elevation data.

3.2. MATERIAL AND METHODS

3.2.1. Algorithm description

The Terrain Analysis in GEE (TAGEE) package use calculation methods adapted to spheroidal angular grids, i.e. the DEM can be referenced in a geographical coordinate system, e.g., the World Geodetic System (WGS84). The following paragraphs briefly describes the calculation methods performed by TAGEE package. The readers are referred to [8] for the mathematical concepts of geomorphometry, a

historical overview of the progress of digital terrain modelling, and the notion of the topographic surface and its limitations.

3.2.1.1. Topographic surface

The land topography can be approximated by a topographic surface defined by a continuous, single-valued bivariate function (Equation 1) [8]:

$$z = f(x, y) \tag{1}$$

where z is elevation (meters), and x and y are the coordinates in geographical coordinates (degrees).

The local morphometric variables are functions of the partial derivatives of elevation. Using the Evans-Young method, the function $z = f(x, y)$ is expressed as the second-order bivariate Taylor polynomial (Equation 2).

$$z = \frac{rx^2}{2} + \frac{ty^2}{2} + sxy + px + qy + u \tag{2}$$

where r, t, s, p and q are the partial derivatives, and u is the residual term.

Differently from a digital elevation model projected on a plane square grid, where the partial derivatives of terrain are estimated by finite differences, the processing and analysis of a spheroidal equal angular DEM must consider the spheroidal geometry. In such case, a grid spacing with approximately equal linear units along meridians and parallels exists only at the Equator. To estimate the parameters of a spheroidal grid, a 3x3 moving window must retrieve both the geometry elements and the elevation values of the window nodes (Figure 1).

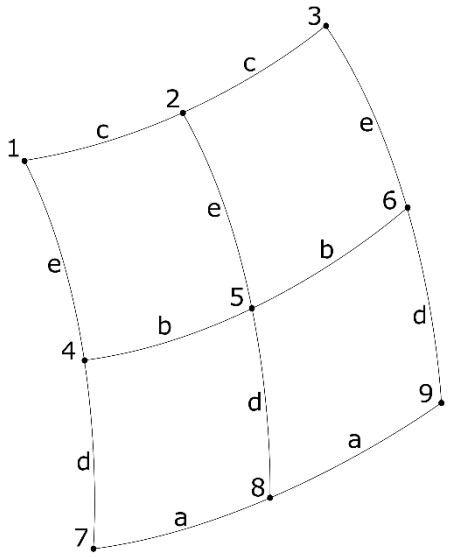

Figure 1. A 3x3 spheroidal equal angular grid with linear geometries a, b, c, d, and f, and nine elevation nodes. Adapted from [8].

3.2.1.2. Terrain parameters: neighbour elevations and geometries

The elevation values of a 3x3 moving window are estimated by convolution kernels. For geometries, the Haversine formula is used to determine the great-circle distances between two neighbour nodes within the spheroidal window, given their latitude and longitude geographical positions (Equations 3, 4, 5):

$$j = \sin^2(\frac{\Delta\phi}{2}) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2(\frac{\Delta\lambda}{2}) \quad (3)$$

$$k = 2 \cdot atan2(\sqrt{j}, \sqrt{(1-j)}) \quad (4)$$

$$l = R \cdot k \quad (5)$$

where ϕ_1 is latitude for the first given node in radians, ϕ_2 is the latitude for the second given node in radians, λ_1 is the longitude for the first given node in radians, λ_2 is the longitude for the second given node in radians, $\Delta\phi$ and $\Delta\lambda$ are the respective differences of latitude and longitude between the given nodes, and R is the mean radius of Earth equals to 6371000 meters. The linear distance l is given in meters.

Knowing the latitude and longitude of the window nodes (Figure 1), the Haversine formula allows the calculation of linear distances of a, b, c, d and e, which are used with the neighbor elevation values (from z_1 to z_9) to calculate the partial derivatives of terrain.

3.2.1.3. Terrain derivatives

To estimate the first and second-order partial derivatives r, t, s, p and q, the polynomial model is fitted by least squares and results in the following estimations (Equations 6, 7, 8, 9, 10) [8]:

$$p = \frac{a^2cd(d+e)(z_3-z_1)+b(a^2d^2+c^2e^2)(z_6-z_4)+ac^2e(d+e)(z_9-z_7)}{2[a^2c^2(d+e)^2+b^2(a^2d^2+c^2e^2)]} \quad (6)$$

$$\begin{aligned} q = \; & \frac{1}{3de(d+e)(a^4+b^4+c^4)} \\ & \cdot\{[d^2(a^4+b^4+b^2c^2)+c^2e^2(a^2-b^2)](z_1+z_3) \\ & -[d^2(a^4+c^4+b^2c^2)-e^2(a^4+c^4+a^2b^2)](z_4+z_6) \\ & -[e^2(b^4+c^4+a^2b^2)-a^2d^2(b^2-c^2)](z_7+z_9) \\ & +d^2[b^4(z_2-3z_5)+c^4(3z_2-z_5)+(a^4-2b^2c^2)(z_2-z_5)] \\ & +e^2[a^4(z_5-3z_8)+b^4(3z_5-z_8)+(c^4-2a^2b^2)(z_5-z_8)] \\ & -2[a^2d^2(b^2-c^2)z_8+c^2e^2(a^2-b^2)z_2]\} \end{aligned} \quad (7)$$

$$r = \frac{c^2(z_1+z_3-2z_2)+b^2(z_4+z_6-2z_5)+a^2(z_7+z_9-2z_8)}{a^4+b^4+c^4} \quad (8)$$

$$\begin{aligned} s = \; & \{c(a[^2(d+e)+b^2e](z_3-z_1) \\ & -b(a^2d-c^2e)(z_4-z_6) \\ & +a[c^2(d+e)+b^2d](z_7-z_9)\} \\ & \cdot\frac{1}{2[a^2c^2(d+e)^2+b^2(a^2d^2+c^2e^2)]} \end{aligned} \quad (9)$$

$$\begin{aligned} t = \; & \frac{2}{3de(d+e)(a^4+b^4+c^4)} \\ & \cdot\{[d(a^4+b^4+b^2c^2)-c^2e(a^2-b^2)](z_1+z_3) \\ & -[d(a^4+c^4+b^2c^2)+e(a^4+c^4+a^2b^2)](z_4+z_6) \\ & +[e(b^4+c^4+a^2b^2)+a^2d(b^2-c^2)](z_7+z_9) \\ & +d[b^4(z_2-3z_5)+c^4(3z_2-z5)+(a^4-2b^2c^2)(z_2-z_5)] \\ & +e[a^4(3z_8-z_5)+b^4(z_8-3z_5)+c^4-2a^2b^2)(z_8-z_5)] \\ & -2[a^2d(b^2-c^2)z_8-c^2e(a^2-b^2)z_2]\} \end{aligned} \quad (10)$$

where the parameters a, b, c, d and e are the linear distances calculated from the Haversine formula (Equations 3, 4, 5), and the z values are elevation values from the neighbors of a moving window (Figure 1).

3.2.1.4. Terrain attributes

Local attributes, such as slope, aspect and curvatures, are calculated from the partial derivatives of terrain [8]. The slope gradient (G, Equation 11) is a flow attribute that relates to the velocity of gravity-driven flows. For measuring the direction, the slope aspect is used (A, Equations 12 and 13). Additionally, one can calculate the amount that a slope is faced to the North or East, resulting in the Northernness (A_N, Equation 14) and Easternness (A_E, Equation 15) derived from the aspect. The remaining flux attributes that can be calculated from the first and second-order partial derivatives are the horizontal (k_h, Equation 16) and vertical curvatures (k_v, Equation 17). While the horizontal curvature relate if a lateral flow converges (k_h < 0) or diverges (k_h > 0), the vertical curvature measures the relative acceleration (k_v > 0) and deceleration (k_v < 0) of a gravity-driven flow.

$$G = \arctan\sqrt{p^2 + q^2} \tag{11}$$

$$\begin{aligned} A = \quad & -90[1 - \text{sign}(q)](1 - |\text{sign}(p)|) + 180[1 + \text{sign}(p)] \\ & -\frac{180}{\pi}\text{sign}(p)\arccos\left(\frac{-q}{\sqrt{p^2 + q^2}}\right) \end{aligned} \tag{12}$$

$$\text{sign}(x) = \begin{cases} 1 & \text{for} x > 0 \\ 0 & \text{for} x = 0 \\ -1 & \text{for} x < 0 \end{cases} \tag{13}$$

$$A_N = \cos A \tag{14}$$

$$A_E = \sin A \tag{15}$$

$$k_h = -\frac{q^2 r - 2pqs + p^2 t}{(p^2 + q^2)\sqrt{1 + p^2 + q^2}} \tag{16}$$

$$k_v = -\frac{p^2 r + 2pqs + q^2 t}{(p^2 + q^2)\sqrt{(1 + p^2 + q^2)^3}} \tag{17}$$

Differently from flow attributes, which are gravity field-specific variables, form attributes are related to principal sections of terrain [8]. The mean curvature (H, Equation 18) is a half-sum of any two orthogonal normal sections and represents two accumulation mechanisms of gravity-driven flows with equal weights: convergence and relative deceleration. Among the class of form attributes, the Gaussian curvature (K, Equation 19) is a product of maximal (k_{max}) and minimal (k_{min}) curvatures. The two principal curvatures calculate the highest and lowest curvature for a given point of the

topographic surface. The maximal curvature (k_{max}, Equation 20) is useful for mapping rigdes (k_{max} > 0) and closed depressions (k_{max} < 0). Likewise, the minimal curvature (k_{min}, Equation 21) is useful for identifying hills (k_{min} > 0) and valleys (k_{min} < 0) across the topographic surface. With the results of mean and Gaussian curvatures, a landform classification can be generated after [27] proposing the continuous form of the Gaussian classification [8,28]. Instead of providing categorical values, the shape index (SI, Equation 22) ranges from -1 to 1 and map convex (SI > 0) and concave (SI < 0) landforms.

$$H = -\frac{(1+q^2)r - 2pqs + (1+p^2)t}{2\sqrt{(1+p^2+q^2)^3}} \tag{18}$$

$$K = \frac{rt - s^2}{(1+p^2+q^2)^2} \tag{19}$$

$$k_{max} = H + \sqrt{(H^2 - K)} \tag{20}$$

$$k_{min} = H - \sqrt{(H^2 - K)} \tag{21}$$

$$SI = \frac{2}{\pi}\arctan\frac{H}{\sqrt{H^2 - K}} \tag{22}$$

3.2.2. Package description

Calculation methods presented in this paper were developed using the JavaScript programming interface available as the online code editor of GEE. TAGEE was developed by different modules of calculation, similarly to what was describe in Methods. The first module, calculateParameters, uses convolution kernels and the Haversine formula to retrieve elevation values and the spheroidal geometries of a 3x3 moving window. In this module, a digital elevation model and a square polygon representing the bounding box (min. Longitude, min. Latitude, max. Longitude, and max. Latitude, in the WGS84 coordinate reference system) are required as input parameters to run. The bounding box is used both in this module and others for generating images with constant values and restrict the calculations to the study area. The first module returns an image with 14 bands, i.e. the neighbor elevation values (from Z_1 to Z_9) and the distances (a, b, c, d and e) (Figure 1).

Once the basic parameters (elevation and distances) were established, the partial derivatives of terrain are calculated with the calculateDerivatives module. This second module requires the returned parameters from calculateParameters and also the bounding box of the study region. The second module adds the partial derivatives (r, t, s, p and q) as new bands to the previous image. Then, terrain attributes are calculated by the module calculateAttributes (Figure 2).

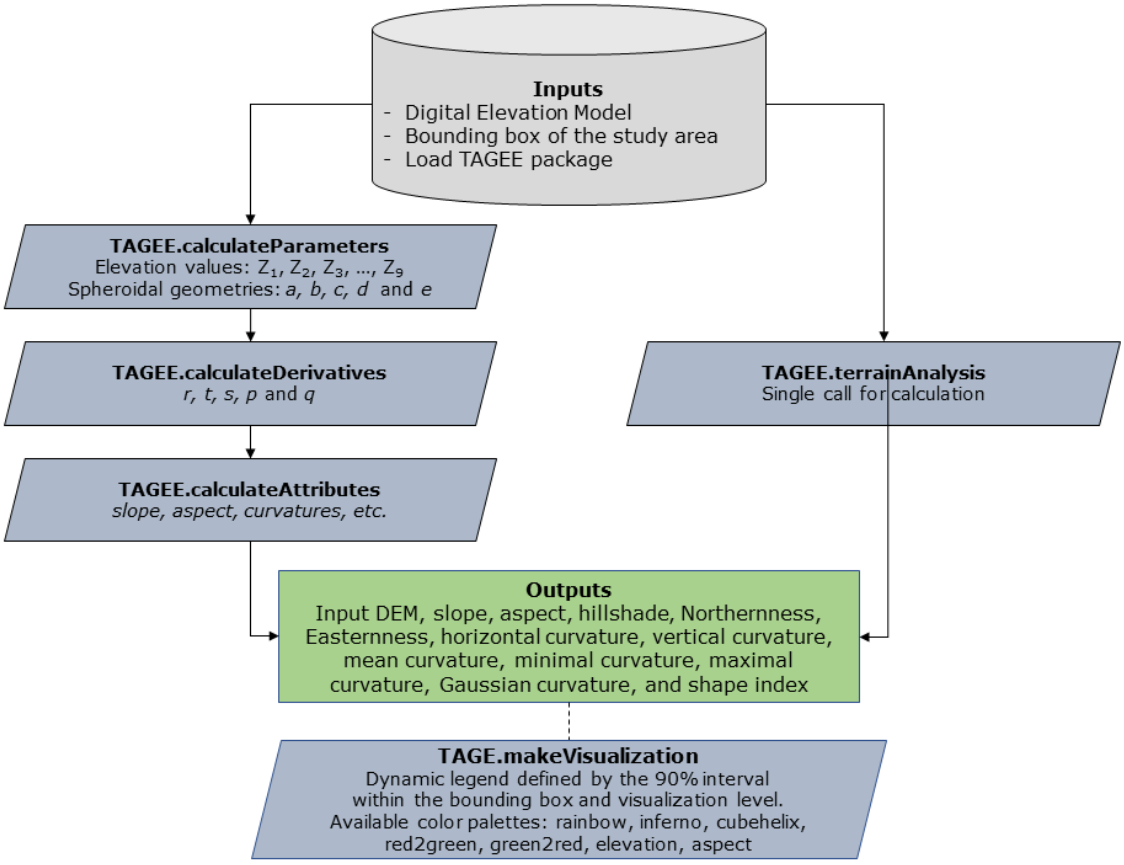

Figure 2. TAGEE modules for calculating terrain parameters, derivatives and attributes.

Terrain attributes can also be calculated by a single function, without calling the intermediate modules. The final output, for both alternatives (Figure 2), is a multi band object containing the same data properties of the digital elevation model (resolution, data type and coordinate reference system) with 13 bands (Table 1). The final attributes can be used for further modeling inside GEE or thematic mapping.

The package has an additional feature that makes easier the visualization of terrain attributes. As the range of attribute values and the pixel resolution may vary according to the visualization level (zoom), which impacts the estimated geometries

and elevation neighbour values, a module called *makeVisualization* automatically calculates the dynamic legend defined by the 0.05 and 0.95 percentiles within the bounding box. In addition, different color palettes for making the map legend are available in TAGEE: rainbow, inferno, cubehelix, red2green, green2red, elevation, aspect.

Table 1. Attributes of terrain, with their units and description, calculated by TAGEE package.

Attribute	Unit	Description
Elevation	meter	Height of terrain above sea level
Slope	degree	Slope gradient
Aspect	degree	Compass direction
Hillshade	dimensionless	Brightness of the illuminated terrain
Northernness	dimensionless	Degree of orientation to North
Easternness	dimensionless	Degree of orientation to East
Horizontal curvature	meter	Curvature tangent to the contour line
Vertical curvature	meter	Curvature tangent to the slope line
Mean curvature	meter	Half-sum of the two orthogonal curvatures
Minimal curvature	meter	Lowest value of curvature
Maximal curvature	meter	Highest value of curvature
Gaussian curvature	meter	Product of maximal and minimal curvatures
Shape Index	dimensionless	Continuous form of the Gaussian classification

3.2.3. Statistical evaluation

We performed the evaluation of TAGEE attributes by comparing the aspect and slope derived from two available functions of GEE (ee.Terrain.aspect and ee.Terrain.slope) on a near-global scale. For this task, we used the Pearson correlation analysis with the SRTM DEM 30m, which contains elevation in meters limited to an area between about 60° north latitude and 56° south latitude. It is important to mention that for the currently available terrain functions of GEE, the local gradient is computed using the 4-connected neighbors of each pixel, differently from the proposed method of TAGEE, which uses a 3x3 pixel window and also considers the spheroidal geometries in its calculation. Thus, minimal differences between the calculation methods are expected to occur. This analysis was performed in GEE and, in addition to Pearson's correlation, we calculated the relative mean absolute error (MAE) between the outputs. The relative MAE is estimated by calculating the mean absolute difference between two rasters and standardizing the result to the range (maximum minus minimum values) of the reference raster.

