UFRRJ

INSTITUTO DE AGRONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CIÊNCIA DO SOLO

TESE

Caracterização de Solos e Avaliação da Vulnerabilidade de Ambientes no Parque Nacional de Itatiaia, Brasil

Elias Mendes Costa

2019

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA
CIÊNCIA DO SOLO

CARACTERIZAÇÃO DE SOLOS E AVALIAÇÃO DA VULNERABILIDADE DE AMBIENTES NO PARQUE NACIONAL DE ITATIAIA, BRASIL

ELIAS MENDES COSTA

Sob a orientação da Professora
Lúcia Helena Cunha dos Anjos

e Coorientação da Professora
Helena Saraiva Koenow Pinheiro

e da Pesquisadora
Laura Poggio

Tese submetida como requisito parcial para obtenção do grau de **Doutor** no Programa de Pós-Graduação em Agronomia, Área de Concentração em Ciência do Solo

Seropédica, RJ
Fevereiro de 2019

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

Costa, Elias Mendes, 1989-
C 837c Caracterização de Solos e Avaliação da
Vulnerabilidade de Ambientes no Parque Nacional de
Itatiaia, Brasil / Elias Mendes Costa. - Seropédica,
2019.
121 f.

Orientadora: Lúcia Helena Cunha dos Anjos.
Coorientadora: Helena Saraiva Koenow Pinheiro.
Coorientadora: Laura Poggio.
Tese (Doutorado). -- Universidade Federal Rural do Rio de Janeiro,
Pós-Graduação em Agronomia - Ciência do Solo, 2019.

1. Pedometria. 2. Mapeamento Digital de Solos. 3.
Levantamento de Solo. 4. Funções do Solo. I. Cunha dos
Anjos, Lúcia Helena, 1957, orient. II. Saraiva Koenow Pinheiro,
Helena, coorient. III. Poggio, Laura, coorient.
IV Universidade Federal Rural do Rio de Janeiro.
Pós-Graduação em Agronomia - Ciência do Solo. V. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA - CIÊNCIA DO SOLO

ELIAS MENDES COSTA

Tese submetida como requisito parcial para obtenção do grau de **Doutor**, no Programa de Pós-Graduação em Agronomia, área de concentração em Ciência do Solo.

TESE APROVADA EM 19/02/2019

_____
Lúcia Helena Cunha dos Anjos. Ph.D. UFRRJ
(Presidente)

_____
Marcos Bacis Ceddia. Dr. UFRRJ

_____
Alessandro Samuel Rosa. Dr. UTFPR

_____
César da Silva Chagas. Dr. Embrapa Solos

_____
Elpídio Inácio Fernandes Filho. Dr. UFV

DEDICATÓRIA

Aos meus pais Marly Mendes Costa e João Raulino Mendes,
aos meus irmãos Joabe, Jaqueline e Janaína,
aos meus sobrinhos Bernardo e Elisa
a minha fonte de inspiração, meu filho Arthur e minha companheira Luzilene
aos familiares, amigos,
e a meus tutores.

Dedico

AGRADECIMENTOS

Primeiramente a Deus por conceder a vida e todos os recursos para a sua existência além de ter colocado pessoas maravilhosas em todas as fases de minha vida.

Ao comitê de orientação Lúcia Helena Cunha dos Anjos, Helena Saraiva Koenow Pinheiro e Laura Poggio, pela paciência e por todos os ensinamentos que tornou esse trabalho possível, e que ainda me inspira na busca pelo conhecimento da "nova" pedologia. Obrigado pela confiança, espero ter feito um bom trabalho.

Aos meus pais Marly Mendes Costa e João Raulino Mendes, que não mediram esforços para que eu pudesse conquistar meus sonhos. Aos meus irmãos Joabe, Jaqueline e Janaína e meus sobrinhos, Bernardo e Elisa por me proporcionarem grandes momentos juntos, sempre com diversão e descontração e por manter nossa família unida.

Ao meu bem maior, meu filho Arthur Souza Mendes. Apesar da pouca idade muito me ensinou sobre a vida, e por ele tento ser a cada dia uma pessoa melhor.

A minha companheira Luzilene Maria de Souza por ser importante coadjuvante no incentivo e na busca dos ideais, das conversas, dos conselhos, além de segurar a barra na minha ausência, sei que passamos difíceis momentos.

As Minhas Avós Eva e Maria Madalena sei que são muito orgulhosas de mim.

Aos professores e amigos Marcos Gervasio Pereira, Mauro Antunes, Wagner de Souza Tassinari, Alessandro Samuel Rosa, Elpídio Inácio Fernandes Filho, Marcos Bacis Ceddia e aos pesquisadores Ademir Fontana, Waldir de Carvalho Junior e César da Silva Chagas e ao Pós doutorando Sidnei Julio Beutler. Muito aprendi com vocês sobre Pedologia, Pedometria Sensoriamento Remoto e Modelagem Espacial.

Aos amigos Yuri Andrei Gelsleichter e Alessandro Samuel Rosa pelas provocações que me levaram a usar o sistema Linux. Muito aprendi desse novo mundo do *free systems.*

Aos amigos Alexandre Medeiros, Gisonley Lopes, Deyvid Maranhão, André Geraldo, Ademir Fontana, Robson Marcondes, Sidinei Beutler, Helena Pinheiro, Yuri Gelsleichter, Vanessa Freo, Roger Mejia. Sem vocês a coleta de dados seria impossível. Serei eternamente grato pelo esforço que todos fizeram. Passamos por condições extremas, mas com a força de todos e no controle de Deus tudo deu certo e agora só lembranças... Mas quem sabe um dia voltamos!!!!

A todos os alunos e estagiários do Laboratório de Gênese e Classificação do Solo, pelo apoio, amizade e experiências em especial a Robson Marcondes e Yuri Gelsleichter que me ajudaram nas análises de laboratório. Aos funcionários do Departamento de Solos, em especial a Maria Helena pelo apoio com as análises e pela amizade. Ao motorista da UFRRJ, Moraes por nos conduzir até a área de coleta de solos e ao coordenador do Programa de Pós-Graduação Everaldo Zonta me muito "pepino" resolveu para mim.

Agradecimento especial a todos os professores e aos colegas de turma pelos seus ensinamentos e experiências compartilhadas em especial Wilk, Lívia e Barbara.

Aos amigos de rural, alojamento em especial os Kappas

A CNPq pela concessão da bolsa de estudos no Brasil (processo 141391/2015-4), a CAPES pela concessão da bolsa de estudos no exterior (processo 88881.135776/2016-01) e a FAPERJ pelo auxílio financeiro deste projeto no edital E_34 / 2014 - PENSA RIO de apoio ao estudo de assuntos relevantes e estratégicos para o RJ.

A toda a coordenação do Parque Nacional de Itatiaia em especial Marcelo Motta, Léo Nascimento e Gustavo Tomzhinski por todo apoio prestado e pelo fornecimento da base cartográfica disponível no parque.

Ao Ministério do Meio Ambiente pelo fornecimento das imagens RapidEye

A todos os Pesquisadores e estudantes do Instituto James Hutton em especial Laura Poggio e Alessandro Gimona.

A EMBRAPA Solos com quem temos parceria e onde parte das análises de laboratório foram realizadas, sob a supervisão de Ademir Fontana.

Em fim a Rural, UFRRJ, onde passei meus últimos 11 anos serei eternamente grato a esta instituição e todos que aqui trabalham.

Tenho certeza que esqueci algumas pessoas, a final de contas recebi muita ajuda. Nesse estudo uma das poucas certezas que tenho, é que sem essas pessoas com quem convivi todos esses anos, não teria conseguido concluir este trabalho.

Muito obrigado.

BIOGRAFIA

Elias Mendes Costa nasceu no município de São Francisco - MG, no dia 02 de novembro de 1989. Em 2004 concluiu o ensino fundamental na Escola Estadual Sebastiana Pereira da Silva no Povoado Santana de Minas (Jiboia), Município São Francisco-MG, onde estudou desde o maternal. Em 2007 concluiu o ensino médio no Colégio Prisma, Montes Claros- MG. Em 2008 ingressou no curso de graduação em Engenharia Agronômica da Universidade Federal Rural do Rio de Janeiro (UFRRJ), concluindo-o em 2013. Durante o período de graduação foi bolsista de Iniciação Científica (PROIC e AGRISUS) também exerceu monitoria da disciplina de Física do Solo. Em março de 2013 ingressou no Mestrado do Programa de Pós-graduação em Agronomia - Ciência do Solo (PPGA-CS), na UFRRJ, concluindo-o em fevereiro de 2015. Ingressou no Curso de Especialização em Estatística Aplicada (*latu sensu*) da UFRRJ em maio de 2016 concluindo-o em novembro de 2017. Em março de 2015 ingressou no Doutorado do PPGA-CS, na UFRRJ, concluindo-o em fevereiro de 2019. Durante o doutorado fez estágio no exterior, Doutorado Sanduíche de maio de 2017 a abril de 2018 no The James Hutton Institute, Aberdeen, Escócia

RESUMO GERAL

COSTA, Elias Mendes. **Caracterização de solos e avaliação da vulnerabilidade de ambientes no Parque Nacional de Itatiaia, Brasil.** 2019. 121f. Tese (Doutorado em Agronomia - Ciência do Solo). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2019.

O conhecimento dos solos e suas propriedades é essencial para o planejamento ambiental em sistemas naturais especialmente em unidade de conservação como o Parque Nacional de Itatiaia (PNI). O PNI apesar da importância ecologia e de preservação não tem informações sobre seus solos em nível de detalhe que possa dar suporte a pesquisas e ao plano de manejo. Buscando entender o processo envolvendo a gênese e distribuição dos solos no ambiente montanhoso do PNI e fatores que envolvem a vulnerabilidade ambiental nessa região o presente estudo foi desenvolvido. Os objetivos foram desenvolver uma base de dados num ambiente SIG com informação sobre os solos (classes e atributos), vegetação, relevo, geologia e (covariáveis ambientais) para apoiar ações de investigação interdisciplinar, programas de educação ambiental e plano do manejo do parque. Ainda avaliar a vulnerabilidade ambiental integrando informações do ambiente físico com conhecimento de especialistas para conciliar a demanda de uso público com a conservação dos ecossistemas. Para tanto foi feita amostragem, coleta, descrição, caracterização, classificação e mapeamento dos solos e foi preparado uma base de dados com todas as covariáveis ambientais de posse dos dados, métodos robustos de mapeamento digital de solos foram testados a fim de se otimizar o desempenho dos algoritmos para a predição de atributos de solo e avaliação de incerteza. Por fim, dados da revisão de literatura, abordagem participativa e conhecimento especializado e variáveis biofísicas produzidas nas etapas anteriores foram incorporadas em uma rede de crença Bayesiana (BBN, inglês) para predizer a vulnerabilidade ambiental, bem como para produzir a incerteza associada. Os resultados produzidos foram suficientes para preencher a lacuna da falta de informação sobre solos no PNI e entender os fatores relacionados a relação solo paisagem do PNI e são úteis para diversos fins. Algoritmos como o Modelos Aditivos Generalizados (GAM) com seleção de covariáveis baseado no modelo *scorpan* são eficientes em predizer atributos do mesmo utilizando limitado número de pontos. E apesar da complexidade da área de estudo, BBN conseguiu produzir um resultado significativo da distribuição espacial da vulnerabilidade ambiental e se mostrou uma abordagem alternativa menos subjetiva do que os convencionais métodos de avaliação da vulnerabilidade ambiental.

Palavras-chave: Pedometria. Mapeamento Digital de Solos. Levantamento de Solo. Funções do Solo.

GENERAL ABSTRACT

COSTA, Elias Mendes. **Soil characterization and evaluation of environments vulnerability in Itatiaia National Park, Brazil**. 2019. 121p. Thesis (Doctor in Agronomy-Soil Science). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2019.

Knowledge of soils and their properties is essential for environmental planning in natural systems especially in a conservation unit such as the Itatiaia National Park (INP). The INP, despite the importance of ecology and preservation, does not have information on its soils in detail that can support research and management plan. Aiming to understand the process involving the genesis and soils distribution in the mountainous environment of the INP and factors that involve the environmental vulnerability in this region the present study was developed. The objectives were to develop a database in a GIS environment with information on soils (classes and attributes), vegetation, relief, geology and geomorphology and to produce (environmental covariates) to support interdisciplinary research actions, environmental education programs and plan of park management. To further evaluate environmental vulnerability by integrating information from the physical environment with expert knowledge to reconcile public use demand with ecosystem conservation. In order to do so, sampling, collection, description, characterization, classification and mapping of soils was prepared and a database was prepared with all the environmental variables of data ownership, robust methods of digital soil mapping were tested in order to optimize the performance of the algorithms for the prediction of soil attributes and uncertainty evaluation. Finally, data from the literature review, participatory approach and specialized knowledge and biophysical variables produced in the previous steps were incorporated into a Bayesian belief network (BBN) to predict environmental vulnerability as well as to produce associated uncertainty. The results produced were sufficient to fill the gap in the lack of information on soils in the INP and to understand the factors related to the landscape soil relationship of the INP and are useful for several purposes. Generalized Additive Model Algorithms (GAM) with covariates selection based on the *scorpan* model are efficient in predicting attributes of the same using a limited number of points. And despite the complexity of the study area, the BBN was able to produce a significant result of the spatial distribution of environmental vulnerability and proved to be an alternative approach less subjective than conventional methods of assessing environmental vulnerability.

Keywords: Pedometrics. Digital Soil Mapping. Soil Survey. Soil Functions

LIST OF TABLES

LIST OF FIGURES

SUMMARY

1 GENERAL INTRODUCTION

Among all Brazilian biomes, the Atlantic Forest stands out for its importance in preserving biodiversity and as source and protection of water resources for the states along the coast of Brazil. It contributes to 7 of the 9 largest river basins in the country. In the case of the Itatiaia National Park (INP), which is located in the southeast portion of the Atlantic Forest, its preservation gains strength, since the park area holds the sources of 12 important regional watersheds. These waters drain into the basins of the Rio Grande, which is an affluent of the rivers Paraiba do Sul and Paraná. The Paraiba do Sul is the main river of Rio de Janeiro state and is responsible for supplying water for most of the cities of Rio de Janeiro and São Paulo States.

INP was selected to be a conservation unit due to the ecological relevance, as a site for integral protection of nature, preservation of natural resources, especially endemic species, and springs of important rivers of the Brazilian south-eastern region. It is the first national park of Brazil, and due to the scenic beauty, it has a high tourist potential. The park has been used intensively for scientific research on fauna and flora, and it presents many opportunities for environmental education and leisure activities.

The INP can be considered an "island of conservation", since it is located between Rio de Janeiro and São Paulo, the two largest metropolises in the country, which intensifies the degree of vulnerability and the relevance of this area. The analysis of environmental vulnerability is related to the susceptibility of the area, which is linked to natural factors such as relief type, geological material, vegetation cover, soils and climate and can be aggravated by anthropic factors, in the case of INP public use and/or agricultural/urban use pressure within its limits and boundaries. These factors can exacerbate erosive processes in the environments that are naturally fragile. It is important to reconcile preservation and public use, in addition to the guarantees of continuous provision of ecosystem services provided by INP, for example, water production and protection of biodiversity. This goal is even more strained in the upper part of INP, the plateau.

The geotechnology techniques applied to landscape analysis and environmental modelling are potentially useful both to produce information of soils, in the case of digital soil mapping (DSM) techniques, and the spatial environmental vulnerability analysis. These tools deliver useful quantitative information, of less subjectivity, with associated errors and uncertainty.

Thus, this study aimed to understand aspects that involve environmental vulnerability and the relation of environmental and anthropic factors related to the soil's degradation, which environmental factors are more related to the properties of the soils and how they vary in space in this singular condition that is the upper part of the INP. Also, investigate the best approach (soil sampling and variable selection) and survey techniques to predict soil classes and attributes, with an uncertainty assessment, in areas with a difficult access in the park.

To that end, three scientific hypotheses have been raised: the first is that the use of environmental covariates associated with modern DSM techniques can be applied for selection of soil samples, better understanding of the soil-landscape relationship and for the soil mapping. The second that powerful MDS techniques can improve spatial prediction results even using limited number of points in a poorly accessible area. And the third is that Physiographic variables associated with models that use specialist knowledge and participatory process, allow better evaluation of environmental vulnerability

To answer these hypotheses, the thesis is organized in three chapters, with the following titles and objectives:

Chapter 1: The Itatiaia National Park - soils and physiography database. The objectives, to develop a database in a Geographic Information System (GIS) environment with information on the soils, satellite images, digital elevation model, terrain attributes, map and land cover, geology and geomorphology. Also, to produce information on soil and environmental covariates to support this thesis (subsequent chapters), interdisciplinary research actions, environmental education programs and INP management plan.

Chapter 2: Mapping soil properties in a poorly accessible area. Case study - Itatiaia national park: as objective creating maps of soil properties 2 and 3D, using as examples, pH, carbon content and cation exchange capacity at fine resolution (25 m) with associated spatial uncertainty. Further elaborate a sampling strategy that balances accessibility, costs, area and environmental covariates; and model soil properties with a limited number of available samples.

