Revista Brasileira de Ciência do Solo

Division – Soil Use and Management | Commission – Soil and Water Management and Conservation

Hydropedological digital mapping: machine learning applied to spectral VIS-IR and radiometric data dimensionality reduction

Priscilla Azevedo dos Santos[(1)*], Helena Saraiva Koenow Pinheiro[(2)], Waldir de Carvalho Júnior[(3)], Igor Leite da Silva[(4)], Nilson Rendeiro Pereira[(3)], Silvio Barge Bhering[(3)], and Marcos Bacis Ceddia[(5)]

[(1)] Universidade Federal Rural do Rio de Janeiro, Departamento de Petrologia e Geotectônica, Programa de Pós-Graduação em Modelagem e Evolução Geológica, Instituto de Geociências, Seropédica, Rio de Janeiro, Brasil.
[(2)] Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, Instituto de Agronomia, Seropédica, Rio de Janeiro, Brasil.
[(3)] Empresa Brasileira de Pesquisa Agropecuária, Embrapa Solos, Centro Nacional de Pesquisa de Solos, Jardim Botânico, Rio de Janeiro, Brasil.
[(4)] Universidade Federal Rural do Rio de Janeiro, Departamento de Matemática, Programa de Pós-Graduação Lato Sensu em Estatística Aplicada, Instituto Pitágoras, Seropédica, Rio de Janeiro, Brasil.
[(5)] Universidade Federal Rural do Rio de Janeiro, Departamento de Agrotecnologias e Sustentabilidade, Instituto de Agronomia, Seropédica, Rio de Janeiro, Brasil.

*** Corresponding author:**
E-mail: priscilla.agricart@hotmail.com

Received: November 22, 2022

Approved: June 05, 2023

How to cite: Santos PA, Pinheiro HSK, Carvalho Júnior W, Silva IL, Pereira NR, Bhering SB, Ceddia MB. Hydropedological digital mapping: machine learning applied to spectral VIS-IR and radiometric data dimensionality reduction. Rev Bras Cienc Solo. 2023;47:e0220149
https://doi.org/10.36783/18069657rbcs20220149

Editors: José Miguel Reichert and Marcos Gervasio Pereirao .

ABSTRACT: Pedosphere-hydrosphere interface accounts for the association between soil hydrology and landscape, represented by topographic and Remote Sensing data support and integration. This study aimed to analyze different statistical radiometric and spectral data selection methods and dimensionality reduce environment-related data to support the classification of soil physical-hydric properties, such as soil basic infiltration rate (bir) and saturated hydraulic conductivity (Ksat); as well as to act in data mining processes applied to hydropedological properties digital mapping. Accordingly, research integrated information from Visible to Infrared (VIS-IR) spectral indices and Sentinel's 2A mission Multispectral Instrument (MSI) sensor bands, terrain numerical modeling and aerogeophysics set to model soil-water content in two soil layers (0.00-0.20 m and 0.20-0.40 m). Pre-processed data were subjected to statistical analysis (multivariate and hypothesis tests); subsequently, the methods were applied (variation inflation factor - VIF, Stepwise Akaike information criterion – Stepwise AIC, and recursive feature elimination - RFE) to mine covariates used for Random Forest modeling. Based on the results, there were distinctions and singularities in spectral and radiometric data selection for each adopted method; the importance degree, and contribution of each one to soil physical-hydric properties have varied. According to the applied statistical metrics and decision-making criteria (highest R^2 and lowest RMSE / MAE), the chosen methods were RFE (0.00-0.20 m layers) and Stepwise AIC (0.20-0.40 m layers) - both concerned with the assessed variables (bir and Ksat). This approach captured the importance of environmental variables and highlighted their potential use in hydropedological digital mapping at Guapi-Macacu watershed.

Keywords: geoprocessing, hydropedology, applied statistics, radiometry, remote sensing.

INTRODUCTION

Remote Sensing techniques significantly contributed to the understanding and supporting environmental phenomena on Earth's surface, especially in correct environmental resources management and sustainability studies (Brenner and Guasselli, 2015). The ability to collect information from different electromagnetic spectrum wavelengths (multiband) acquired by remote sensors is the main reason for this contribution, as it allows differentiating terrestrial targets and covering large areas (Brenner and Guasselli, 2015). Therefore, soil basic infiltration ratio (vib) and saturated hydraulic conductivity (Ksat) could be map and modeled according to spectral and topographic relationships extracted from products obtained from remote sensing.

Also, a significant advancement in the environmental geophysics field is notable, which focuses on hydrogeological research accounting for combining airborne data gathered by geophysical sensors with field data related to soil and rocks lithology, mineralogy, physics, and chemistry. These studies can assess and model water resources distribution in groundwater reservoirs in sedimentary watersheds (Novakowski et al., 2006; Madrucci et al., 2008; Kirsch, 2009; Lee et al., 2012; Lee and Lee, 2015; Pires and Miranda, 2017).

However, there's still a lack of studies concerning geophysical maps potential integration in water favorability (determining areas presenting the greatest groundwater occurrence potential) and in hydropedological dynamics, as well as research treating such data as input variables in Digital Soil Mapping (DSM). This scenario has inspired the search for an in-depth evaluation of the potential of this tool in soil-water modeling studies.

The literature brings several references about topographic attributes selection and classification to map soil properties. These studies follow geomorphometric Digital Elevation Model (DEM) covariates to depict soil/landscape ratio in a given study site (McKenzie and Austin, 1993; Wilson and Gallant, 2000; Böhner and Selige, 2006; Oliveira et al., 2017; Santos et al., 2019). Nowadays, digital soil mapping applied to soil properties remains mainly based on soil morphometric parameters; however, given the advancements in remote sensing applied to spatial-spectral resolutions and data availability, adopting indices based on the spectral bands association in predictive spatial modeling techniques, such as machine learning, has become a common reality in the target variables quantification (Chagas, 2006; Pinheiro, 2012; Cunha, 2013; Zhang et al., 2017).

Accordingly, although remote sensing using in digital modeling has grown, it is also possible that spectral indices' individual influence remains poorly explored when it comes to soil physical-hydric attributes modeling. Studies have focused on soil physical-chemical properties mapping, climatic regionalization and forest species differentiation based on spectral indices (Carvalho Junior et al., 2011; Pinheiro et al., 2019; Rajah et al., 2019).

Consistent data collection and integration are fundamental requirements for modeling aiming to reflect a representative assessment of the water resources state (Baalousha, 2010). However, the crucial point remains in treating these data to transform them into reliable information. The dynamic attributes of spatial variability related to physical-hydraulic aspects are generally controlled by various properties and variables in soils, mainly conditioned by the natural landscape (excluding atypical or extreme landscape alterations), with different effects depending on the analyzed depth (Manzione and Castrignanò, 2019).

Linked to this, factors such as difficulty in measuring dynamic physical-hydric variables in the field (recognition of the area; high costs in carrying out field campaigns; team displacement in large areas for data coverage; sluggishness in sample collections in one or more depth levels), the collected data absence for these variables in Brazil traditional technical soils surveys (hydropedological tests non-performance in the field and laboratory analysis) and restrictions naturally imposed by the study region (accessing regions infeasibility with large slopes and areas with dense forest), limit these properties mapping (Ottoni, 2005; Oliveira et al., 2017).

Considering that soil physical-hydric parameters are influenced by genetic soil properties associated with morphology, texture, and soil physics such as: particle size and aggregation degree, total porosity, soil density, soil surface cover type, profile moisture, organic matter amount, among others (Reichert et al., 1992; Everts and Kanwar, 1993; Bertoni et al., 2017), and also due to study area spatial variability (Klar, 1984); evaluate the potential spectral and radiometric data inputs that contribute to these properties quantification in digital pre-modeling can make the mapping process more robust and reliable in the study area.

Few studies in Brazil focus on modeling the surface and subsurface behavior of water in the soil, particularly considering variable pre-selection protocols and data dimensionality reduction. Most current studies explore statistical methods and data analysis techniques for modeling physical-hydric properties, but they do not significantly explore the input variables of these models (Granata et al., 2022; Yamaç et al., 2022). Furthermore, they rely on established knowledge of soil morphology properties (such as texture, porosity, soil density, and particle size) as the main input variables in the modeling process (McKeague et al., 1982; Chapuis, 2012).

Studies like Manzione and Castrignanò (2019), Granata et al. (2022) and Yamaç et al. (2022), on the other hand, focus on exploring models based on regressive analysis, geostatistics and machine learning using Ksat measurement best practices. However, they do not discuss the importance of variable pre-selection in the pre-modeling phase and fail to identify potential interrelated attributes that could be used to predict physical-hydraulic attributes beyond the previously mentioned soil properties. McBratney et al. (2003) suggest, among other topics, studies on potential environmental covariates for applications in DSM as an open topic for further discussions within the academic community.

On the contrary, certain studies, such as Fathololoumi et al. (2021), discuss a variable reduction in environmental covariates utilized in a soil moisture prediction Digital Soil Mapping (DSM) approach, incorporating satellite images and morphological covariates. However, this study solely relies on autocorrelation and collinearity analysis for variable reduction. It fails to explore model protocols within DSM to pre-select these variables (for instance, statistical hypothesis test applied in data set modeled by regression models like the cubist model used) or provide justification for their inclusion (previous environmental covariates analysis to understand all potential data that can be used to represent the area dynamics, not only the common explored ones in DSM). It is worth noting that the only topic addressed in the study is the feature importance obtained throughout the modeling process and predictive model performance step.

Adequate variability assessment of these properties can effectively aid in understanding the processes that cause attributes spatial variation, enabling proper water resources management in river basins and correct soil management. However, Atkinson and Tate (2000) and Gotway and Young (2002) infer that many statistical problems arise when integrating different data obtained at different scales and characterized by different support and uncertainty.

Conducting studies to define the most important input variables to represent soil physical-hydric properties variability and to rule out modeling issues associated with multicollinearity inflation and autocorrelation in data sets is essential, especially those models based on regressive analyses-based models to integrate physical-hydric attributes, in a robust and refined way (O'Hagan and McCabe, 1975; Andrews, 1991; Carvalho et al., 1999; Mihola and Bílková, 2014). Thus, some methods can optimize input variables selection and reduce dimensionality in the assessed data set, ensuring robustness and refinement in the modeling process.

Considering the Guapi-Macacu watershed as the study area target, the knowledge developed in these Digital Soil Mapping (DSM) stages enables sustainable management of water resources in the Rio de Janeiro State. This approach prevents issues such as aquifer depletion, reduction of groundwater levels and riverbeds, biodiversity loss, and negative impacts on agriculture due to water scarcity or reduced water availability.

Inspired by such a scenario, the present study aims to: (1) analyze different statistical methods based on multivariate analysis applied to select and reduce dimensionality, using topographic and remote sensing data related to vegetation, soil, and geology; and (2) map soil water properties spatial variability in Guapi-Macacu watershed in Rio de Janeiro, Brazil, based on Random Forest (RF) classifier and on using the most relevant variables selected in the previous data mining stage.

MATERIALS AND METHODS

Study area

Guapimirim-Macacu river basin was selected for the current study, which is located within Guanabara Bay Hydrographic Region V (HR-V) domain in Rio de Janeiro State Metropolitan Region (Figure 1). It encompasses Itaboraí, Guapimirim and Cachoeiras de Macacu counties political-managerial limits. The basin has 1,250.78 km² of capture-area and 199.2 km of perimeter. Its relief is featured by elevation ranging 0-2,254 m, which was assessed through the region's Digital Elevation Model (DEM), at 20 m spatial resolution (Figure 1).

The prevailing climate in the region is rainy tropical type, with dry winter. Mean annual temperature is close to 23 °C and mean annual rainfall ranges from 1,200 to 2,600 mm due to mountain buttresses (HWA et al., 2010). The region stands out for housing the Atlantic Forest biome, with tropical forest fractions and *Mar de Morros* environment vegetation characteristic with transition to Coastal Lowland (Pinheiro, 2015).

Regions geology is featured by Gráben da Guanabara: an elongated valley with plane bottom, with geological faults forming a sequence of sedimentary decomposition caused by the tectonic activity that had started in the Tertiary, at Macau Sedimentary Basin formation (Ferrari, 2001). Depositional events that have taken place in the region determine its geomorphological features, presenting alluvial, river and lacustrine features (Pinheiro, 2012).

Region soils present sedimentary deposition environment features given the local geomorphology's lacustrine and river influence. The basin is surrounded by valleys and mountains due to *Mar de Morros* environment's transition to high-altitude fields, reaching altitude up to 2,254 m, and leading to soils presenting great pedogenetic variability and taxonomic diversity such as: Ultisols, Cambisols, Gelisols, Oxisols and Neosols (Pinheiro, 2012).

Macacu database and hydropedological tests

The pedological database provides the collected samples morphological descriptions and physical-chemical analyses. They are an important part of the study since it focuses on featuring the basins soil. Data were collected in the basin area based on the Conditioned Latin Hypercube Sample (cLHS) technique. This technique allows using computer resources to infer the selected points in the area. These resources are highly representative of Guapi-Macacu River basin's environmental features (Mckay et al., 2000; Minasny and McBratney, 2006; Carvalho Júnior et al., 2014).