Similarly, we compared the results from TAGEE with terrain attributes calculated by the System for Automated Geoscientific Analyses (SAGA) GIS version 2.3.2 [12]. In this case, we downloaded from GEE the 30 m SRTM DEM together with the resulting 12 attributes calculated by TAGEE, all covering the Mount Ararat (located between 44.2° and 44.5° E, and 39.6° and 39.8° N). The Mount Ararat was selected due its high variability of landforms and the availability of published maps from previous works [8,29], allowing the visual comparison of spatial patterns. The Mount Ararat SRTM-DEM was processed in SAGA GIS using the "Slope, Aspect, Curvature" from the Morphometry module of Terrain Analysis. The calculation method was the "Evan (1979)" based on 6 parameters and 2nd order polynomials, similarly to TAGEE calculation method. The comparison was performed by calculating the Pearson's correlation coefficient (*r*) and the relative MAE, where the aspect, slope, horizontal curvature and vertical curvature from TAGEE were compared with aspect, slope, tangential, and profile curvature from SAGA GIS, respectively, following the equivalence described in [8].

3.3. RESULTS AND DISCUSSION

The statistical analysis revealed a significant correlation ($p < 0.01$) of the TAGEE outputs with equivalent terrain attributes calculated from GEE and SAGA GIS (Table 2). The slope estimated over a near-global extent reached a correlation of 0.98 (error of 2%) between TAGEE and functions of GEE, while the aspect resulted in a Pearson's *r* of 0.89 (13% of error). The lower correlation of aspect can be associated to its dimension nature, i.e., a circular variable, as well as to the differences of calculation methods between TAGEE and GEE. Despite the small differences, TAGEE revealed the same spatial patterns and allowed the estimation of additional attributes at the global scale, such as the Northernness, horizontal and vertical curvatures (Figures 3 A, B and C, respectively). The main mountain ranges of the Earth, such as the Rocky Mountains in North America, Andes in South America, Alps in Europe, Himalayas, and Tibetan plateau in Asia, etc., present the highest curvatures calculated by TAGEE. Conversely, the plains and flat surfaces had the lowest estimates for both curvatures. The degree of orientation to North (Figure 3 A) also depict the main landforms of the Earth.

Table 2. Comparison of TAGEE attributes with outputs from GEE and SAGA GIS algorithms.

Attribute	Region	Reference	Pearson's *r*	rMAE[1]
Aspect	Near global SRTM DEM 30 m	GEE	0.89*	13%
Slope	Near global SRTM DEM 30 m	GEE	0.98*	2%
Aspect	Mount Ararat SRTM DEM 30 m	SAGA GIS	0.96*	4%
Slope	Mount Ararat SRTM DEM 30 m	SAGA GIS	0.98*	3%
Horizontal curvature	Mount Ararat SRTM DEM 30 m	SAGA GIS	0.98*	4%
Vertical curvature	Mount Ararat SRTM DEM 30 m	SAGA GIS	0.98*	4%

* Significant for $p < 0.01$; [1] relative mean absolute error.

TAGEE was developed in GEE to take advantage of the high-performance computing of the platform. As the cloud-based interfaces have created much enthusiasm and engagement in the remote sensing and geospatial fields, many processing algorithms have been adapted to make substantive progress on global challenges involving the processing of big geospatial data [30]. In this sense, GEE is providing petabytes of publicly available remote sensing imagery and other ready-to-use products. The high-speed parallel processing of GEE servers and the libraries of operators and machine learning algorithms available by Application Programming Interfaces (APIs) in popular coding languages, such as JavaScript and Python, are enabling users to discover, analyze and visualize geospatial big data without needing access to supercomputers [30]. Within this framework, TAGEE supports the development of customized terrain analysis with different elevation data across large geographical extents.

When TAGEE outputs were compared to those from SAGA GIS (Table 2), the statistical evaluation resulted in a significant and high correlation for the slope, horizontal and vertical curvatures of terrain (Pearson's *r* of 0.98, with an error difference of 3 and 4%). Aspect from TAGEE and SAGA GIS had an inferior correlation coefficient, but the result was higher than the aspect from the algorithm of GEE. The region of Mount Ararat was also used to visually compare the slope, horizontal and vertical curvatures, calculated from both TAGEE and SAGA GIS (Figure 4). The 3D visualizations revealed a high similarity between both maps, but some small

differences can be visualized by the color intensity. This is the case of the slope of the Mount Ararat calculated by TAGEE (Figure 4 A), which had a higher intensity compared to the slope of SAGA GIS (Figure 4 B). A slightly higher intensity for the vertical curvature calculated by SAGA GIS was also evident on an edge of the Mount Ararat (Figure 4 F). Despite small, these visual differences confirm the relative error of both methods (Table 2). In addition, the spatial patterns of aspect, slope and curvatures from TAGEE presented a high correspondence with the terrain maps of Mount Ararat available in [8,29], reinforcing the confidence of TAGEE calculation method.

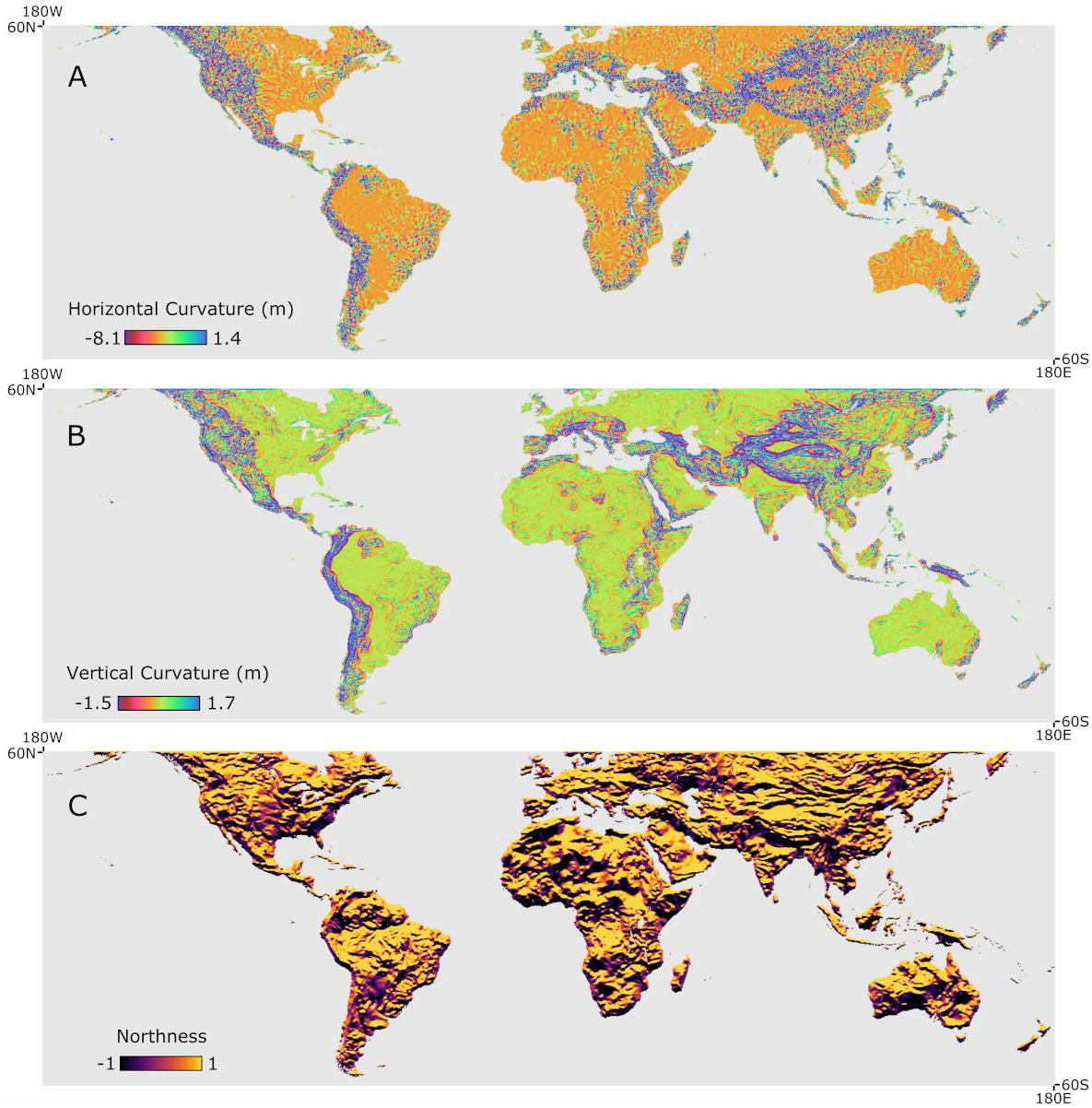

Figure 3. Example of terrain attributes calculated from TAGEE package and 1 arc-second SRTM DEM, displayed for the near-global extent at the visualization level 3 (~20 km pixel resolution): horizontal curvature (A), vertical curvature (B) and Northernness (C).

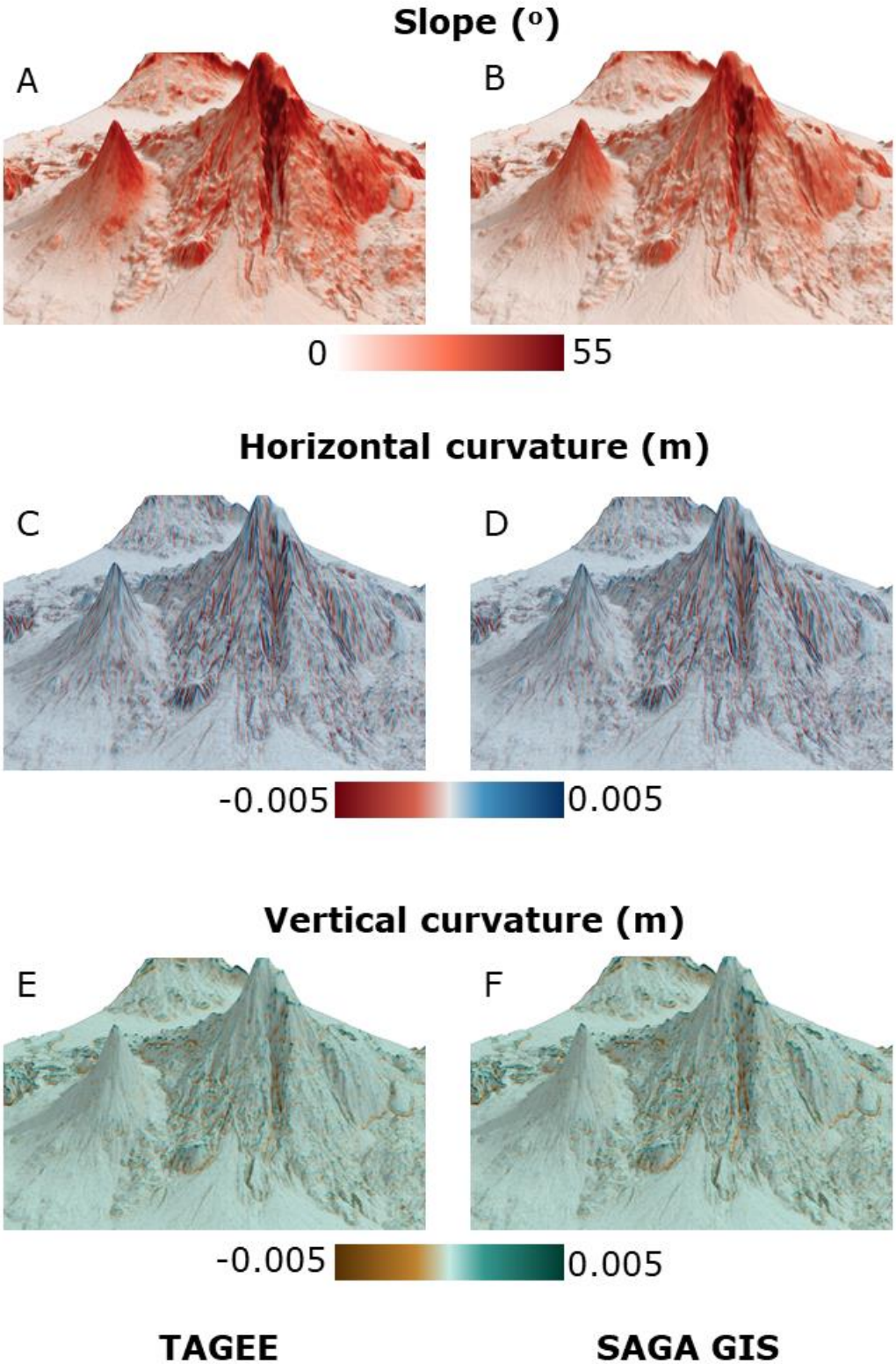

Figure 4. 3D visualizations of terrain attributes produced near Mount Ararat: slope, horizontal and vertical curvature from TAGEE (a, c, and e, respectively) and SAGA GIS (b, d, and f, respectively). 3D maps are displayed with a vertical exaggeration of 2.

In this work, the TAGEE algorithm was developed to consider spheroidal geometries in its calculation method. This approach diverges from the techniques available in traditional GIS, where TAGEE considers the great circle distances of the DEM defined by Latitude and Longitude positions. Common GIS software, such as SAGA GIS, requires the projection of the DEM to ensure the elevation data has the same pixel size. However, as identified by [25], some researchers continue to apply square-grid algorithms to spheroidal equal angular DEMs, which can lead to substantial computational errors in models of morphometric variables. The small relative errors (Table 2) between TAGEE and GEE or SAGA GIS could be linked to the differences in their calculation methods.

Finally, some limitations of TAGEE can also be noted. Only local morphometric variables can be calculated by the package, which includes flux and form attributes. Non-local attributes, such as specific catchment area, were not implemented due to the absence of a general analytical theory, which is still little developed [29], and due to the recursion processing that is still challenging within GEE [17]. Furthermore, a novel method became available to handle major problems of terrain analysis, which includes the approximation of DEM, generalization and denoising, and the computation of morphometric variables. The universal spectral analytical method based on high-order orthogonal expansions using the Chebyshev polynomials were developed by [31] to handle the aforementioned issues into an integrated framework, but was not implemented in this work.

3.4. CONCLUSIONS

The proposed package (TAGEE) can calculate terrain attributes using the high-performance platform of GEE with an accuracy equivalent to traditional GIS. The approach of using spheroidal geometries does not require the projection of input elevation data for terrain attributes calculation. The comparison between algorithms demonstrated that TAGEE estimates terrain slope and aspect similarly to the available functions of GEE. The advantage of TAGEE over the currently available functions is that additional outputs can be produced, such as curvatures and shape index, which can be useful for environmental mapping and modelling studies. In addition, a good agreement was also found when TAGEE was compared to equivalent outputs from

SAGA GIS, reaching a Pearson's correlation coefficient between 0.96 and 0.98, and differences between 3-4 %. Thus, TAGEE becomes a feasible tool for making terrain analysis of big geospatial data, which can be customized to any spatial resolution and scaled up to the global extent.

ACKNOWLEDGMENTS

The authors are grateful to the Geotechnologies in Soil Science (GEOCIS) group. This research was funded by São Paulo Research Foundation, grants number 2014/22262-0 and 2016/01597-9.

REFERENCES

1. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrol. Process. 1991, 5, 3–30, doi:10.1002/hyp.3360050103.
2. Amatulli, G.; Domisch, S.; Tuanmu, M.-N.; Parmentier, B.; Ranipeta, A.; Malczyk, J.; Jetz, W. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci. Data 2018, 5, 180040, doi:10.1038/sdata.2018.40.
3. Pike, R.J. Geomorphometry -diversity in quantitative surface analysis. Prog. Phys. Geogr. Earth Environ. 2000, 24, 1–20, doi:10.1177/030913330002400101.
4. Bogaart, P.W.; Troch, P.A. Curvature distribution within hillslopes and catchments and its effect on the hydrological response. Hydrol. Earth Syst. Sci. 2006, 10, 925–936, doi:10.5194/hess-10-925-2006.
5. Alexander, C.; Deák, B.; Heilmeier, H. Micro-topography driven vegetation patterns in open mosaic landscapes. Ecol. Indic. 2016, 60, 906–920, doi:10.1016/j.ecolind.2015.08.030.
6. Oliveira, S.; Pereira, J.M.C.; San-Miguel-Ayanz, J.; Lourenço, L. Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. Appl. Geogr. 2014, 51, 143–157, doi:10.1016/j.apgeog.2014.04.002.
7. McGuire, K.J.; McDonnell, J.J.; Weiler, M.; Kendall, C.; McGlynn, B.L.; Welker, J.M.; Seibert, J. The role of topography on catchment-scale water residence time. Water Resour. Res. 2005, 41, doi:10.1029/2004WR003657.