Chapter 3: Spatial Bayesian belief network: a participatory approach for mapping environmental vulnerability at the Itatiaia national park: as objectives, to assess soil vulnerability in the INP, integrating information from the physical environment with expert knowledge to reconcile the demand for public use with conservation of ecosystems. In addition, to reduce the subjectivity of the traditional process of environmental vulnerability analysis, incorporating specialized knowledge and literature review with a quantitative/probabilistic approach called the Bayesian belief network (BBN).

2 LITERATURE REVIEW

2.1 Ecotourism, Soil and Environmental Degradation and Vulnerability Analysis

Among the studies in Brazil that deal with ecotourism and its relationship with the landscape and environmental conservation, it is highlighted Oliveira et al. (2007), who identified landscape units using remote sensing and terrain attributes. In this study, information was generated for ecotourism planning in the *Serra dos Órgãos* National Park, RJ. According to the authors, the analysis of the heterogeneity of the landscape is fundamental for the planning of ecotourism because it allows to estimate the optimal relation between the conservation and the tourist alternatives. Landscape research for ecotourism integrates the different natural components (relief, climatic conditions, soil, vegetation cover, etc.) and evaluates their interrelations with the characteristics of the tourist destination. Soils are understood as a component that influences the landscape and is influenced by other components such as relief, climate, parent material, organisms (fauna and flora) and time.

In order to understand the effect of ecotourism and soil degradation on environmental preservation areas (EPA) in Brazil, Figueiredo et al. (2010) studied soil compaction as a pedogeomorphological indicator for erosion in trails of a conservation area, with a case study in the National Park of Serra do Cipó, MG. Sena et al. (2014) studied the degradation of soils along an attraction trail of the geotouristic monument of the Serra de São José, Tiradentes, MG. According to these authors, the highest rates of soil erosion were a direct function of the slope; and soil compaction rates were higher within the trail and less marked at the edges, corroborating results from Figueiredo et al. (2010). Oliveira et al. (2013) studied the soil quality of the trails of Cerrado State Park, Paraná, regarding the effect of people's traffic on soil physical attributes, to guide future actions to plan the use and occupation of the area in a way that would not affect the preservation of the ecosystem.

Barbosa et al. (2015) studied aspects of the environmental degradation of a recreational trail in Serra do Lenheiro, São João del-Rei, MG. According to the authors, inappropriate recreational use has contributed to a greater intensification of the degradation processes and factors such as slope and pedological characteristics of the trail are important for the analysis of the data and also to determine the local fragility.

At the international level, we highlight the work of Tomczyk (2011), who studied the Geographic Information System (GIS) evaluation and environmental sensitivity modelling of recreational trails in Gorce National Park, Poland. According to the author, the understanding and modelling of the factors related to the degradation of forest tracks and roads are crucial for park managers, and he proposes a methodology using basically two variables: vulnerability of plants community to trampling and vulnerability of soils to erosion processes, to assess the spatial distribution of areas with varying degrees of environmental sensitivity to the impact on the trail.

For the recreational planning in trails in protected areas of the Gorce National Park, Poland (Tomczyk and Ewertowski, 2013a) proposed the application of regression trees and geographic information system, using data of soils, geology, geomorphology, relief and information of the type and intensity of recreational use for modelling. A new method was proposed for detailed surveys of surface dynamics in small tracks under study. It involved the analysis of the spatial aspect of the microscale surface change, the quantification of soil loss and depositional processes, and the method used the topographic survey and different digital elevation models (Tomczyk and Ewertowski, 2013b).

In a study of soil loss on recreational trails at a US National Park Service, Olive and Marion (2009) found, through the use of the adjusted regression model, that the drainage, position, degree of slope and the alignment angle of the trail added to administrative factors were determinant in the soil loss. D'Antonio et al. (2013) assessed the impacts of visitors to the

parks and protected areas located within Rocky Mountain National Park, in the North Central region of the Colorado state, through a combined social and ecological approach. The authors concluded that social and biophysical data provide important information to park managers and an integrated approach increases the usefulness of the data.

Barros et al. (2013) indicate the lack of studies on ecological recreation in South America, especially those on informal trails. In a survey carried out in the largest protected area of the Southern Hemisphere, the Aconcagua Provincial Park located in the South Andean ecoregion (Argentina), the authors verified impacts on the area such as damage to vegetation and soil erosion resulting from the creation of informal trails formed by passers-by and animals associated with the lack of regulation and management in the trails. They pointed as a solution to the problem the limitation of the propagation of informal trails by the administrators.

The analysis of environmental vulnerability is linked to the susceptibility of the system to undergoing interventions, or of being changed. The destabilization of the system may have as inductors both natural processes and human actions (Sporl, 2007). From the point of view of environmental vulnerability analysis methodologies, there are two main approaches in Brazil. One defined by Ross (1994), where it is empirically described how to analyse the fragility of natural and anthropogenic environments. Another approach is by Crepani et al. (2001) that using the idea of basic territorial units describe the analysis of vulnerability applied to economic-ecological zoning and land use planning. The main difference between the two approaches is that the geological material is not directly considered in the methodology proposed by Ross. Variants or applications of these methodologies were made by Adami et al. (2012); Ross (2012); Manfré et al. (2013); Valle et al. (2016); Choudhary et al. (2017); Calderano Filho et al. (2018).

2.2 Machine Learning Techniques Applied to Digital Soil Mapping

According to some authors, the 1970s and 1990s were the most fruitful time for Brazilian soil surveys (Araújo Filho and Jacomine, 2014; Nolasco-Carvalho et al., 2013), where there were major investments in the sector, especially with the creation of the most ambitious project of Brazil until now the "Radam" (Radar of the Amazon) that later came to be called "RadamBrasil" (Araújo Filho and Jacomine, 2014). From this time to the present, several transformations occurred in the methods and mainly in the materials used to produce soil maps. In the XXI century, with the large expansion of computing (computing power), the ease to access data such as digital elevation models (DEM), satellite imagery and in some cases geology, geomorphology and climate data sped the work of soil surveying and made it more dynamic. For example, one of the challenges pointed out in 2004, the spatial prediction algorithms (Hengl and Heuvelink, 2004), today can be considered as an advance in the digital soil mapping (DSM) and it opened an infinity of possibilities.

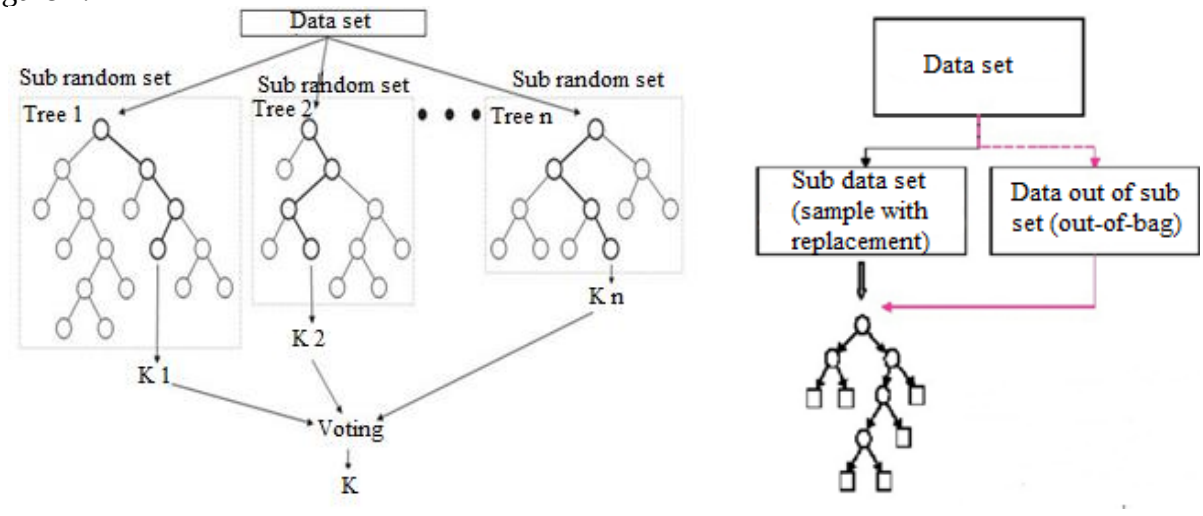
Algorithms such as Random Forest, Artificial Neural Networks, Support Vector Machines, Regression Trees, Decision Trees, Generalized Additives, Linear Regression among others are examples of machine learning methods, also called automatic learning. Automatic learning is a subfield of artificial intelligence dedicated to the development of techniques that allow the computer to learn, that is, extract rules and patterns from large data sets, and this is the reasoning it is known as inductive (Monard and Baranauskas, 2003).

For this study, we focused on the most popular and/or recent techniques for DSM as the regression, widely used in many different areas of science, nonlinear models such as the generalized additive model (GAM), which is relatively recent in the soil science and does not have so many works related to thematic. And one of the most recently developed and easy to use, the Random Forest (RF) has a great potential for learning and generalization (Meier et al., 2018; Jeune et al. 2018), and it can be used both for predicting soil attributes and classes.

In the RF case, it's learning and hierarchical form consist of a set of classification trees (for categorical variables) or random regression (continuous variables) (Breiman, 2001). In both cases, large numbers of trees are generated within the algorithm and then aggregated to obtain a single value or prediction class. For example, as a value, soil attributes can be predicted.

When the objective is to predict continuous variables (for example soil properties) the prediction is an average of the results of individual trees, while in the classification, categorical variables (for example soil classes), the predicted value is represented by the tree with the majority of the votes on the correct classification (Grimm et al., 2008). Under the training procedure, each RF tree algorithm is constructed based on a subset, different from the original data identified as bootstrap. This subset of predictor variables is randomly selected across all variables, and that variable that provides the best division, according to a function or goal, is used to divide each node. At the next node, the procedure is repeated; this procedure avoids overfitting (Cavazzi et al., 2013). The use of bootstrap-type sampling in RF modelling allows the out-of-bag (OOB) subassembly to be used to estimate general errors (Yang et al. 2016).

The RF depends only on three user-defined parameters: the number of trees in the forest, the minimum number of data points on each terminal node (nodesize), and the number of variables used to produce each tree (mtry). The values indicated by the literature are ntree = 500 nodesize = 5 and mtry = (one-third of the total number of predictors). However, other values can be tested with the smallest error, for which the RF provides error estimates using the so-called OOB data (which is a portion of the data not used in the bootstrap subset (Grimm et al., 2008; Mutanga et al., 2012; Rad et al., 2014; Taghizadeh-mehrjardi et al., 2015; Were et al., 2015; Yang et al., 2016). A scheme showing the RF formation procedure is presented in the Figure 1.



(a) (b)

Figure 1.(a) Random Forest general architecture. (b) Subset bootstrap and out-of-bag in Random Forest. Source: Adapted from Nguyen et al. (2013)

The linear regression is used in several areas of knowledge and it consists of modelling linear relations between a target variable and predictor variables (Hastie et al., 2009). The general equation of multiple linear regression is given by Equation 1:

$$Y_i = \beta_0, \beta_1 X_1, \beta_2 X_2, \beta_3 X_3 \ldots \beta_i X_i + \varepsilon \qquad \text{Eq. 01}$$

where Y_i is the value of the dependent variable (target) in the *i-th* observation, β_0, β_1, β_2, β_3 ... β_i are parameters, X_1, X_2, X_3 ... X_i are known constants, that is, the value of the predictor

variables in the i-th observation, ε is the random error with mean E {ε} = 0 and variance σ^2 {εi} = σ^2, and i = 1, ..., n (Hastie et al., 2009).

In soil science, it is used for spatial prediction of several soil properties (Chagas et al., 2016; Hengl et al., 2015; Pinheiro et al., 2018; Samuel-Rosa et al., 2013; Silva et al., 2017) as well as for the generation of functions called pedotransfer functions used to estimate attributes that are difficult to obtain from data that are easier to obtain (McBratney et al., 2002; Medeiros et al., 2014; Rodríguez-Lado et al., 2015; Souza et al., 2016). In addition, the regression models are relatively simple, and their interpretation is facilitated because it is possible to construct prediction equations. It is still possible to evaluate the contribution of each variable to adjustment of the regression model, as well as to select (for example Stepwise method) the variables most relevant to the model. It does not require large computational capacity comparing to the computational requirement of non-parametric models, usually more complex.

Although linear models are widely used to describe natural phenomena, in some cases these models do not work well simply because in real life the effects between covariates and target variables are generally nonlinear (Wood, 2006; Hastie et al., 2009). To capture non-linear relationships between covariates and a given phenomenon, GAM was developed, which is an extension of the generalized linear model (Hastie et al., 2009). The general equation of the GAM is given by Equation 2:

$$E(Y|X_1, X_2, ..., X_i) = \alpha + f_1(X_1) + f_2(X_2) + (X_i) \quad \text{Eq. 2.}$$

As usual X_1, X_2, And X_i represents the predictors (covariables) and Y the output (target variable), α is a parameter to be estimated, *f* is unspecified (non-parametric) smoothing functions (Hastie et al., 2009).

To model nonlinear relationships, GAM uses a smoothing function (*f*). There are different smoothing functions, but the most used are based on the spline function and its variants (Wood, 2006). GAM models, although more flexible than linear models, have maintained their interpretability, which differentiates them from, for example, methods such as Artificial Neural Networks, Support Vector Machine, Random Forest, and other machine learning methods (Hastie et al., 2009). In addition to modelling non-linear relationships, additive models may have limitations for variable selection in a large number of covariates and small sample size, which are mainly related to the number of degrees of freedom in the model (Poggio et al., 2013). The algorithm adjusts all covariates that are not feasible or desirable when a large number is available (Hastie et al., 2009). In this sense, it is worth using selection strategies, which can vary in the gradual selection (forward or backward) using stepwise selection (Poggio et al., 2013, de Brogniez et al., 2015, Chartin et al., 2017), or Recursive Feature Elimination (RFE) (Jeong et al., 2017) or other possible approaches (Marra et al., 2011). In the field of soil science, the use of GAM models is relatively recent, and despite the great potential, it is a still underused technique. In relation to the use of GAM to predict soil attributes, we highlight the studies of Poggio et al. (2010, 2013), Poggio and Gimona (2014, 2017a, 2017b), by Brogniez et al. (2015) and Jeong et al. (2017).

2.3 Models Based on Bayesian Inference

Bayesian inference is based on the *Bayes' theorem* that involves a prior (or a priori) distribution of a variable which may be based on observed data, some theoretical reason or on the investigator's judgment about the likely behavior of the variable and is fundamental for methods related to data mining and more currently digital soil and environmental mapping (Xiong et al., 2015; Poggio et al., 2016; Huang et al., 2017).Although there are a multitude of models, including some combinations with classic statistical models for this study, we will consider the method of Bayesian Belief Network (BBN) that is a multivariate statistical model

for a set of variables (nodes) (Aguilera et al., 2011) and can be used in various disciplines such as social, economic and environmental aspects. Due to the powerful theory of probability involved, BBNs are able to deal with a wide range of problems (Aguilera et al., 2011). The probability distribution of a node X is determined by the realized states of its preceding or parent nodes, using the conditional probability P(X|parents(X)) described in Bayes' theorem Equation 3.

$$P(X|parents(X)) = P(parents(X)|X) * P(X) / (P(parents(X)) \quad \text{Eq.3}$$

BBNs are also known as probabilistic graphical models that represent variables and their dependencies by specifying probabilistic relationships and when applied to spatial data are called spatial BBN (Gonzalez-Redin et al., 2016). The BBN approach can capture and structure available knowledge and rationalize complex interactions where empirical data are limited or poorly compatible and processes are complex or uncertain (Aalders et al., 2011). The model consists of nodes, spatial or not, a set of links representing the relationship between nodes and a set of conditional probability tables (CPT), the strength of the probabilistic relationships between the different variables and their states is defined by CPT.

The modelling of the BBN is a useful tool to integrate a participatory process with qualitative and quantitative information and spatial data (Celio et al., 2014, Meyer et al., 2014, Landuyt et al., 2015, Bashari et al., 2016; Gonzalez-Redin et al., 2016). In addition, BBN's ability to take into account the uncertainties and their propagation makes BBN a very useful tool (Marcot et al., 2012, Landuyt et al., 2015, Huang et al., 2017).

3 CHAPTER I:

THE ITATIAIA NATIONAL PARK SOILS AND PHYSIOGRAPHY DATABASE

3.1 RESUMO

Ambientes como a "Serra da Mantiqueira" onde se encontra o Parque Nacional de Itatiaia (PNI) apresentam elevada complexidade de solos, pois possuem grande variação geológica, geomorfológica, relevo, clima e vegetação. Como nesses ambientes as relações nos fatores envolvidos com a gênese do solo são ainda mais complexas é relevante a sua caracterização detalhada. O desafio, porém, é que a informação disponível é reduzida e dispersa em materiais de acesso restrito, aumentando a dificuldade de organizar essas informações em uma base de dados. Além disso o PNI possui acessibilidade muito limitada, em especial na sua parte alta, o que dificulta por exemplo o levantamento de solos. Na tentativa de suprimir a falta de informação de solos e pela necessidade da criação de uma base de dados em ambiente SIG que possa dar suporte a pesquisa, a programas de educação ambiental e ao plano de manejo do parque foi feita uma caracterização detalhada do ambiente usando as modernas ferramentas pedométricas como suporte. O levantamento e a caracterização do solo, apoiados pelas técnicas quantitativas de análise de dados do solo, juntamente com a criação de uma base de dados em ambiente SIG fornecem informações para pesquisas futuras, bem como para apoiar a tomada de decisão. A caracterização mostrou que os solos do PNI são predominantemente rasos, com altos teores de C, N e H + AL, elevada CEC, baixos valores de pH e de densidade do solo, e possuem alta capacidade de armazenar água. Esses solos são altamente vulneráveis à degradação, especialmente por erosão, compactação e deslizamento de encostas. Alguns solos identificados no PNI não foram antes relatados na literatura. Assim, pode-se afirmar que as informações produzidas neste estudo são potencialmente úteis para diversas pesquisas multidisciplinares e, em particular, para melhorar o plano de gestão do PNI, bem como para a avaliação da vulnerabilidade ambiental da parte alta do parque e orientar normas de acesso do público.