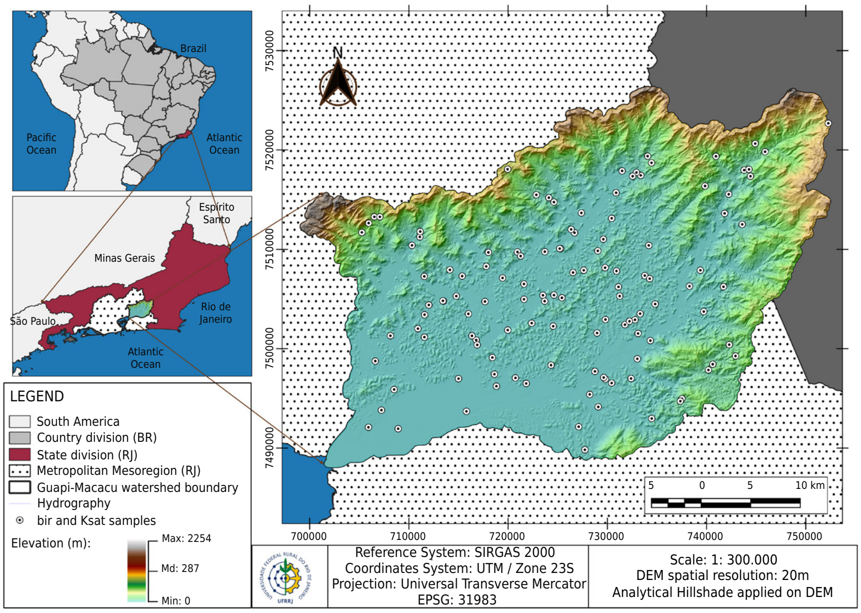

Figure 1. Study area location map and spatial distribution of the 122 points associated with hydropedological information about the assessed data set.

Therefore, a hundred and twenty-two (122) sampling points were collected in the Guapi-Macacu river basin area, originated as follows: fourteen (14) were obtained from pedological survey with profiles described in Carvalho Filho et al. (2003) (Extra Rio, Rio 9, PRJ 1, PRJ 2 id points) and a hundred and eight (108) points from pedological survey (ninety-nine obtained by the cLHS technique and nine using traditional survey method) with profiles described in Carvalho Júnior et al. (2014).

Soil basic infiltration rate (bir) data were collected through hydropedological survey campaigns (two in total, which occurred in September 2019 and February 2020), which were carried out *in situ* in the watershed; these campaigns used the Guelph permeameter, model 2800K1, by Soil Moisture, whose associated-measurement method was elaborated by Reynolds and Elrick (1985), and enhanced at University of Guelph, in Canada, back in 1985 (Elrick et al., 1989). The application of Darcy’s equation’s, and their evolutions (Darcy, 1856; Richards, 1931; Reynolds and Elrick, 1985) allowed finding the saturated water conductivity (Ksat) values, which were estimated based on bir by using formulas based on known parameters estimated in the laboratory.

The Ksat values can be transformed based on the measured bir data, according to the equipment’s parameters (water column, water load and water column diameter), soil features (porosity, genesis and storage coefficient), on Darcy’s equation and of their evolutions (Darcy, 1856; Richards, 1931; Reynolds and Elrick, 1985). In total, thirty-six (36) data values for both physical-hydric attributes were measured in the study site. The measured data were separated into four levels, one of them regarding the soil water condition (hydromorphic and non-hydromorphic) and the other the layer analyzed (surface, ranging 0.00-0.20 m in depth, and subsurface, ranging 0.20-0.40 m in depth).

Finally, data were subjected to pedotransfer functions calibration based on intrinsic soil features (particle size composition, soil density, water pH, water dispersed clay, porosity, particle density, organic carbon, and T-value). The remaining points (83 total) were estimated by pedotransfer functions applied in soils vertical modeling in the basin. This estimate resulted in 122 points with the surface (0.00-0.20 m) and subsurface (0.20-0.40 m) information about the analyzed properties (Figure 1). These points encompassed the analyzed data set in the current research. The previously described stages were conducted by the authors during analysis performed prior to the current study.

Terrain numerical modeling: covariables extraction and association with basin's landscape

SAGA GIS (Conrad et al., 2015), an Geographic Information System (GIS) focused on geo-scientific analysis, was used to plot topographic maps aimed at depicting surface morphometric parameters (primary and secondary covariables) deriving from hydrologically consisted digital elevation model of orthometric altitudes (altimetry) applicable to topographic surface aggregated to vector elements in the basin, such as vegetation cover. Vector data used to generate DEM resulted from Instituto Brasileiro de Geografia e Estatística (IBGE) Geosciences repository in partnership with Secretaria Estadual do Ambiente (SEA), constituting a continuous cartographic database of Rio de Janeiro State's planialtimetric survey, at 1:25,000 scale.

Basin's political-managerial region limit mask was acquire from geo-spatial data available at Instituto Estadual do Ambiente (GeoINEA) page for data clipping in the target areas, which were delimited based on compiled data from a series of topographic maps at 1:50,000 scale, provided by National Cartographic System (NCS). The basin's limit was also standardized based on datum and coordinate systems used to treat IBGE's vector data. The next step involved proceeding with the DEM extraction by using *TopoToRaster* interpolator in ESRI ArcGIS 10.6 environment (Redlands, 2011). This tool was designed to create digital elevation models with hydrological consistency.

Vector data were interpolated at 20 m spatial resolution to create regions DEM. Subsequently, a hydrologic consistency analysis based on Flow Accumulation, Flow Direction and Fill tools was applied over DEM, turning the model into a Hydrologically Consisted Digital Elevation Model (HCDEM). Accordingly, the assessed region's elevation map was obtained (Figure 1).

Terrain covariables were carefully selected to design the basin's landscape. This process considered pedological and landscape environmental elements (soil, vegetation, hydrology, and geology) observed in it. Radiometric matrix data and their features are shown in table 1.

The primary and secondary selected attributes (36 attributes) presented significant expressiveness in representing the landscape's environmental properties and active processes, and it has featured the soil formation factors, mainly relief. Therefore, the basin-representative covariables selection was a priority because it concerned the continuous surfaces extraction based on *Hydrology*, *Lighting*, *Visibility*, *Morphometry*, *Slope Stability* and *Channels* tools application, as well as on SAGA GIS *Terrain Analysis* modulus, which can quantify mountains eco-systemic variables (altitude fields) and region's native biome (Atlantic Forest biome).

Aerogeophysical data acquisition and processing of Sentinel-2A optical images

The Brazil's Geological Survey Geosciences system (GeoSGB), an online repository maintained by Companhia de Pesquisa de Recursos Minerais (CPRM), provided the magnetometry and gamma spectrometry aerogeophysical data used in the current study (CPRM, 2012). This system provides Aerogeophysical Survey Projects aerial bases. These projects are conducted by CPRM and other institutions, and the database holds airborne sensor data used to detect magnetometric, gamma spectrometric, gravimetric, and radiometric parameter values, among other data about Brazil national territory.

Raw aeromagnetometric and gamma spectrometric digital data were collected in GeoSGB aerogeophysical projects database, at 500 m distancing between flight lines. These data were processed in Geosoft Oasis Montaj software. Aeromagnetometric data were treated to correct common aerial survey and equipment issues (parallax error correction, diurnal variation removal, profile leveling and micro-levelling, and geomagnetic field trend surface definition called International Geomagnetic Reference Field IGRF). These procedures were performed by CPRM in its final processing report (CPRM, 2012).

Table 1. Terrain covariables concerning the topographic physical-hydric and thermodynamic DEM's primary and secondary attributes

Type	Terrain attributes	Physical process	Description
Pr	Digital Elevation Model (DEM) or Elevation (elev)	M	Landscape dynamics and its phenomena.
P	Slope, Aspect, Analytical Hillshade (AH), Profile Curvature (ProfC); Planar Curvature (PlanC)		Water flow dynamics; vegetation; geomorphology; sun radiation intensity; Failures in geological structures, soil particles and sediment deposition.
S	Tangential Curvature (TanC), Total Curvature (TotalC), Morphometric Features (MF), Topographic Position Index (TPI), Wind Exposition Index (WEI), Terrain Surface Convexity (TSC), Longitudinal Curvature (LonC), General Curvature (GC), Landforms (LF), Terrain Surface Texture (TST), Convergence Index (CI), Generalized Surface (GS), Morphometric Protection Index (MPI)		Rate of lateral accumulation in the landscape; water soil retention and storage; soil features; landscape morphometric parameters derivation approach.
S	Catchment Area (CA), Catchment Slope (CS), Multiresolution Index of Valley Bottom Flatness (MRVBF), Multiresolution Index of The Ridge Top Flatness (MRRTF), Terrain Classification Index for Lowlands (TCIL), Flow Accumulation (FA), Stream Power Index (SPI), Topographic Wetness Index (TWI), SAGA Wetness Index (SWI), Valley Depth (VD), Vertical Distance to Channel Network (VDCN), Channel Network Base Level (CNBL), Flow Path Length (FPL), Euclidian Distance (ED)	CH	Sediment deposits identification and featuring; saturation zones prediction; Hydrological studies on water flow and runoff; water concentration in connectivity channels.
S	LS Factor (LSF), Wetness Index (WI)	E	Erosion estimates for watersheds; sediments transport; rock permeability.
S	Geomorphons (Gm)	VI	Landscape classification based on landforms; region's geomorphology description.

Pr: Principal; P: Primary; S: Secondary; M: Morphometry; CH: Channels and hydrology; E: Stability; VI: Visibility and light. Adapted from Beven and Kirkby (1979), Zevenbergen and Thorne (1987), Moore et al. (1991), Köthe et al. (1996), Montgomery and Dietrich (1994), Bock et al. (1996, 2007), Guisan et al. (1999), Wilson and Gallant, (2000), Breuer (2001), Böhner et al. (2002), Florinsky et al. (2002), Yokoyama et al. (2002), Gallant and Dowling (2003), Romano and Chirico (2004), Böhner and Selige (2006), Iwahashi and Pike (2007), Seibert and McGlynn (2007), Stepinski and Jasiewicz (2011), Jasiewicz and Stepinski (2013), Gerlitz et al. (2015), Oliveira et al. (2017).

Corrected tabular data were spatialized in software for georeferencing and for the Anomalous Magnetic Field's (AMF) radiometric bi-directional grid creation using tridimensional coordinates (x, y, z) and magnetometric value measured by IGRF. Values were interpolated at maximum of 1/5 of flight lines to avoid post-processing magnetometric signature losses (*MagMap* module). The lacking grid values were also filled through dummy interpolation based on square method (Grid and Image modulus). The anomalous magnetic field (AMF) map was generated by the radiometric grid, at spatial resolution of

100 m, and plotted based on IGRF correction (Fourier fast change calculation to domain frequency). Subsequently, analytical signal amplitude (ASA) filter was applied to find the variation rate (gradient) on the three axes (x, y, z) and to reduce AMF's interference on magnetometric domain identification, since it could be associated with aquifer rocks and mineralogy.

Reduction to pole (RTP) was simulated based on the derivative of axis x, by plotting the ASA map, because Rio de Janeiro State is in a low-altitude region. This simulation allowed to highlight the geologic profile limits (most superficial mining). Finally, spectral signature variation rate was expressed in shorter wavelength to scour smaller bodies (mineralogy and smaller rocks) represented in ASA map. Slope (-35.64), declination (-21.69) and the corrected amplitude (-54.36) were IGRF's parameters necessary to reduce aerial survey data (January 8, 2012). The two first parameters were automatically calculated by Oasis Montaj, based on aerial survey's data input.

As CPRM's recommendations (CPRM, 2012), the minimum curvature interpolator was used for referred gamma data to create Uranium (U), Thorium (Th) and Potassium (K) grids. Also, the interpolation was based on one-fifth of flight lines to equate with aeromagnetometric data's spatial resolution (100 m). Grid lacked information found was corrected based on the square method (Grid and Image modulus) through dummy interpolation. Finally, a radiometric ternary map was plotted in red, green, and blue composition (RGB) depending on radioelements bands (R-potassium, G-thorium, B-uranium) represented by the ternary triangle, using Oasis Montaj Grid and Image modulus.

Multispectral images from MSI/Sentinel-2 mission were selected based on data proximity and availability in Copernicus Sentinel Hub repository to meet the hydropedological data collection period (September 2019 and February 2020). The images acquisition dates were August 2, 2019 (for the first field survey) and February 10, 2020 (for the second field survey). The conducted treatment and processing were based on transforming radiance information into surface reflectance and allowing spatial resolution matching between sensor's bands by transforming the 10 m bands resolution into 20 m ones to make pixels compatible for band algebra performance (index calculations). The spectral bands used in analysis were Sentinel-2A mission B2 to B8A, B11 and B12 bands to identify possible relational patterns with the physical-hydric attributes (ESA, 2020).

Spectral indices were calculated using visible-infrared (VIS-IR) bands. The spectral indices selection related to vegetation, soil and geology was carefully made, considering possible associations with the study variables, as shown in table 2.

All spectral data treatment and processing procedures were carried out through statistical routine implemented in RStudio environment (Rstudio Team, 2020), based on sen2R package used to process Sentinel-2A images. The indices were also calculated in this environment by using images from nine spectral bands (VIS-IR) from Sentinel 2A mission based on bands mathematics.

Modeling: selection methods and data dimensioning reduction, random forest, and validation

Initially, the input data underwent analysis using four statistical adjustment techniques: autocorrelation analysis, multi-collinearity analysis, principal components analysis, and hypothesis tests. These techniques were employed to apply methods, minimize errors, and avoid information loss.

Subsequently, statistical methods based on regression analysis, multivariate analysis, and machine learning were used to select and reduce the data dimensionality. Specifically, the Variation Inflation Factor (VIF), Stepwise Akaike Information Criterion (Stepwise AIC), and Recursive Feature Elimination (RFE) were applied. These analyses were implemented using routines in the RStudio environment.