8. Florinsky, I. V. Digital terrain analysis in soil science and geology; Academic Press, 2016; ISBN 9780128046326.
9. USGS EROS USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) Void Filled Available online: https://doi.org/10.5066/F7F76B1X (accessed on Apr 4, 2020).
10. JAXA EORC ALOS Global Digital Surface Model “ALOS World 3D - 30m (AW3D30)” Available online: https://www.eorc.jaxa.jp/ALOS/en/aw3d30/index.htm (accessed on Apr 4, 2020).
11. Liu, X.; Hu, G.; Chen, Y.; Li, X.; Xu, X.; Li, S.; Pei, F.; Wang, S. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. Remote Sens. Environ. 2018, 209, 227–239, doi:10.1016/j.rse.2018.02.055.
12. Olaya, V.; Conrad, O. Geomorphometry in SAGA. In Developments in Soil Science; Elsevier, 2009; pp. 293–308.
13. Miliaresis, G. The Landcover Impact on the Aspect/Slope Accuracy Dependence of the SRTM-1 Elevation Data for the Humboldt Range. Sensors 2008, 8, 3134–3149, doi:10.3390/s8053134.
14. Bindzárová Gergeľová, M.; Kuzevičová, Ž.; Labant, S.; Gašinec, J.; Kuzevič, Š.; Unucka, J.; Liptai, P. Evaluation of Selected Sub-Elements of Spatial Data Quality on 3D Flood Event Modeling: Case Study of Prešov City, Slovakia. Appl. Sci. 2020, 10, 820, doi:10.3390/app10030820.
15. Xia, J.; Yang, C.; Li, Q. Building a spatiotemporal index for Earth Observation Big Data. Int. J. Appl. Earth Obs. Geoinf. 2018, 73, 245–252, doi:10.1016/j.jag.2018.04.012.
16. Danielson, J.J.; Gesch, D.B. Global multi-resolution terrain elevation data 2010 (GMTED2010); U.S. Geo-logical Survey Open-File Report 2011–1073, 2011;
17. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 2017, 202, 18–27, doi:10.1016/J.RSE.2017.06.031.
18. Abernathey, R.; Paul, K.; Hamman, J.; Rocklin, M.; Lepore, C.; Tippett, M.; Henderson, N.; Seager, R.; May, R.; Del Vento, D. Pangeo NSF Earthcube Proposal Available online: https://figshare.com/articles/Pangeo_NSF_Earthcube_Proposal/5361094.

19. mundialis GmbH & Co. KG actinia - geoprocessing in the cloud Available online: https://actinia.mundialis.de/.
20. Hansen, M.C.; Potapov, P. V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S. V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science (80-.). 2013, 342, 850–853, doi:10.1126/science.1244693.
21. Donchyts, G.; Baart, F.; Winsemius, H.; Gorelick, N.; Kwadijk, J.; van de Giesen, N. Earth's surface water change over the past 30 years. Nat. Clim. Chang. 2016, 6, 810–813, doi:10.1038/nclimate3111.
22. Crowley, M.A.; Cardille, J.A.; White, J.C.; Wulder, M.A. Multi-sensor, multi-scale, Bayesian data synthesis for mapping within-year wildfire progression. Remote Sens. Lett. 2019, 10, 302–311, doi:10.1080/2150704X.2018.1536300.
23. Demattê, J.A.M.; Safanelli, J.L.; Poppiel, R.R.; Rizzo, R.; Silvero, N.E.Q.; Mendes, W. de S.; Bonfatti, B.R.; Dotto, A.C.; Salazar, D.F.U.; Mello, F.A. de O.; et al. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. Sci. Rep. 2020, 10, 4461, doi:10.1038/s41598-020-61408-1.
24. USGS EROS GTOPO30 - Global 1-km digital raster data derived from a variety of sources Available online: https://doi.org/10.5066/F7DF6PQS.
25. Florinsky, I. V. Spheroidal equal angular DEMs: The specificity of morphometric treatment. Trans. GIS 2017, 21, 1115–1129, doi:10.1111/tgis.12269.
26. Brainerd, J.; Pang, A. Interactive map projections and distortion. Comput. Geosci. 2001, 27, 299–314, doi:10.1016/S0098-3004(00)00108-4.
27. Koenderink, J.J.; van Doorn, A.J. Surface shape and curvature scales. Image Vis. Comput. 1992, 10, 557–564, doi:10.1016/0262-8856(92)90076-F.
28. Gauss, C.F. General investigations of curved surfaces of 1827 and 1825; Princeton University Library: Princeton, NJ, 1902;
29. Florinsky, I. V An illustrated introduction to general geomorphometry. Prog. Phys. Geogr. Earth Environ. 2017, 41, 723–752, doi:10.1177/0309133317733667.
30. Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 2020, 164, 152–170, doi:10.1016/j.isprsjprs.2020.04.001.

31. Florinsky, I. V.; Pankratov, A.N. A universal spectral analytical method for digital terrain modeling. Int. J. Geogr. Inf. Sci. 2016, 30, 2506–2528, doi:10.1080/13658816.2016.1188932.

4. LEVERAGING THE APPLICATION OF EARTH OBSERVATION DATA FOR MAPPING AND MONITORING CROPLAND SOILS IN BRAZIL

ABSTRACT

Brazilian agriculture has been playing an important role in the global supply of food and fiber. As the cropland expansion took place in the last decades and led to significant land-use change and environmental impacts, detailed information about soils became fundamental for the planning and sustainable use of the agricultural lands. Thus, considering the demand for spatially explicit information and the lack of detailed sources for understanding the current conditions of cropland soils in Brazil, we aimed to perform high-resolution mapping of key topsoil attributes using spectral and terrain features extracted from Earth observation data (EOD). With the resulting information, we also aimed at performing a general examination of the main agricultural regions and estimate the total organic carbon stocks on croplands soils. For this, we prepared environmental predictors from the historical collection of Landsat data and the digital elevation model from Shuttle Radar Topographic Mission at the cloud-based platform of Google Earth Engine. The extracted spectral and terrain features with a spatial resolution of 30 m were calibrated with georeferenced soil samples (n = 5097) for predicting the topsoil content (0-20 cm) of clay, sand, silt, cation exchange capacity, pH, soil organic carbon (SOC) and SOC stock. Ensembles of bootstrapped and random regression trees were used to deliver both the predicted and uncertainty estimates across the Brazilian croplands. The spectral features in the form of a bare soil composite containing six multispectral bands had a good agreement with the reference spectral patterns from the soil dataset, being fundamentally important for estimating clay and sand contents. Prediction models of clay, sand, SOC content, and SOC stocks had the best performance metrics, achieving a R2 ranging from 0.44 to 0.74 and ratio of performance to the interquartile range higher than 1.5. The predicted maps revealed the variability of topsoil among the cropped areas, suggesting that the agriculture expansion took place on sandy soils. The SOC stock map provided consistent estimates compared to previous datasets available in Brazil but revealed additional information about the spatial distribution at the local and regional scales. Thus, this study supports the proposition that EOD is a valuable source for extracting environmental features for mapping and monitoring cropland soils at finer resolutions, assisting the evaluation of soil spatial distribution and the historical agriculture expansion in Brazil.

Keywords: Google Earth Engine, Machine Learning, Soil Spectral Library, Remote Sensing, Bare Soil Composite.

4.1. INTRODUCTION

In the last four decades, Brazilian agriculture had an annual average growth of 4% and became the fourth greatest global producer of commodities (Chaddad, 2016). Such increase is associated to the agricultural expansion in Cerrado (Brazilian Savannah) and Amazônia biomes (Dias *et al.*, 2016; Morton *et al.*, 2016), leading to

uncontrolled land use change and significant environmental impacts (Fonseca *et al.*, 2019; Martinelli *et al.*, 2010; Noojipady *et al.*, 2017; Phalan *et al.*, 2013; Stabile *et al.*, 2020; Zinn, Lal e Resck, 2005). This situation becomes more critical because most of the Brazilian territory does not have detailed information about the soil spatial distribution (Polidoro *et al.*, 2016), which hinders the planning and sustainable use of the current agricultural lands. Along with this, as many studies projected a growing demand for food, fiber, and energy for the next decades (Godfray *et al.*, 2010; Tilman *et al.*, 2011). Thus, understanding the current cropland soil distribution and their capacities ensures the meeting of future global demands.

Cropland soil attributes are very dynamic in depth, space and time compared, for example, to soils with natural vegetation in protected areas, where the SOC levels are in a steady-state (not contributing to carbon emissions) and the soil quality indicators are in optimal levels (Cherubin *et al.*, 2015; Ondrasek *et al.*, 2019; Zinn, Lal e Resck, 2005). Agriculture intensification can induce many alterations to soil attributes and cause modifications of their physical, chemical, and biological conditions, therefore threatening the provision of its functions. In this sense, the protection, restoration, and sustainable use of soils to combat land degradation have been promoted by the United Nations as part of the Sustainable Development Goals (Tóth *et al.*, 2018). Naturally, advanced processing frameworks become fundamental to map and monitor cropland soils (Tziolas *et al.*, 2020).

Earth-Observation Data (EOD) have been used for monitoring croplands and soils over large geographical extents (Azzari *et al.*, 2019; Cao *et al.*, 2019; Picoli *et al.*, 2018). This happens due to the availability of more than 40 years of EOD acquisition coupled with the efforts of open-access policies for distributing data (Drusch *et al.*, 2012; Wulder *et al.*, 2016). The four-decade Landsat and the Moderate Resolution Imaging Spectroradiometer (MODIS) historical collections are examples of open-access catalogs that allow recovering the spatial and temporal patterns of agricultural areas. Moreover, the recent availability of cloud-based processing interfaces and the widespread of machine learning algorithms have promoted the digital mapping covering large geographical areas. Therefore, soil mapping and monitoring can be benefited by the ability of EOD to provide concise and detailed information about soils. These can reduce costs and the laborious sampling campaigns that aim to characterize the soil spatial variability (Angelopoulou *et al.*, 2019; Chabrillat *et al.*, 2019).

A myriad of studies have already demonstrated the potential of spectral reflectance from the visible (VIS), near-infrared (NIR) and shortwave infrared spectral regions (SWIR), 0.4-2.5 µm, to estimate soil mineralogical, physical and chemical attributes (Demattê et al. 2019; Dotto et al. 2016; Ji et al. 2016; Moura-Bueno et al. 2019; Stevens et al. 2013; Viscarra Rossel et al. 2016). Reflectance of VIS-NIR-SWIR allows to make qualitative and quantitative analysis for describing soil properties, being possible to satisfactorily predict several soil attributes in laboratory conditions, such as inorganic carbon, organic carbon (SOC), clay, sand, extractable iron, cation exchange capacity (CEC), and pH (Soriano-Disla *et al.*, 2014; Viscarra Rossel *et al.*, 2016). Reflectance measurements from multispectral imagery (e.g. Landsat), on the other hand, can limit soil analysis mainly due to sensor and acquisitions characteristics, and due the topsoil conditions happening under the scene acquisition (Ben-Dor *et al.*, 2009; Lagacherie *et al.*, 2008). Similarly, the continuous cover of vegetation and other objects is another issue that hinders the direct measurements of soil reflectance using satellite images. Nevertheless, reflectance patterns from satellite images still can help to depict the soil spatial variability over large geographical areas (Chabrillat *et al.*, 2019).

Recent literature showed that bare surfaces detected by multi-temporal satellite imagery can be aggregated into soil composites to improve the evaluation of topsoil properties and cropland dynamics (Demattê et al. 2018; Diek et al. 2017; Roberts, Wilford, and Ghattas 2019; Rogge et al. 2018). The Geospatial Soil Sensing System (GEOS3, Demattê et al. 2018) detects bare surfaces over a collection of Landsat images and aggregates the images into a Synthetic Soil Image (SYSI). The method has been used in different regions in Brazil and Europe for mapping soil attributes due to the strong correlation of spectral patterns with soil texture and mineralogy (Fongaro *et al.*, 2018; Gallo *et al.*, 2018; Mendes *et al.*, 2019; Poppiel *et al.*, 2019; Safanelli, Chabrillat, *et al.*, 2020). Rogge et al. (2018) developed the automated Soil Composite Mapping Processor (SCMaP), a similar framework for retrieving and mapping bare soils in Germany. The SCMaP method has presented a high correlation with existing soil maps, lithology, and land use patterns. Diek et al. 2017 proposed the Barest Soil Composite to map the Swiss Plateau and Europe, demonstrating that such information is not only useful for soil mapping, but can also be integrated into land management initiatives. More recently, Roberts, Wilford, and Ghattas (2019) presented a similar method to estimate the spectral response of bare surfaces using the full collection of Landsat images across Australia.

Therefore, the mapping and monitoring of soil attributes that are commonly used as key criteria to land suitability evaluation, agricultural recommendations, or soil classification, such as SOC, particle size distribution, and soil fertility components (e.g. pH and CEC), can be fostered with machine learning coupled with environmental features extracted from EOD. In this way, spectral features in the form of bare soil composites, as well as the topographic attributes calculated from digital elevation models (DEM), are sound options to be employed in digital soil mapping (DSM). Indeed, terrain features have a strong influence on process that form soils and have been extensively used in DSM, while bare soil composites can be considered as a direct measurement of soils from satellite sensors (McBratney, Mendonça Santos e Minasny, 2003).

Thus, considering the demand for spatially explicit information and the lack of detailed sources for understanding the current conditions of cropland soils in Brazil, we aimed to perform a 30-meter DSM of key topsoil attributes using spectral and terrain features extracted from EOD. With the resulting maps, we also aimed to perform a general examination of the soil attributes distribution among the main agricultural regions and estimate the total SOC stock of croplands. Therefore, the results of this work can support several other studies, such as the analysis of agriculture expansion and intensification, land suitability assessment, soil monitoring, and others.

4.2. MATERIAL AND METHODS

4.2.1. Cropland soils in Brazil

According to the last census of agriculture performed in Brazil (IBGE, 2018), agriculture is composed by 7.9 Mha of permanent crops, 55.4 Mha of temporary crops, 46.7 Mha of natural pastures and 111.8 Mha of cultivated pastures. This information was surveyed for the reference crop year from October 2016 to September 2017. Despite the periodicity of the survey and the investigation on agricultural establishments and activities, the spatially explicit location of croplands is not available from this data collection. Thus, the Copernicus Global Land Cover layers (CGLS, Collection 2) was used for making the study constrained to the croplands mapped in 2015 with a spatial resolution of 100 m (Buchhorn et al., 2020).

CGLS collection 2 was produced based on extracted features from PROBA-V satellite observations, which included the base reflectance, time series harmonics, vegetation indicators, and descriptive & data density statistics. The training of the land cover classification algorithm was made with more than 141.000 points and the final maps were independently validated using around 20.000 points (Buchhorn et al., 2020). The CGLS collection 2 has reached an overall accuracy of around 80%, and a moderate accuracy of 70-80% for croplands, when considering the discrete land cover classification. In Brazil, the CGLS collection 2 has spatially identified 65.6 Mha of croplands, which is close to the estimates from the census of agriculture when considering the permanent and temporary croplands (i.e. 63.3 Mha).

The soil types that are found in Brazil are mostly weathered-leached soils derived from sedimentary and igneous rocks with loamy sand and clayey texture, such as Ferralsols and Acrisols (Demattê et al. 2019; Gomes et al. 2019). Many of the soils in the southern and southeastern regions in Brazil have long supported the production of food, fiber, and energy (Figure 1A, contour 1). The subtropical climate and Atlantic Forest and Pampa biomes predominate in the southern region, while the tropical climate starts from the southeastern towards Midwest and north, with the occurrence of Cerrado (Savannahs) and Amazon biomes (Figure 1A, contour 2). In the last decades, commercial agricultural production has expanded over extensive areas of Cerrado and some parts of the Amazon forest, and over the last frontier located in the federal states of Maranhão, Tocantins, Piauí, and Bahia (MATOPIBA, contour 3 in Figure 1A) (Buol, 2009; Dias et al., 2016; Martinelli et al., 2010; Spera, 2017).

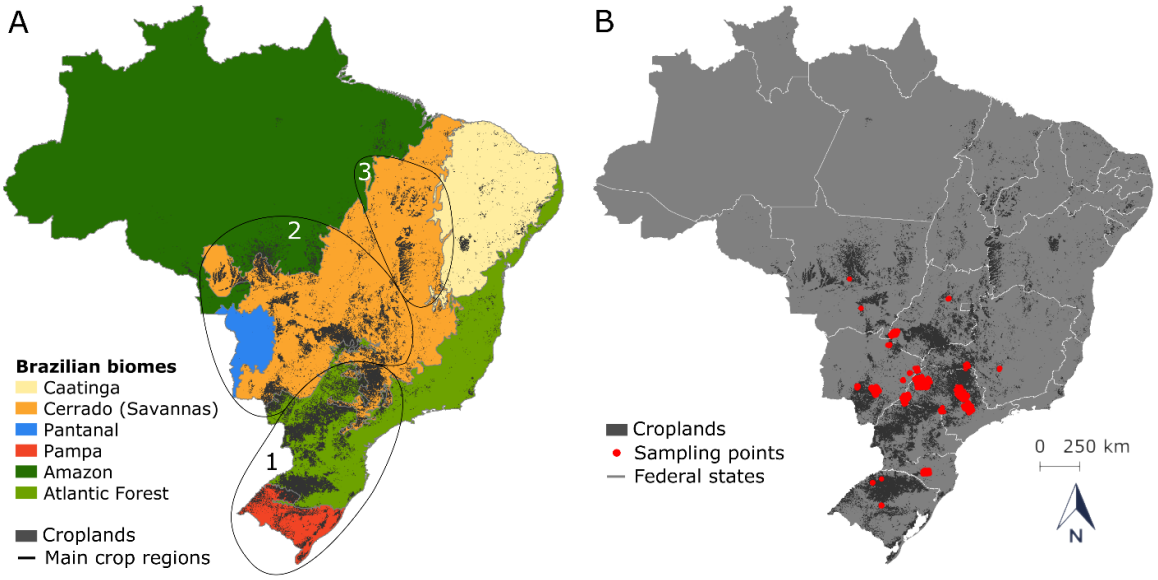

Figure 1. (A) Location of biomes, croplands distribution (derived from Collection 2 of Copernicus Global Land Cover products) and main cropland regions in Brazil: 1) south and southeast; 2) Midwest; and 3) the region that composes the federal states of Maranhão, Tocantins, Piauí and Bahia (MATOPIBA). Gray colors represent missing information, and the light gray lines are the boundaries of federal states. (B) Observations gathered from previous surveys used for calibrating prediction models of soil attributes.