Palavras-chave: Banco de dados de solos. Levantamento de solos. Plano de manejo.

3.2 ABSTRACT

Areas such as the "Serra da Mantiqueira", where the National Park of Itatiaia (INP) is located, present high soil complexity, as they have great geological, geomorphological, relief, climate and vegetation variation. In these environments, the relationships of the factors involved with soil genesis are even more complex, and its detailed characterization is highly relevant. The challenge, however, is that the available information is reduced and dispersed in restricted access materials, increasing the difficulty of organizing this information in a database. In addition, the INP has very limited accessibility, especially in the upper part, which makes it difficult, for example, to survey the soils. In an attempt to overcome the lack of soil information and the need to create a database in a GIS environment that can support research, environmental education programs and the park management plan, a detailed characterization of the environment was made using modern pedometrics tools. Soil survey and characterization, supported by quantitative soil data analysis techniques, together with the creation of a GIS database provide information for future research as well as to support decision making. The characterization showed that INP soils are predominantly shallow, with high levels of C, N and H + Al, high CEC, low pH and soil density, and high-water storage capacity. And these soils are highly vulnerable to degradation, especially by erosion, compaction and land sliding. Some soils identified in the INP were not previously reported in the literature. Thus, it can be stated that the information produced in this study is potentially useful for several multidisciplinary kinds of research and, in particular, to improve the INP Management Plan, as well as to assess the environmental vulnerability of the upper part of the park and guide public access rules.

Keywords: Soil database. Soil survey. Management plan.

3.3 INTRODUCTION

The Itatiaia National Park (INP) is the first Brazilian national park. The area was chosen to be preserved because of its broad geological, geomorphological, hydrological and vegetation variation, making it an area rich in diversity, and a great potential for tourism due to its singular and attractive landscapes (Barreto et al., 2013). Several studies have been carried out in the INP since its creation, with emphasis on floristic composition and description of endemic species on the INP ecosystems. Among the pioneers are the bulletins 05 and 08, from 1965, of the Ministry of Agriculture Forest Service, which are a major contribution to the knowledge on the flora of INP (Brade, 1956). More recently, many other studies of the flora of the park were developed (Lima and Guedes-Bruni, 2004; Morim and Barroso 2007; Barberena et al, 2008; Silva Neto and Peixoto, 2012; and Mezabarba et al., 2013). As for the characterization of the geological, geomorphological and soil resources, few studies were done in the park. Among the most relevant, the geological and geomorphological mapping of the Brazilian Foundation for Sustainable Development, in the 1: 50,000 scale (Santos et al., 2000).

Of the 29 INP Bulletins published (until the time of the search) since 1949 (the year the first one was published), none brings detailed information about the soils of the park. The vast majority of articles and news refer to research on fauna and flora, as well as the public usage of INP. In this sense, soil characterization and its distribution in the space are important to the implementation of the INP Management Plan and to subside other studies in the park. As an example, studies on forest resilience, ecology, climate changes, nutrients cycling, evaluation of ecotourism impact or risks of degradation. Also, to contribute to the definition of areas that have to be of limit to the public, among others.

The soil studies for the INP Management Plan (Almeida et al., 2011) used information from Rio de Janeiro and Minas Gerais maps (Carvalho Junior et al., 2000; Fernandes Filho et al., 2010), and the identification of soil classes is based on outdated soil classification. Both, the INP Management Plan (Barreto et al., 2013) and the Bulletin 18 of the Ministry of the Environment (Aximoff et al., 2014) bring generalized information about the soils, in a scale of 1: 500,000, which does not allow adequate subsidy to INP's planning regarding the vulnerability of soils and environments. Rodrigues (2011), which mapped soils in a small part of the park's area, and Soares et al. (2016) that studied the genesis of *Organossolos* in the upper part of the INP, show soil classes that were not reported in the previous studies. Thus, a more detailed soil survey is essential to guide environmental researches in the INP, and for the park management plan, including to establish fire prevention plans and trail conservation.

On the evaluation of impacts of public activities on the park's trails, the following studies stand out: Magro and Barros (2004); Barros and Magro (2007), Iwamoto and Rodrigues (2011), Richter e Souza (2013), Freire and Lemos (2014). Also, relevant works on burnings and fires in the INP (Aximoff and Rodrigues, 2011, Tomzhinski et al., 2011, Tomzhinski et al., 2012 and Sousa et al., 2015).

The INP has limited access, due to steep slopes, dense forest cover in the forested areas or by rocky outcrops in the altitude fields of the plateau region (Barreto et al., 2013). The INP landscape makes it an excellent case study for digital soil mapping (DSM), in order to produce a viable result at a lower cost than conventional methods. Thus, the usage of DSM tools, ranging from optimization of the sampling site (Minasny and McBratney, 2006; Roudier et al., 2012; Stumpf et al., 2016) to the covariate selection in powerful predictive algorithms (Beguin et al., 2017; Chagas et al., 2017; Jeong et al., 2017) was proposed in this work.

The general objective of this chapter was to compile and organize the soils database and physiographic variables to subsidize the subsequent chapters. As specific objectives:

a) To develop a database in a GIS environment with information on the soils, satellite images, digital elevation model, terrain attributes, land use map and coverage, geology and geomorphology.

b) To produce pedological and environmental covariate information to support this thesis (subsequent chapters), interdisciplinary research actions, environmental education programs and the INP management plan.

3.4 MATERIAL AND METHODS

3.4.1 Study area characterization and covariates description

3.4.1.1 Location

The INP has an area of 225.54 km^2 and it is located in the *Serra da Mantiqueira*, at the boundary of Minas Gerais (MG) and Rio de Janeiro (RJ) States (Barreto et al., 2013). According to Tomzhinski et al. (2012), the INP can be divided into three broad areas: the "lower part", which comprises the southern part of the park, the "upper part" of the plateau (**Erro! Fonte de referência não encontrada.**) and Visconde de Mauá in the east side. The upper part, our study case, has about 164.01 km^2.

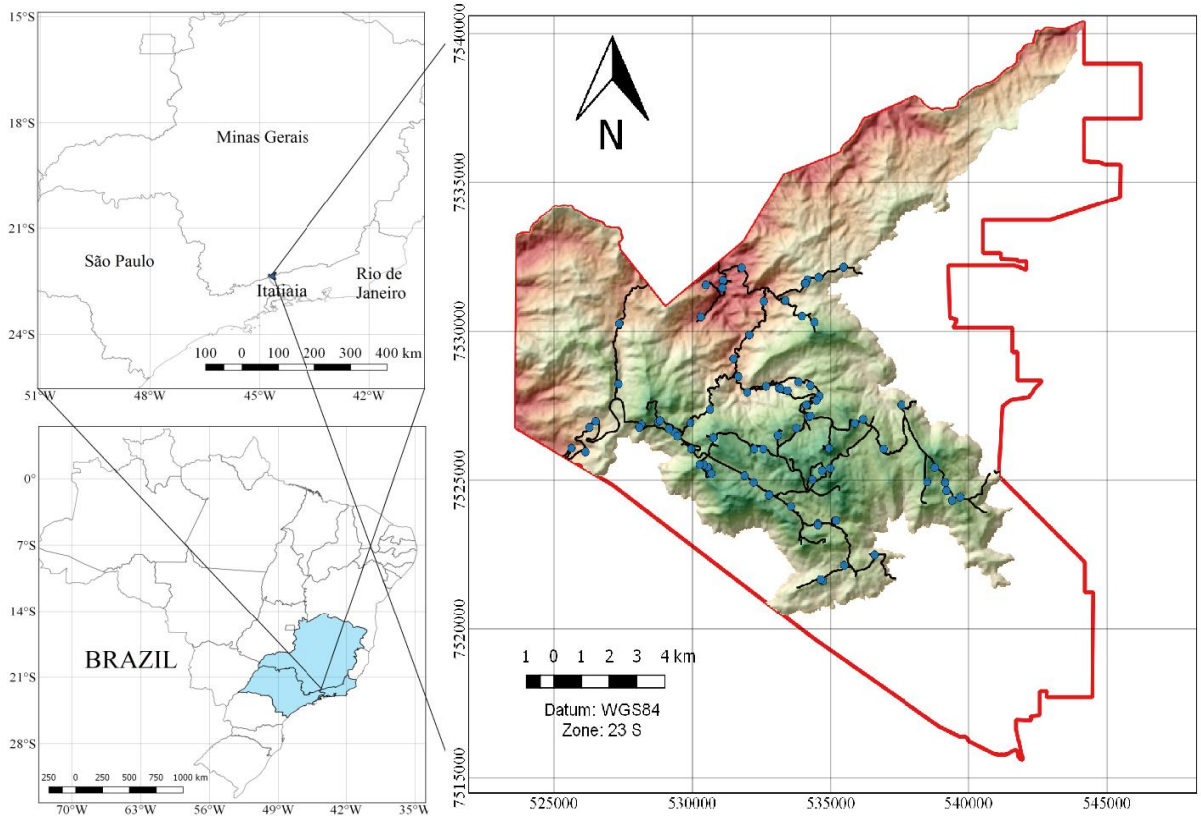


Figure 2. The red polygon marks the total area of the INP in the south-eastern region of Brazil and the area in relief corresponds to the upper part of the park. The major roads and trails are in black. Blue points are soil sampling points selected according to cLHS method (Minasny and McBratney, 2006).

The division into two regions is used as reference by park managers, where the "lower part" covers the areas of Posto 1, Visitor Centre, Mirante do Último Adeus, Serrinha, Três Picos, Abrigo Macieiras and Maromba; while the "upper part" includes the areas of plateau, Posto 3, Abrigo Rebouças, Morro do Couto, Pedra do Camelo, Pedra Cabeça de Leão, Picos das Agulhas Negras e Abrigo Massena (Barreto et al., 2013).

3.4.1.2 Climate

The climate of the INP varies according to the elevation, that influences mainly the temperature since the higher the elevation the lower the temperature (Figure 3). Besides that, the mountain range conditions specific orographic processes that will influence the amount and

intensity of rainfall. In the INP management plan, Barreto et al. (2013), the climate is described as mesothermic with a mild summer and a rainy season in the summer (Cwb), and mesothermic with a mild summer without a dry season (Cfb). The Cwb type occurs mainly in the upper part of the park, usually above 1,600 m altitude; while the Cfb climate characterizes the lower part (Alvares et al., 2013; Barreto et al., 2013).

However, the predominant climate in the park is the humid subtropical zone, where the averages vary from 15 to 18 ºC. In the area of the park located in Minas Gerais State, the climate is denominated mesothermic with average temperatures below 10 ºC, with three months of drought per year. In the Alto dos Brejos and Serra Negra regions there is a mild-mesothermic climatic influence range, where the average temperatures vary between 10 ºC and 15 ºC and there is practically no dry season (Barreto et al., 2013). There is no map of climatic variables spatial distribution, rainfall and temperature in high spatial resolution for INP. However, a 30-year average data set, from 1970 to 2000, is available worldwide (Fick and Hijmans, 2017; Poggio et al., 2018) with 1 km spatial resolution and those were interpolated to 25 m using bilinear interpolation. As pointed above, the upper part of INP has the lowest temperature and higher rainfall (Figure 3).

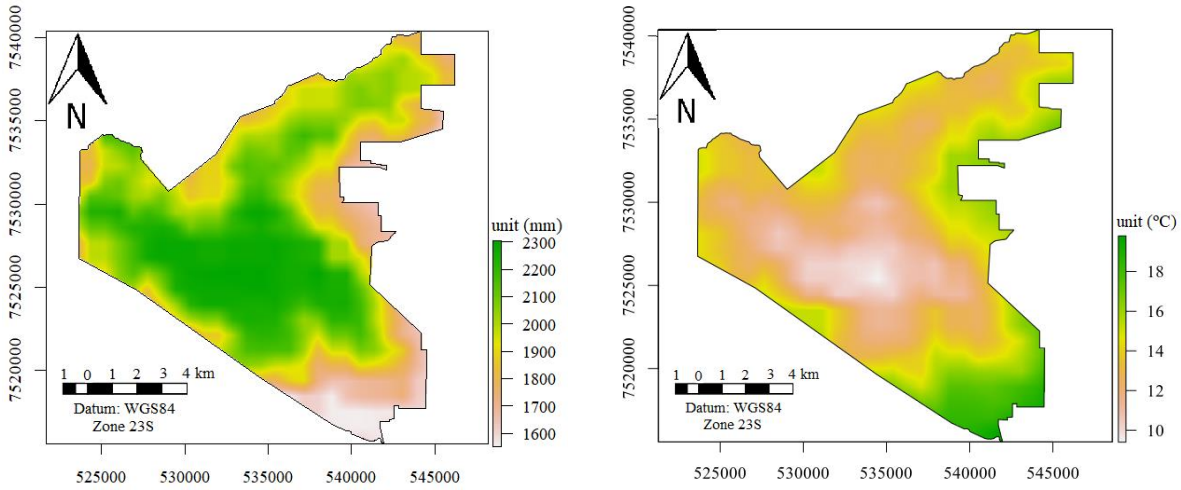


Figure 3. Average annual rainfall value (left) and mean temperature (right) over a period of 30 years. Adapted from Fick and Hijmans (2017).

3.4.1.3 Geology and geomorphology

Information of geology and geomorphology was obtained through the survey realized by the Brazilian Foundation for the Sustainable Development, authored by Santos et al. (2000) in a 1: 50,000 scale. Subsequently, Barreto et al. (2013) added specific information in the INP Management Plan. According to these sources, the following types of parent materials occur in the INP: magmatic breccia, homogeneous gneisses, alkaline granite, nepheline syenite, quartz syenites, alluvium sediments and colluvial sediments (**Erro! Fonte de referência não encontrada.**).

Those authors report that homogeneous gneiss rock is predominant in the northern part of the park and are mainly composed of orthoclase, plagioclase, quartz, biotite and hornblende minerals. The predominant rocks in the park are composed by nepheline, syenites and foyaite that are predominantly constituted by micropertite, albite, nepheline and sodalite minerals. The quartz syenite occurs mainly in the central portion of the INP, having as main constituent minerals the micropertite and the quartz. Alkaline granite consists essentially of micropertite and quartz and can be found in the vicinity of the "*Abrigo Rebouças*" and the "*Prateleiras*". The magmatic breccia, also found in the central region of the park, has a feldspathic nature and

also contains minerals such as chlorite, pyrite, magnetite, calcite, sericite, apatite and biotite. Colluvium sediments are predominantly composed of blocks and boulders of alkaline rocks. The oldest depositions form hills with many boulders on their slopes. Alluvial sediments correspond to the fluvial plains that are filled by sandy and clayey sediments rich in organic matter, which corresponds to flooded areas that often are constituted by peatlands (Santos et al., 2000).

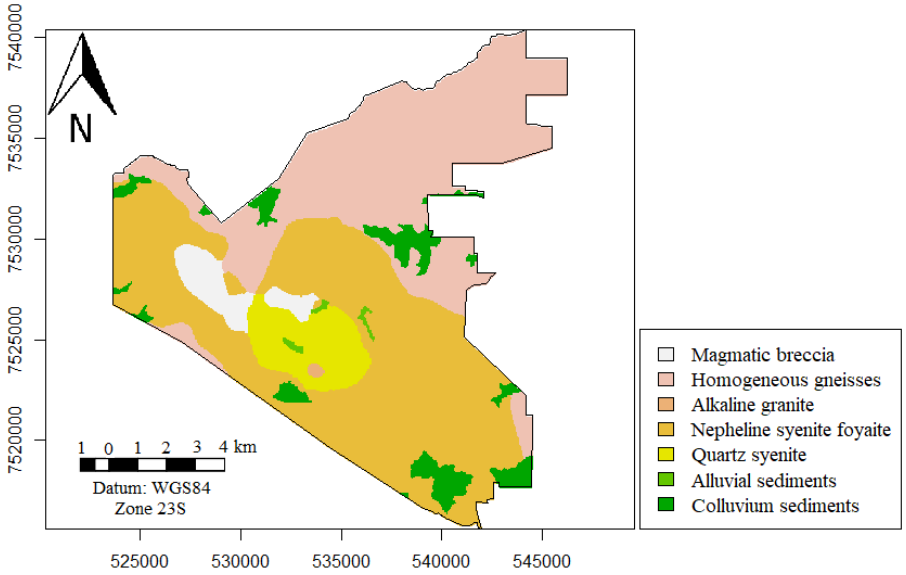


Figure 4. Geological map of the Itatiaia National Park. Adapted from Santos et al. (2000).