Table 2. Vegetation indices deriving from multispectral sensor's spectral bands of the Sentinel-2 mission

Index	Indices formula based on MSI Sentinel-2 bands	Reference
Enhanced Vegetation Index (EVI)	$2.5x\left[\frac{(\rho_{band\,8A}-\rho_{band\,4})}{(\rho_{band\,8A}+6x\,\rho_{band\,4}-7.5x\,\rho_{band\,2}+1)}\right]$	Huete et al. (2002); Hunt et al. (2011)
Clay Minerals Ratio (CM); Ferrous Minerals Ratio (FM); Ferruginous Regolith Ratio (FR); Iron Oxide Ratio (IO)	$\frac{\rho_{band\,11}}{\rho_{band\,12}};\frac{\rho_{band\,11}}{\rho_{band\,8A}};\frac{\rho_{band\,8A}}{\rho_{band\,3}};\frac{\rho_{band\,4}}{\rho_{band\,2}}$	Segal (1982); Drury (1987); Rowan and Mars (2003)
Grain Size Index (GSI)	$\frac{(\rho_{band\,4}-\rho_{band\,2})}{(\rho_{band\,4}+\rho_{band\,2}+\rho_{band\,3})}$	Perera et al. (2005); Xiao et al. (2006)
Normalized Difference Red-Edge Index (NDRE)	$\frac{(\rho_{band\,8A}-\rho_{band\,5})}{(\rho_{band\,8A}+\rho_{band\,5})}$	Clevers and Gitelson (2013)
Normalized Difference Vegetation Index (NDVI)	$\frac{(\rho_{band\,8A}-\rho_{band\,4})}{(\rho_{band\,8A}+\rho_{band\,4})}$	Rouse et al. (1973)
Normalized Difference Water Index (NDWI)	$\frac{(\rho_{band\,3}-\rho_{band\,8A})}{(\rho_{band\,3}+\rho_{band\,8A})}$	Brenner and Guasselli (2015)
Non-Linear Index (NLI)	$\frac{(\rho_{band\,8A}{}^{2}-\rho_{band\,4})}{(\rho_{band\,8A}{}^{2}+\rho_{band\,4})}$	Goel and Qin (1994)
Soil Adjusted Vegetation Index (SAVI)	$1.5x\left[\frac{(\rho_{band\,8A}-\rho_{band\,4})}{(\rho_{band\,8A}+\rho_{band\,4}+0.5)}\right]$	Huete (1988)
Transformed Difference Vegetation Index (TDVI)	$1.5x\left[\frac{(\rho_{band\,8A}-\rho_{band\,4})}{\sqrt{(\rho_{band\,8A}{}^{2}+\rho_{band\,4}+0.5)}}\right]$	Bannari et al. (2002)
Visible Atmospherically Resistant Index (VARI)	$\frac{(\rho_{band\,3}-\rho_{band\,4})}{(\rho_{band\,3}+\rho_{band\,4}-\rho_{band\,2})}$	Gitelson et al. (2001); Hunt et al. (2011)

Adapted from Sentinel-Hub Repository Satellite Indices (Sinergise, 2020).

The VIF method was used to address multicollinearity issues by quantifying the increase in variance of an estimated regression coefficient due to collinearity. This method consists of the quotient between a model's variance with several terms and the model's variance with a single term. It quantifies this variance severity based on ordinary least squares and provides an index to measure to which extent the variance (the square of the estimate's standard deviation) of an estimated regression coefficient increases due to collinearity (Daniel and Wood, 1999).

Stepwise AIC is an automatic and iterative method used to select variables based on regression analysis. It evaluates each variable's contribution to the model and determines whether it should be added or subtracted based on a pre-specified criterion (AIC). Based on a pre-specified criterion, AIC can interactively add or remove a variable from a set of explanatory variables. It is done by estimating the relative amount of information lost by a given model, in case an important variable is taken out of the analysis (Taddy, 2019; McElreath, 2020).The RFE is a selection method used in machine learning models and Data Mining universe. It aims to select predictor variables that best fit the desired model, whether regression, multivariate, or machine learning. The RFE optimizes the model by selecting the most relevant predictor variables (Blum and Langley, 1997; Bradley et al., 1998). Accordingly, RFE results in predictor variables selection to better optimize the model one wishes to set (Svetnik et al., 2004).

The predictors selection based on importance classifications follows a Backward Selection approach, often used in Random Forest models, aiming to integrate physical-water attributes in a robust and refined way. This approach leverages multicollinearity principles to limit the number of predictors and select variables based on their importance in decision trees, and, therefore, it also helps select the variables to be used in the final model by reducing predictors in target scores.

To reduce data dimensionality, Principal Component Analysis (PCA) was applied. Abdi and Williams (2010) defined PCA as a multivariate technique to describe data observed through several inter-correlated quantitative dependent variables that aim to extract important information. The technique represents the data as orthogonal variables called principal components, enabling the patterns and similarities identification among observations and variables as multi-dimensional vectors by grouping input variables responses without losses.

The Random Forest algorithm was chosen to classify water resources given its easy implementation and robustness (Blum and Langley, 1997; Bradley et al., 1998). Predictive property values obtained from the Random Forest model (quality parameters) were used to determine the method with the best performance in estimating physical-hydric data.

The input data were previously separated into training (70 %) for the Random Forest models implementation, and testing (30 %) for quality validation purpose. Quality criteria set for models encompassing the selected and reduced variables (from the database and based on the implemented methods) resulted from careful statistical analysis, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), determination coefficient (R^2), and Random Forest model importance evaluation. This statistical approach was used to assess the model's accuracy and precision under bir/Ksat values estimated (Schaap and Leij, 1998; Schaap, 2004). In addition, the model with the best prediction performance and highest correctness degree was selected based on the validation results, because validation implementation used 30 % of the non-trained sampling data.

The selection methods and data dimensionality reduction allowed predicting bir and Ksat values for the Macacu basin entire area grid. The surfaces mapped based on both attributes resulted from data separation criterion into two soil layers: 0.00-0.20 and 0.20-0.40 m. The final resulting cartographic products included four maps estimating bir and Ksat attributes at two soil layers (surface 0.00-0.20 m and subsurface 0.20-0.40 m) in the basin, based on the Random Forest classification.

The analyses were conducted using applied statistical routines in R language and RStudio software environment (R Core Team, 2020; Rstudio Team, 2020). Packages such as *Raster*, *sp*, *sf*, *shapefiles*, *openxlsx*, *Sen2R* (Sentinel-2A images treatment and process), *caret*, *corrplot*, *usdm*, *FactoMineR*, *randomForest*, and *ggplot2* (graphic plotting) were utilized for data processing (manipulate geo-spatial data in tabular, matrix and vector formats), selection methods development, image processing, classification, and visualization. The cartographic products were produced using open-source software Quantum GIS (QGIS) (QGIS Development Team, 2020). The described methodological steps summary is presented in the figure 2 flowchart.

RESULTS AND DISCUSSION

Autocorrelation and principal component analysis

Autocorrelation method performs an initial examination of data behavior through meticulous analysis conducted by the researcher, considering prior knowledge about the database and utilizing the Pearson's correlation matrix. Inman (1994) suggested that Pearson's determination coefficient values (R) above 0.6 indicate a moderate to strong correlation (either positive or negative). Values lower than 0.6 indicate weak correlation (either positive or negative), while null values (zero) indicate correlation absence. Figure 3 shows Pearson's correlation matrix for the evaluated data.

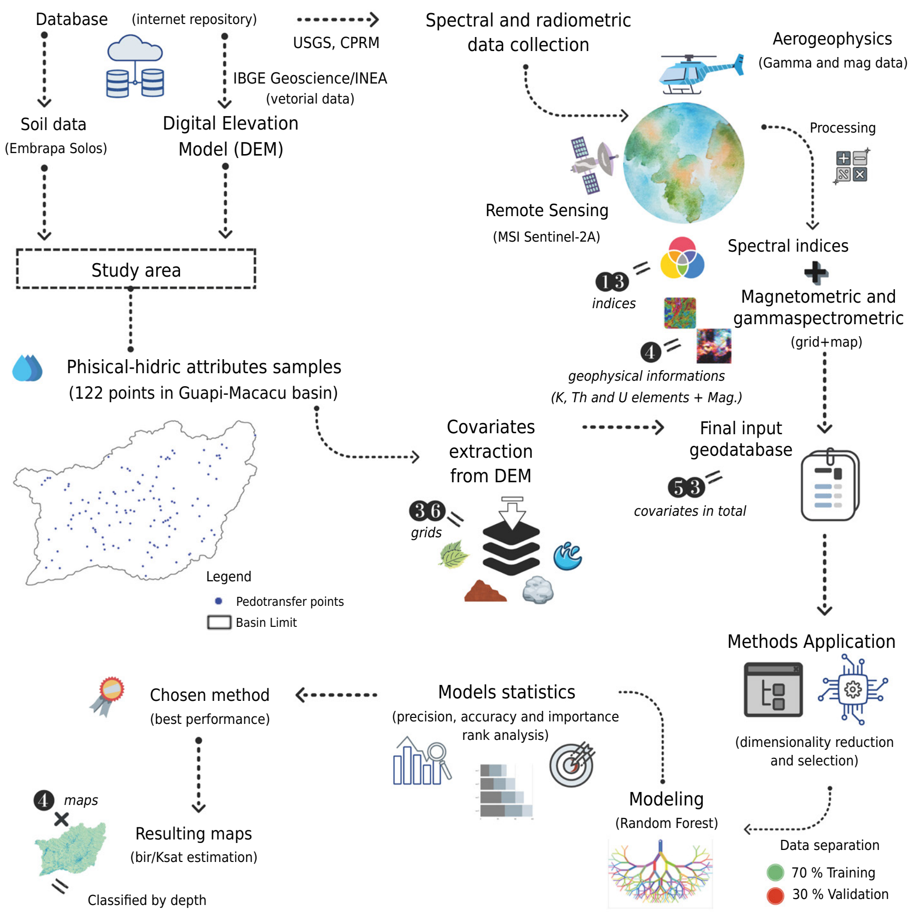

Figure 2. Proposed methodology flowchart.

Based on the correlation matrix, spectral indices demonstrated a moderate correlation with Ksat and bir attributes. This correlation can be either positive or negative. The following spectral indices show a positive correlation: Clay Minerals, EVI, Ferruginous Regolith, NDRE, NDVI, NLI, SAVI, VARI, and TDVI. On the other hand, Ferrous Minerals, GSI, and Iron Oxide exhibit a negative correlation. Regarding the terrain covariables, DEM morphometry and primary variables (Slope, Aspect, and Elevation) exhibited a weak to moderate positive correlation ($0.4< R <0.6$). Similarly, hydrology channel covariables (such as MRRTF and MRVBF) presented a weak to moderate, albeit inverse, correlation. In other words, as physical-water attribute values increase, these variables values decrease. This analysis aligns with the local relief features and suggests effective water drainage, primarily driven by surface and/or vertical runoff throughout the basin. This finding should be thoroughly investigated in the prediction map. Visibility, light, and hydrology covariables showed a weak to moderate correlation ($0.2< R <0.4$) when compared to the overall dataset.

Spectral data within the matrix (Figure 3) exhibit interdependencies that may indicate the multicollinearity or high dimensionality presence, wherein multiple data points represent identical or highly similar features. Principal Component Analysis (PCA) was employed to assess the spectral data contribution, addressing this issue. The PCA analysis accounted variance of over 90 % in both principal components, enabling a comprehensive understanding of the model through a two-dimensional spectral variables analysis (Figure 4a).

Upon principal components spatial observation, it becomes apparent that the variables (Figure 4b) are separated into three distinct groups based on their spectral characteristics. Notably, the IO and GSI indices contributed proportionally less to group 1 in physical-hydric variables explanation. Figure 4b illustrates the variables separation into group 1 (10 variables), group 2 (3 variables), and group 3 (10 variables). It is worth noting that two variables from group 1, namely the IO and GSI indices, warrant closer examination during the modeling stage. These results underscore the spectral data significance derived from remote sensing in modeling, as they account for most of its variability. It is advisable to reduce the dimensionality of the data associated with these variables (Figure 4) to mitigate issues arising from multicollinearity.

Further discussions about the variables analyzed in this step can be found in dos Santos et al. (2019) and Santos et al. (2020) studies. The Guapi-Macacu basin soil attributes characterization is wider discussed by Santos et al. (2022) and can be integrated with the analysis carried out by the cited authors.