4.2.2. Soil observations

Georeferenced soil observations were selected from former field surveys performed across agricultural regions by the Geotechnologies in Soil Science group and collaborators (Bellinaso, Demattê, and Romeiro 2010; Demattê et al. 2019). All sampling points (n = 5097) had their soil physical and chemical attributes analyzed for the 0-20 cm layer. The analysis followed the Brazilian soil analysis standards (Teixeira et al., 2017), gathering information about clay, silt, sand contents (g kg-1); soil organic carbon (SOC, g kg-1); pH determined in water solution (pHH2O) and the cation exchange capacity (CEC, mmolc kg-1). Soil bulk density (BD), in our study, was estimated using a pedotransfer function that required the clay and SOC contents, which was calibrated for Brazilian soils and reached an R2 value of 0.63 and standard error of 0.11 g cm-3 (Benites et al., 2007). Based on the SOC content and BD estimates, we calculated the SOC stock (kg m-2) for the 0-20 cm layer (Gomes et al., 2019).

For the same soil observations, reflectance measurements were taken under laboratory conditions using a Spectroradiometer FieldSpec 3 (ASD, Boulder, CO, USA). The spectra were acquired between 350 and 2500 nm, covering the visible, near-infrared, and shortwave infrared (VIS-NIR-SWIR) spectral range. The soil reflectance data are part of the Brazilian Soil Spectral Library project (BSSL) and were used to evaluate the reflectance of the bare soil composite derived from EOD (Demattê et al. 2019).

4.2.3. Earth Observation Data (EOD)

The Landsat collection, which included Landsat 4 Thematic Mapper (TM, from 1982 to 1993), Landsat 5 TM (from 1984 to 2012), Landsat 7 Enhanced Thematic Mapper Plus (ETM+, from 1999 to present), and the Landsat 8 Operational Land Imager (OLI, from 2013 to present), were used to extract soil features using the GEOS3 algorithm, which includes the bare soil composite and the frequency of soil exposure (Demattê et al. 2018; Safanelli et al. 2020). Only higher-level products (Tier 1) of surface reflectance processed by the LEDAPS (Landsat 4, 5 and 7) and LASRC (Landsat 8) algorithms of USGS (USGS, 2018a; b) were defined as inputs. During the processing, the quality assessment bands from the images were used to remove cloudy and cloud shadow pixels. To combine all the Landsat data into a single collection, the bands of Landsat 4, 5, 7, and 8 were harmonized into a common name using the specific band number positioned in equivalent spectral regions (Table 1 in Appendix B): blue (~450-520 nm), green (~520-600 nm), red (~630-690 nm), NIR (~760-900 nm), SWIR1 (~1550-1750 nm) and SWIR2 (~2080-2350 nm).

GEOS3 flags soil pixels based on a set of spectral indices and classification thresholds (Demattê et al. 2018). Each image from the harmonized image collection is filtered with bare soil masks and latter aggregated (reduced) into a single image using the median estimate. The soil mask combines the Normalized Difference Vegetation Index (NDVI), Normalized Burn Ratio 2 (NBR) and soil spectral tendency (Safanelli et al. 2020). The soil pixels were flagged when having an increased reflectance from blue to SWIR1, and the spectral indices falling between -0.05 and 0.30 for NDVI, and -0.15 and 0.15 for NBR2. These thresholds were slightly modified from previous studies to avoid misrepresenting the variability of soils found over the large geographical extent of Brazil (Fongaro et al. 2018; Poppiel et al. 2019). In order to evaluate the median-

aggregated bare soil image, namely the Synthetic Soil Image (SYSI), the Pearson's correlation coefficient was calculated between the resampled reflectance from the BSSL with the topsoil reflectance SYSI. Additionally, the mean and standard deviation of the absolute residuals estimated by the difference between SYSI and resampled BSSL reflectance were used as additional evaluation.

Considering that for a given pixel location the soil surface can be exposed more than once due to tillage operations of agriculture, the relative frequency of soil exposure was estimated using the bare soil masks produced for SYSI. The frequency of exposures was estimated by dividing the number a pixel location was flagged as bare to the total number of clear conditions that happened along the satellite image collection, i.e. disregarding the cloudy or cloud shadow pixels. The resulting product is named as soil frequency (SF) and was used for visually comparing its spatial distribution with the cropland maps of soil attributes, as it can give a hint about the disturbances caused by tillage operations on soil surface. The SYSI and SF were produced in the cloud-based platform of Google Earth Engine (GEE) to take advantage of the high-performance and distributed computing for processing the big collections of Landsat data. For making the statistical analysis and spatial prediction, SYSI and SF results were split in smaller tiles of 1x1+0.01 degree (edge buffer).

Another important set of features extracted from EOD and used as predictor of soil attributes were terrain attributes, which were also estimated in GEE using a calculation method adapted to spheroidal equal angular grids (Safanelli et al. 2020). Using a 3x3 moving window over a digital elevation model (DEM), the first and second partial derivatives of terrain were calculated to estimate flux and form attributes. Thus, slope, northernness, easternness, vertical curvature, horizontal curvature, and shape index were calculated from the terrain derivatives. The Shuttle Radar Topography Mission (SRTM) DEM with a spatial resolution of 1 arc-second (approximately 30 m) was used as input data for terrain analysis.

4.2.4. Prediction of soil attributes

Reflectance bands of SYSI and terrain attributes were used as environmental features for calibrating prediction models of clay, sand, silt, SOC, pH, CEC and SOC stock. At each soil sample location, the values of environmental features were extracted using the bilinear sampling method, which considers a few neighbor pixels

for the coordinate intersection. An ensemble of bootstrapped and random regression trees from the Python's scikit-learn library was used as prediction algorithm (Pedregosa *et al.*, 2011). Instead of fitting single regression trees, we emulated the algorithm of Random Forests (Breiman, 2001) by generating bootstrapping trees to later aggregate them by the mean (Equation 1) and standard deviation values (Equation 2). Random Forests have been employed and the standard deviation of the ensemble was used do estimate the uncertainty of the spatial prediction, which was defined by the confidence interval (95% probability level) standardized to the mean at the pixel level (Equation 3).

$$\hat{f}_p = \frac{1}{B}\sum_{b=1}^{B} f_b(X_p) \qquad \text{Equation 1}$$

$$S_p = \sqrt{\frac{\sum_{b=1}^{B}(f_b(X_p) - \hat{f}_p)^2}{B-1}} \qquad \text{Equation 2}$$

$$CR_{std} = \frac{\left(\hat{f}_p + 1.96\,\frac{S_p}{\sqrt{B}}\right) - \left(\hat{f}_p - 1.96\,\frac{S_p}{\sqrt{B}}\right)}{\hat{f}_p} \qquad \text{Equation 3}$$

where X is the set of environmental features for pixel p; B is the forest size determined by the number of bootstrapping samples b; f_b is the regression tree fitted to each bootstrapped sample b; $\hat{f}_p$ is the mean of pixel p; and S_p is the standard deviation of location p.

The optimal number of bootstrapped trees (forest size), number of features to be sampled in tree splits and tree size (depth) was defined with a grid search of hyperparameters seeking to minimize overfitting during calibration. The amplitude of values tested for forest size was 30, 60, 100, 200 and 500 trees. The amount of 3, 5, 8, 11 and 13 predictors were investigated to be randomly used in tree splits. For the minimum number of observations at leaves, which defined the individual tree size or depth, the values of 10, 20, 30, 40, 50, 100, 200, and 500 observations were tested.

The optimal model for each soil attribute was determined by the minimum averaged Root Mean Square Error (RMSE, Equation 4) from the calibration set after testing all the combinations of hyperparameters. The calibration set was composed by the sampled observations from the bootstrapping, while the remaining observations

that were hold out were used for testing, resulting in an out-of-bootstrapped validation. The predictive model's accuracy was evaluated using the RMSE (Equation 4). The Coefficient of Determination (R2, Equation 5) was calculated to assess the explained variance of the prediction models, and the Ratio of Performance to Interquartile range (RPIQ, Equation 6) was calculated to assess the consistency between the predicted values with the variability of soil observations. The final evaluation metrics were calculated by averaging the statistics from the out-of-the-bootstrapped samples (testing observations), reporting both the mean and the standard deviation.

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right)} \quad \text{Equation 4}$$

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}} \quad \text{Equation 5a}$$

$$SS_{residuals} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad \text{Equation 5b}$$

$$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{Equation 5c}$$

$$RPIQ = \frac{IQR_y}{RMSE_{\hat{y}}} \quad \text{Equation 6}$$

where y is the vector of measured values, $\hat{y}$ is the vector of predicted values, $\bar{y}$ is the mean of vector y, and IQR is the interquartile range.

For the visualization, all the produced maps were masked by the cropland's location identified by the CGLS collection 2 referenced to the year of 2015. CGLS collection 2 has identified 65.6 Mha in Brazil, but the bare soil composite produced by GEOS3 represented around 94% of this amount. Yet, GEOS3 still provided significant coverage on cropland soils that allowed to assess the predicted distribution of key soil attributes among the main agricultural regions and calculate the total SOC stock considering the 0-20 cm soil layer.

Additionally, the predicted total SOC stock was compared with available maps from previous studies. A reference SOC stock for the 0-20 cm soil layer was prepared from the predictions of Gomes et al. (2019), once the original study provided estimates for 0-5, 5-15 and 15-30 cm. The stock for 0-20 cm from Gomes et al. (2019) was estimated using the successive layers and a weighted sum (Simó et al., 2014). Other

two reference datasets were prepared from Vasques et al. (2017), which is part of the Global Soil Organic Carbon Map from FAO's Global Soil Partnership, and the SoilGrids 2.0 (Hengl et al., 2017). These two datasets have originally produced estimates of SOC stock for the 0-30 cm soil layer. The SOC stocks maps from these three sources were masked by the croplands location from the CGLS collection 2, which allowed both the calculation of the total SOC stock for croplands soils, and the comparison of the spatially predicted patterns over a reference site located within -54.495 and -48.591 degrees of Longitude, and -24.970 and -21.200 degrees of Latitude.

4.3. RESULTS

4.3.1. Characterization of soil data

Soil samples used in this study have a highly skewed distribution for the particle sizes (Figure 2). The mean value of clay, silt and sand is respectively 213, 99 and 689 g kg^{-1}, revealing that most of the samples belong to sandy textured soils. The SOC content has a mean value of 10 g kg^{-1} with standard deviation of 5 g kg^{-1}, while pH has the most normally distribution around the mean value of 5.6. The CEC has a mean of 44 $mmol_c$ kg^{-1} and follows the same pattern distribution of SOC, confirming the importance of organic matter for cation exchange capacity in tropical soils (Cherubin *et al.*, 2015; Ondrasek *et al.*, 2019). Soil bulk density has a mean value of 1.4 g cm^{-3} and can be notably related by soil texture and organic carbon, while the SOC stock has a mean value of 2.5 kg m^{-2} with a standard deviation of 1.1 kg m^{-2}. Despite the occurrence of highly skewed distribution of soil attributes, ensemble machine learning algorithms does not require rigid statistical assumptions about the distribution and can provide robust estimations for the predicted values (Hengl *et al.*, 2018).

Attributes	n	Min.	Mean	SD	Max.
Clay (g kg^{-1})	4847	10	213	172	830
Silt (g kg^{-1})	4847	10	99	97	714
Sand (g kg^{-1})	4847	10	689	237	969
SOC content (g kg^{-1})	4534	1	10	5	48
pH H_2O	4483	4.0	5.6	0.6	8.5
CEC ($mmol_c$ kg^{-1})	4549	6	44	30	695
Density (g cm^{-3})	4462	0.8	1.4	0.1	1.6
SOC stock (kg m^{-2})	4462	0.3	2.5	1.1	8.4

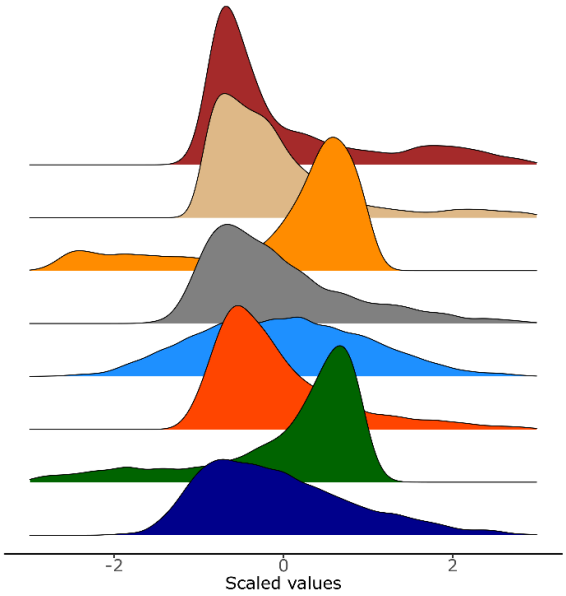

Figure 2. Summary statistics and distribution plots (scaled by mean and standard deviation) of soil samples used in this study. SOC: soil organic carbon; CEC: cation exchange capacity; Min.: minimum value; Max.: maximum value.

4.3.2. Environmental features extracted from EOD

When the SYSI reflectance was compared to resampled reflectance of soil data, the statistical correlation analysis resulted in coefficients varying from 0.68 to 0.86 (Table 1). The lowest correlation was identified for blue band with an estimated coefficient of 0.68, while the remaining bands had at least 70% of correlation. The residuals of the comparison were higher for SWIR2, with a mean and standard deviation of 0.08 and 0.06 decimals of reflectance, respectively (Table 1). In addition, the correlation between reflectance residuals and SF could indicate if regions with many exposures of soils provide a better estimate for the median reflectance. Nevertheless, the correlation of SF was insignificant and low to the reflectance residuals (Table 1). Therefore, the moderate to good correlation and low residuals of SYSI demonstrates the capacity of GEOS3 in separating bare soil pixels from the other spectral patterns present in satellite images, such as vegetation, crop residues and water bodies.

Table 1. Pearson's correlation coefficient between the synthetic soil image (SYSI) and the resampled reflectance of soil spectral library (SSL). Mean and standard deviation of the residuals calculated from the two reflectance sets (resampled laboratory and SYSI), and Pearson's correlation of residuals and soil frequency (SF).

Band[1]	ρ[2]	MAR[3]	SDAR[4]	ρ_{SF}[5]
Blue	0.66	0.02	0.02	0.18
Green	0.73	0.02	0.02	0.16
Red	0.70	0.03	0.02	0.17
NIR	0.78	0.04	0.03	0.19
SWIR1	0.87	0.06	0.05	0.20
SWIR2	0.86	0.09	0.06	0.26

[1]SWIR1, first shortwave infrared band; SWIR2, second shortwave infrared band. [2]Pearson's correlation between the SYSI and the resampled reflectance of SSL. All correlations are significant at 95% probability. [3]Mean of the absolute residuals between SYSI and resampled reflectance of SSL. [4]Standard deviation of the absolute residuals between SYSI and resampled reflectance of SSL. [5]Pearson's correlation between the SF and the absolute residuals of SYSI and resampled reflectance of SSL.

The shades of SYSI demonstrates a diversity of surface colors over the Brazilian territory (Figure 3A). The dark gray background represents missing information of bare soil exposure and is mostly associated to continuous cover of tropical forests or other land cover. As the land use changes on time and historical bare soil exposures can exist in places different from the current croplands, up to 36% of the territory had at least one bare soil exposure to compose the SYSI visualization. In turn, the results are more consistent if we consider the bare soil reflectance in regions dominated by croplands, where the crop cultivation makes the topsoil layer (0-20 cm) more homogeneous due to tillage practices. It is worthy to note that across the Cerrado biome (Savannas, Figure 1A), GEOS3 retrieved a considerable proportion of bare soils when the fractional cover of vegetation was probably low during the historical period of the image collection. The soil frequency confirms these patterns (Figure 3B), where a high frequency is found in regions close to croplands (Figure 1A), and a low to intermediate level of exposure is noticed over other areas within the Cerrado biome (Figure 3B). Semi-arid areas also had a higher frequency of soil exposures.