As for the geomorphology, the main features are: fluvial plains with a slope of less than 2% and average elevation of 2,300 to 2,400 m above mean sea level; talus deposits with slopes from 10 to 35 %; mountains with elevation from 900 to 2300 m and slopes higher than 45%; rocky outcrops and escarpments with 2200 to 2700 m and slopes higher than 50 % (Figure 5) (Santos et al., 2000).

The fluvial plains mapped in the INP (1: 50,000 scale) correspond to the flatlands along the Campo Belo, Aiuruoca and Preto rivers. They are depositional zones or stream terrace systems that seasonally become flooded and allow the accumulation of organic matter. The talus bodies are depositional ramps associated with the bottoms of valleys and the foothills of steep slopes. The mountain is the dominant geomorphological form in the park, a degradation system of relief that has a high slope and more developed soils than in the rocky outcrops and escarpments. Erosional processes of high intensity occur in this form, with ravines, gullies and mass movements (Santos et al., 2000).

The escarpments are characterized by rocky outcrops and a rugged and high rocky massif occurring in the central and highest part of the park, comprising the *Serras das Prateleiras, Negra, Lambari* and *Itatiaia*, where the *Agulhas Negras* peak is located. Due to the presence of exposed rocks, fewer sediments are produced and soils are dominantly shallow, and large boulders are often observed (Santos et al., 2000).

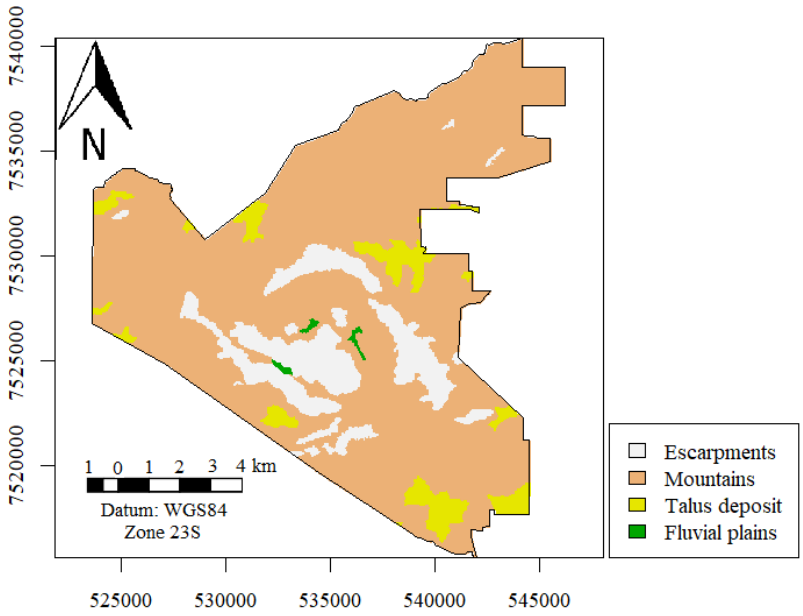


Figure 5. Geomorphological map of the Itatiaia National Park. Adapted from Santos et al. (2000).

In order to use the geology and geomorphology maps, which were in PDF format, they were georeferenced using the INP limits provided by the park administration as a reference base. After the georeferencing the maps were vectorized and transformed into a raster file in .tif format. The maps were generated in the same spatial resolution of the DEM (25 m), since these were initially produced in the same cartographic scale, which is the IBGE database used to generate DEM (1: 50,000). All the procedures for geology and geomorphology maps were made using ArcGis 10.2.2 software (ESRI, 2015).

3.4.1.4 Soils

The previous information about distribution of soils of the INP results from a generalized map where the soil mapping units are dominantly associations (Barreto et al., 2013, Aximoff et al., 2014). As a result, the current INP management plan reports only four soil classes, with the *Cambissolos Humicos* (Cambisols) predominant and occurring in the slope areas of the mountains geomorphological unit. In the steeper slopes and elevated areas, *Neossolos Regolíticos* (Regosols) alternate with rock outcrops, and are associated with the geomorphological units described as a mountain and rocky outcrops and escarpments. *Argissolos* (Lixisols or Acrisols) and *Cambissolos* (Cambisols) can be found in the mountain geomorphological unit. In the lower part of the park, due to conditions more favourable to weathering and pedogenetic processes, *Latossolos* (Ferralsols) can be found.

Soil data of the park management plan and in the Bulletin Number 18 of the Ministry of the Environment (MMA) were compiled from the 1: 500,000 soil survey of the state of Rio de Janeiro (Carvalho Filho, 2000), and the Map of Soils of the State of Minas Gerais with the expanded legend and scale of 1: 650,000 (Fernandes Filho et al. 2010).

Other soil classes, such as that of the *Organossolos* (Soares et al., 2016), were found that were not identified in the soil survey used for the management plan of the park. In Minas Gerais, Rodrigues (2011) described 4 profiles that were classified in 3 soil orders according to the Brazilian Soil Classification System - SiBCS (Santos et al., 2018): *Cambissolo Húmico Distrófico típico* (Umbrisols), *Cambissolo Húmico Distrófico latossólico* (Umbrisols),

Organossolo Háplico Hêmico típico (Histosols), and *Neossolo Litólico Húmico típico* (Leptosols).

The level of soil information, as a result of the common soil survey methodologies employed at the time the INP management plan was developed, was one of the reasons why the DSM was essential to this work.

3.4.1.5 Digital elevation model and terrain attributes

The Digital Elevation Model (DEM), with a spatial resolution of 25 m, was generated from the contour lines with 20 m equidistance and hydrography extracted from the plani-altimetric charts, both in the 1:50,000 scale used (Figure 6). The sheets used were SF-23-ZA-I-2 "*Alagoa*", SF-23-ZA-I-3 "*Passa Quatro*" and SF-23-ZA-I-4 "*Agulhas Negras*". They were obtained in vector format from the cartographic base of Brazilian Institute of Geography and Statistics (IBGE). The dataset was provided by the INP administration.

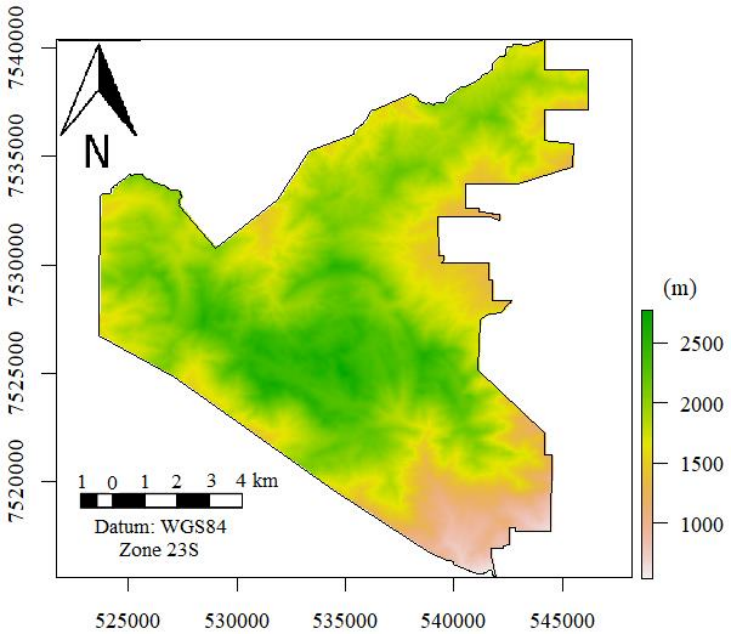


Figure 6. Digital elevation model of the Itatiaia National Park.

Terrain attributes that allegedly have high relation with soils genesis and their properties and/or commonly used in the literature in DSM projects were extracted from the DEM, using R software (R Core Team, 2018) and *RSAGA* package (Brenning et al., 2008). They were:

a) *Elevation* (Acronym: DEM, unit: m): represents the elevation relative to the reference plane, the sea;
b) *Slope* (Acronym: Slope, unit: %): which affects the velocity of surface and subsurface flows, soil loss and soil erosion;
c) *Aspect* (Acronym: Aspect, unit: degrees): attribute representing the exposure faces, represented by values in degrees ranging from 0 to 360 °;
d) *Northernness* (Acronym: Northernness, unit: degrees): indicates the direction of the slope relative to the northern. Northernness =abs(180°−Aspect);
e) *Plan curvature* (Acronym: Plan_curv, unit: m^{-1}): the shape of the hillside on the horizontal plane (concave, rectilinear or convex);
f) *Profile curvature* (Acronym: Prof_curv, unit: m^{-1}): the shape of the hillside on the vertical plane (concave, rectilinear or convex);

g) Convergence index (Acronym: Convergence, unit: %): the general shape of the hillside in all directions (concave, rectilinear or convex);
h) *Catchment area* (Acronym: Cat_area, unit: m^{-2}): it is related to the volume of flooding that reaches a certain cell;
i) *Topographic wetness index* (Acronym: TWI, unit: dimensionless): describes a tendency for a cell to accumulate water;
j) *LS factor* (Acronym: LS_factor, unit: dimensionless): attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE);
k) *Relative slope position* (Acronym: RSP, unit: dimensionless): represents the relative slope position based on the base channel network;
l) Channel network distance (Acronym: CHND, unit: m): altitude above the channel network (CHNB- original elevation);
m) Channel network base level (Acronym: CHNB, unit: m): interpolation of a channel network base level elevation

3.4.1.6 Remote sensing images and indexes

Two scenes from the RapidEye sensor (2011) were used. They have 12-bit radiometric resolution, 6.5m spatial resolution, and were orthorectified to 5m spatial resolution (RapidEye, 2012). To reconcile the spatial resolution of the image with that of the DEM, the image was interpolated to a resolution of 25 m using the value of the neighbouring 5 pixels to calculate a mean value. The images were atmospherically corrected using the 6S (Second Simulation of Satellite Signal in the Solar Spectrum) model (Vermote et al., 1997) to convert radiance at the satellite level into a physical variable, surface reflectance and remove the atmosphere effect (Antunes et al., 2014). Processing details are in Costa et al. (2016) and the image´s characteristics are listed in **Erro! Fonte de referência não encontrada.**

Table 1. General technical characteristics of RapidEye satellite.

Characteristics	**Information**	
Number of Satellites	5	
Orbit Altitude	630 km in Sun-synchronous orbit	
Equator Crossing Time	11:00 am local time	
Sensor Type	Multi-spectral push-broom imager	
Spectral Bands Ground	Band Blue (1) Green (2) Red (3) Red-Edge (4) NIR (5)	Spectrum band (nm) 440 – 510 520 – 590 630 – 685 690 – 730 760 – 850
Ground sampling distance (nadir)	6.5 m	
Pixel size (orthorectified)	5.0 m	
Swath Width	77 km	
Satellite life expectancy	7 years	
Revisit time	Daily (off-nadir) / 5.5 days (at nadir)	
Horizontal Datum	WGS84	
Camera dynamic range	12 bits	

Font: (RapidEye, 2012)

For image atmospheric correction were used the following input parameters (Table 2).

Table 2. Input parameters used in 6S Model

Parameters	Scene 1	Scene 2
Image date	02/07/2011	16/08/2011
UTC time decimal	11.120833	11.143889
Gas model	Tropical	Tropical
Aerosol Model	Continental	Continental
Longitude of the centre of the scene	-44.65035º	-44.64981º
Latitude of the centre of the scene	-22.35798º	-22.57479º
Average altitude	1470 m	1470 m
Sun-Earth Distance	1,01668	1,01263
Solar Zenital Angle	42,12º	38,84º
Visibility	25 km	25 km
Total number of rows and columns	25000000 (5000 R x5000 C)	25000000 (5000 R x5000 C)

After the atmospheric correction there were obtained the normalized difference vegetation index (NDVI) and the soil adjusted vegetation index (SAVI) by using arithmetic operations (equations 1 and 2, respectively) in the *raster* package (Hijmans, 2016) of R software (R Core Team, 2018).

$$\text{NDVI} = \frac{\rho\text{nir} - \rho\text{red}}{\rho\text{nir} + \rho red} \quad \text{Eq.4}$$

$$SAVI = \frac{(1 + L)(\rho\text{nir} - \rho\text{red})}{\rho\text{nir} + \rho red + L} \quad \text{Eq.5}$$

Where ρnir is the radiant flux reflected in the near infrared, represented by the band 5 of the RapidEye sensor, ρred is the radiant flux reflected in the red, represented by the band 3. The constant L can present values from 0 to 1, varying according to the own biomass; the reference values of L are (Huete 1988):

L = 1 (for low vegetation densities)
L = 0.5 (for medium vegetation densities)
L = 0.25 (for high vegetation densities)

3.4.2 Soil sampling selection: important aspects for the Itatiaia National Park

One of the most important steps for digital soil mapping (DSM) is the selection of sampling points (Carvalho Júnior et al., 2014). For conventional soil mapping, the prospecting method and sampling frequency will depend on the survey level of detail (survey objective) (IBGE, 2015). For example, in the third edition of the Pedology Technical Manual (IBGE, 2015) the method indicated for an exploratory survey is the extrapolation, generalization, and correlation, with a few if any number of field observations; plus, the sampling frequency is of a complete profile per class of predominant soil in the association. For a detailed survey, the prospecting method includes planned transects with field checks along toposequences or free walking and relating to geomorphic surfaces; where the sampling frequency is of a complete profile and two complementary ones per soil class at the lowest taxonomic levels (usually subgroup and phases).

However, for the DSM this sampling scheme is not applied, since it does not allow for a representative statistical estimation. In DSM it is necessary to use statistically robust sampling

strategies to reduce subjectivity and consequently to make possible to calculate prediction errors (Minasny and McBratney, 2007). In this sense some studies were developed about sampling optimization for the DSM (Minasny and McBratney, 2006; Roudier et al., 2012, Cambule et al., 2013; Carvalho Júnior et al. 2014, Clifford et al., 2014, Ließ, 2015; Brus et al., 2015).

In this study, the technique known as conditioning Latin hypercube (cLHS) was chosen. This method was proposed by Minasny and McBratney (2006), and it is built on sampling based on the use of ancillary data. It is considered a robust tool for allocation of sampling points by using a set of auxiliary covariates, which can be categorical or continuous, such as terrain attributes derived from the DEM, geological and geomorphological data, satellite images and their derived indices. It has as assumption that the auxiliary data (covariables) must be able to capture all the variation of geo-environmental characteristics (Minasny and McBratney, 2006).

Works by Minasny and McBratney (2006); Minasny and McBratney (2007); Roudier et al. (2012); Carvalho Júnior et al. (2014) show how cLHS can be efficient to allocate points in DSM studies. However, a limitation observed by Roudier et al. (2012) and confirmed by Kidd et al. (2015) is that cLHS can present points allocation in inaccessible areas, considering only the distribution of available covariables, thus presenting operational limitation due to access restrictions. Trying to solve this problem, Roudier et al. (2012) proposed a methodology considering the cost of access. In turn, Carvalho Júnior et al. (2014) proposed a methodology using as a rule of spatial restriction a buffer (in the case of this study, 100 m wide) along the main roads and access roads, besides the exclusion of urban areas, conservation units and bodies of water; thus, configuring the effective area of study.

Still, in this sense, Cambule et al. (2013) proposed a sampling methodology for DSM in areas that are not accessible, as it happens in the INP. In this methodology, the authors proposed the sampling in an area of greater accessibility that is representative of the total area (accessible and not accessible) for the construction, application and validation of a predictive model for the accessible area, and later the application and validation of the model in the area that is not accessible.

The approach used in this work follows the principles proposed by Minasny and McBratney (2006), with the selection of points by the cLHS method using auxiliary variables; however, considering the access costs (Roudier et al., 2012). As a rule of constraint, 3 buffers sizes were created in relation to roads and trails, as proposed by Carvalho Júnior et al. (2014), being tested the distances of 100, 200 and 400 m. These values were chosen to reconcile what is feasible to be sampled and to obtain a good representation of the environmental characteristics of the total area (upper part of INP), considering it has predominantly areas of difficult access. Also, the trails pass through great outcrops of rock and can reach more than 15 km of extension as it is the case of the crossing "*Rancho Caído*" and the trail that leads to "*Pedra Cabeça de Leão*" going from "Rebouça" shelter. In this particular case this was the furthest point from which it is only possible to reach on foot and lasted two days to reach and required to camp along the trail, since it was not possible to go to the sampling point and return on the same day. Considering that a point could be allocated to any region occupying an area of 1913.37 ha (buffer 100 m) it would be even more difficult to reach the farthest point allocated in any region of the 200 and 400 m buffers, that have areas respectively of 3468.56 and 5877.81 ha.

Throughout the procedure of selection of the points, the software R (R Core Team, 2018) and the *clhs* package (Roudier et al., 2012) were used. The auxiliary variables used were: geology, elevation, slope, northernness, and soil-adjusted vegetation index. For the allocation of the sample points, a collection with 80 points was chosen, which includes complete profiles and extra soil samples. The number of iterations used in the algorithm was equal to 5000.

The auxiliary variables (environmental covariates) used were chosen to represent to the maximum the factors involved in the soil genesis with the minimum as possible covariables

(model parsimony). Thus, based on the pedological knowledge and from the study area the variables selected were: geology, which is related to the factor source material; elevation, which is related to the relief factor and directly influences the climatic factor in the INP; slope, which is related to the relief factor and soil removal rates; northernness, which is related to the relief factor and influences the amount of light that the exposure face receives which influences the soil moisture and organic matter decomposition rates; soil- adjusted vegetation index, that represents the organism factor explained by the soil cover/land use

The representativity analysis of the effective area tested (buffer 100, 200 and 400 m) was done qualitatively and quantitatively. The qualitative form followed the evaluation of the histograms of the continuous (auxiliary) environmental variables, as proposed by Cambule et al. (2013). The quantitative evaluation consisted of descriptive statistics of the total area and buffers. Data analysis was performed using the Lilliefors test, after that the buffers data set and the total area were tested.