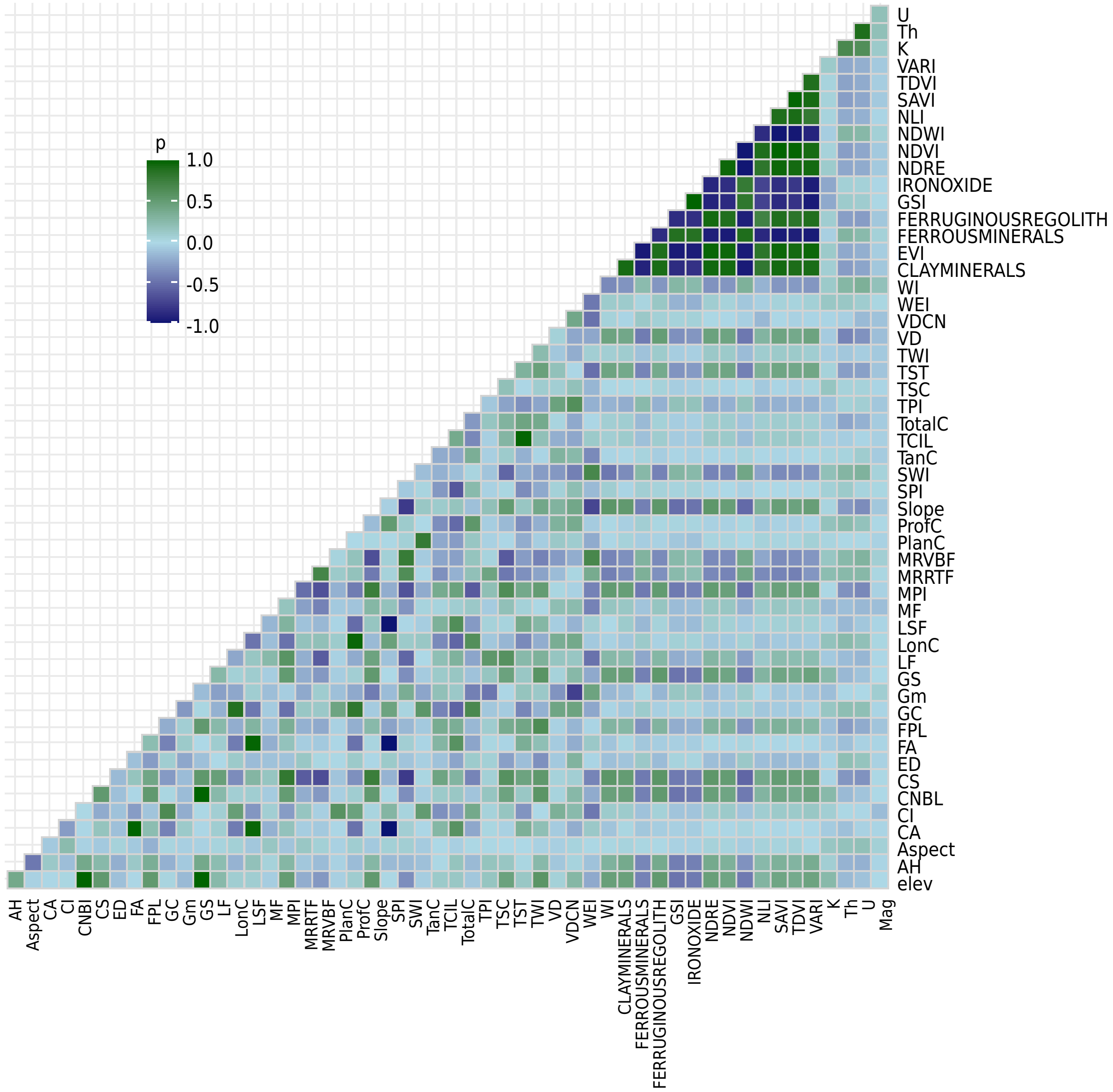

Figure 3. Pearson's correlation matrix about the set of assessed covariables (radiometry, geophysics, and spectral indices), whose ρ refers to Karl Pearson's correlation coefficient.

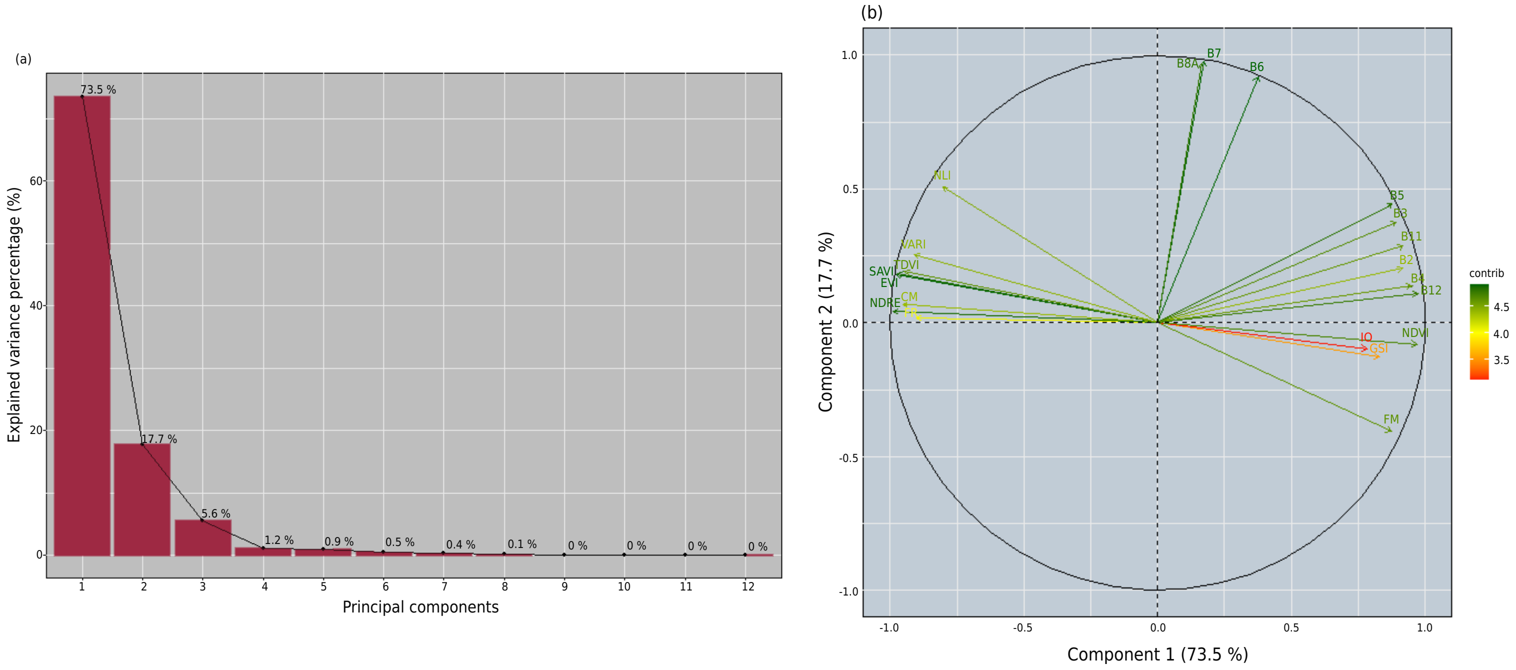

Figure 4. Spectral variables are represented based on the PCA model's explanatory principal components, wherein (a) is the principal components' importance in PCA; (b) is the principal components' separation in PCA, with each contribution expressed in the right legend.

Correlation and dimensionality analyses play a crucial role in multivariable studies. However, they serve as preliminary steps before a specific method application, to explore and identify patterns in the study site's data. But they do not account for factors such as homoscedasticity (trends absence in error variances), multicollinearity, and normality. Neglecting these considerations can have a detrimental impact on drawing direct conclusions about the study outcomes. Addressing this, statistical requirements demanded by the chosen methods can be met by applying criteria such as the *Breusch-Pagan* test (Breusch, 1978), *Shapiro-Wilk* test (Shapiro and Wilk, 1965), and *Durbin-Watson* test (Durbin and Watson, 1950; Watson and Durbin, 1951; King, 1992); to the databases. These tests help adjust accordingly the data, ensuring compliance with the necessary statistical assumptions.

Multicollinearity, as indicated by Mansfield and Helms (1982), can introduce various undesirable effects on coefficients estimated through multiple regression analysis. Hence, it is essential for researchers to possess the skills to identify its presence. In this study, three distinct methods (VIF, RFE and Stepwise AIC) were examined. These methods aimed to select the appropriate variables for each assessed attribute, considering their respective depths (surface or subsurface). Furthermore, they aimed to evaluate the quality of the estimates obtained for these attributes using the Random Forest machine learning classifier. By employing these methods, researchers can make informed decisions regarding variable selection and assess the accuracy of attribute estimation achieved through the modeling process.

Hydropedological data spatial modeling

Based on database separation categorized according to soil layer (0.00-0.20 and 0.20-0.40 m) and the variables under investigation (bir and Ksat), response variables were selected by calibrating three different spectral data dimensionality reduction methods and analysis. Each applied method (Table 3) yielded different output variables definitions (explanatory) that displayed significant positive or negative correlations (at a significance level of α = 5 %) with the physical-hydric variables.

Based on the resulting findings (Table 3), most methods employed resulted in more than ten explanatory variables selection. Furthermore, these methods demonstrated a multicollinearity lack and strong interference in data correlation. Focusing on bir and Ksat estimates within the depth range of 0.00 to 0.20 m, it can be observed that the RFE selection method yielded the lowest number of response variables composing the model, with 11 and 14 variables estimated total, respectively. However, it is important to note that the methods employed different approaches in selecting covariates based on the four adopted criteria, thereby incorporating either the topographic or radiometric aspects of the study site. Figures 5a to 5d provide an analysis based on the RFE method to determine the optimal variables number that would result in the lowest residue in the Random Forest model (as indicated by the lowest RMSE).

Dashed lines in graphics depicted in figures 5a, 5b, 5c and 5d represent the optimal variables number required to achieve the lowest Root Mean Squared Error (RMSE) values when utilizing the Random Forest model. This criterion is specifically applied in analysis based on RFE method. Among the four attributes modeled, only the model based on $\text{vib}_{0.00\text{-}0.20\text{ m}}$ attribute demonstrated lowest RMSE value with the smallest variables number (11) after 99 iterations (Figure 5a). The remaining three attributes - $\text{Ksat}_{0.00\text{-}0.20\text{ m}}$, $\text{Ksat}_{0.20\text{-}0.40\text{ m}}$, and $\text{vib}_{0.20\text{-}0.40\text{ m}}$ (Figures 5b to 5d) were required 27, 14, and 37 variables, respectively,

Table 3. Covariables selected after applying the proposed methods

Method	Selected covariables [(1)]	Total
$\textit{Ksat}_{0.00\text{-}0.20\ m}$		
VIF	AH; Aspect; CI; CS; ED; FA; FPL; Gm; GS; LF; LSF; MF; MRRTF; PlanC; TanC; TotalC; TPI; TSC; TST; TWI; VD; VDCN; WEI; WI; Ferruginous Regolith; Iron Oxide; NLI; K; U; Mag	30
RFE	CS; GS; CNBL; NDWI; elev; LSF; Ferruginous Regolith; SWI; Clay Minerals; TSC; SAVI	11
Stepwise AIC	elev; AH; Aspect; CA; CNBL; FPL; Gm; GS; LF; LonC; MF; MPI; MRVBF; ProfC; SWI; TanC; TCIL; TPI; TST; TWI; VD; WI; EVI; Ferrous Minerals; GSI; Iron Oxide; NDRE; NDVI; NDWI; SAVI; TDVI; VARI; K; Th; U; Mag	36
$\textit{Ksat}_{0.20\text{-}0.40\ m}$		
VIF	MRRTF; MRVBF; PlanC; ProfC; Slope; TanC; TCIL; TotalC; TSC; VD; VDCN; WEI; Iron Oxide; NLI; K; U; Mag	17
RFE	Mag; Ferrous Minerals; VARI; MRRTF; K; LSF; MPI; elev; SWI; GSI; AH; Iron Oxide; CNBL; GS; VD; EVI; NDVI; Slope; Clay Minerals; MF; GC; NLI; TDVI; SAVI; Aspect; WEI; NDRE	27
Stepwise AIC	elev; CA; FA; FPL; GC; GS; LonC; LSF; MF; MPI; ProfC; Slope; SPI; TCIL; TotalC; TPI; TSC; TST; TWI; VD; Ferrous Minerals; Ferruginous Regolith; GSI; Iron Oxide; NDWI; NLI; SAVI; TDVI; VARI; Th; U	31
$\textit{Bir}_{0.00\text{-}0.20\ m}$		
VIF	AH; Aspect; CI; CS; ED; FA; FPL; Gm; GS; LF; LSF; MF; MRRTF; PlanC; TanC; TotalC; TPI; TSC; TST; TWI; VD; VDCN; WEI; WI; Ferruginous Regolith; Iron Oxide; NLI; K; U; Mag	30
RFE	CNBL; GS; elev; VD; NDWI; CS; NLI; Clay Minerals; TDVI; TanC; NDRE; Ferruginous Regolith; SAVI; NDVI	14
Stepwise AIC	elev; CS; FA; FPL; GC; Gm; GS; LF; LSF; MF; MPI; MRRTF; MRVBF; ProfC; SWI; TanC; TotalC; TPI; TST; TWI; VD; WEI; WI; EVI; Ferruginous Regolith; GSI; Iron Oxide; NDRE; NDVI; NDWI; NLI; SAVI; VARI; K; Th; U	36
$\textit{Bir}_{0.20\text{-}0.40\ m}$		
VIF	MRRTF; MRVBF; PlanC; ProfC; Slope; TanC; TCIL; TotalC; TSC; VD; VDCN; WEI; Iron Oxide; NLI; K; U; Mag	17
RFE	LSF; Mag; CS; VD; GS; CNBL; elev; MRRTF; TotalC; SWI; TCIL; U; TWI; GSI; Slope; WEI; MRVBF; AH; Clay Minerals; NLI; VARI; Th; NDVI; K; NDRE; TDVI; EVI; FPL; SAVI; SPI; TPI; LF; Iron Oxide; MPI; Aspect; PlanC; Ferrous Minerals	17
Stepwise AIC	elev; AH; CA; CI; FA; FPL; GS; LonC; LSF; MRRTF; MRVBF; PlanC; ProfC; Slope; TanC; TCIL; TotalC; TPI; TST; TWI; GSI; Iron Oxide; NDVI; NDWI; TDVI; VARI	26

VIF: Variance Inflation Factor; RFE: Recursive Feature Elimination; AIC: Akaike Information Criterion. [(1)] Acronyms whose meanings are shown in table 1.