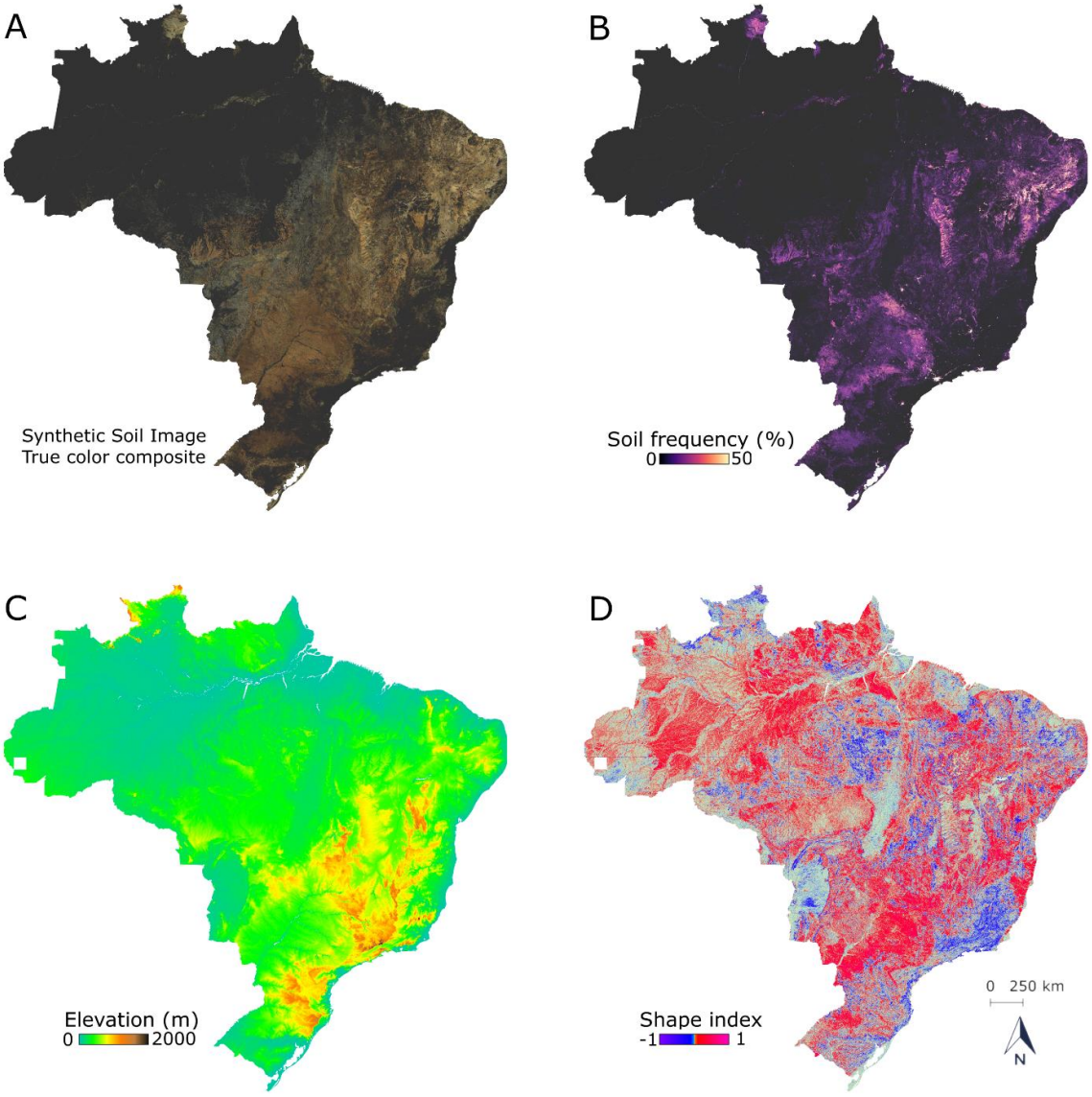

Figure 3. Synthetic Soil Image (SYSI) with the true color composition. (B) Bare soil frequency (SF) from 1982 until 2018. (C) Digital elevation model (DEM) from the SRTM. (D) Shape index calculated from the SRTM DEM. All the environmental features have 30 m resolution. Dark background represents unmapped area.

The diversity of landforms is also noticeable within the Brazilian territory (Figure 3C and 3D). From the seven terrain attributes used for calibrating prediction models, only the elevation and shape index are represented in Figure 3. The SRTM DEM of 30 m reveals the predominance of landscape with less than 1000 m of elevation (Figure 3C). Higher elevated areas are mostly situated in the south and southeastern regions, where the temperature decreases because of topography and

can impact some soil process. In north region, the Amazon forest prevails over the lowlands of the Amazon basin (Figure 1A). The shape index (SI) of terrain (Figure 3D) also depict the territory in three different landforms: convex (SI > 0), flat (SI around 0) and concave (SI < 0). Therefore, considering the visualization level, two major basins can be depicted by the shape index, i.e. the Amazon basin in northwest and the Paraná basin in southeast and south.

4.3.3. Prediction of soil attributes

When the soil attributes were calibrated by machine learning models with the environmental features extracted from EOD, the prediction performance was higher for clay, sand and SOC, reaching a RMSE of 87.31, 131.46 and 3.76 g kg^{-1}, respectively, and a R^2 and RPIQ of at least 0.50 and 1.52, respectively (Table 2). Other attributes had an inferior accuracy for all evaluation parameters, and a low deviation (determined from the bootstrapped ensemble) was identified for all the models, regardless its performance. The poor performance for pH (R^2 of 0.07) is possibly linked to the variability caused by crop management, such as lime applied to soils. Furthermore, the poor performance of pH coincided with a high number of observations in the leaves during the model calibration, where the predictions were averaged from 500 samples in the terminal leaves, being not capable of estimating a precise value. The best prediction models were optimized with 30 and 200 trees and a small number of samples in its leaves (40-50), which resulted in more precise estimates evidenced by the evaluation parameters (Table 2).

Table 2. Evaluation metrics and optimal hyperparameters of ensembled bootstrapped regression trees used to map agricultural soils in Brazil. The evaluation metrics are provided with their mean and standard deviation (in parenthesis).

Attribute[1]	FS[2]	NRF[3]	MSL[4]	[5]R^2	RMSE[6]	RPIQ[7]
Clay (g kg^{-1})	30	13	50	0.74 (0.01)	87.31 (2.08)	2.08 (0.09)
Sand (g kg^{-1})	200	13	40	0.69 (0.02)	131.46 (3.61)	1.63 (0.09)
Silt (g kg^{-1})	30	13	50	0.37 (0.04)	76.52 (2.69)	1.08 (0.04)
SOC (g kg^{-1})	30	8	40	0.50 (0.03)	3.76 (0.13)	1.52 (0.06)
CEC ($mmol_c$ kg^{-1})	100	11	50	0.27 (0.04)	25.33 (2.18)	1.03 (0.09)
pH H_2O	30	11	500	0.07 (0.01)	0.57 (0.01)	1.54 (0.07)
SOC stock (kg m^{-2})	30	13	100	0.43 (0.03)	1.02 (0.04)	1.51 (0.04)

[1]SOC: soil organic carbon, CEC: cation exchange capacity, pH H2O: soil pH determined in water solution, SOC stock: soil organic carbon stock for the 0-20 cm layer; [2]FS: forest size, i.e. number of bootstrapped trees; [3]NRF: hyperparameter number of random features tested in each tree split; [4]MSL: hyperparameter minimum samples at leaves; [5]R^2: coefficient of determination; [6]RMSE: root mean squared error; [7]RPIQ: ratio of performance to interquartile range.

The most important variable for predicting clay and sand content was SWIR2, reaching a mean permutation importance of more than 150% (Figure 4). Since the clay model relied almost exclusively on this spectral band, the permutation importance indicates that shuffling SWIR2 values could increase the prediction error by 1.5 folds (Breiman, 2001). For soil texture components (clay and sand), the importance of SWIR2 varied from around 130 to almost 200%, confirming its consistency at each individual tree that compose the forest ensemble. For SOC prediction, the red, NIR, SWIR2 and elevation had a significant contribution, with a mean importance of at least 20% for each feature (Figure 4). However, elevation was the most consistent feature for the individual trees. Among the relevant spectral features considered for SOC prediction, SWIR2 had the highest variation with estimates fluctuating from around 20% to 80%. The red and NIR bands, on the other hand, varied between 10 and 40 and 60%, respectively.

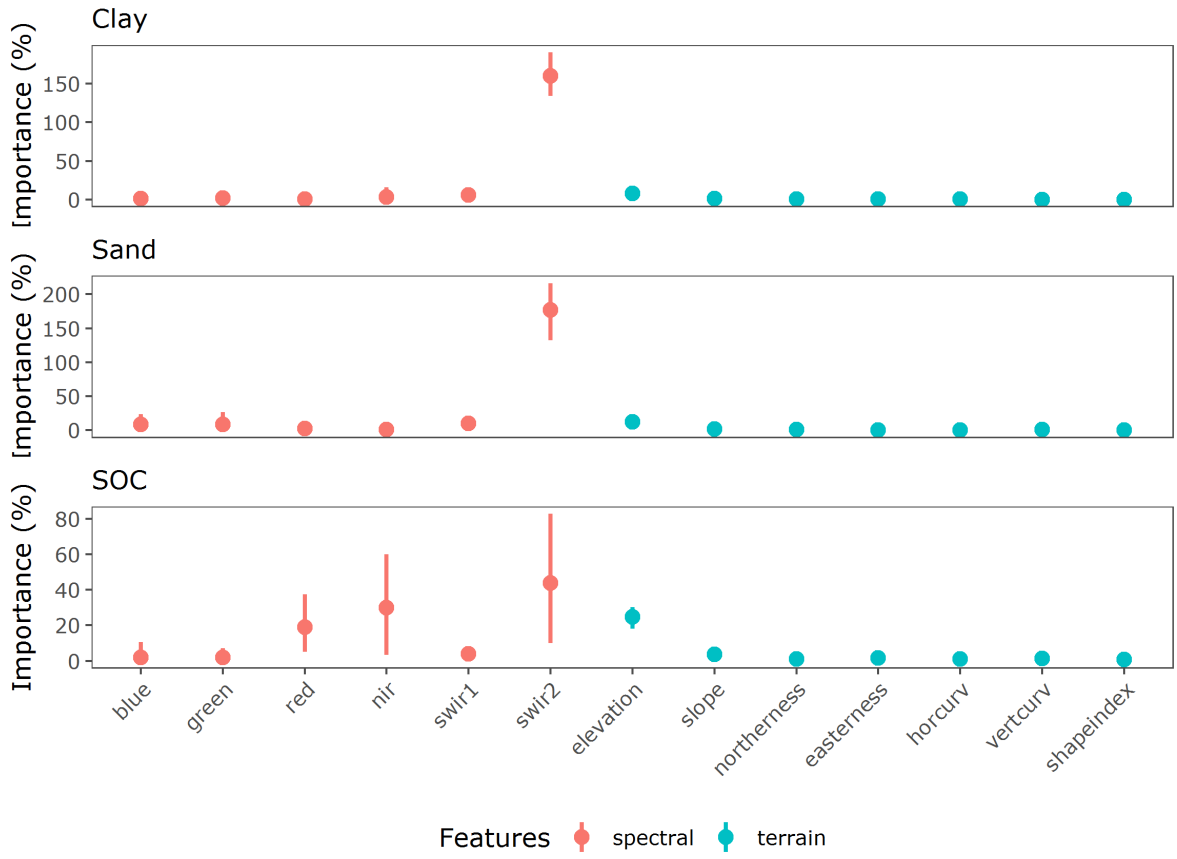

Figure 4. Variable importance for soil properties prediction (clay, sand and SOC), categorized by spectral (from SYSI) and terrain features. The points represent the mean values, while the line ranges are displayed by the minimum and maximum importance estimated from the ensemble of bootstrapped trees.

4.3.4. Cropland soils

The clay attribute had the best validation parameters after employed to machine learning. As the clay content is almost inversely proportional to the sand content in tropical soils, the clay map was used as proxy for the soil texture. Higher clay contents were mostly predicted in the south and southeast regions in Brazil, with other significant sites in the Midwest as well (Figure 5A). Soils with lower clay contents were predicted on croplands of the Midwest and MATOPIBA (Figure 1A), in the center of the southeast region, and the southmost part of the country. A simple classified version of the clay map suggests that clayey (clay > 350 g kg^{-1}), medium texture (150 g kg^{-1} < clay <= 350 g kg^{-1}) and sandy soils (clay <= 150 g kg^{-1}) are equally distributed across the Brazilian territory, representing 35%, 32% and 33%, respectively. The maps

of uncertainty (Figure 5B) reveals that higher estimates exist on areas with low clay content and where a quick transition between distinct soil textures can happen.

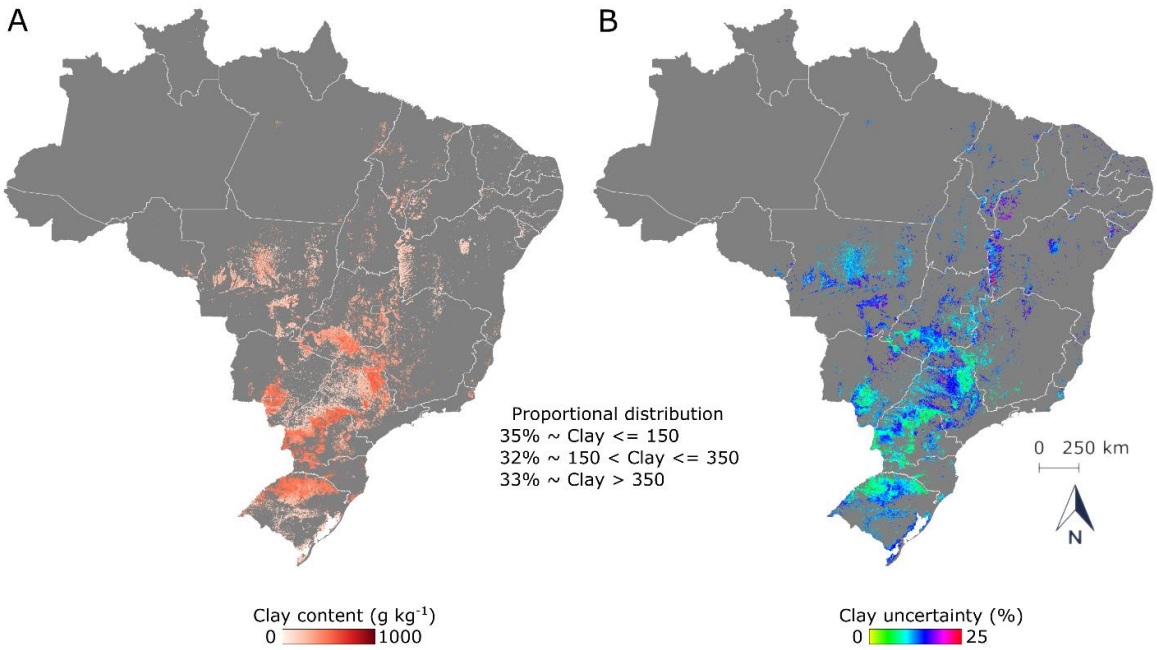

Figure 5. Predicted clay content (A) and its uncertainty (B) for the 0-20 cm soil layer across the croplands in Brazil. Gray background represents unmapped area, and the light gray lines are the boundaries of federal states.

The spatial distribution of SOC revealed the presence of significant stocks in the south and southern regions of Brazil (Figure 6A). The croplands with high SOC stocks are associated to soils with higher clay content (Figure 5A) and to elevated terrains (Figure 3C). Conversely, predicted SOC stocks for the Midwest and MATOPIBA regions (Figure 1A) were moderate to low. The highest uncertainty seems to be linked to the elevation gradient (Figure 3C) and to the frequency of bare soil exposures (Figure 3B). The predicted total SOC stock for the 0-20 cm layer across the Brazilian croplands is 2.04 Pg (CI 95%: 1.94, 2.14).

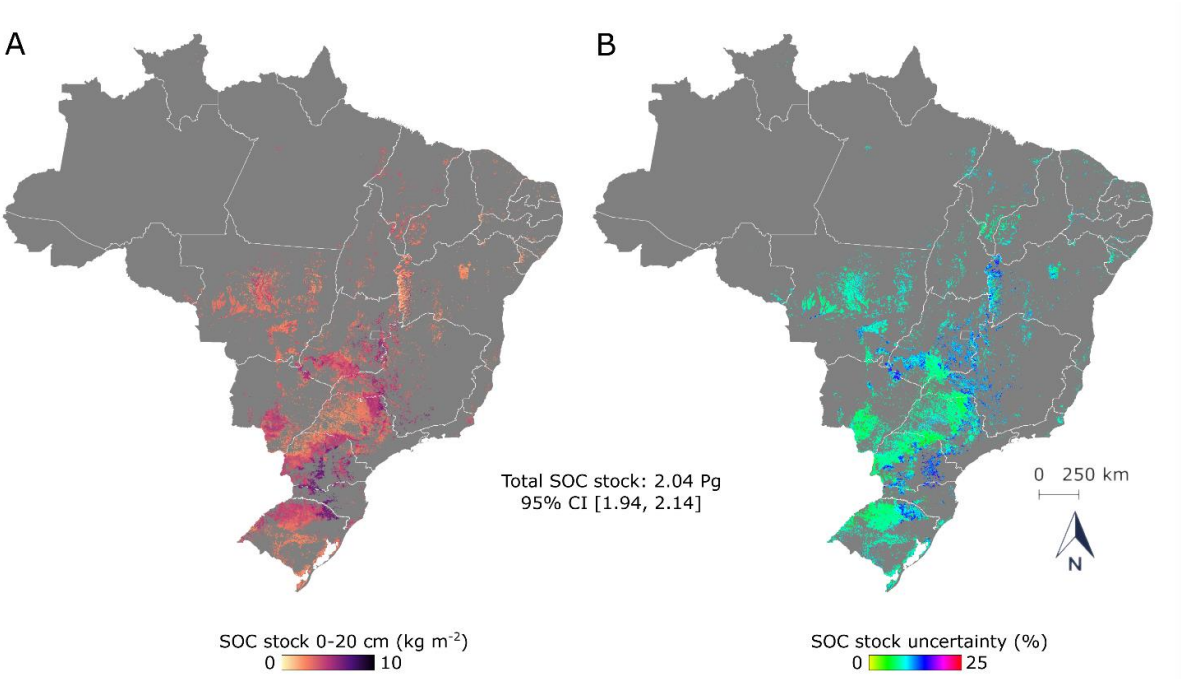

Figure 6. Predicted SOC stock (A) and its uncertainty (B) for the 0-20 cm soil layer across the croplands in Brazil. Gray background represents unmapped area, and the light lines are the boundaries of federal states.