3.4.3 Soil's description, analysis, classification and mapping

In the horizon morphological description, the following attributes were evaluated: thickness, colour, mottling (if present), texture, structure, consistency and transition between horizons; as well as the general description of the landscape and the profile of the conditions of the source material, relief and slope, stoniness, drainage, among other characteristics, according to the Manual of Description and Soil Collection in the Field (Santos et al., 2015).

The soil samples were passed through a 2.00 mm mesh screen, obtaining the sample for laboratory analyses. In this material, pH (H_2O), calcium (Ca), magnesium (Mg), potassium (K), sodium (Na), phosphorus (P), potential acidity (H + Al) were analysed, and base sum (S value) (Ca + Mg + K + Na), cation exchange capacity (CEC) or T value and base saturation (V%) (S/T) were calculated according to the methodology in Donagemma et al. (2011). Soil bulk density (BD), sample taken with the aid of cutting ring, soil moisture (SM) at the time of sampling and soil particle size were also determined. The percentage of fine gravel and coarse gravel were also quantified in relation to the fine material as recommended by the Santos et al (2015). Since many of the selected profiles are organic types, specific analyses were made as recommended by the Brazilian Soil Classification System – SiBCS (Santos et al., 2018).

Based on the morphological, chemical and physical data of the horizons of each profile, the soils were classified based on SiBCS (Santos et al., 2018) and made the equivalence to the classes (Table 3) in the World Reference Base for Soil Resources - WRB (IUSS Working Group WRB, 2015).

To map the soil spatial variation a Random Forest (RF) model was calibrated using all covariates described, from climate, geology and geomorphology, terrain attributes and remote sensing data, for covariate selection the Recursive Feature Elimination was used, this is a robust method to select covariates in RF (Jeong et al., 2017; Jeune et al., 2018; Meier et al., 2018). The algorithm performs a backward selection. When the full model is created, a measure of variable importance is computed and shows the ranks of predictors from most to least important ones and those more important to the model are selected (Kuhn, 2017). To access the map accuracy the leave one out cross validation was carried out using the total observations, n=107.

Table 3. Soil map units, soil classes and number of soil pits surveyed per map unit

MU[(1)]	Taxonomic unit		n[(4)]	n[(5)]
	IUSS-WRB[(2)]	SiBCS[(3)]		
RL	Leptosols	*Neossolos Litólicos (Histíco típico, Distrófico típico* and *Húmico típico)*	5	3
RR	Regosols	*Neossolos Regolíticos (Distrófico leptico, Húmico lítico* and *Húmico típico)*	6	1
CI	Folic Umbrisols	*Cambissolos Histícos (Distrófico típico* and *Distrófico Leptofragmentário)*	8	1
CH1	Cambic Umbrisols	*Cambissolos Húmicos (Distrófico saprolítico)*	14	3
CH2	Umbrisols+Ferrasols	*Cambissolos Húmicos (Distrófico típico, Alumínico típico* and *Distrófico latossólico) + Latossolo Bruno Distrófico húmico*	9	0
CX	Cambisols + Rhodic Acrisols	*Cambissolo Háplico Tb Ditrófico típico + Argissolo Vermelho Amarelo Distrófico nitossólico*	2	1
OO1	Folic Leptic Histosols	*Organossolos Fólicos (Hêmico lítico, Sáprico lítico* and *Hêmico fragmentário)*	17	1
OO2	Folic Histosols	*Organossolos Fólicos (Hêmico típico, Sáprico típico and Sáprico cambissólico)*	8	6
OX	Histosols	*Organossolos Háplicos (Hêmico típico* and *Sáprico típico)*	2	5
Rock	Rock outcrop	Rock outcrop	9	6
Total			80	27
Total			**107**	

(1) Map units. (2) International soil classification system (IUSS Working Group WRB, 2015). (3) Brazilian Soil Classification System (Santos et al., 2018) (4) n = number of soil data selected by the cLHS approach; (5) n = number of soil data from legacy data and extra soil points collection

3.4.4 Pedometric tools for quantitative soil characterization of the Itatiaia National Park database

For the quantitative characterization of the soil database, the legacy data and those obtained during the development of this work were combined, totalizing 90 profiles and 359 horizons. For the soil profiles analysis, the morphological, physic and chemical characteristics of the horizons were considered. In order to analyse the data, the Algorithm for Quantitative Pedology (AQP), developed by Beaudette et al. (2013), was used. This algorithm was developed to address some of the difficulties associated with the processing of soil information, such as visualization, aggregation and data classification of soil profiles (Beaudette et al., 2013) from the profile perspective (vertical variability). Although the AQP method was presented in 2013, few studies have been developed with this technique for analysis of soil databases. In Brazil, the most recent are from Pinheiro et al. (2016, 2018).

The main functions used for data analysis were: plotting functions that seek to show the vertical distribution of soil properties, sketch plot; and functions such as boxplot, which gives an idea of the behaviour of soil properties and functions for data harmonization, and enables to standardize the soil depths as defined by GlobalSoilMap (Arrouays et al., 2014).

3.5 RESULTS AND DISCUSSION

3.5.1 Representativeness of the accessible area for soil sampling

According to the descriptive statistics (Table 4), in general, all buffers tested were able to represent the variation pattern of the covariates used as auxiliaries in the selection of sampling points by the cLHS method.

Table 4. Descriptive statistics for the variables: elevation, slope, northernness, and SAVI for buffers of 100, 200 and 400 m, and total area.

Variables	Unit	Area	Min	1º Qu	Mean	3º Qu	Max	SD
Elevation	m	buffer 100	1456	2019	2206	2418	2772	277.91
		buffer 200	1455	2027	2206	2420	2776	276.29
		buffer 400	1446	2033	2201	2413	2776	267.21
		Total area	1446	1903	2083	2278	2776	249.67
Slope	%	buffer 100	0.1	18.5	33.1	45.6	151.2	18.7
		buffer 200	0.1	21.0	35.3	47.9	156.7	19.4
		buffer 400	0.1	23.5	37.1	49.4	156.7	19.0
		Total area	0.1	26.0	39.1	51.15	156.7	18.6
Northernness	degrees	buffer 100	0.0	58.8	100.1	144.0	180.0	51.2
		buffer 200	0.0	54.36	98.3	134.5	180.0	52.0
		buffer 400	0.0	50.4	96.45	142.3	180.0	52.6
		Total area	0.0	46.0	95.6	144.2	180.0	54.2
SAVI	dime	buffer 100	0.01	0.40	0.52	0.66	0.91	0.16
		buffer 200	0.02	0.40	0.53	0.67	0.91	0.16
		buffer 400	0.00	0.41	0.52	0.68	0.91	0.16
		Total area	0.00	0.46	0.60	0.73	0.93	0.16

Note: SD= Standard deviation; Qu= Quartile; dime= dimensionless

For the elevation attribute a variation from 1446 to 2776 m is observed (Table 4) and both the maximum and minimum values were included in the 400 m buffer. However, the difference of these values in relation to the 100 m buffer can be considered small, only 10 m in relation to the minimum elevation value and 4 m in relation to the maximum value. As for the average value, the three buffers presented similar values, with 5 m difference in buffers 1 and 2 in relation to 4, and 123 m in relation to the total area. The standard deviation was larger in the 100 m buffer and greater than the standard deviation of the total area by 28.24 m.

The slope followed the same pattern of elevation at which all buffers were able to pick up the variation pattern. The minimum value and the total area were 0.1% and the maximum of 156.7% for buffers 200 and 400 and total area. Differing only 5.5% in relation to buffer 100. The mean slope for total area was 39.1%, higher by 2 percentage points than the 400 m buffer and by 6 percentage points than the 100 m buffer. The standard deviation value that most approached the total area (18.6%) was that of the 100 m buffer (18.7%).

As for northernness, both the total area and the buffers had pixels on the north-south line, i.e. the exposure index equal to 0 (minimum value), and maximum exposure values of 180 degrees, which is equivalent to the values of 0 and 360 degrees (Samuel-Rosa et al., 2015)

For the SAVI, virtually all the results of the descriptive statistics were the same or very similar between the buffers and the total area. The maximum variation for the study area was 0.0 to 0.93 with an average value of 0.6. The values closest to 0 are associated to rock outcrops and / or shaded area, while values closer to 0.93 are associated with dense vegetation cover as in forest areas. The vegetation cover of altitude field is associated to values around 0.25 to 0.45.

As with the descriptive statistics, the analysis of the results of the histograms indicates that the areas considered accessible (buffers 100, 200 and 400 m) were representative of the total area (upper part of the INP, Figure 2). In general, the variables have a great similarity between the buffers, presenting almost the same distribution.

The elevation histogram, apparently, shows a normal distribution, especially for total area, with a frequency of values of more than 2,100 m for total area and 2,400 m in the accessible area (all buffers) (Figure 7). The accessible area presented greater asymmetry and toward the left, because the arithmetic mean is less than the median which, in turn, is less than the mode. This pattern differed from the total area, in which the distribution is symmetrical (normal) because the mean "is close" with the median and the mode. However, the maximum and minimum elevation values were represented similarly. This characterizes some representativeness of the elevation data obtained with the buffer and elevation data of the total area.

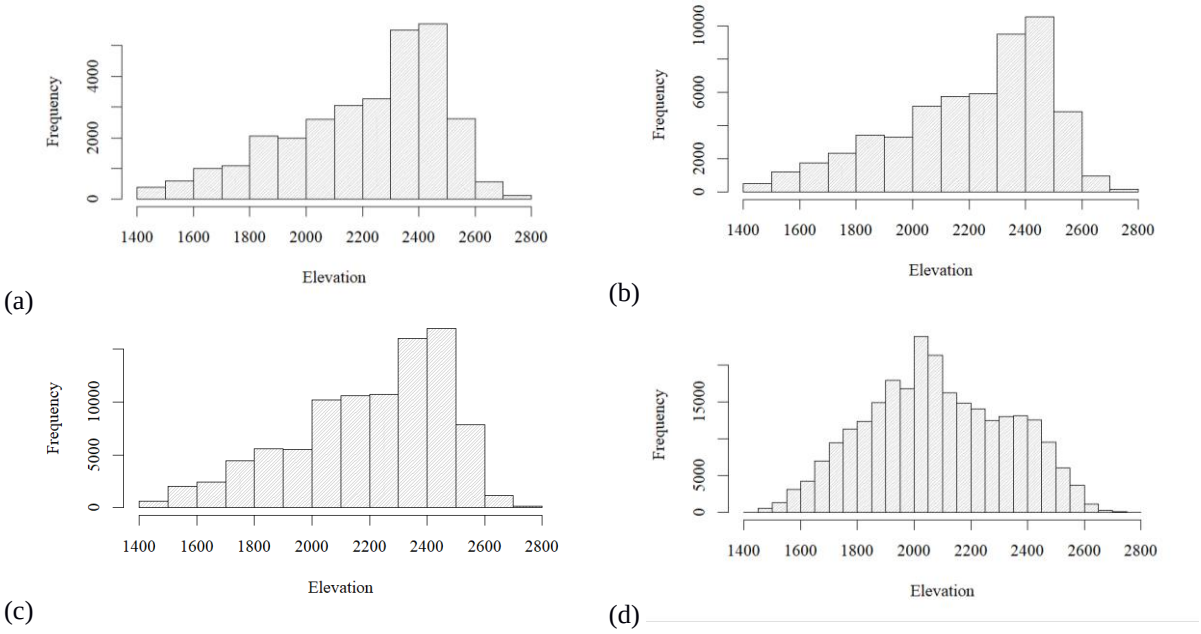


Figure 7. Histogram with frequency distribution of elevation (m) for buffers of 100 (a), 200 (b), 400 m (c) and total area (d).

Similar to elevation, the slope presented a distribution that approximates the normal distribution (visually), being more symmetrical in the total data of the area (higher part of the INP), but with high similarity between this and the buffers. The slope varied from 0 to a little more than 156.7%, with a predominance of values (higher frequency) in the range of 30-40%, values close to the mean values for the buffers and total area (Table 4).

By the analysis of the histogram (Figure 9) it is evident that the northernness does not present normal distribution and, in general, the buffers presented distribution similar to total

area. Also, with a higher frequency of values with exposure above 140 degrees in all cases, for buffers and total area.

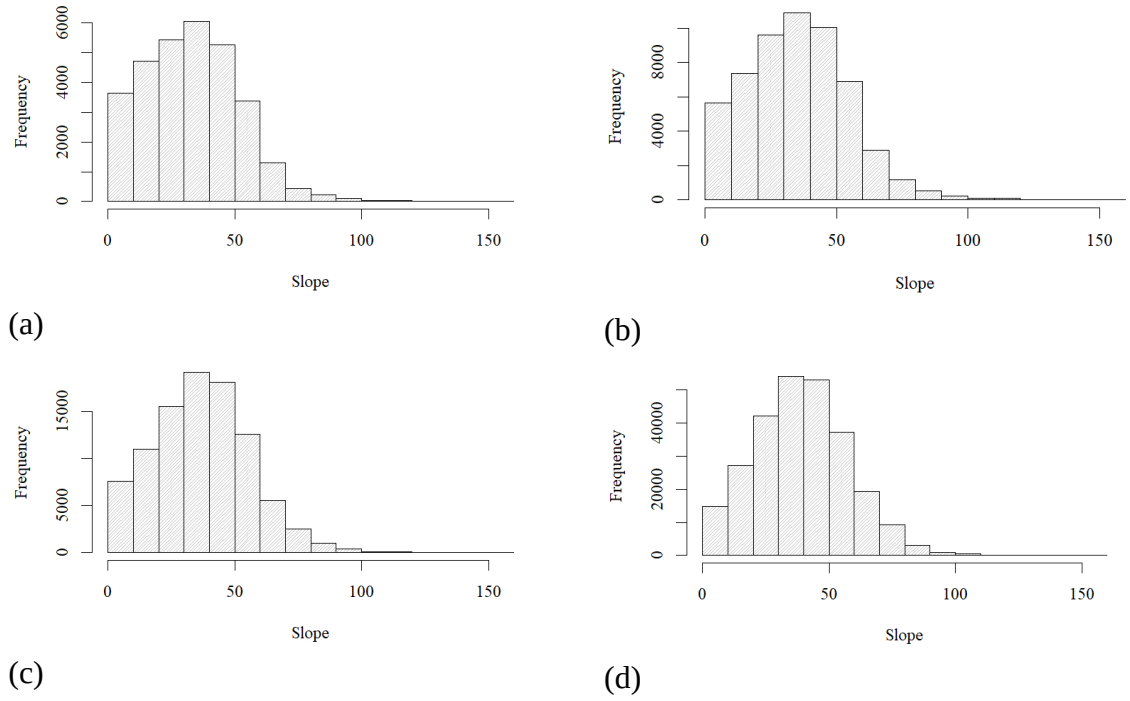


Figure 8. Histogram with slope frequency distribution (%) for buffers of 100 (a), 200 (b), 400 m (c) and total area (d).

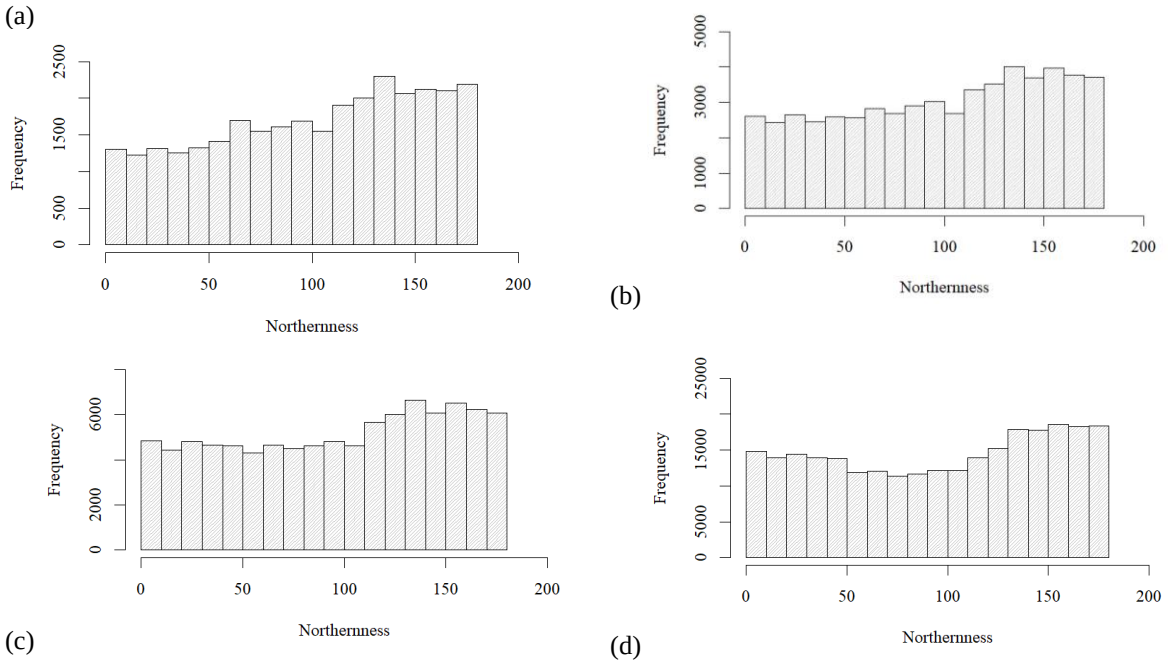


Figure 9. Histogram with frequency distribution of the northernness (degrees) for buffers of 100 (a), 200 (b), 400 m (c) and total area (d).