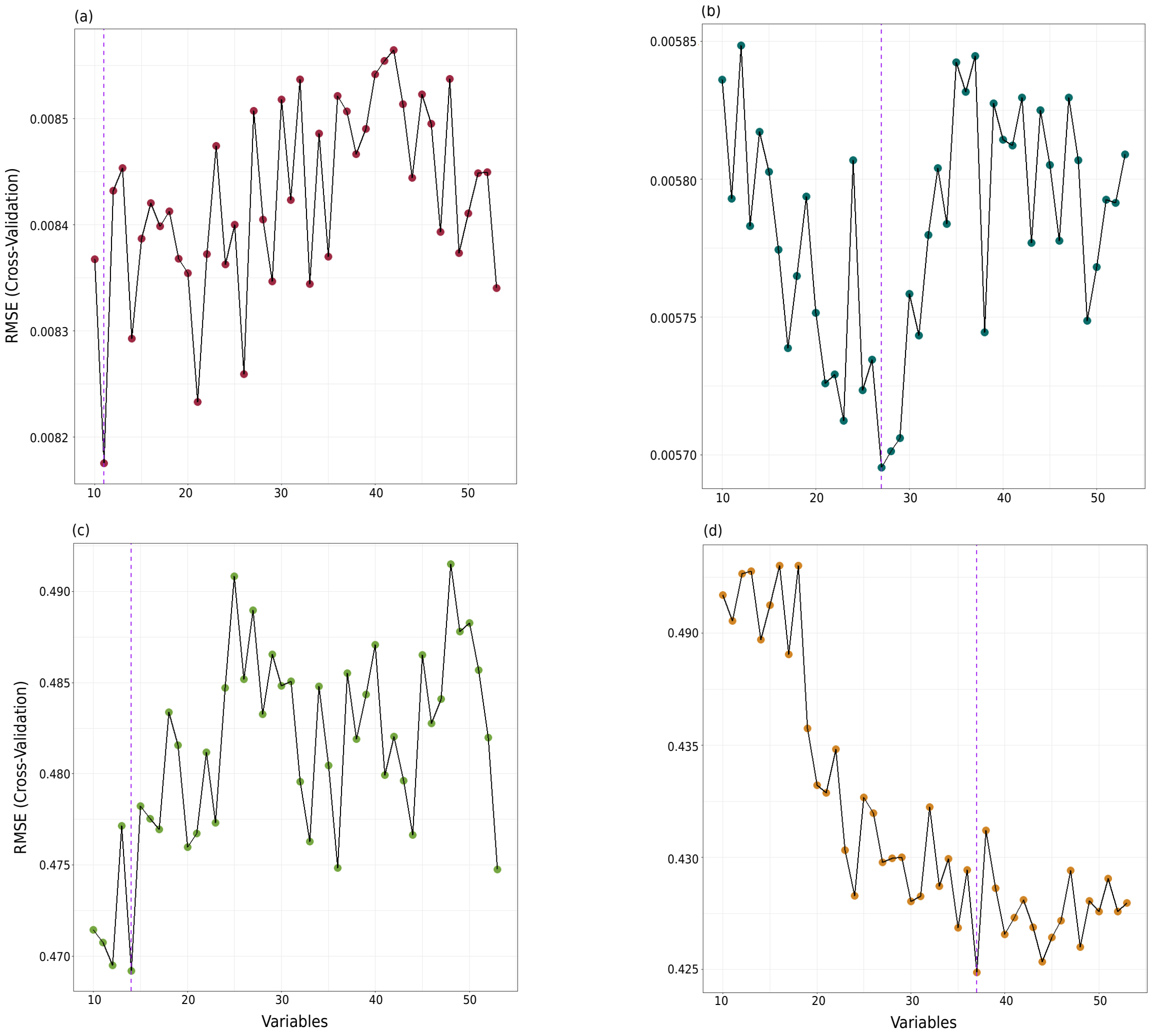

Figure 5. The RFE method graphics (number of explanatory variables versus RMSE adjustment) of the assessed physical-hydric attributes and their respective layers: (a) $Ksat_{0.00\text{-}0.20\ m}$; (b) $Ksat_{0.20\text{-}0.40\ m}$; (c) $bir_{0.00\text{-}0.20\ m}$; (d) $bir_{0.20\text{-}0.40\ m}$. The dashed line highlights the reach of the optimum number of variables selected through RFE.

to achieve the ideal adjustment in RF model. Notably, the RFE model applied to $vib_{0.20\text{-}0.40\ m}$ demanded the largest variables number in comparison to its explanatory set when compared to the other two implemented methods. Specifically, it required one more explanatory variable than the Stepwise AIC method evaluated for $Ksat_{0.00\text{-}0.20\ m}$ and $vib_{0.00\text{-}0.20\ m}$.

Breusch-Pagan, Durbin-Watson and Shapiro-Wilk tests were applied to the database. These tests yielded significant values necessary for the null hypothesis (H_0) acceptance in favor of the alternative hypothesis (H_1) at a 5 % significance level (α = 5 %, thus, β = 95 % confidence level). Shapiro-Wilk test was utilized to assess data normality residuals (H_0: normality of residuals versus H_1: non-normality of residuals). Breusch-Pagan test was employed to assess variances in residues homoscedasticity (H_0: equal variances - homoscedasticity versus H_1: different variances - heteroscedasticity). Lastly, Durbin-Watson test was applied to examine the correlation presence among residuals, which serves as a multicollinearity indicator (H_0: autocorrelation among residuals equals zero - residues independence versus H_1: autocorrelation among residuals different from zero - residues dependence). The test stage is crucial in ensuring that statistical assumptions will be fulfilled, especially considering that VIF and Stepwise models are based on regression analysis principles. The *p*-values obtained from tests were within the required interval in all methods (*p*-value < α = 5 %), indicating that no data corrections or treatments were necessary. The trained methods quality performance using 70 % of the database, which was previously separated through RF model, is shown in table 4.

Table 4. Random Forest model quality in physical-hydric data training stage based on the applied selection methods (70 % of the database)

Assessed variable	Applied method	Quality metrics evaluated for Random Forest		
		$RMSE_{training}$	$MAE_{training}$	$R^2_{training}$
Ksat Z= 0.00-0.20 m	VIF	0.0054	0.0034	0.8755
	RFE	0.0049	0.0032	0.8731
	Stepwise AIC	0.0050	0.0035	0.8948
Ksat Z= 0.20-0.40 m	VIF	0.0025	0.0019	0.8679
	RFE	0.0025	0.0019	0.8657
	Stepwise AIC	0.0025	0.0020	0.8663
bir Z = 0.00-0.20 m	VIF	0.2743	0.1834	0.9336
	RFE	0.3038	0.1980	0.8130
	Stepwise AIC	0.2733	0.1879	0.9004
bir Z= 0.20-0.40 m	VIF	0.2713	0.1867	0.7276
	RFE	0.2710	0.1871	0.7306
	Stepwise AIC	0.2767	0.1962	0.7220

Values based on Random Forest model crossed evaluation applied to data separated at training.

All models reached an accuracy higher than 70 % (R^2 >0.70) in predicting the target attributes (Ksat and bir) at 0.00-0.20 and 0.20-0.40 m soil layers (Table 4). These predictions were assessed using cross-validation criteria, indicating well-calibrated models within the RF method. The Random Forest model adjustment, across all three proposed methods, resulted in a RMSE error stability of approximately 300-350 random trees generated for the assessed physical-hydric variables (Figures 6 and 7). Among these RF models, the ones that exhibited the best accuracy in variables prediction, based on RFE method, are highlighted in colors in figures 6 and 7, with RMSE stability starting at 300 trees.

The Random Forest model's performance evaluation (response) during validation, where the trained models were compared to test values (30 % of the separated database), revealed significant discrepancies between predicted values (by model estimation) and the actual values (measured *in situ*). This discrepancy is further supported by low R-squared (R^2_{test}) values obtained during the testing phase, which were below 20 % for the implemented methods based on criteria analysis (depths and independent variables), except for the bir variable assessed for the 0.00-0.20 m layer which recorded higher R^2 values. The R^2 values for bir variable assessed for the 0.00-0.20 m layer ranged from 28 to 48 % (Table 5). These results highlight challenges in accurately predicting the target variables using the implemented methods.

The Random Forest model's quality assessed the internal validation (Table 5) and revealed overall lower values for the applied evaluation metrics (RMSE, MAE and R^2). These results indicate that the adjusted models have less than 20 % explanatory variability for Ksat and bir attributes in the assessed regions. This finding can be attributed to the lack of data values or their low significance, specifically in the elevation values within the basin's lowland. This low significance is likely due to this region's planar curvature (Figure 8) and its proximity to Guanabara Bay, which experiences constant sedimentation processes as water from the basin flows out through the river mouth. The observed result may also be associated with training data overfitting or random selection of validation values. To solve this problem, increasing the estimated variables sample size while considering criteria such as soil classification and measurement points homogeneity in the study site can be a potential solution.

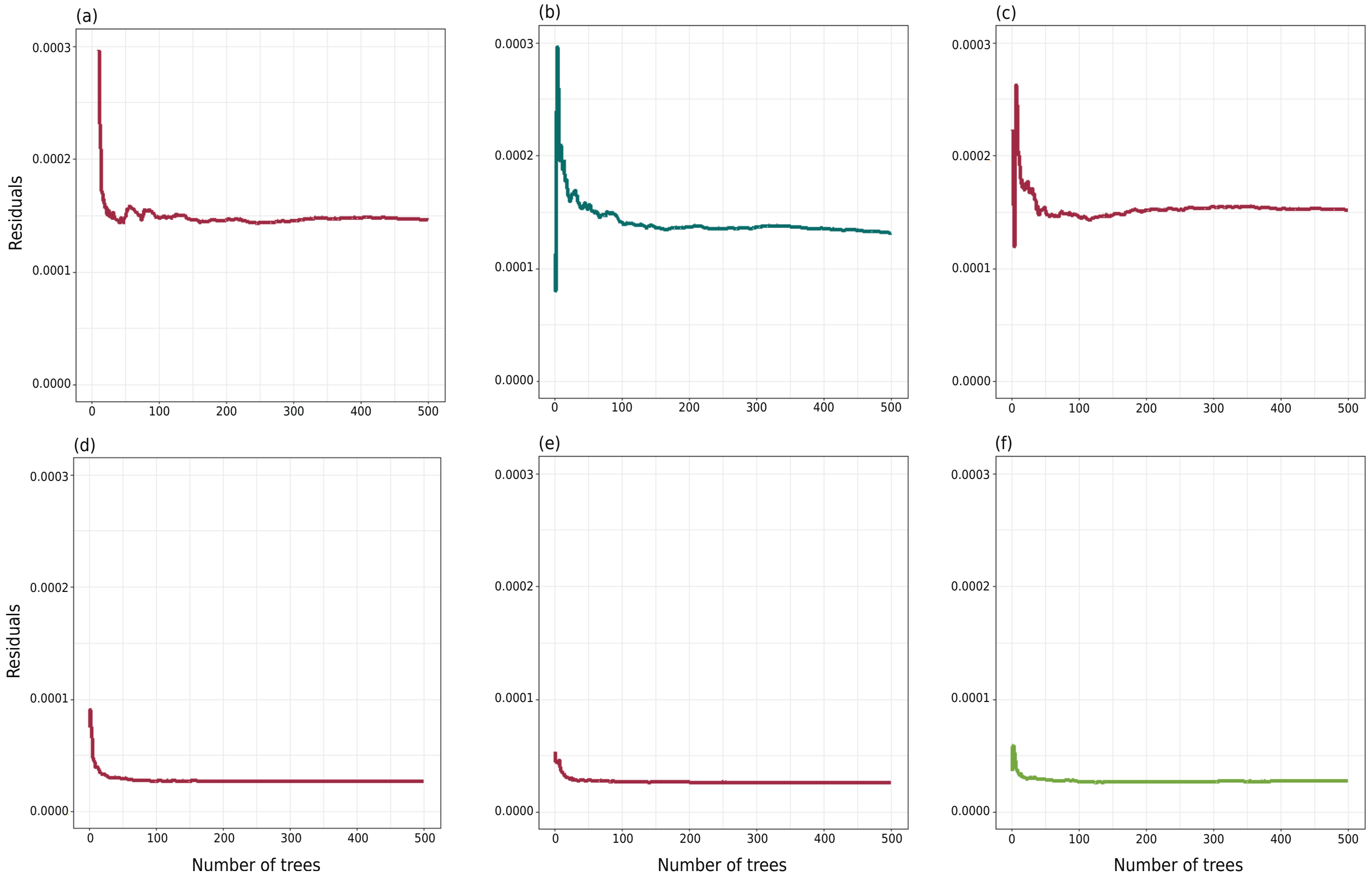

Figure 6. Random Forest model adjustment graphics evidencing the amount of trees versus error minimization recorded for Ksat variable, based on the assessed layer: (a-c) Ksat$_{0.00\text{-}0.20\ m}$; (d-f) Ksat$_{0.20\text{-}0.40\ m}$; and the three analyzed methods: (a,d) VIF; (b,e) RFE; (c,f) Stepwise AIC.

Despite careful survey planning, it was not feasible to collect data in regions characterized by high human interference levels and dense forest cover. These landscape features are present in Cachoeiras de Macacu, Guapi-Açu, Guapimirim, and Itaboraí counties, which are influenced by the Guapi-Macacu basin. These factors can significantly impact the Ksat and bir values estimation for the 122 sampled points using pedotransfer models, as indicated in previous research (Santos et al., 2022). As a result, there may be inherent randomness in errors and discrepancies arising from external mapping sources. However, it is worth mentioning that the 122 sample points estimated achieved high accuracy and precision in the mapping process.

Considering the region's sedimentary composition, it is expected that the physical-hydric attributes exhibit varying values in areas with higher clay content, particularly dispersed clay. These areas are characterized by high activity, resulting from a balance between micro and macropores that favor capillarity and water retention. Additionally, the organic matter presence in the soils promotes the colloid formation, increasing the water particles specific contact surface and further enhancing water retention.

Conversely, regions with high sand content and thick, poorly structured materials exhibit lower values for these physical-hydric attributes. This is explained by the macropores prevalence, which reduces the soil's capacity to transfer water through capillarity and retain it long enough for aquifer recharge.

Moreover, the region with mineral soils in the basin (feldspar and associated quartz fractions), as indicated by the 1:50,000 scale geological map (DRM, 2019), exhibits higher soil density, leading to reduced infiltration processes and water capillarity.