In addition to the produced maps for the Brazilian territory, the comparison of the predicted SOC stock with other available maps revealed some differences of the total stock estimate and the spatial distribution at finer scales (Figure 7). The total SOC stock of croplands estimated from the dataset of Gomes et al. (2019) (0-20 cm), Vasques et al. (2017) (0-30 cm) and SoilGrids 2.0 for the 0-30 cm layer (Hengl et al., 2017) were 2.16, 2.87, and 2.98 Pg, respectively. The distinct results can be linked to the spatial patterns found at the local to regional level (Figure 7). The predicted map from this study yielded more spatially resolved transitions of SOC stocks, but also the lowest SOC stocks in places where the estimates are predominantly lower among all maps (Figure 7A). Some SOC pools become more evident at the top left and top right of the represented region when compared to map from Gomes et al. (2019) (Figure 7B). Furthermore, the SOC pools is quite different at the central right and bottom part of the same represented region when compared to Figure 7B. Despite the soil layer is different for the top and bottom maps (Figure 7), the predicted map of this study seems to be more compatible to the map from Vasques et al. (2017) (Figure 7C). The map from SoilGrids 2.0 (Hengl et al., 2017) (Figure 7D) provides the higher local estimates of SOC stock compared to all available maps, as also demonstrated by the total estimate masked for the Brazilian croplands. Overall, the predicted SOC stocks from

this study followed the major patterns presented in other similar studies, corroborating the presence of significant SOC pools at the south of the represented region, but with more spatially resolved transitions (Figure 7).

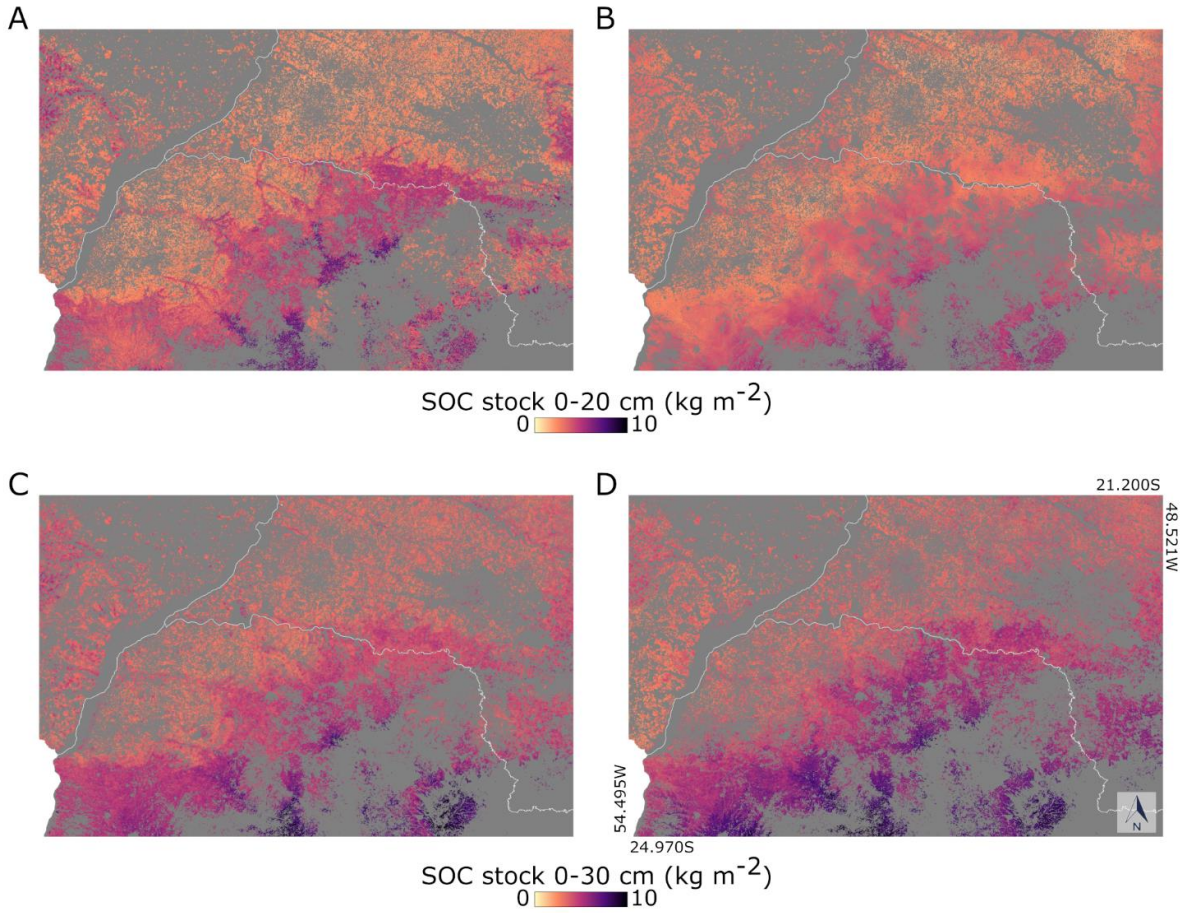

Figure 7. Comparison of the predicted SOC stock for the 0-20 cm soil layer on croplands (A) with other available maps: (B) SOC stock 0-20 cm from Gomes et al. (2019), (C) SOC stock 0-30 cm from Vasques et al. (2017), and (D) SOC stock 0-30 cm from SoilGrids 2.0 (Hengl et al., 2017). Gray background represents unmapped area (i.e. not croplands), and the light gray lines are the boundaries of federal states.

4.4. DISCUSSION

4.4.1. Environmental features extracted from EOD and soil data

A bare soil composite has been produced across the Brazilian territory based on the analysis of 36 years of Landsat EOD (Figure 3A). This approach allowed correlating the spectral patterns coupled with terrain attributes to predict topsoil attributes at 30 m resolution. GEOS3 produced for the first time a bare soil composite

representing 94% of the Brazilian croplands, which had a good agreement with the reference spectra of the BSSL (Table 1). The correlation coefficients demonstrated that, despite the challenges of multitemporal and multispectral satellite imagery for quantitative analysis of the soil surface, the resulting bare soil composite had a satisfactory association with the reference reflectance patterns measured in laboratory conditions.

The median statistics was used to attenuate extreme reflectance measurements when aggregating bare soil exposures from the multitemporal Landsat collection (Demattê et al. 2018; Poppiel et al. 2019; Safanelli et al. 2020). The multidate aggregation has been used to attenuate the effects of temporally dynamic aspects of the soil surface, such as soil rugosity, crust formation, moisture level, and even some misidentification of soil exposure from the historical EOD collection. Yet, the 30-m spatial resolution still might be affected by several factors, making difficult a complete attenuation of field conditions in automated processing systems. This strategy can also affect the prediction performance of dynamic soil attributes, e.g. pH, CEC and even SOC (Table 2). In turn, soil attributes that do not easily change by management practices or weather conditions correlates with stable landscape features present in the bare soil composite (Fongaro *et al.*, 2018; Lobell *et al.*, 2010). Previous studies have demonstrated the ability of bare soil composite for mapping key soil attributes due to the strong relationship with particle size distribution and surface minerals, which are thereby linked to underlying parent materials (Gallo *et al.*, 2018; Roberts, Wilford e Ghattas, 2019; Rogge *et al.*, 2018).

The prediction of soil attributes is also influenced by the quality of field observations and sampling campaign. Field data used in this study were gathered from previous works that sampled soils in transects or grid schemes using positioning systems at local to regional scales (Bellinaso, Demattê, and Romeiro 2010; Demattê et al. 2019). Some alternative datasets provide soil information with better spatial distribution in Brazil (Cooper *et al.*, 2005; Samuel-Rosa *et al.*, 2020), but they have issues with their positioning accuracy and became incompatible to the target resolution of this study. In addition to this aspect, the unbalanced distribution of soil attributes can potentially hamper the spatial predictions (Figure 1). Another limitation from the dataset used in this study was the estimated values for soil bulk density using an external pedotransfer function (Benites *et al.*, 2007), which was employed for the SOC stock determination. As the soil dataset had insufficient information about coarse fragments,

SOC stock was uncorrected for this factor as well. Therefore, despite the fact that the availability of soil data has always been a limitation in Brazil (Polidoro *et al.*, 2016), it is important to be aware of these potential issues on the interpretation of the results and future work.

4.4.2. Prediction models

The customized algorithm based on bootstrapped and random regression trees yielded moderate to good performance for clay, sand, SOC and SOC stock (Table 2). Ensemble and tree-based algorithms, especially Random Forest, have been outperforming other machine learning algorithms in DSM studies (Gomes *et al.*, 2019; Hengl *et al.*, 2015; Nussbaum *et al.*, 2018). The prediction method adopted in this study delivered both the mean and standard deviation when aggregating the bootstrapped trees, which allowed to spatially predict the uncertainty, similar to the method explored in other studies (Guevara *et al.*, 2018; Padarian, Minasny e McBratney, 2017; Viscarra Rossel e Chen, 2011).

Although spectral features are not depicted in multispectral EOD (Gomez et al., 2018), the reflectance intensity and the band differences were particularly important for predicting clay and sand contents (Figure 4). Soil reflectance is determined by inorganic and organic carbon, particles size distribution, mineralogy, and other factors, which affects the overall intensity and the spectral patterns of soils (Ben-Dor e Banin, 1995; Chabrillat *et al.*, 2019; Viscarra Rossel *et al.*, 2016). Terrain attributes, widely explored in DSM studies (Costa, Samuel-Rosa, and Anjos 2018; Guo et al. 2019; Moura-Bueno et al. 2016; Nussbaum et al. 2018; Poppiel et al. 2019; Sena et al. 2020), had an essential contribution particularly for SOC (Figure 4), possibly due to the known effects of elevation on terrestrial temperature, and consequently, to the SOC dynamic.

The application of EOD for predicting soil texture and related attributes over large geographical areas in Brazil have been already explored (Poppiel et al. 2019; Poppiel et al. 2020). Recently, a study have demonstrated the potential of bare soil composites for predicting clay and calcium carbonates in cropped areas across the European extent (Safanelli et al. 2020). Several other studies, developed in relatively small areas, have also demonstrated the value of bare soil composites or other spectral features extracted from EOD for mapping soil particle distribution, which achieved from

moderate to good performance in their predictions (Fongaro *et al.*, 2018; Gallo *et al.*, 2018; Gasmi *et al.*, 2019; Gomez *et al.*, 2018).

For predicting SOC-related attributes with EOD, there are relatively more studies already developed in Brazil (Gomes *et al.*, 2019; Guevara *et al.*, 2018; Vasques *et al.*, 2017) and in other places (Liang *et al.*, 2019; Pouladi *et al.*, 2019; Siewert, 2018; Tziolas *et al.*, 2020; Zhang *et al.*, 2020). The prediction model developed in this study for the SOC content had a satisfactory performance with a RMSE of around 4 g kg^{-1} and R^2 of 0.50, with a slightly small accuracy for SOC stock (Table 2). The results were almost identical to the mapping of Zeraatpisheh et al (2019) using Random Forests (RMSE of 3.3 g kg^{-1} and R^2 of 0.55), which combined terrain attributes and reflectance features for mapping the SOC content in central Iran. Additionally, when we compared the estimated total SOC stock with other datasets produced for the Brazilian territory, the results were relatively consistent (Figure 7). The estimated total SOC stock of croplands resulted in 2.04 Pg, while the other datasets yielded 2.16 Pg for the same 0-20 cm layer (Gomes *et al.*, 2019), and 2.87 Pg (Vasques *et al.*, 2017) and 2.98 Pg (Hengl *et al.*, 2017) for 0-30 cm.

Although it is difficult to make a fair comparison with the aforementioned studies, Guevara et al (2018) have reported a lower prediction performance ($R^2 < 0.2$) compared to this study (R^2 of 0.44, RMSE of 1.02 kg m^{-2}, and RPIQ of 1.51), despite the large availability of samples in their study. The results from Gomes et al. (2019) yielded similar SOC pools for the cropped area in Brazil, but the authors reported a performance of R^2 of 0.33 and RMSE from 0.56 to 1.19 kg m^{-2} for the topsoil layers. Similarly, Vasques et al. (2017) have reached a mapping performance for the 0-30 cm SOC stock with a R^2 of 0.24 and RMSE of 0.5 kg m^{-2}. The contrasting results with the previous studies can be associated to many factors, but mostly to the set of soil observations and the environmental predictors. Differently from the other works, this study has employed for the first time a bare soil composite coupled with terrain attributes for predicting the SOC stock on Brazilian croplands at a resolution of 30 m, revealing some peculiar details on the SOC spatial variability at the local and regional scales (Figure 7).

4.4.3. Cropland Soils

Legacy soil maps have long supported the interpretation of agricultural expansion in Brazil. Most of the Brazilian soil maps date from 1970-80 and were produced with low cartographical detail (scale lower than 1:500.000). This information helped the development of the country in several areas such as agriculture, mining, and environmental management. Nowadays, stakeholders are requesting detailed maps to promote more efficient and sustainable use of the land. In this sense, the country is trying to put in practice a new project of soil survey and mapping named 'Programa Nacional de Solos do Brasil' – PronaSolos (Polidoro *et al.*, 2016). This project will certainly demand EOD and digital technologies due to restriction on resources for mapping the large Brazilian territory. Therefore, the results produced in this study support the application of EOD for mapping and monitoring cropland soils, contributing with a brief overview of the historical agriculture expansion in Brazil as well.

Most of the clayey soils are situated in the south and southeast region (Figure 1A, contour 1) due to the underlying lithology of basalt deposits (Figure 5A). Soils with lower clay content are evident in the southernmost and eastern parts, where sedimentary settings prevail. The south and southeast region is the place of the first concentration of croplands in Brazil, which happened before 1970 after the growing interest in agriculture commodities for export (Dias *et al.*, 2016; Freitas, de e Landers, 2014). In the Midwest (contour 2 in Figure 1A), soils have a high clay content at the bottom of the region, where the underlying basalt materials are common, with sandy soils occurring towards north (Figure 5A). This region has the second agricultural frontier that expanded from 1960/1970s over the Brazilian Savannah and the lower part of the Amazonian biome (Figure 1A). Nowadays, this region incorporates the major production of soybean, cotton, corn, and cattle ranch, which conducted Brazil to top levels of agricultural commodities production and exports (Buol, 2009; Dias *et al.*, 2016; Morton *et al.*, 2016). The last and more recent expansion of Brazilian agriculture happened in the states of Maranhão, Tocantins, Piauí and Bahia (MATOPIBA) (Spera, 2017) after the 2000s (Figure 1A, contour 3), where soils with lower clay content are predominant.

Although the climate conditions and natural vegetation of the south and southeast of Brazil are distinct from the Midwest and MATOPIBA regions, we noticed

that the distribution of clay content is also different (Figure 6). Our results demonstrate that the historical agricultural expansion took place on areas dominated by sandy soils, expanding from the coast towards inside the country. This pattern corroborates to the development and propagation of new technologies for soil management and plant nutrition, such as the application of lime and fertilizer inputs, no-tillage, and biological nitrogen fixation (Buol, 2009; Döbereiner, 1997; Freitas, de e Landers, 2014). It is also worthy to note that most of the croplands in Brazil are positioned in plateaus or flatter terrain that make possible the mechanization of agriculture (Figure 3). Additionally, our results suggest the current croplands are almost equally distributed in distinct soil texture compositions (Figure 5).

In Brazil, the SOC stocks in the upper 20 cm are low compared to other countries with temperate climate. SOC stock values ranging from 2 to almost 7 kg m^{-2} were reported in literature for soils located in the Amazon region and Cerrado, but high values reaching up to 10 kg m^{-2} can be found in the southeast and south region (Batlle-Bayer, Batjes e Bindraban, 2010; Zinn, Lal e Resck, 2005). This amplitude of estimates resembles the spatially explicit information identified in our study, with higher pools located in the croplands of the south and southeast regions, similarly to what was identified in other studies (Figure 7). Yet, as our results provides an detailed overview about the spatial distribution of SOC stocks across the Brazilian croplands, it is necessary to explore in-depth evaluations about the organic carbon changes, integrating for example, dynamic modeling systems with geospatial data (Zinn, Lal e Resck, 2005).