For SAVI (Figure 10) despite the similar distribution among the data sets, the buffers had a higher frequency of values in the range of 0.4 and around 0.7, approaching a bimodal frequency. However, in the total area this effect was less pronounced, and the frequency of

values was higher in the range of 0.7. Among the buffers, there was little difference and the highest frequency occurred in the range of 0.4.

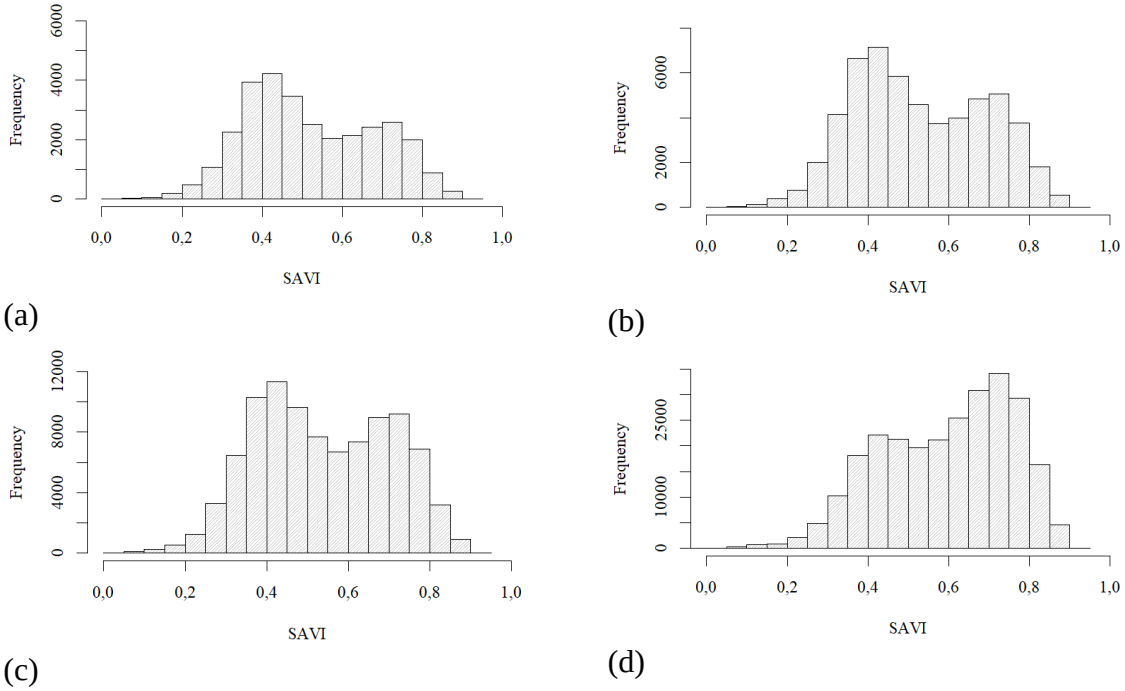


Figure 10. Histogram with the frequency distribution of SAVI (dimensionless) for buffers of 100 (a), 200 (b), 400 m (c) and total area (d).

As shown in Table 4 and Figures 7, 8, 9 and 10, the 100 m buffer does not differ significantly from the others, 200 and 400 m, and it implies in a much smaller area to survey (1913 hectares) compared to the 200 m (3468 hectares) and 400 m (5877 hectares) buffers. Thus, considering that and taken in account the limited access in the upper part of the INP, the 100 m was chosen as accessible and representative area for selection of the 80 sampling points.

3.5.2 Quantitative analysis for soil characterization

The shallow soils and/or with more fine and coarse gravel represented almost 50 and 75%, respectively, of the samples, and they are often associated with rock outcrops. On the other hand, the more developed and deeper soils did not show coarse fragments in the samples (Table 5).

Both the bulk density and the soil moisture content at the time the sample was taken showed a great variation in values. The soil moisture value was around 1.53% for the sampling in the driest season and horizons with a lower content of organic matter. The values were much higher (up to 600% for horizons of organic constitution under conditions of impeded drainage and taken in the wet season). In the same way, the density varied from 0.12 $Mg.m^{-3}$ (organic constitution horizon) to 1.45 $Mg.m^{-3}$ (mineral constitution horizon), with an average value of around 0.77 $Mg.m^{-3}$, which is much lower than the average for soils in Brazil.

In general, the soils of the INP have pH values from the maximum of 5.72 to the minimum of 3.24; base saturation from the maximum of 46.6%, that is, all the soils are dystrophic and, in some horizons, the value does not reach 1%. The potential acidity (H + Al) is up to 67.16 $cmolc.dm^{-3}$, close to the maximum for the T value (69.01 $cmolc.dm^{-3}$), that is, despite the high cation exchange capacity, most exchangeable ions are Al^{+3} and H^{+}.

The base contents are low, with maximum of 2.55 cmolc.dm^{-3} for Ca, 1.67 cmolc.dm^{-3} for Mg, and Na and K not reaching 0.8 and 1.26 cmolc.dm^{-3}, respectively. Therefore, the sum of bases (S value) reaches a maximum of 4.24 cmolc.dm^{-3}.

Table 5. Descriptive statistics of the soil dataset

Variable	Unit	Min	1º Qu	Mean	3º Qu	Max
SM	%	1.53	28.04	61.83	72.63	662.96
BD	Mg.m-3	0.12	0.54	0.77	0.97	1.45
Fine gravel	%	0.00	0.00	1.56	0.00	42.00
Coarse gravel	%	0.00	0.00	4.54	5.00	75.00
pH	----	3.24	4.24	4.51	4.77	5.72
Ca	cmolc.dm^{-3}	0.00	0.00	0.16	0.20	2.55
Mg	cmolc.dm^{-3}	0.00	0.30	0.55	0.65	1.67
Al	cmolc.dm^{-3}	0.00	1.00	2.12	2.70	9.20
Na	cmolc.dm^{-3}	0.00	0.02	0.04	0.04	0.80
K	cmolc.dm^{-3}	0.00	0.04	0.13	0.19	1.26
P	mg.dm^{-3}	0.00	1.51	7.38	9.33	97.52
H+Al	cmolc.dm^{-3}	2.3	6.05	12.80	16.33	67.16
C	g.kg^{-1}	2.40	16.25	64.25	95.62	294.80
N	g.kg^{-1}	0.00	0.929	3.89	6.28	10.85
S value	cmolc.dm^{-3}	0.20	0.52	0.88	1.08	4.24
T value	cmolc.dm^{-3}	3.00	6.71	13.68	17.57	69.01
V%	%	0.95	4.37	8.20	9.60	46.46

Note: SM: soil moisture; BD: bulk density; pH(H2O): pH in water-saturated soil paste (1:2.5); Ca^{2+}: Calcium; MG^{2+}: Magnesium; Al^{3+}: Aluminium; Na^{+}: Sodium; K^{+}: Potassium; P: phosphorus; H+Al: potential acidity; C: total carbon; N total nitrogen; S: sum of bases (Ca, Mg, K, and Na); T: cation exchange capacity V: base saturation

The properties of most INP soils are associated to the parent material, generally of acid nature, and the relief, predominantly mountainous and with high elevations that influence the climate and organisms leading to conditions favouring the accumulation of organic matter (Soares et al., 2016). The large amount of organic material accumulated, which in terms of soil organic carbon can reach almost 300 g.kg^{-1}, influences directly other attributes, such as the high potential acidity, associated with the low pH values, and the high CEC due to a large number of phenolic and carboxylic radicals from the organic matter. The high levels of organic matter also explain the low bulk density and the high-water storage capacity of some horizons.

The plots in standardized sketches for some soil attributes, such as pH, carbon, CEC and bulk density give us an idea of the distribution, both in depth and in the different horizons, in each type of soil (Figures 11, 12, 13, 14, 15). For example, except for soils with imperfect or more limited drainage, organic soils, which are found in the valleys, there is a tendency to increase the pH (Figure 11) and bulk density (Figure 14) with soil depth. Also, a decrease in the organic carbon (Figure 12), nitrogen and H+Al (not shown here), and CEC (Figure 13) content with soil depth. An exception for this pattern is the profile 62, with a variation of carbon, pH and CEC contents, probably due to different events of deposition of organic material and mineral sediments.

An important function of the AQP is the option of plotting standardized sketches with Munsell colours staining the soil horizons or layers according to the variations of attributes. It is possible to create a colour vector with the hue, value and chrome information of each horizon. In general, the more vivid the colouring the greater influence of the hue, for the subsurface horizons of mineral soils. The AQP also was important to differentiate better drained *Organossolos* (Histosols), with a mineral subsurface horizon closer to the surface, from those with a thicker organic horizon above the mineral layer; since the organic matter "masks" the soil hue, that would be more expressed in the mineral horizons, as can be seen in Figure 15.

In general, the surface horizons with higher organic matter have low values of value and chroma, that is, they are darker in colour. Mineral soils are generally deeper and are covered predominantly by the Atlantic Forest, differing from the *Organossolos* (Histosols), usually shallow soils with sparse vegetation of grasses and shrubs typical of the altitude fields (Figure 15), varying with drainage and landscape position. There is a predominance of yellow and red-yellowish colorations for the mineral horizons. In addition to the climate and relief, the parent material influences this colouration, since it is poor in iron and the humid conditions favour formation of goethite as iron oxide.

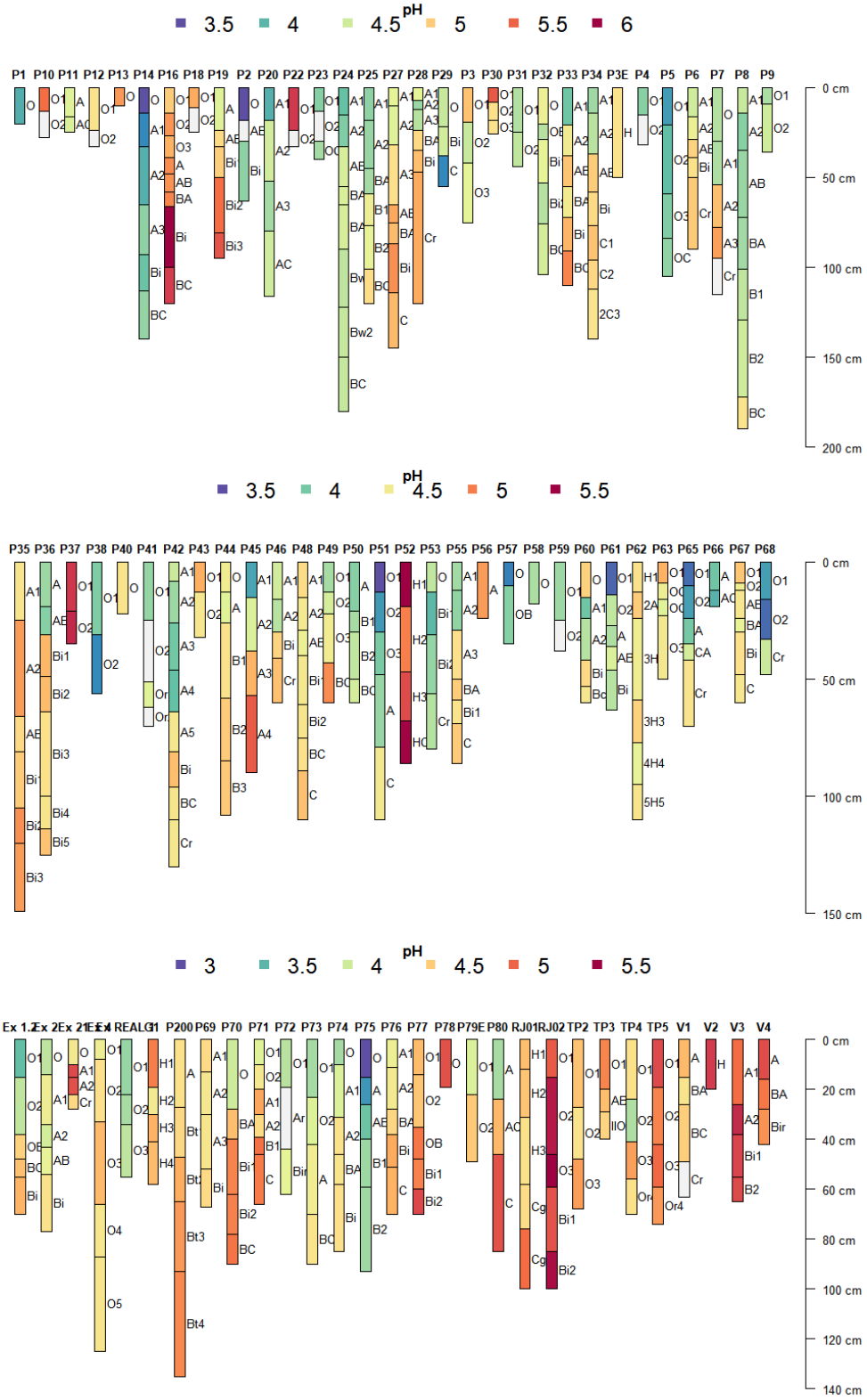


Figure 11. pH values for the soil profile collection from the upper part of INP.

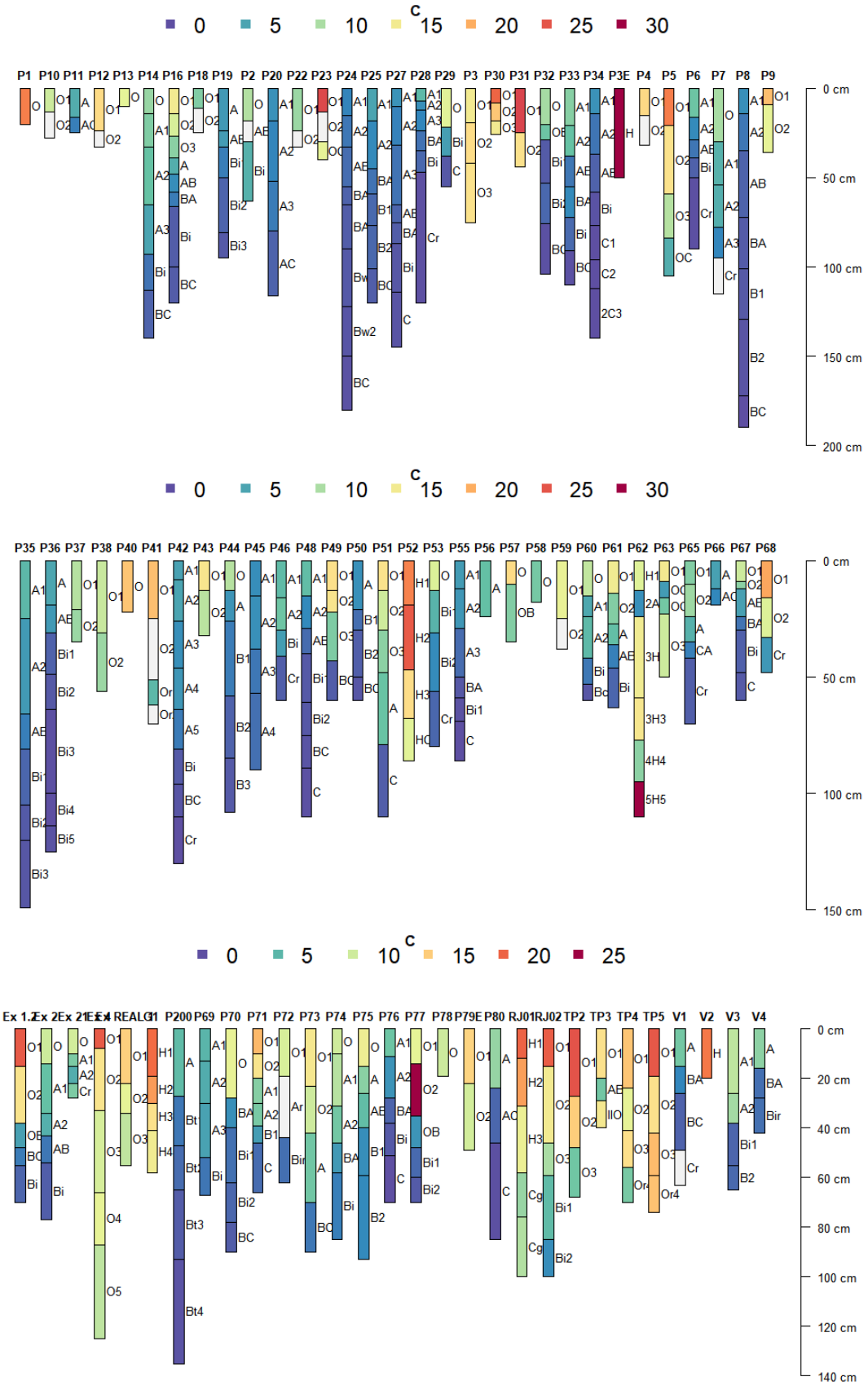


Figure 12. Total soil carbon content (%) for the soil profile collection from the upper part of INP.