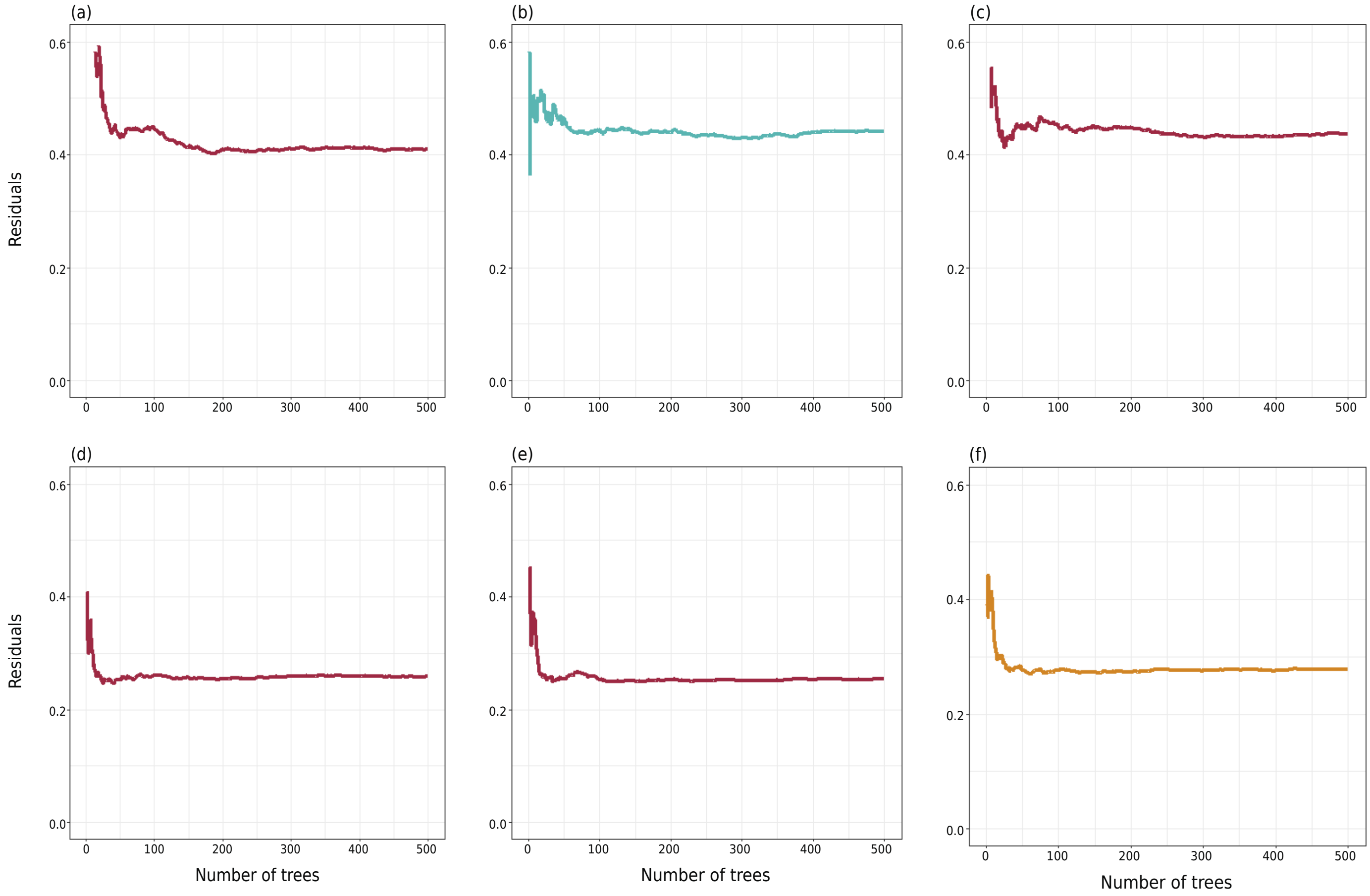

Figure 7. Random Forest model adjustment graphics evidencing the amount of trees versus error minimization recorded for Ksat variable, based on the assessed layer: (a-c) bir$_{0.00\text{-}0.20\ m}$; (d-f) bir$_{0.20\text{-}0.40\ m}$; and the three analyzed methods: (a,d) VIF; (b,e) RFE; (c,f) Stepwise AIC.

Table 5. Random Forest model quality in physical-hydric data validation test stage (30 % of the database)

Assessed variable	Applied method	Quality metrics evaluated for Random Forest		
		RMSE$_{test}$	MAE$_{test}$	R$^2_{test}$
Ksat Z= 0.00-0.20 m	VIF	0.0083	0.0071	0.0525
	RFE	0.0065	0.0056	0.1433
	Stepwise AIC	0.0084	0.0076	0.0461
Ksat Z= 0.20-0.40 m	VIF	0.0077	0.0045	0.0199
	RFE	0.0074	0.0044	0.0375
	Stepwise AIC	0.0069	0.0042	0.1088
bir Z = 0.00-0.20 m	VIF	0.4586	0.3594	0.4057
	RFE	0.4171	0.3394	0.4735
	Stepwise AIC	0.4799	0.3910	0.2801
bir Z= 0.20-0.40 m	VIF	0.3704	0.2771	0.1439
	RFE	0.3617	0.2748	0.1629
	Stepwise AIC	0.3531	0.2747	0.2003

Values based on Random Forest model validation of data separated at the test stage.

Values highlighted in the map follow bir and Ksat variables interval scale (Figure 8). Both variables exhibited lower values in the transition zone between profiles (surface to subsurface). This observation suggests a reduced capacity for infiltration and capillary flow, indicating lower water retention in the deeper basin soil profiles.

Conversely, the mountainous region surrounding the basin, encompassing Petrópolis, Teresópolis, and Nova Friburgo counties, displays the highest bir and Ksat values in subsurface profiles. This can be attributed to the dense vegetation, specifically the Atlantic Forest biome. These finding evidences these areas as priority sites with significant potential for water resource conservation within the basin. It is important to recognize that soil degradation and deforestation in these regions can contribute to water scarcity issues. Therefore, the preservation and restoration of these sites are crucial for maintaining water availability in the basin.

Considering the geological characteristics of the herein-assessed region, the Center-Northwestern fraction of the basin exhibits notable features in the aeromagnetic map, indicating a higher concentration of water due to the presence of deeper rocks. This specific area encompasses the Rio de Janeiro Petrochemical Complex (COMPERJ), which represents the region with the highest urban density and includes the flooded lowland areas of the basin. Consequently, these regions display distinct spectral responses. The airborne gamma spectrometry map highlights elevated levels of Potassium in the Center-Northwestern and Center-Southeastern sections and a notable Potassium and thorium combination (high concentration values) on the surface, particularly in other regions of the basin.

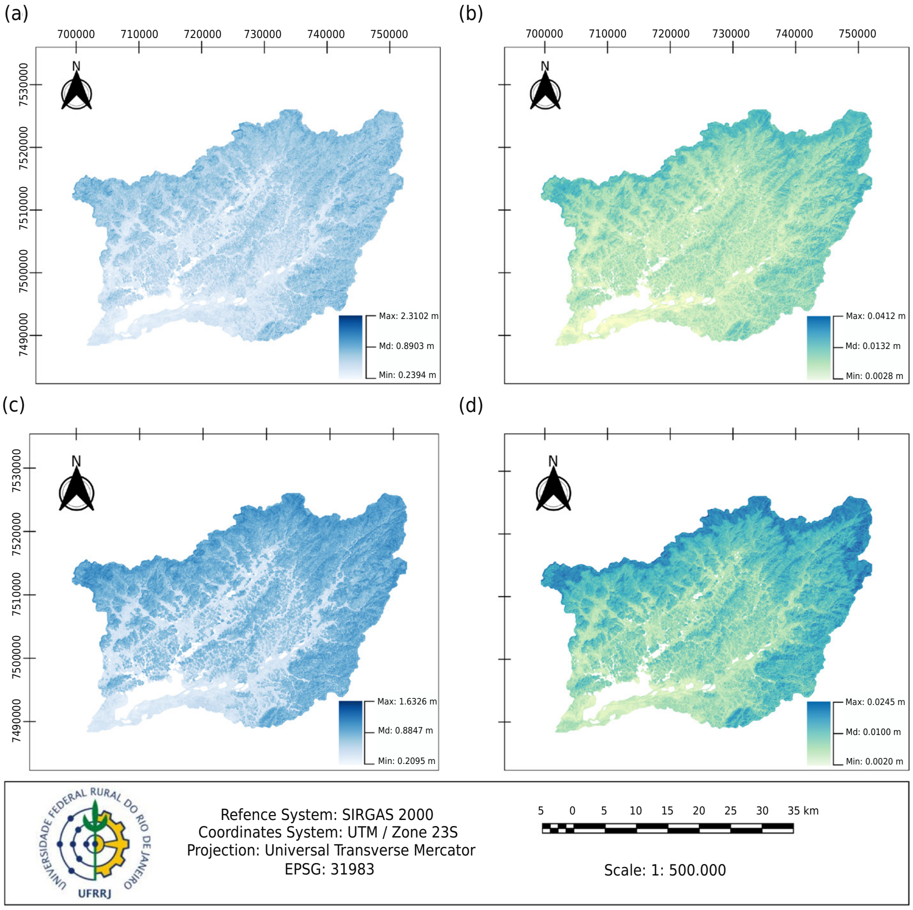

Figure 8. Bir and Ksat variability maps in Guapi-Macacu river basin, Rio de Janeiro, estimated through the Random Forest model and based on the respective depths and selection methods: (a) $\text{bir}_{0.00\text{-}0.20\text{ m}}$ – RFE; (b) $\text{Ksat}_{0.00\text{-}0.20\text{ m}}$ – RFE; (c) $\text{bir}_{0.20\text{-}0.40\text{ m}}$ – Stepwise AIC; (d) $\text{Ksat}_{0.20\text{-}0.40\text{ m}}$ – Stepwise AIC. Validation scale of 1:500,000.

The flight conducted by CPRM revealed a dummy correlation effect, explicitly observed in the COMPERJ restricted area. This effect impacted on data interpolation and showed noise in the middle portion of the basin, characterized by a faint linear feature (tenuous lineament) extending towards the East-Western basin part. Nevertheless, this procedure was crucial to feasible the modeling process, as it required a complete matrix structure for value prediction using the Random Forest algorithm.

CONCLUSIONS

The three evaluated methods demonstrated distinct selection responses, resulting in a reduction of 53 initial input variables during the pre-modeling stage. Specifically, the number of variables was reduced to 30, 11 and 36 for Ksat at Z = 0.00-0.20 m; 17, 27 and 31 for Ksat at Z = 0.20-0.40 m; 30,14 and 36 for bir at Z = 0.00-0.20 m; 17, 37 and 26 for bir at Z = 0.20-0.40 m, respectively. The Iron Oxide index was the most frequently selected variable by the applied methods among all analyzed data set variables, except for RFE applied to bir and Ksat attributes in soil surface layers (0.00-0.20 m), where this variable was not returned.

The VIF and RFE methods yielded the smallest number of explanatory variables for the 0.00-0.20 m ($RFE_{Ksat0.00\text{-}0.20\ m}$ = 11 variables; $RFE_{bir0.00\text{-}0.20\ m}$ = 14 variables) and 0.20-0.40 m ($VIF_{Ksat0.20\text{-}0.40\ m}$ = $VIF_{bir0.20\text{-}0.40\ m}$ = 17 variables) assessed variables layers, respectively. This reduction in data dimensionality by removing redundant variables contributed to more informative components for the overall predictive models. Consequently, the final models adopted for data modeling and spatialization (mapping), based on separation per soil profile and study variable, were VIF and RFE for Ksat and Stepwise AIC and RFE for bir, respectively, considering both surface and subsurface variables profiles. These models played a crucial role in selecting and reducing input data dimensions at the pre-modeling stage, especially for water resources digital mapping in the studied basin's soils.

In terms of modeling quality, RFE and Stepwise AIC methods consistently yielded the best results for Random Forest models, regardless of the study variable (RFE: $Ksat_{0.00\text{-}0.20\ m}$ = 14.33 % and $bir_{0.00\text{-}0.20\ m}$ = 47.35 %; Stepwise AIC: $Ksat_{0.20\text{-}0.40\ m}$ = 10.88 % and $bir_{0.20\text{-}0.40\ m}$ = 20.03 %) and layer depth (applied to layers 0.00-0.20 m and 0.20-0.40 m). The explained variability by these models ranged 10.88-47.35 %, with the highest values achieved for bir at a depth of 0.00-0.20 m. However, it is worth noting that the RF models showed better adjustment for Ksat in both surface and subsurface profiles compared to bir, due to the higher spatial variability of the latter, which affects the algorithm modeling process.

The results provided a deeper comprehension of saturated water conductivity (Ksat) and basic soil infiltration velocity rate variability in the study area. The spectroradiometric and topographic data derived from Digital Elevation Models (DEM) integration led to more robust models, as evidenced by the high-quality digital soil physical-hydric attributes representation. Radiometry played a significant qualitative role in analyzing the soil particle sizes composition at the surface level. However, there were limitations that hindered a comprehensive analysis and prevented the determination of the potential of these data for numerical modeling using machine learning algorithms.

The approach herein adopted proved valuable in enhancing inherent relationship understanding among surface spectroradiometry, topography, soil composition, and water response at the soil surface, involving all the assessed databases. It raised important considerations regarding the selection of variables, showing how tenuous variables selection based on the Ksat and bir classification process is in the overall digital soil mapping context.