4.4.4. Further considerations

Although we have demonstrated the potential of spectral and terrain features extracted from EOD for mapping agricultural soils, additional environmental information and soil datasets, coupled with other machine learning frameworks, can boost the prediction performance and provide more precise estimates. The availability of adequate soil data is an issue faced in this and other studies developed in Brazil, and some initiatives have been fostered in recent years seeking to gather and make public available soil datasets, despite their limitations (Samuel-Rosa *et al.*, 2020). Furthermore, the land cover map used to mask cropped areas can also introduce uncertainty in the analysis. The MapBiomas project (https://mapbiomas.org/en) is an

initiative that have been developing land cover maps for the whole Brazilian territory at 30-m resolution and have the potential to be explored in future studies with dynamic mapping of soil properties.

4.5. CONCLUSIONS

Exploring the full and open-access collection of Landsat surface reflectance data was decisive for retrieving the multispectral features of soils at 30 m resolution. The cloud-based interface of the Google Earth Engine was also important to calculate terrain attributes, which coupled with the multispectral features from the bare soil composite, legacy soil observations, and a machine learning algorithm, made possible the mapping of soil attributes across the Brazilian croplands. The spectral and terrain features extracted from Earth Observation data allowed the calibration of prediction models of clay, sand, SOC content and SOC stock with satisfactory accuracy, reaching an R^2 ranging from 0.44 to 0.74, and a RPIQ of higher than 1.5.

The estimated total SOC stock (0-20 cm) for the Brazilian croplands is 2.04 Pg (CI 95%: 1.94, 2.14). The SOC stock mapping produced a consistent result with previous datasets available in Brazil. However, the produced map allowed us to depict the spatial variability of SOC stock from the local to the regional level. Furthermore, our results suggest that croplands are proportionally distributed in regions with different soil texture, but the historical expansion happened towards sandy soils. Thus, this study supports the proposition that EOD is a valuable source for extracting environmental features for mapping and monitoring cropland soils at finer resolutions, assisting the evaluation of soil spatial distribution and the historical agriculture expansion in Brazil.

ACKNOWLEDGMENTS

This study had financial support from the São Paulo Research Foundation (FAPESP), grants numbers 2016/01597-9 and 2014/22262-0. The authors are also grateful to Geotechnology in Soil Science Group (https://esalqgeocis.wixsite.com/english).

REFERENCES

Angelopoulou, T. *et al.* (2019) “Remote Sensing Techniques for Soil Organic Carbon Estimation: A Review,” *Remote Sensing*, 11(6), p. 676. doi: 10.3390/rs11060676.

Azzari, G. *et al.* (2019) “Satellite mapping of tillage practices in the North Central US region from 2005 to 2016,” *Remote Sensing of Environment*, 221, pp. 417–429. doi: 10.1016/j.rse.2018.11.010.

Batlle-Bayer, L., Batjes, N. H. and Bindraban, P. S. (2010) “Changes in organic carbon stocks upon land use conversion in the Brazilian Cerrado: A review,” *Agriculture, Ecosystems & Environment*, 137(1–2), pp. 47–58. doi: 10.1016/j.agee.2010.02.003.

Bellinaso, H., Demattê, J. A. M. and Romeiro, S. A. (2010) “Soil Spectral Library and Its Use in Soil Classification,” *R. Bras. Ci. Solo*, 34(3), pp. 861–870. doi: 10.1590/S0100-06832010000300027.

Ben-Dor, E. *et al.* (2009) “Using Imaging Spectroscopy to study soil properties,” *Remote Sensing of Environment*. Elsevier Inc., 113, pp. S38–S55. doi: 10.1016/j.rse.2008.09.019.

Ben-Dor, E. and Banin, A. (1995) “Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm),” *International Journal of Remote Sensing*, 16(18), pp. 3509–3528. doi: 10.1080/01431169508954643.

Benites, V. M. *et al.* (2007) “Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil,” *Geoderma*, 139(1–2), pp. 90–97. doi: 10.1016/j.geoderma.2007.01.005.

Breiman, L. (2001) “Random Forests,” *Machine Learning*. Kluwer Academic Publishers, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Buchhorn, M. *et al.* (2020) “Copernicus Global Land Cover Layers—Collection 2,” *Remote Sensing*, 12(6), p. 1044. doi: 10.3390/rs12061044.

Buol, S. W. (2009) “Soils and agriculture in central-west and north Brazil,” *Scientia Agricola*, 66(5), pp. 697–707. doi: 10.1590/S0103-90162009000500016.

Cao, B. *et al.* (2019) “Spatial modeling of litter and soil carbon stocks on forest land in the conterminous United States,” *Science of The Total Environment*, 654, pp. 94–106. doi: 10.1016/j.scitotenv.2018.10.359.

Chabrillat, S. *et al.* (2019) “Imaging Spectroscopy for Soil Mapping and Monitoring,” *Surveys in Geophysics*, 40(3), pp. 361–399. doi: 10.1007/s10712-019-09524-0.

Chaddad, F. (2016) *The Economics and Organization of Brazilian Agriculture*. Elsevier. doi: 10.1016/C2014-0-00991-4.

Cherubin, M. R. *et al.* (2015) “Sugarcane expansion in Brazilian tropical soils—Effects of land use change on soil chemical attributes,” *Agriculture, Ecosystems & Environment*, 211, pp. 173–184. doi: 10.1016/j.agee.2015.06.006.

Cooper, M. *et al.* (2005) “A National Soil Profile Database for Brazil Available to International Scientists,” *Soil Science Society of America Journal*, 69(3), p. 649. doi: 10.2136/sssaj2004.0140.

Costa, E. M., Samuel-Rosa, A. and Anjos, L. H. C. dos (2018) “Digital elevation model quality on digital soil mapping prediction accuracy,” *Ciência e Agrotecnologia*, 42(6), pp. 608–622. doi: 10.1590/1413-70542018426027418.

Demattê, J. A. M. *et al.* (2018) “Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images,” *Remote Sensing of Environment*, 212, pp. 161–175. doi: 10.1016/j.rse.2018.04.047.

Demattê, J. A. M. *et al.* (2019) “The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges,” *Geoderma*, 354, p. 113793. doi: 10.1016/j.geoderma.2019.05.043.

Dias, L. C. P. *et al.* (2016) “Patterns of land use, extensification, and intensification of Brazilian agriculture,” *Global Change Biology*, 22(8), pp. 2887–2903. doi: 10.1111/gcb.13314.

Diek, S. *et al.* (2017) “Barest Pixel Composite for agricultural areas using landsat time series,” *Remote Sensing*. Multidisciplinary Digital Publishing Institute, 9(12), p. 1245. doi: 10.3390/rs9121245.

Döbereiner, J. (1997) “Biological nitrogen fixation in the tropics: Social and economic contributions,” *Soil Biology and Biochemistry*, 29(5–6), pp. 771–774. doi: 10.1016/S0038-0717(96)00226-X.

Dotto, A. C. *et al.* (2016) “Potential of spectroradiometry to classify soil clay content,” *Revista Brasileira de Ciencia do Solo*, 40, pp. 1–8. doi: 10.1590/18069657rbcs20151105.

Drusch, M. *et al.* (2012) “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” *Remote Sensing of Environment*. Elsevier, 120, pp. 25–36. doi: 10.1016/J.RSE.2011.11.026.

Fongaro, C. *et al.* (2018) “Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images,” *Remote Sensing*, 10(10), p. 1555. doi: 10.3390/rs10101555.

Fonseca, M. G. *et al.* (2019) “Effects of climate and land-use change scenarios on fire probability during the 21st century in the Brazilian Amazon,” *Global Change Biology*, 25(9), pp. 2931–2946. doi: 10.1111/gcb.14709.

de Freitas, P. L. and Landers, J. N. (2014) “The Transformation of Agriculture in Brazil Through Development and Adoption of Zero Tillage Conservation Agriculture,” *International Soil and Water Conservation Research*, 2(1), pp. 35–46. doi: 10.1016/S2095-6339(15)30012-5.

Gallo, B. *et al.* (2018) “Multi-Temporal Satellite Images on Topsoil Attribute Quantification and the Relationship with Soil Classes and Geology,” *Remote Sensing*, 10(10), p. 1571. doi: 10.3390/rs10101571.

Gasmi, A. *et al.* (2019) “Surface soil clay content mapping at large scales using multispectral (VNIR–SWIR) ASTER data,” *International Journal of Remote Sensing*, 40(4), pp. 1506–1533. doi: 10.1080/01431161.2018.1528018.

Godfray, H. C. J. *et al.* (2010) “Food Security: The Challenge of Feeding 9 Billion People,” *Science*, 327(5967), pp. 812–818. doi: 10.1126/science.1185383.

Gomes, L. C. *et al.* (2019) “Modelling and mapping soil organic carbon stocks in Brazil,” *Geoderma*, 340, pp. 337–350. doi: 10.1016/j.geoderma.2019.01.007.

Gomez, C. *et al.* (2018) “Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios,” *Remote Sensing of Environment*, 204, pp. 18–30. doi: 10.1016/j.rse.2017.10.047.

Guevara, M. *et al.* (2018) “No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America,” *SOIL*, 4(3), pp. 173–193. doi: 10.5194/soil-4-173-2018.

Guo, Z. *et al.* (2019) “Selection of terrain attributes and its scale dependency on soil organic carbon prediction,” *Geoderma*, 340, pp. 303–312. doi: 10.1016/j.geoderma.2019.01.023.

Hengl, T. *et al.* (2015) “Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions,” *PLOS ONE*, 10(6), p. e0125814. doi: 10.1371/journal.pone.0125814.

Hengl, T. *et al.* (2017) “SoilGrids250m: Global gridded soil information based on machine learning,” *PLOS ONE*. Edited by B. Bond-Lamberty. Office for official publications of the European Communities, 12(2), p. e0169748. doi: 10.1371/journal.pone.0169748.

Hengl, T. *et al.* (2018) “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables,” *PeerJ*, 6, p. e5518. doi: 10.7717/peerj.5518.

IBGE (2018) *Census of Agriculture 2017*. Available at: https://www.ibge.gov.br/en/statistics/economic/agriculture-forestry-and-fishing/21929-2017-2017-censo-agropecuario-en.html?edicao=21928&t=o-que-e (Accessed: March 10, 2020).

Ji, W. *et al.* (2016) “Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions,” *Soil and Tillage Research*, 155, pp. 492–500. doi: 10.1016/j.still.2015.06.004.

Lagacherie, P. *et al.* (2008) “Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements,” *Remote Sensing of Environment*, 112(3), pp. 825–835. doi: 10.1016/j.rse.2007.06.014.

Liang, Z. *et al.* (2019) “High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling,” *Science of The Total Environment*, 685, pp. 480–489. doi: 10.1016/j.scitotenv.2019.05.332.

Lobell, D. B. *et al.* (2010) “Regional-scale Assessment of Soil Salinity in the Red River Valley Using Multi-year MODIS EVI and NDVI,” *Journal of Environmental Quality*, 39(1), pp. 35–41. doi: 10.2134/jeq2009.0140.

Martinelli, L. A. *et al.* (2010) “Agriculture in Brazil: impacts, costs, and opportunities for a sustainable future,” *Current Opinion in Environmental Sustainability*, 2(5–6), pp. 431–438. doi: 10.1016/j.cosust.2010.09.008.

McBratney, A. ., Mendonça Santos, M. . and Minasny, B. (2003) “On digital soil mapping,” *Geoderma*, 117(1–2), pp. 3–52. doi: 10.1016/S0016-7061(03)00223-4.

Mendes, W. de S. *et al.* (2019) “Is it possible to map subsurface soil attributes by satellite spectral transfer models?,” *Geoderma*. doi: 10.1016/j.geoderma.2019.01.025.

Morton, D. C. *et al.* (2016) “Reevaluating suitability estimates based on dynamics of cropland expansion in the Brazilian Amazon,” *Global Environmental Change*, 37, pp. 92–101. doi: 10.1016/j.gloenvcha.2016.02.001.

Moura-Bueno, J. M. *et al.* (2019) “Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions,” *Geoderma*. Elsevier, 337, pp. 565–581. doi: 10.1016/J.GEODERMA.2018.10.015.

Moura-Bueno, J. M. J. M. *et al.* (2016) “Assessment of Digital Elevation Model for Digital Soil Mapping in a Watershed with Gently Undulating Topography,” *Revista Brasileira de Ciência do Solo*, 40. doi: 10.1590/18069657rbcs20150022.

Noojipady, P. *et al.* (2017) “Forest carbon emissions from cropland expansion in the Brazilian Cerrado biome,” *Environmental Research Letters*, 12(2), p. 025004. doi: 10.1088/1748-9326/aa5986.

Nussbaum, M. *et al.* (2018) “Evaluation of digital soil mapping approaches with large sets of environmental covariates,” *SOIL*, 4(1), pp. 1–22. doi: 10.5194/soil-4-1-2018.

Ondrasek, G. *et al.* (2019) “Biogeochemistry of soil organic matter in agroecosystems & environmental implications,” *Science of The Total Environment*, 658, pp. 1559–1573. doi: 10.1016/j.scitotenv.2018.12.243.

Padarian, J., Minasny, B. and McBratney, A. B. (2017) “Chile and the Chilean soil grid: A contribution to GlobalSoilMap,” *Geoderma Regional*, 9, pp. 17–28. doi: 10.1016/j.geodrs.2016.12.001.

Pedregosa, F. *et al.* (2011) “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Phalan, B. *et al.* (2013) “Crop Expansion and Conservation Priorities in Tropical Countries,” *PLoS ONE*. Edited by S. G. Willis, 8(1), p. e51759. doi: 10.1371/journal.pone.0051759.

Picoli, M. C. A. *et al.* (2018) “Big earth observation time series analysis for monitoring Brazilian agriculture,” *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, pp. 328–339. doi: 10.1016/j.isprsjprs.2018.08.007.

Polidoro, J. C. *et al.* (2016) *Programa Nacional de Solos do Brasil (PronaSolos)*. 1st ed. Rio de Janeiro, RJ: Embrapa Solos.

Poppiel, R. R. *et al.* (2019) “Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil,” *Remote Sensing*, 11(24), p. 2905. doi: 10.3390/rs11242905.

Poppiel, R. R. *et al.* (2020) “Soil Color and Mineralogy Mapping Using Proximal and Remote Sensing in Midwest Brazil,” *Remote Sensing*, 12(7), p. 1197. doi: 10.3390/rs12071197.

Pouladi, N. *et al.* (2019) “Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging,” *Geoderma*, 342, pp. 85–92. doi: 10.1016/j.geoderma.2019.02.019.

Roberts, D., Wilford, J. and Ghattas, O. (2019) “Exposed soil and mineral map of the Australian continent revealing the land at its barest,” *Nature Communications*, 10(1), p. 5297. doi: 10.1038/s41467-019-13276-1.

Rogge, D. *et al.* (2018) “Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014),” *Remote Sensing of Environment*. Elsevier, 205, pp. 1–17. doi: 10.1016/J.RSE.2017.11.004.

Safanelli, J. L., Chabrillat, S., *et al.* (2020) “Multispectral Models from Bare Soil Composites for Mapping Topsoil Properties over Europe,” *Remote Sensing*, 12(9), p. 1369. doi: 10.3390/rs12091369.

Safanelli, J. L., Poppiel, R. R., *et al.* (2020) “Terrain Analysis in Google Earth Engine: A Method Adapted for High-Performance Global-Scale Analysis,” *ISPRS International Journal of Geo-Information*, 9(6), p. 400. doi: 10.3390/ijgi9060400.

Samuel-Rosa, A. *et al.* (2020) “Open legacy soil survey data in Brazil: geospatial data quality and how to improve it,” *Scientia Agricola*, 77(1). doi: 10.1590/1678-992x-2017-0430.

Sena, N. C. *et al.* (2020) “Analysis of terrain attributes in different spatial resolutions for digital soil mapping application in southeastern Brazil,” *Geoderma Regional*, 21, p. e00268. doi: 10.1016/j.geodrs.2020.e00268.

Siewert, M. B. (2018) “High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: a case study in a sub-Arctic peatland environment,” *Biogeosciences*, 15(6), pp. 1663–1682. doi: 10.5194/bg-15-1663-2018.

Simó, I. *et al.* (2014) “Modelling Soil Organic Carbon stocks using a detailed soil map in a Mediterranean mountainous area,” in Arrouays, D. et al. (eds.) *GlobalSoilMap: Basis of the global spatial soil information system*. 1st ed. London: CRC Press, p. 421. doi: https://doi.org/10.1201/b16500.

Soriano-Disla, J. M. *et al.* (2014) “The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties,” *Applied Spectroscopy Reviews*, 49(2), pp. 139–186. doi: 10.1080/05704928.2013.811081.

Spera, S. (2017) “Agricultural Intensification Can Preserve the Brazilian Cerrado: Applying Lessons From Mato Grosso and Goiás to Brazil’s Last Agricultural Frontier,” *Tropical Conservation Science*, 10, p. 194008291772066. doi: 10.1177/1940082917720662.

Stabile, M. C. C. *et al.* (2020) “Solving Brazil’s land use puzzle: Increasing production and slowing Amazon deforestation,” *Land Use Policy*, 91, p. 104362. doi: 10.1016/j.landusepol.2019.104362.