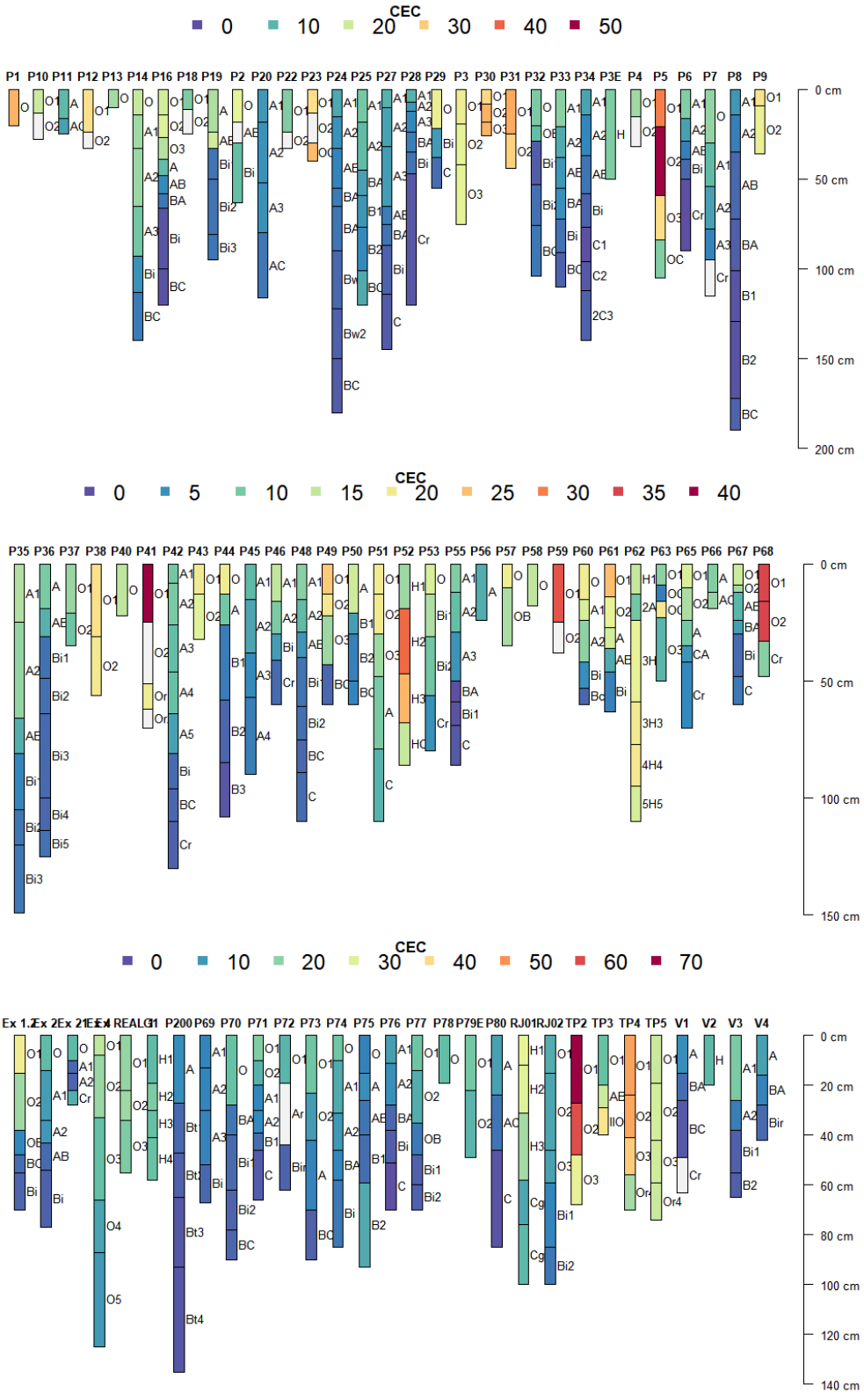


Figure 13. Cation exchange capacity (g.dm-3) for the soil profile collection from the upper part of INP.

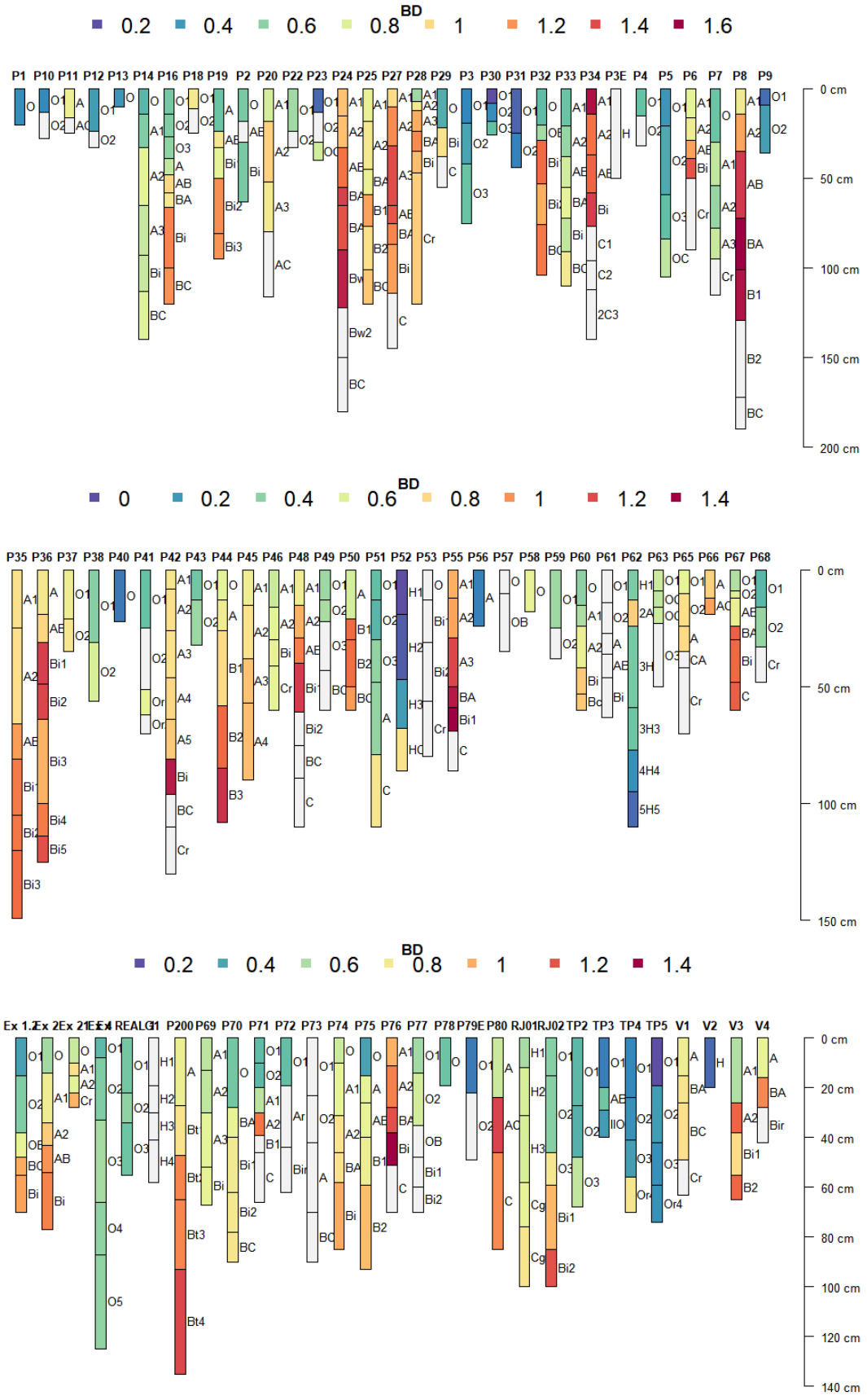


Figure 14. Bulk density (Mg.m-3) for the soil profile collection from the upper part of INP.

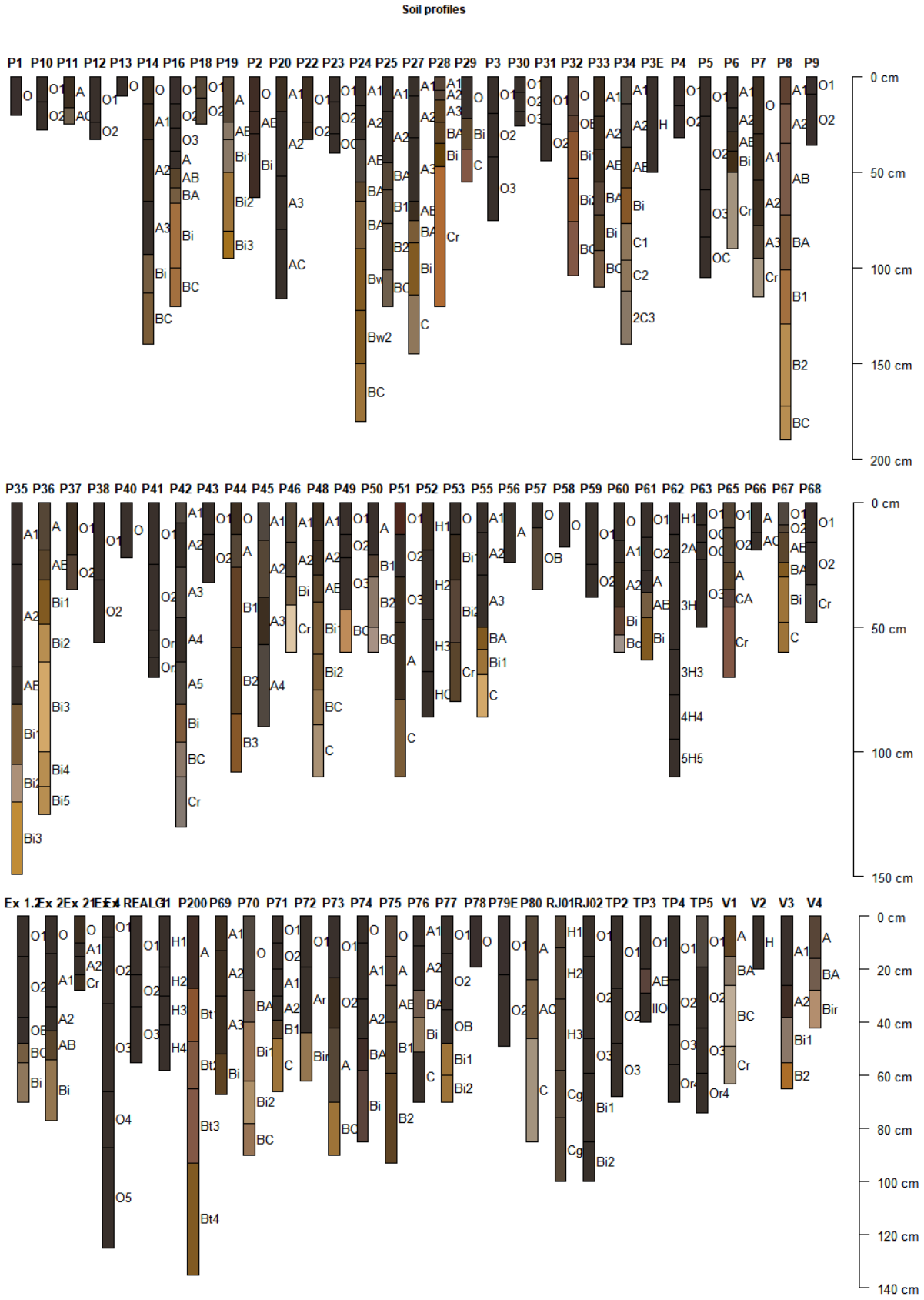


Figure 15. Plotting the sketches standardized according to the Munsell coloration for the soil profile collection from the upper part of INP.

By grouping all the horizons of the different profiles in a boxplot (Figure 16) some conclusions can be drawn regarding some soil properties. For example, the phosphorus content does not depend on the type of horizon, with high values in the organic and in the C horizons, both types showing high and low values of P, in other words, high amplitude variation.

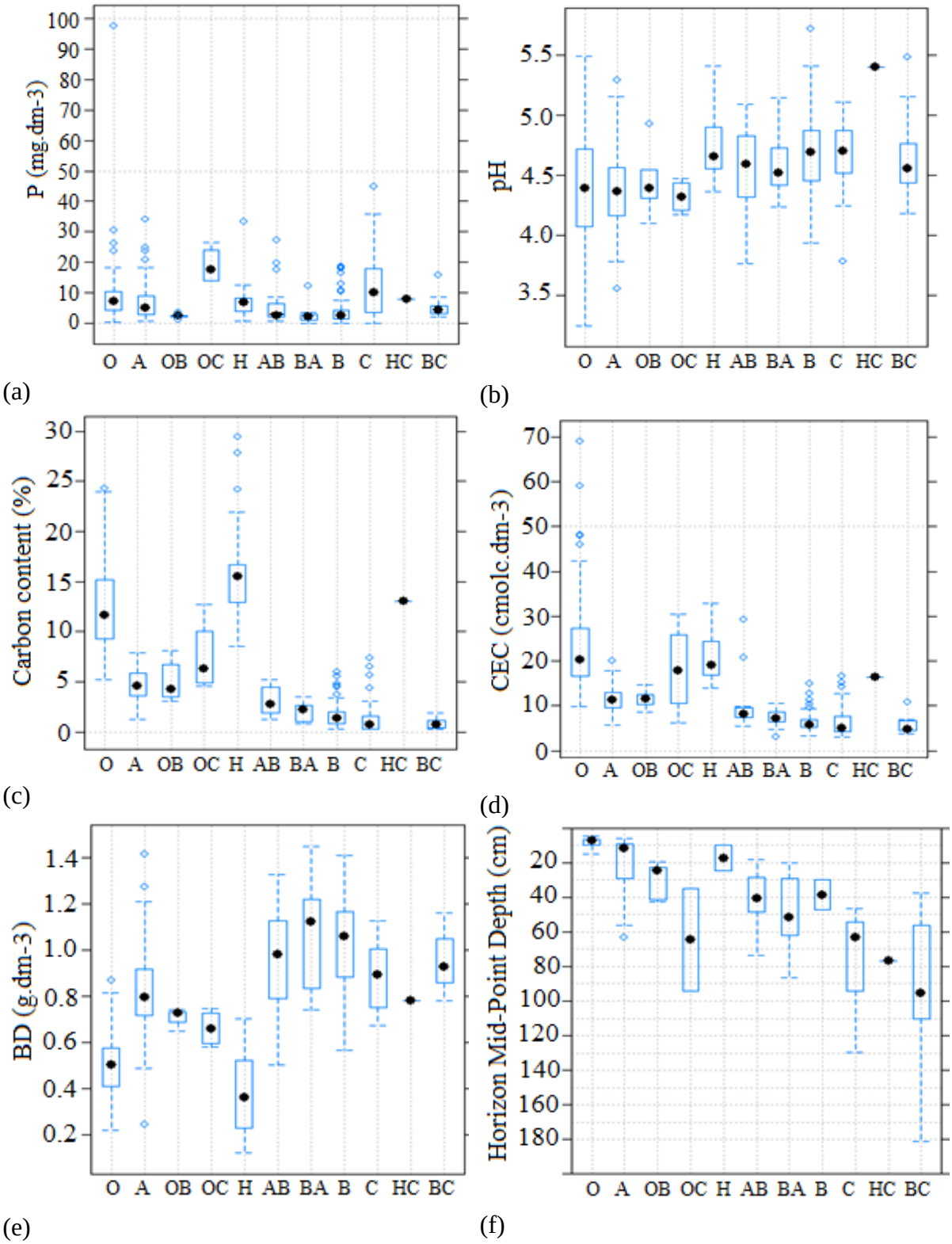


Figure 16. Boxplot for values of phosphorus (a), pH (b), carbon (c), CEC (d), BD (e) content and midpoint depth of soils (f) for each soil horizon.

Another interesting pattern is that among the organic horizons the H horizons had an average carbon content and CEC larger than the O horizons and reach deeper in the soil (Figure 16). As expected, the mineral horizons presented lower levels of C and CEC. Corroborating with the carbon content, the lowest values of bulk density were for the horizons O and H. The highest averages were for the horizons B and BA. As for the pH, it is observed that the organic matter is not the only factor influencing the acidity and, consequently, the pH values. For example, the O horizon has high amplitude of pH variation, from less than 3.5 to 5.5.

The correlation between contents of carbon and terrain attributes showed positive tendencies regarding elevation (correlation of 0. 59) and negative (correlation of -0,44; -0,35; -0,21; -0,40; -0,49 with bands 1, 2, 3, 4 and 5 respectively) for spectral bands data (Figure 17).