ACKNOWLEDGMENTS

The authors acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil, by the 001-finance code, for providing the master's scholarship resource. The authors also like to thank the Federal University of Rio de Janeiro (UFRRJ), for guaranteeing a public, free, and quality education; and the Graduate Program in Modeling and Geological Evolution (PPGMEG) and Brazilian Petroleum S.A. (Petrobras) for the project financial support. Furthermore, the authors are grateful to Embrapa Solos National Center for Soil Research (CNPS) for the partnership maintained with UFRRJ Soils Department and for providing equipment and support transport for field surveys. Finally, we would like to emphasize our special thanks to the Mineral Resources Directorate (DRM) for providing Guapi-Macacu river basin geological maps at a scale of 1:50,000, allowing the analyzes of the region.

AUTHORS CONTRIBUTIONS

Conceptualization: Priscilla Azevedo dos Santos (equal), Helena Saraiva Koenow Pinheiro (equal), Waldir de Carvalho Júnior (equal), and Igor Leite da Silva (equal).

Data curation: Priscilla Azevedo dos Santos (equal), Helena Saraiva Koenow Pinheiro (equal), Waldir de Carvalho Júnior (lead), Nilson Rendeiro Pereira (equal), and Silvio Barge Bhering (equal).

Formal Analysis: Priscilla Azevedo dos Santos (lead) and Igor Leite da Silva (equal).

Funding acquisition: Helena Saraiva Koenow Pinheiro (lead) and Waldir de Carvalho Júnior (equal).

Investigation: Priscilla Azevedo dos Santos (equal), Helena Saraiva Koenow Pinheiro (equal) and Waldir de Carvalho Júnior (equal).

Methodology: Priscilla Azevedo dos Santos (equal), Helena Saraiva Koenow Pinheiro (equal), Waldir de Carvalho Júnior (equal), Igor Leite da Silva (equal), Nilson Rendeiro Pereira (equal), Silvio Barge Bhering (equal), and Marcos Bacis Ceddia (supporting).

Project management: Helena Saraiva Koenow Pinheiro (lead), Waldir de Carvalho Júnior (equal), Nilson Rendeiro Pereira (equal) and Priscilla Azevedo dos Santos (equal).

Resources: Helena Saraiva Koenow Pinheiro (lead) and Waldir de Carvalho Júnior (equal).

Software and processing: Priscilla Azevedo dos Santos (lead) and Igor Leite da Silva (equal).

Supervision: Helena Saraiva Koenow Pinheiro (equal) and Waldir de Carvalho Júnior (equal).

Validation: Priscilla Azevedo dos Santos (lead), Helena Saraiva Koenow Pinheiro (equal), Igor Leite da Silva (equal) and Waldir de Carvalho Júnior (equal).

Visualization: Priscilla Azevedo dos Santos (lead), Helena Saraiva Koenow Pinheiro (equal) and Igor Leite da Silva (equal).

Writing – original draft: Priscilla Azevedo dos Santos (lead).

Writing – review and editing: Priscilla Azevedo dos Santos (lead), Helena Saraiva Koenow Pinheiro (supporting), Waldir de Carvalho Júnior (supporting), Marcos Bacis Ceddia (supporting), Silvio Barge Bhering (supporting) and Igor Leite da Silva (equal).

REFERENCES

Abdi H, Williams LJ. Principal component analysis. WIREs Comp Stat. 2010;2:433-59. https://doi.org/10.1002/wics.10110.1002/wics.101

Andrews DWK. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica. 1991;59:817-58. https://doi.org/10.2307/2938229

Atkinson PM, Tate NJ. Spatial scale problems and geostatistical solutions: A review. Prof Geogr. 2000;52:607-23. https://doi.org/10.1111/0033-0124.00250

Baalousha H. Assessment of a groundwater quality monitoring network using vulnerability mapping and geostatistics: A case study from Heretaunga Plains, New Zealand. Agr Water Manage. 2010;97:240-6. https://doi.org/10.1016/j.agwat.2009.09.013

Bannari A, Asalhi H, Teillet PM. Transformed difference vegetation index (TDVI) for vegetation cover mapping. In: IEEE International Geoscience and Remote Sensing Symposium; 2002 Jun; Toronto, ON, Canada. Toronto: IEEE; 2002. p. 3053-5. https://doi.org/10.1109/IGARSS.2002.1026867

Bertoni J, Lombardi Neto F. Conservação do solo. 10th ed. São Paulo: Ícone Editora; 2017.

Beven KJ, Kirkby MJ. A physically based, variable contributing area model of basin hydrology. Hydrol Sci B. 1979;24:43-69. https://doi.org/10.1080/02626667909491834

Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell. 1997;97:245-71. https://doi.org/10.1016/S0004-3702(97)00063-5

Bock M, Böhner J, Conrad O, Köthe R, Ringeler A. Methods for creating functional soil databases and applying digital soil mapping with SAGA GIS. In: Hengl T, Panagos P, Jones A, Toth G, editors. Status and prospect of soil information in south-eastern Europe: Soil databases, projects and applications. Luxemburg: Office for Official Publications of the European Communities; 2007. p. 149-62.

Böhner J, Selige T. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. Gött Geogr Abhandlungen. 2006;115:12-27.

Böhner J, Koethe R, Conrad O, Gross J, Ringeler A, Selige T. Soil regionalisation by means of terrain analysis and process parameterisation. Europ Soil Bur. 2002;7:213-22.

Bradley PS, Mangasarian OL, Street WN. Feature selection via mathematical programming. Informs J Comput. 1998;10:209-17. https://doi.org/10.1287/ijoc.10.2.209

Brenner VC, Guasselli LA. Índice de diferença normalizada da água (NDWI) para identificação de meandros ativos no leito do canal do rio Gravataí/RS–Brasil. In: Anais do XVII Simpósio Brasileiro de Sensoriamento Remoto -SBSR; 2015; João Pessoa, PB; Brasil. São José dos Campos: INPE; 2015. p. 3693-9.

Breuer B. Landform modelling of a research area in the Upper Palatinate, Germany, with the SARA software package (System for Automatical Relief Analysis). Z Geomorphol. 2001;45:17-31. https://doi.org/10.1127/zfg/45/2001/17

Breusch TS. Testing for autocorrelation in dynamic linear models. Aust Econ Pap. 1978;17:334-55. https://doi.org/10.1111/j.1467-8454.1978.tb00635.x

Carvalho CGP, Oliveira VR, Cruz CD, Casali VWD. Análise de trilha sob multicolinearidade em pimentão. Pesq Agropec Bras. 1999;34:603-13. https://doi.org/10.1590/S0100-204X1999000400011

Carvalho Filho A, Lumbreras JF, Wittern KP, Lemos AL, Santos RD, Calderano Filho B, Oliveira RP, Aglio MLD, Souza JS, Chaffin CE, Mothci EP, Larach JOI, Conceição M, Tavares NP, Santos HG, Gomes JBV, Calderano SB, Goncalves AO, Martorano LG, Barreto WO, Claessen MEC, Paula JL, Souza JLR, Lima TC, Antonello LL, Lima PC. Levantamento de reconhecimento de baixa intensidade dos solos do Estado do Rio de Janeiro. Rio de Janeiro: Embrapa Solos; 2003. (Boletim de pesquisa e desenvolvimento, 32).

Carvalho Junior W, Chagas CS, Fernandes Filho EI, Vieira CAO, Schaefer CEG, Bhering SB, Francelino MR. Digital soilscape mapping of tropical hillslope areas by neural networks. Sci Agric. 2011;68:691-6. https://doi.org/10.1590/S0103-90162011000600014

Carvalho Júnior W, Chagas CS, Muselli A, Pinheiro HSK, Pereira NR, Bhering SB. Método do hipercubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos. Rev Bras Cienc Solo. 2014;38:386-96. https://doi.org/10.1590/S0100-06832014000200003

Chagas CS. Mapeamento digital de solos por correlação ambiental e redes neurais em uma bacia hidrográfica no Domínio de mar de morros [thesis]. Viçosa, MG: Universidade Federal de Viçosa; 2006.

Chapuis RP. Predicting the saturated hydraulic conductivity of soils: A review. Bull Eng Geol Environ. 2012;71:401-34. https://doi.org/10.1007/s10064-012-0418-7

Clevers JGPW, Gitelson AA. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. Int J Appl Earth Obs. 2013;23:344-51. https://doi.org/10.1016/j.jag.2012.10.008

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci Model Dev. 2015;8:1991-2007. https://doi.org/10.5194/gmd-8-1991-2015

Cunha AM. Seleção de variáveis ambientais e de algoritmos de classificação para mapeamento digital de solos [thesis]. Viçosa, MG: Universidade Federal de Viçosa; 2013.

Daniel C, Wood FS. Fitting equations to Data: Computer analysis of multifactor Data. 2nd. ed. New York: Wiley-InterScience; 1999.

Darcy H. Les fontaines publiques de la ville de Dijon: Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau, ouvrage terminé par un appendice relatif aux fournitures d'eau de plusieurs villes au filtrage des eaux et à la fabrication des tuyaux de fonte, de plomb, de tole et de bitume. France: Victor Dalmont; 1856.

Diretoria de Recursos Minerais - DRM. Carta geológica na escala 1:50.000 dos municípios de Itaboraí, Itaipava, Nova Friburgo, Teresópolis, Petrópolis e Rio Bonito. 2019. In: Santos, P. A.: Mapeamento e modelagem digital da variabilidade tridimensional de atributos físico-hídricos dos solos da bacia do rio Guapi-Macacu - RJ, por estatística multivariada e algoritmos [dissertation]. Seropédica: Universidade Federal Rural do Rio de Janeiro; 2021, p.172-178. Data provided by Official Request Letter No. 07/2019 to the Federal Rural University of Rio de Janeiro (UFRRJ). Available from: https://tede.ufrrj.br/jspui/handle/jspui/6870

Drury SA. Image interpretation in geology. Geocarto Int. 1987;2:48. https://doi.org/10.1080/10106048709354098

Durbin J, Watson GS. Testing for serial correlation in least squares regression: I. Biometrika. 1950;37:409-28. https://doi.org/10.1093/biomet/37.3-4.409

Elrick DE, Reynolds WD, Tan KA. Hydraulic conductivity measurements in the unsaturated zone using improved well analyses. Groundwater Monit Remediat. 1989;9:184-93. https://doi.org/10.1111/j.1745-6592.1989.tb01162.x

European Space Agency - ESA. User guide for Sentinel-2 MSI Processing Levels: Level-2 Products. European Space Agency Signature; 2020 [cited 2020 Feb 28]. Available from: https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/processing-levels/level-2.

Everts CJ, Kanwar RS. Interpreting tension-infiltrometer data for quantifying soil macropores: Some practical considerations. Trans ASAE. 1993;36:423-8. https://doi.org/10.13031/2013.28354

Fathololoumi S, Vaezi AR, Alavipanah SK, Ghorbani A, Saurette D, Biswas A. Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach. Geoderma. 2021;385:114901. https://doi.org/10.1016/j.geoderma.2020.114901

Ferrari AL. Evolução tectônica do Graben da Guanabara [thesis]. São Paulo: Universidade de São Paulo; 2001.

Florinsky IV, Eilers RG, Manning GR, Fuller LG. Prediction of soil properties by digital terrain modelling. Environ Modell Softw. 2002;17:295-311. https://doi.org/10.1016/S1364-8152(01)00067-6

Gallant JC, Dowling TI. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour Res. 2003;39:1347. https://doi.org/10.1029/2002WR001426

Gerlitz L, Conrad O, Böhner J. Large-scale atmospheric forcing and topographic modification of precipitation rates over High Asia – A neural-network-based approach. Earth Syst Dynam. 2015;6:61-81. https://doi.org/10.5194/esd-6-61-2015

Gitelson AA, Merzlyak MN, Zur Y, Stark R, Gritz U. Non-destructive and remote sensing techniques for estimation of vegetation status. In: Proceedings 3rd European Conference on Precision Agriculture; 2001; University of Nebraska. Montpelier, France: Grenier & Blackmore editors; 2001. p. 205-10.

Goel NS, Qin W. Influences of canopy architecture on relationships between various vegetation indices and LAI and Fpar: A computer simulation. Remote Sens Rev. 1994;10:309-47. https://doi.org/10.1080/02757259409532252

Gotway CA, Young LJ. Combining incompatible spatial Data. J Am Stat Assoc. 2002;97:632-48. https://doi.org/10.1198/016214502760047140

Granata F, Di Nunno F, Modoni G. Hybrid machine learning models for soil saturated conductivity prediction. Water. 2022;14:1729. https://doi.org/10.3390/w14111729

Guisan A, Weiss SB, Weiss AD. GLM versus CCA spatial modeling of plant species distribution. Plant Ecol. 1999;143:107-22. https://doi.org/10.1023/A:1009841519580

Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ. 2002;83:195-213. https://doi.org/10.1016/S0034-4257(02)00096-2

Huete AR. A soil-adjusted vegetation index (SAVI). Remote Sens Environ. 1988;25:295-309. https://doi.org/10.1016/0034-4257(88)90106-X

Hunt ER, Daughtry CST, Eitel JUH, Long DS. Remote sensing leaf chlorophyll content using a visible band index. Agron J. 2011;103:1090-9. https://doi.org/10.2134/agronj2010.0395

HWA CS, Hora MAG, Hora AF. Projeto Macacu. Planejamento estratégico da região hidrográfica dos rios Guapi-Macacu e Caceribu-Macacu. Região Hidrográfica Baía de Guanabara. RJ: Universidade Federal Fluminense / Fundação Euclides da Cunha; 2010.