Stevens, A. *et al.* (2013) “Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy,” *PLoS ONE*. Edited by H. Y. Chen, 8(6), p. e66409. doi: 10.1371/journal.pone.0066409.

Teixeira, P. C. *et al.* (2017) “Manual de métodos de análise de solo.” Brasília, DF: Embrapa, 2017. doi: http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1085209.

Tilman, D. *et al.* (2011) “Global food demand and the sustainable intensification of agriculture,” *Proceedings of the National Academy of Sciences*, 108(50), pp. 20260–20264. doi: 10.1073/pnas.1116437108.

Tóth, G. *et al.* (2018) “Monitoring soil for sustainable development and land degradation neutrality,” *Environmental Monitoring and Assessment*, 190(2), p. 57. doi: 10.1007/s10661-017-6415-3.

Tziolas, N. *et al.* (2020) “An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs,” *Remote Sensing of Environment*, 244, p. 111793. doi: 10.1016/j.rse.2020.111793.

USGS (2018a) *Landsat 4-7 Surface Reflectance Code LEDAPS Product Guide*. doi: https://www.usgs.gov/media/files/landsat-4-7-surface-reflectance-code-ledaps-product-guide.

USGS (2018b) *Landsat 8 Surface Reflectance Code LaSRC Product Guide*. doi: https://www.usgs.gov/media/files/landsat-8-surface-reflectance-code-lasrc-product-guide.

Vasques, G. M. *et al.* (2017) *Soil organic carbon stock at 0-30 cm map for Brazil: technical report*. Rio de Janeiro.

Viscarra Rossel, R. A. *et al.* (2016) “A global spectral library to characterize the world’s soil,” *Earth-Science Reviews*. Elsevier, 155, pp. 198–230. doi: 10.1016/j.earscirev.2016.01.012.

Viscarra Rossel, R. A. and Chen, C. (2011) “Digitally mapping the information content of visible-near infrared spectra of surficial Australian soils,” *Remote Sensing of Environment*, 115(6), pp. 1443–1455. doi: 10.1016/j.rse.2011.02.004.

Wulder, M. A. *et al.* (2016) “The global Landsat archive: Status, consolidation, and direction,” *Remote Sensing of Environment*. Elsevier, 185, pp. 271–283. doi: 10.1016/j.rse.2015.11.032.

Zeraatpisheh, M. *et al.* (2019) “Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran,” *Geoderma*, 338, pp. 445–452. doi: 10.1016/j.geoderma.2018.09.006.

Zhang, Y. *et al.* (2020) “Estimating soil organic carbon and pH in Jilin Province using Landsat and ancillary data,” *Soil Science Society of America Journal*, 84(2), pp. 556–567. doi: 10.1002/saj2.20056.

Zinn, Y. L., Lal, R. and Resck, D. V. S. (2005) “Changes in soil organic carbon stocks under agriculture in Brazil,” *Soil and Tillage Research*, 84(1), pp. 28–40. doi: 10.1016/j.still.2004.08.007.

5. FINAL REMARKS

We demonstrated in this study that the collection of Landsat images is a valuable source for extracting environmental features useful for mapping and monitoring cropland soils. The cloud-based platform of Google Earth Engine has become an essential platform for processing and analyzing large-scale geospatial data. In this platform, we were able to develop the algorithms for extracting spectral and temporal features related to soils, and also a develop a toolbox for making terrain analysis seamlessly, which can be adapted to customized needs and scale up to the global scale. With the combination of extracted features from Earth Observation data, legacy soil datasets, and machine learning algorithms, we performed the medium-resolution mapping of cropland soils over the geographical extents of Europe and Brazil.

In chapter 1, we presented a study that explored the development of prediction models of key soil attributes using bare soil composites and the European Land-Use/Land-Cover Area Frame Survey (LUCAS) soil dataset. The bare soil composite based on the median of 37 years of Landsat imagery allowed the prediction of clay and calcium carbonates over croplands with moderate performance. When compared to laboratory-based models, the prediction models based on satellite images were relatively robust. In that study, we confirmed that bare soil composites can be added to digital soil mapping of croplands at finer scales covering large geographical extents. Furthermore, the generation of the bare soil composite using a method originally developed for tropical soils, i.e. GEOS3, with some customizations, provided consistent results for mapping the European extent.

In chapter 2, we developed and made available a package (TAGEE) to calculate terrain attributes using the high-performance platform of GEE. The package was adapted to not require the projection of input elevation data for terrain attributes calculation. The comparison between terrain analysis algorithms demonstrated that TAGEE had an accuracy comparable to other available tools. Thus, TAGEE became available for the geospatial community for making terrain analysis customized to their needs and adapted to global scale analysis. Furthermore, that package was particularly important for preparing additional information for mapping the cropland soils in Brazil.

In chapter 3, the integration of spectral and terrain features with legacy soil observations and a machine learning algorithm made possible the mapping of soil attributes across the Brazilian croplands. The spectral and terrain features extracted from Earth Observation data allowed the calibration of prediction models of clay, sand, SOC content and SOC stock with satisfactory accuracy. With the resulting maps, we were able to estimate the total SOC stock for the 0-20 cm layer, and identify some aspects related to the distribution of soil attributes regarding the main agricultural regions. In summary, that study supported the proposition that EOD is a valuable source for extracting environmental features for mapping and monitoring cropland soils at finer resolutions, assisting the evaluation of soil spatial distribution and the historical agriculture expansion in Brazil.

APPENDIX

APPENDIX A.

Table 1. Harmonization of the specific band numbers to common spectral names.

Common name[1]	Landsat 4 TM[2]	Landsat 5 TM[3]	Landsat 7 ETM+[4]	Landsat 8 OLI[5]
Blue	1 (450-520 nm)	1 (450-520 nm)	1 (450-520 nm)	2 (452-512 nm)
Green	2 (520-600 nm)	2 (520-600 nm)	2 (520-600 nm)	3 (533-590 nm)
Red	3 (630-690 nm)	3 (630-690 nm)	3 (630-690 nm)	4 (636-673 nm)
NIR	4 (770-900 nm)	4 (770-900 nm)	4 (770-900 nm)	5 (851-879 nm)
$SWIR_1$	5 (1550-1750 nm)	5 (1550-1750 nm)	5 (1550-1750 nm)	6 (1566-1651 nm)
$SWIR_2$	7 (2080-2350 nm)	7 (2080-2350 nm)	7 (2080-2350 nm)	7 (2107-2294 nm)

[1]NIR: Near infrared; SWIR1: Shortwave infrared 1; SWIR2: Shortwave infrared 2. [2]Landsat 4 Thematic Mapper (TM). [3]Landsat 5 Thematic Mapper (TM) sensor. [4]Landsat 7 Enhanced Thematic Mapper (ETM+) sensor. [5]Landsat 8 Operational Land Imager (OLI) sensor.

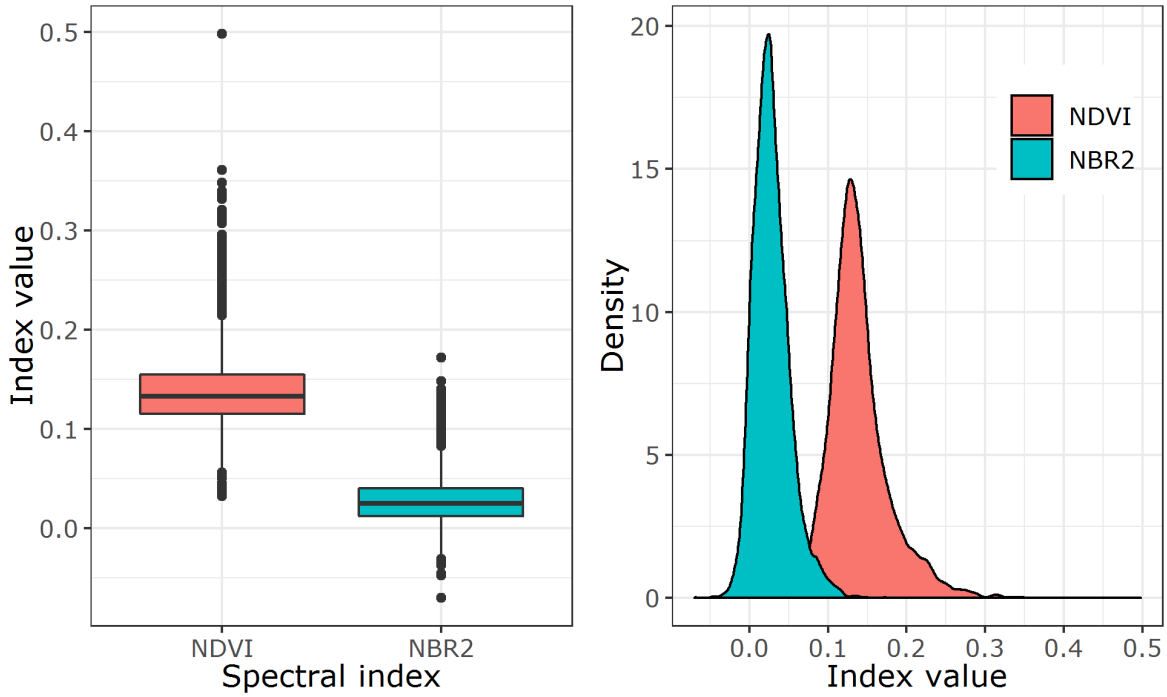

Figure 1. Boxplot and density plot of spectral indices calculated from the convolved reflectance measurements of LUCAS topsoil samples. Normalized Difference Vegetation Index (NDVI) and Normalized Burn Ration 2 index (NBR2) are used to identify potential soil pixels on satellite images.

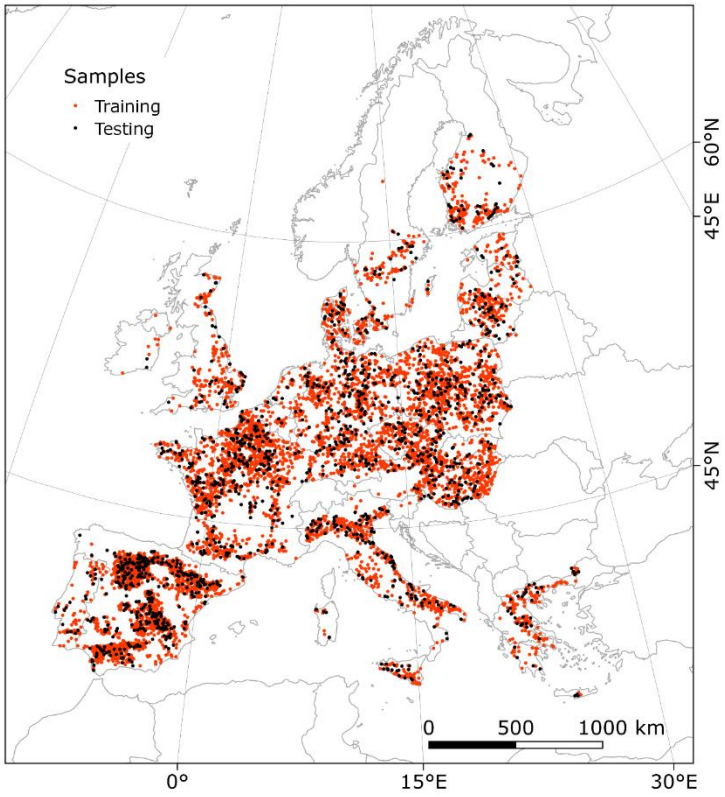

Figure 2. Location of sampling points (n = 7142) used in this study, which are a subset from the LUCAS dataset of 2009 and were split in training (80%) and testing (20%) samples.

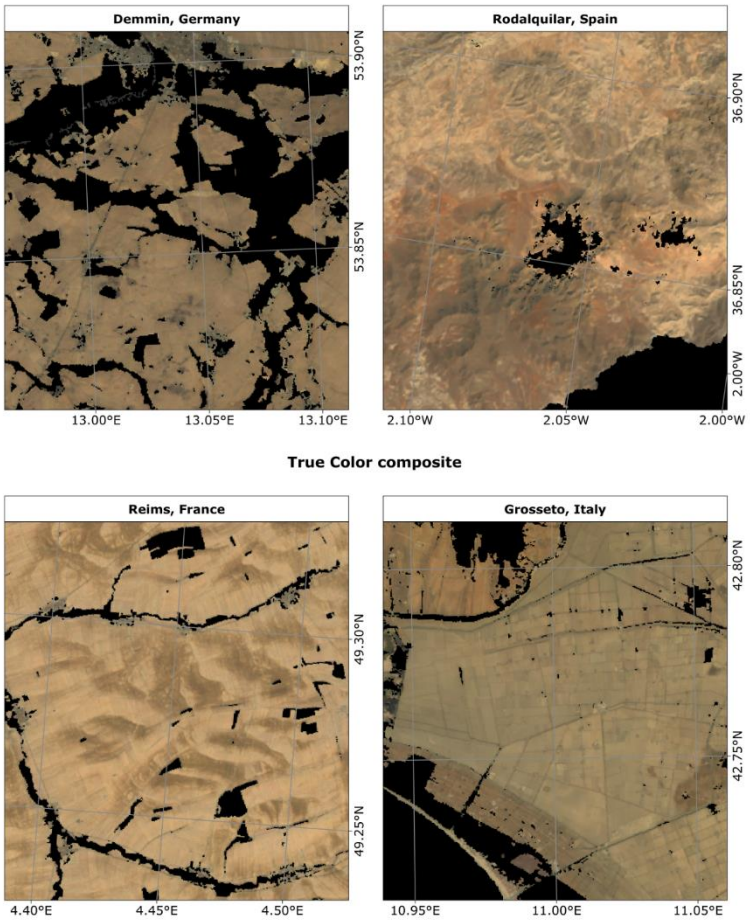

Figure 3. Regional maps of full SYSI represented by the true color composite, for the same sites of Figure 5 (Germany, France, Spain and Italy).

Table 2. Hyperparameters of the best regressions from gradient boosting trees, defined by 10-fold cross-validation of the training set (80%).

Soil attribute[1]	Reflectance source[2]	Seed	LR[3]	NE[4]	MF[5]	MD[6]	MSS[7]	MSL[8]
Clay	Original	1993	0.10	500	1	10	50	20
	Resampled	1993	0.20	500	6	10	50	10
	Framed SYSI	1993	0.10	250	6	5	100	20
	Full SYSI	1993	0.15	500	4	8	200	20
Sand	Original	1993	0.10	250	5	8	50	10
	Resampled	1993	0.10	500	4	10	50	20
	Framed SYSI	1993	0.10	250	4	10	100	5
	Full SYSI	1993	0.15	500	2	10	50	10
SOC	Original	1993	0.10	100	1	10	200	5
	Resampled	1993	0.10	500	6	5	50	20
	Framed SYSI	1993	0.10	100	2	5	100	20
	Full SYSI	1993	0.10	100	4	8	50	20
$CaCO_3$	Original	1993	0.10	100	1	5	100	20
	Resampled	1993	0.20	500	2	8	50	20
	Framed SYSI	1993	0.10	100	4	8	100	20
	Full SYSI	1993	0.10	100	4	10	200	5
PH H_2O	Original	1993	0.10	100	1	10	50	10
	Resampled	1993	0.10	500	6	8	50	20
	Framed SYSI	1993	0.15	250	4	10	100	20
	Full SYSI	1993	0.15	250	4	8	200	20
CEC	Original	1193	0.10	250	1	10	50	20
	Resampled	1993	0.10	500	6	10	50	20
	Framed SYSI	1993	0.10	100	4	8	200	20
	Full SYSI	1993	0.10	500	4	8	200	20

[1]Soil attributes: Soil organic carbon (SOC), calcium carbonate ($CaCO_3$), pH determined in water solution (pH H_2O), and cation exchange capacity (CEC). [2]Synthetic soil image (SYSI). Original and Resampled terms refer to the LUCAS absorbance spectra (450 to 2500 nm) used as reference prediction models, where the original was converted to reflectance and reduced by principal component analysis, and the resampled was converted to reflectance and resampled to the Landsat multispectral bands. Hyperparameters of gradient boosting trees from scikit-learn Python library: [3]learning_rate (LR); [4]n_estimators (NE); [5]max_features (MF); [6]max_depth (MD); [7]min_samples_split (MSS); [8]min_samples_leaf (MSL).

APPENDIX B.

Table 1. Harmonization of the specific band numbers to common spectral names.

Common name	Landsat 4 TM	Landsat 5 TM	Landsat 7 ETM+	Landsat 8 OLI
Blue	1 (450-520 nm)	1 (450-520 nm)	1 (450-520 nm)	2 (452-512 nm)
Green	2 (520-600 nm)	2 (520-600 nm)	2 (520-600 nm)	3 (533-590 nm)
Red	3 (630-690 nm)	3 (630-690 nm)	3 (630-690 nm)	4 (636-673 nm)
NIR	4 (770-900 nm)	4 (770-900 nm)	4 (770-900 nm)	5 (851-879 nm)
SWIR1	5 (1550-1750 nm)	5 (1550-1750 nm)	5 (1550-1750 nm)	6 (1566-1651 nm)
SWIR2	7 (2080-2350 nm)	7 (2080-2350 nm)	7 (2080-2350 nm)	7 (2107-2294 nm)

NIR: Near infrared; SWIR1: Shortwave infrared 1; SWIR2: Shortwave infrared 2.