Inman HF. Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature. Am Stat. 1994;48:2-11. https://doi.org/10.1080/00031305.1994.10476010

Iwahashi J, Pike RJ. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology. 2007;86:409-40. https://doi.org/10.1016/j.geomorph.2006.09.012

Jasiewicz J, Stepinski TF. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology. 2013;182:147-56. https://doi.org/10.1016/j.geomorph.2012.11.005

King ML. Introduction to Durbin and Watson (1950, 1951) testing for serial correlation in least squares regression. I, II. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics. Volume 2: Methodology and distribution. New York: Springer; 1992. p. 229-36. https://doi.org/10.1007/978-1-4612-4380-9_19

Kirsch R. Groundwater geophysics: A tool for hydrogeology. Berlin, Heidelberg: Springer; 2009. https://doi.org/10.1007/978-3-540-88405-7

Klar AE. A água no sistema solo-planta-atmosfera. São Paulo: Livraria Nobel; 1984.

Köthe R, Lehmeier F. SARA-system zur automatischen relief-analyse. User Manual. 2nd ed. Goettingen: Department of Geography, University of Goettingen; 1996. [unpublished].

Köthe, R., Gehrt, E., and Böhner, J.: Automatische Reliefanalyse für geowissenschaftliche Kartierungen, Arbeitshefte Boden, 1, 31–37, 1996. Available from: https://www.researchgate.net/publication/285449046_Automatische_Reliefanalyse_fur_geowissenschaftliche_Anwendungen-_derzeitiger_Stand_undWeiterentwicklungen_des_Programms_SARA.

Lee S, Lee C-W. Application of decision-tree model to groundwater productivity-potential mapping. Sustainability. 2015;7:13416-32. https://doi.org/10.3390/su71013416

Lee S, Song K-Y, Kim Y, Park I. Regional groundwater productivity potential mapping using a geographic information system (GIS) based artificial neural network model. Hydrogeol J. 2012;20:1511-27. https://doi.org/10.1007/s10040-012-0894-7

Madrucci V, Taioli F, Araújo CC. Groundwater favorability map using GIS multicriteria data analysis on crystalline terrain, São Paulo State, Brazil. J Hydrol. 2008;357:153-73. https://doi.org/10.1016/j.jhydrol.2008.03.026

Mansfield ER, Helms BP. Detecting multicollinearity. Am Stat. 1982;36:158-60. https://doi.org/10.1080/00031305.1982.10482818

Manzione RL, Castrignanò A. A geostatistical approach for multi-source data fusion to predict water table depth. Sci Total Environ. 2019;696:133763. https://doi.org/10.1016/j.scitotenv.2019.133763

McBratney AB, Santos MLM, Minasny B. On digital soil mapping. Geoderma. 2003;117:3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. 2nd. ed. New York: Chapman and Hall/CRC; 2020. https://doi.org/10.1201/9780429029608

Mckay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. technometrics. 2000;42:55-61. https://doi.org/10.1080/00401706.2000.10485979

McKeague JA, Wang C, Topp GC. Estimating saturated hydraulic conductivity from soil morphology. Soil Sci Soc Am J. 1982;46:1239-44. https://doi.org/10.2136/sssaj1982.03615995004600060024x

McKenzie NJ, Austin MP. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma. 1993;57:329-55. https://doi.org/10.1016/0016-7061(93)90049-Q

Mihola J, Bílková D. Measurement of multicolinearity using determinants of correlation matrix. Int J Math Sci. 2014;34:1543-9.

Minasny B, McBratney AB. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput Geosci. 2006;32:1378-88. https://doi.org/10.1016/j.cageo.2005.12.009

Montgomery DR, Dietrich WE. A physically based model for the topographic control on shallow landsliding. Water Resour Res. 1994;30:1153-71. https://doi.org/10.1029/93WR02979

Moore ID, Grayson RB, Ladson AR. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrol Process. 1991;5:3-30. https://doi.org/10.1002/hyp.3360050103

Novakowski K, Bickerton G, Lapcevic P, Voralek J, Ross N. Measurements of groundwater velocity in discrete rock fractures. J Contam Hydrol. 2006;82:44-60. https://doi.org/10.1016/j.jconhyd.2005.09.001

O’Hagan J, McCabe B. Tests for the severity of multicolinearity in regression analysis: A comment. Rev Econ Stat. 1975;57:368-70. https://doi.org/10.2307/1923927

Oliveira KD, Kapiche ALAF, Costa TA, Sanches ID. Classificação de atributos topográficos para distinção de propriedades físico-hídricas e termodinâmicas do solo. In: Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR; 2017; Campinas, SP. Galoá; INPE Santos; 2017. p. 3499-506.

Ottoni MV. Classificação físico-hídrica de solos e determinação da capacidade de campo in situ a partir de testes de infiltração [dissertation]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2005.

Perera YY, Zapata CE, Houston WN, Houston SL. Prediction of the soil-water characteristic curve based on grain-size-distribution and index properties. In: Proceedings of Geo-Frontiers Congress 2005 – Advances in Pavement Engineering; 2005 Oct 9; Austin, Texas, United States. Reston, Virgínia: American Society of Civil Engineers; 2005. p. 1-12. https://doi.org/10.1061/40776(155)4

Pinheiro HSK. Métodos de mapeamento digital aplicados na predição de classes e atributos dos solos da bacia hidrográfica do rio Guapi-Macacu, RJ [thesis]. Seropédica: Universidade Federal Rural do Rio de Janeiro; 2015.

Pinheiro HSK. Mapeamento digital de solos por redes neurais artificiais da bacia hidrográfica do rio Guapi-Macacu, RJ [dissertation]. Seropédica: Universidade Federal Rural do Rio de Janeiro; 2012.

Pinheiro HSK, Barbosa TPR, Antunes MAH, Carvalho DC, Nummer AR, Carvalho Junior W, Chagas CS, Fernandes-Filho EI, Pereira MG. Assessment of phytoecological variability by red-edge spectral indices and soil-landscape relationships. Remote Sens. 2019;11:2448. https://doi.org/10.3390/rs11202448

Pires CA, Miranda A. Análise geométrica de lineamentos e suas relações com águas subterrâneas associadas ao Aquífero Guaratiba - Região de Campo Grande e Guaratiba, RJ [monography]. Seropédica: Universidade Federal Rural do Rio de Janeiro; 2017. https://doi.org/10.13140/RG.2.2.18646.91209

QGIS Development Team. QGIS Geographic Information System [software]. 2020.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: http://www.R-project.org/.

Rajah P, Odindi J, Mutanga O, Kiala Z. The utility of Sentinel-2 Vegetation Indices (VIs) and Sentinel-1 Synthetic Aperture Radar (SAR) for invasive alien species detection and mapping. Nat Conserv. 2019;35:41-61. https://doi.org/10.3897/natureconservation.35.29588

Redlands. ArcGIS Desktop: Release 10. Redlands: Environmental Systems Research Institute; 2011.

Reichert JM, Veiga M, Cabeda MSV. Selamento superficial e infiltração de água em solos do Rio Grande do Sul. Rev Bras Cienc Solo. 1992;16:289-98.

Reynolds WD, Elrick DE. In situ measurement of field-saturated hydraulic conductivity, sorptivity, and the α-parameter using the guelph permeameter. Soil Sci. 1985;140:292-302. https://doi.org/10.1097/00010694-198510000-00008

Richards LA. Capillary conduction of liquids through porous mediums. Physics. 1931;1:318-33. https://doi.org/10.1063/1.1745010

Romano N, Chirico GB. The role of terrain analysis in using and developing pedotransfer functions. In: Pachepsky Y, Rawls WJ, editors. Developments in soil science. Netherlands: Elsevier; 2004. vol. 30. p. 273-94. . https://doi.org/10.1016/S0166-2481(04)30016-4. ISSN: 0166-2481

Rouse JW, Haas RH, Schell JA, Deering DW. Monitoring vegetation systems in the great plains with ERTS. In: Freden SC, Mercanti EP, Becker MA, editors. The proceedings of a symposium held by Goddard Space Flight Center at Washington, D.C; Dec 1973. Washington, D.C: Scientific and Technical Information Office, National Aeronautics and Space Administration; 1973. p. 208-317.

Rowan LC, Mars JC. Lithologic mapping in the Mountain Pass, California area using Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data. Remote Sens Environ. 2003;84:350-66. https://doi.org/10.1016/S0034-4257(02)00127-X

RStudio Team. RStudio: Integrated Development Environment for R. Boston: RStudio; 2020.

Santos PA, Pinheiro HSK, Carvalho Junior W, Bhering SB. Análise preliminar da correlação entre parâmetros hidropedológicos e covariáveis ambientais morfométricas e radiométricas como suporte ao mapeamento e modelagem da velocidade de infiltração básica dos solos da bacia hidrográfica do Rio Guapi-Macacu. In: Geosudeste 2019 – Anais do 16 Simpósio de Geologia do Sudeste, 9 Simpósio Nacional de Ensino e História de Ciência da Terra, 20 Simpósio de geologia de Minas Gerais; 2019 Oct; Campinas, São Paulo. São Paulo: Sociedade Brasileira de Geologia, Núcleo São Paulo; 2019. p. 271.

Santos PA, Pinheiro HSK, Junior WC, Pereira NR. Aplicação de ferramentas SIG nas análises geométrica e morfométrica para caracterização hidrológica das Bacias Hidrográficas do Rio Guapi-Macacu, RJ. In: Anais da V Jornada de Geotecnologias do Estado do Rio de Janeiro (V JGEOTEC), 09-12 de novembro de 2020; Niterói, Rio de Janeiro. Rio de Janeiro: Geopartners; 2020. p. 1018-21.

Santos PA, Pinheiro HS, Carvalho Junior W, Pereira NR, Bhering SB, Silva IL. Modeling soils physical-hydric attributes through algorithms for quantitative pedology in Guapi-Macacu watershed, RJ. In: In: II Pedometrics Brazil Annals, 24-27 november 2021. Rio de Janeiro: Embrapa Solos, UFRRJ; 2022. p. 25-8.

Santos PA, Pinheiro HSK, Silva IL. Análise de produtos oriundos de MDE para compreensão dos recursos hídricos na bacia hidrográfica do Rio Guapi-Macacu, RJ: Um estudo em ambiente SIG. In: Anais do II Congresso Alagoano de Engenharia de Agrimensura (CONEAGRI), 02 a 04 de dezembro de 2019; Centro de Ciências Agrárias (CECA), Rio Largo, Alagoas. Maceió: Repositório Institucional da Universidade Federal de Alagoas, Editora EDUFAL; 2019. p. 40-59.

Schaap MG. Accuracy and uncertainty in PTF predictions. Dev Soil Sci. 2004;30:33-43. https://doi.org/10.1016/S0166-2481(04)30003-6

Schaap M, Leij FJ. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil Till Res. 1998;47:37-42. https://doi.org/10.1016/S0167-1987(98)00070-1

Segal D. Theoretical basis for differentiation of ferric-iron bearing minerals, using Landsat MSS Data. In: Proceedings of the 2nd Thematic Conference on Remote Sensing for Exploratory Geology, Symposium for Remote Sensing of Environment; 1982 Dec; Fort Worth, Texas, USA. United States: Department of Energy; 1982. p. 949-51.

Seibert J, McGlynn BL. A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. Water Resour Res. 2007;43:W04501. https://doi.org/10.1029/2006WR005128

Serviço Geológico do Brasil - CPRM. Projeto Aerogeofísico Rio de Janeiro. Relatório Final do Levantamento e Processamento dos Dados Magnetométricos e Gamaespectrométricos. Rio de Janeiro: Prospectors Aerolevantamentos e Sistemas Ltda; 2012. Available from: file:///C:/Users/DeniseM/Downloads/Relatorio%20Final%20Projeto%20Aerogeofisico%20Rio%20de%20Janeiro.pdf.

Shapiro SS, Wilk MB. An Analysis of variance test for normality (Complete Samples). Biometrika. 1965;52:591-611. https://doi.org/10.2307/2333709

Sinergise. Sentinel-Hub Repository Satellite Indices: Index database for Sentinel-2 Satellite (Sentinel-2 RS indices); 2020 [cited 2020 Dec 23]. Available from: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/indexdb/.

Stepinski TF, Jasiewicz J. Geomorphons - a new approach to classification of landforms. Proc Geomorph. 2011;2011:109-12.

Svetnik V, Liaw A, Tong C, Wang T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T, editors. Multiple classifier systems. Berlin, Heidelberg: Springer; 2004. p. 334-43. https://doi.org/10.1007/978-3-540-25966-4_33

Taddy M. Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. New York: McGraw-Hill Education; 2019.

Watson GS, Durbin J. Exact tests of serial correlation using noncircular statistics. Ann Math Stat. 1951;22:446-51. https://doi.org/10.1214/aoms/1177729592

Wilson JP, Gallant JC. Terrain analysis: Principles and applications. New York: Wiley; 2000.

Xiao J, Shen Y, Tateishi R, Bayaer W. Development of topsoil grain size index for monitoring desertification in arid land using remote sensing. Int J Remote Sens. 2006;27:2411-22. https://doi.org/10.1080/01431160600554363

Yamaç SS, Negiş H, Şeker C, Memon AM, Kurtuluş B, Todorovic M, Alomair G. Saturated hydraulic conductivity estimation using artificial intelligence techniques: A case study for calcareous alluvial soils in a semi-arid region. Water. 2022;14:3875. https://doi.org/10.3390/w14233875

Yokoyama R, Shirasawa M, Pike RJ. Visualizing topography by openness: A new application of image processing to digital elevation models. Photogramm Eng Rem S. 2002;68:257-66.

Zevenbergen LW, Thorne CR. Quantitative analysis of land surface topography. Earth Surf Process Landforms. 1987;12:47-56. https://doi.org/10.1002/esp.3290120107

Zhang T, Su J, Liu C, Chen W-H, Liu H, Liu G. Band selection in sentinel-2 satellite for agriculture applications. In: Poceedings of the 23rd International Conference on Automation and Computing (ICAC), 7-8 september 2017. Huddersfield, United Kingdom: IEEE; 2017. p. 1-6. https://doi.org/10.23919/IConAC.2017.8081990