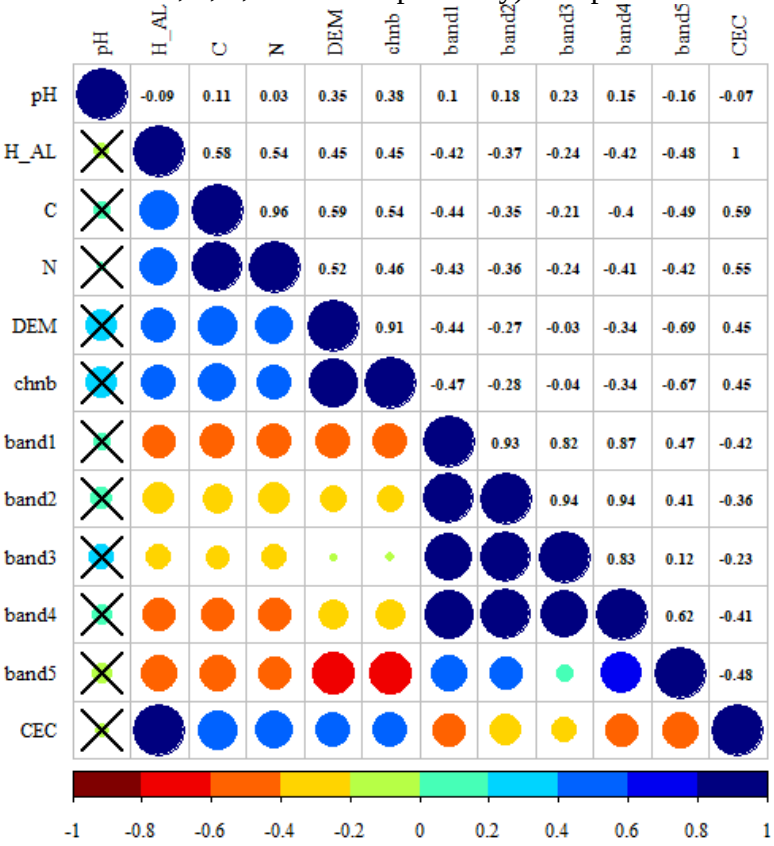


Figure 17. Matrix of correlation between some environmental covariates and some soil properties. Correlations with "X" are not significant at 5% of confidence**.**

This result indicates that with increasing elevation, which in turn conditions the climate (lower temperatures and higher precipitation Figure 3), there is an increase in the accumulation of organic material due to its slow decomposition. Besides that, a reduction of the reflectance values coincides with the decrease of the forest vegetal coverage and its substitution by high-altitude fields.

Another high and positive correlation was the total carbon contents with the channel network base level (correlation of 0.54) (Figure 17). This is probably due to the relief, the second most important factor that conditions the accumulation of organic matter. In this case accumulation of organic material is conditioned by its slow decomposition due to prevalent anaerobic conditions, such as in figure 16, where the highest C and N contents were observed for the H horizons formed in this climatic and relief conditions.

The same result was observed for CEC and H + Al contents, in which they presented a high correlation with carbon content (correlation of 0.59 and 0.58, CEC and H+Al respectively) and consequently high correlation (positive) with elevation (correlation of 0.45 for both, CEC

and H+Al) and negative with spectral bands of the RapidEye sensor (correlation of -0.42; -0.36; -0.21; -0.41; -0.48 between the CEC and bands 1, 2, 3, 4 and 5 respectively) and (correlation of -0.42; -0.37; -0.24; -0.42; -0.48 between the H+Al and bands 1, 2, 3, 4 and 5 respectively on Figure 17. This high relation between carbon and CEC can also be observed on figures 16 and 17, where the pattern was similar for these attributes in the different types of horizons.

3.5.3 Soil types, landscape aspect and spatial distribution

Most of the profiles were classified as *Organossolos* (Histosols) n=39 (Table 3). The deeper organic soils are found in flat areas and dominantly from by coluvial-alluvial sediments and organic material deposits (Histosols with a histic horizon in the WRB, ex. profile G1 - Figure 18). The drainage is imperfect or impeded and have high levels of carbon and total nitrogen. The *Organossolos* (Histosols) on slopes are generally shallower, well-drained or moderately well-drained, and have lower carbon and nitrogen content. Both types are under high altitude fields coverage. The better drained *Organossolos* (Histosols), in many places, had a folic horizon of little more than 20 cm, and directly in contact with rock (Figure 18 and 19).

Another soil classes, without a subsurface horizon and with shallow depth over the rock, is the *Neossolo Regolítico* (Regosol), map unit RR (Figure 20), and the *Neossolo Litólico* (Leptosols), map unit RL. In some cases, the profile has a surface organic horizon but it is less than 20 cm depth. These soils are often associated with the rocky outcrops. The dominant vegetation is of high-altitude fields.

Figure 18. Typical mineral soils on the Atlantic forest (upper left); soil profile with an organic surface horizon and deep mineral subsurface horizon (upper right); soil with shallow organic horizon over the rock and on the slope (bottom left); soil with thick organic horizon in a flat valley area (bottom right).

There are also deep soils with high organic matter content on surface and mineral subsurface horizons, generally with a yellowish colour (Figure 18 and 19). The *Argissolos* (Lixisols and Acrisols) and *Cambissolos* (Cambisols), map unit CX, are associated with forest coverage, have lower carbon content, CEC, are more developed than the *Neossolos Litólicos* (Regosols) and *Neossolos Regolíticos* (Leptosols), and have higher pH values. Some soils have a high carbon content in the surface horizons, but they still meet the mineral soil criteria, and most were classified as Umbrisols (Folic Umbrisols, Cambic Umbrisols and Umbrisols), respectively in the map units - CI, CH1 and CH2 (Figure 20).

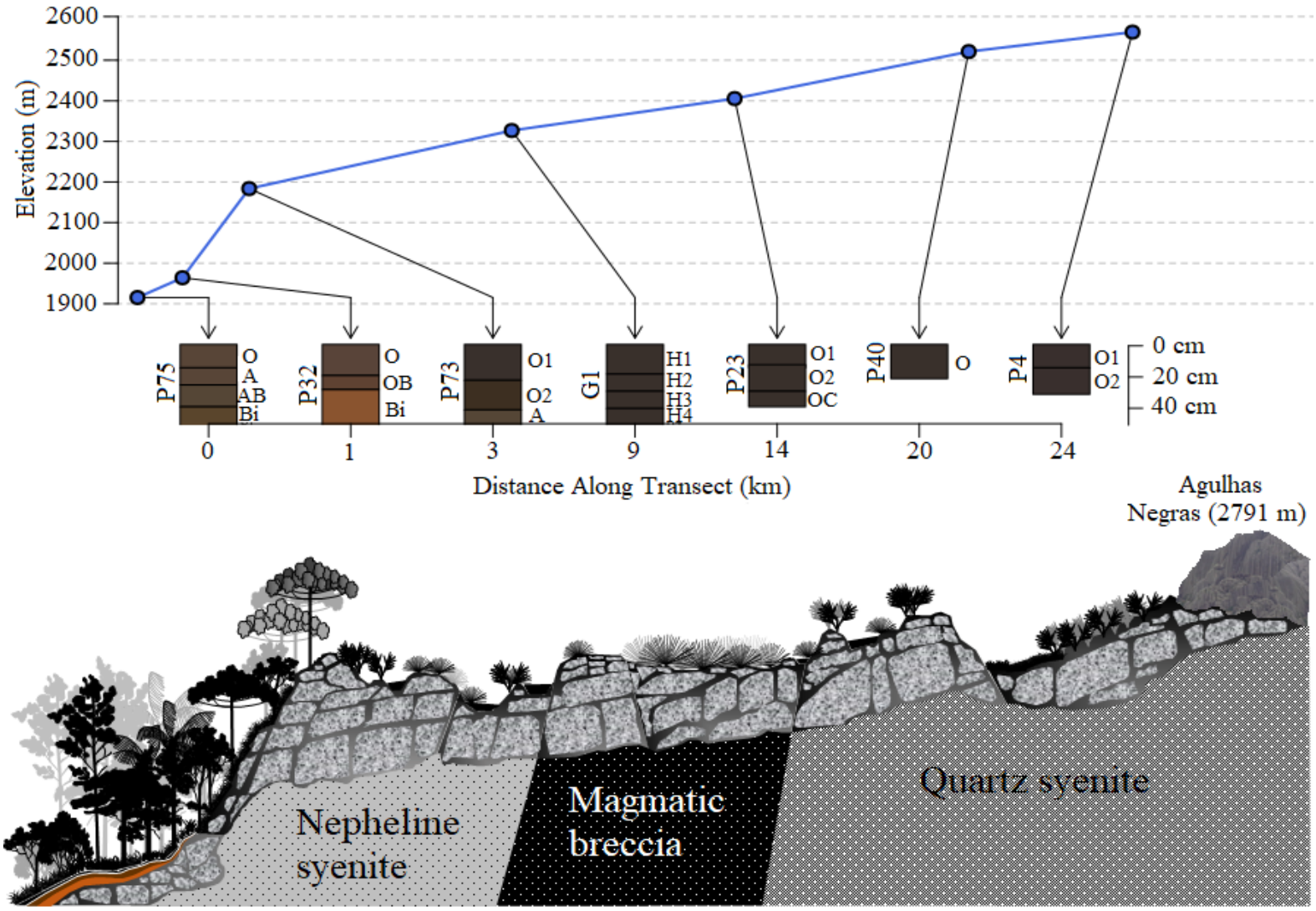


Figure 19. Schematic distribution of vegetation, geology and soils along an transect in the INP (Made by Orlando Carlos Huertas Tavares)

Twelve covariates for soil mapping were selected by the recursive feature elimination algorithm, ranked by importance, and they were: 4 morphometric covariates (elevation, channel network base level, channel network distance and relative slope position); 3 from the images (bands 1, SAVI and NDVI); 2 from climatic maps (rainfall and temperature), 2 from spatial information (Latitude [Y] and a polygon of second order of X and Y [XY]), plus 1 from parent material and relief information associated (geomorphology).

Soils distribution in the upper part of the INP shows a predominance of profiles with a high organic matter content, shallow depth and located in high slopes (Figure 20). Although, these soils are very fragile they were not reported in the survey for the INP's management plan, nor in most of the previous studies of soils in the park, except for Soares et al. (2016) and Silva Neto et al. (2018) that had only point descriptions without spatial information or spatial distribution of these soils.

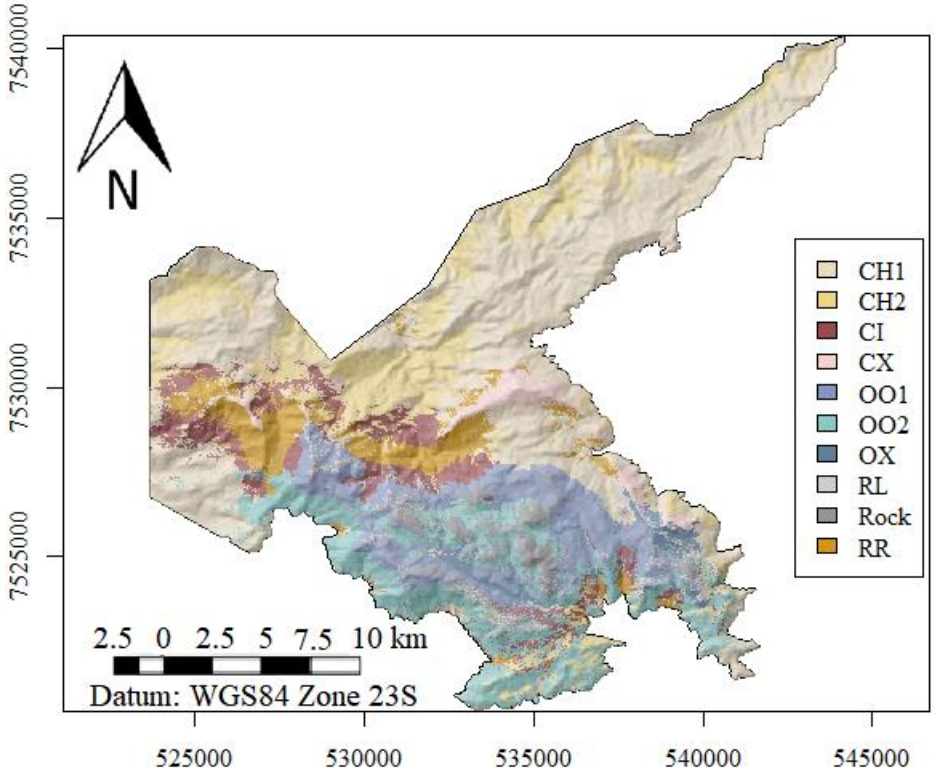


Figure 20. Current and more detailed soil map for the upper part of INP using the DSM techniques

In terms of importance for the final model the top 5 covariates were Y, DEM, XY, rainfall and SAVI. This result shows that the soil formation factors that influence most the soil classes discrimination and their spatialization are the space position (Y and XY), relief (elevation), climate (rainfall) and organisms (vegetal cover, SAVI).

When the fitted model is evaluated the results (using leave-one-out cross validation [LOO-CV] approach) show an overall accuracy of 0.41 and Kappa of 0.32 with a confidence interval of 0.31-0.51 (Table 6). That can be considered as a reliable result in terms of soil classes prediction for a very complex mountainous area such as the INP plateau.

These results are comparable to those of Jeune et al. (2018) and Meier et al. (2018) that found values of Kappa index from 0.42 to 0.47 respectively using RF algorithm. In addition, the RF has a strong ability for learning and generalizing soil data of larger geographic areas (Jeune et al., 2018). Thus, pedometric methods should be seriously considered as a

complementary approach to conventional methods for mapping complex mountainous tropical areas with limited access (Meier et al., 2018).

In terms of percentage of predicted area for each soil mapping unit the largest one is constituted by the association of CH1 and CH2 (Cambic Umbrisols Umbrisols + Ferrasols, respectively), where the *Cambissolo Húmico* (SiBCS) is the dominant class. The total area is of 8470.75 hectares that represents more than half of the total studied area (51.64%) (Table 6).

The organic soils (OO1, OO2 and OX) occupy an area of 4148.38 hectares (25.29% of the total area). Although the Histosols with better drainage are dominant, 1.79% of the total area (294.25 hectares) is constituted of imperfect or impeded drainage (OX) soils, classified as *Organossolos Háplicos* (SiBCS).

Table 6. Area and percentage in relation to the total area of the soil map units in the upper part of the INP**.**

Map unit	Area (ha)	Area %
CH1	5812.81	35,44
CH2	2657.94	16,20
CI	1285.25	7,84
CX	498.438	3,04
OO1	1647.44	10,04
OO2	2206.69	13,45
OX	294.25	1,79
RL	184.61	1,13
Rock	653.13	3,98
RR	1163.13	7,09
Total	**16403.69**	**100,00**

The less developed soils *Neossolo Litólico* (Leptosols) (RL) and *Neossolo Regolítico* (Regosols) (RR) represent 1.13% and 7.09% of the total area respectively, and in addition to rock outcrops (3.98% of the total area) correspond to 2000.87 hectares, that is, 12.20% of the total area and are mainly concentrated in the central area of the upper part of INP (Figure 20).

The *Cambissolos Húmicos* (Folic Umbrisols) (CI) represent less than 8% of the total area and *Cambissolos Háplicos* (Cambisols) (CX) slightly more than 3%. The latter mainly predicted in the upper part of the Serra Negra.

When analysing the confusion matrix (Table 7), it is observed that both the CX and CI soil map units were erroneously classified in all soil samples. In the case of the CX it is explained by the low representativity (only three soil samples); while the CI is usually occupying similar landscape position with other map units that also have elevated carbon contents in the surface, for example OO1 had 4 profiles classified in that unit when in truth they belonged to CI unit.

The classes CH1 and CH2 had a reasonable accuracy (0.47 and 0.44, respectively), where the greatest confusion was between the two units, which have similar characteristics (soil attributes and landscape position) differing mainly in soil depth. For the class OO1 that had 0.56 of accuracy of the predictions, the biggest confusion was with RL and Rock classes. This is explained because this class has a lower depth of soil compared to OO2, which leads to being predicted together with classes that have short soil depth, example RL, or even as a rock outcrop (Rock class). On the other hand, the OO2 class, which had greater confusion (0.36 accuracy), was more confounded with OO1 and CI. Both OO classes have similar soil properties and they differ mainly in soil depth. Also, CI and OO2 have similar organic matter content in the surface

and depth, where the main difference is that CI presents mineral subsurface horizons at a shallower depth and/or superficial horizon of smaller thickness compared to OO2.

Table 7. Confusion matrix of soil classification using Random Forest with LOO-CV.

		Reference classes										
		OO1	**CH2**	**OO2**	**OX**	**RR**	**RL**	**ROCK**	**CX**	**CI**	**CH1**	**Total**
Predicted classes	**OO1**	10	0	3	2	0	1	4	0	4	2	26
	CH2	0	4	0	0	0	0	0	1	1	3	9
	OO2	0	0	5	2	0	1	0	0	2	2	12
	OX	1	0	2	2	0	2	0	0	0	1	8
	RR	0	0	0	0	4	0	0	1	1	1	7
	RL	2	0	0	0	0	1	0	0	0	0	3
	ROCK	2	0	1	1	0	3	10	0	0	0	17
	CX	0	1	0	0	1	0	0	0	0	0	2
	CI	2	1	3	0	1	0	0	0	0	0	7
	CH1	1	3	0	0	1	0	1	1	1	8	16
	Total	18	9	14	7	7	8	15	3	9	17	**107**
Overall accuracy												**0.41**
Confidence interval (95%)												**(0.31-0.51)**
Kappa												**0.32**

For the less developed mineral soils RR and RL, there was a considerable difference in the accuracy; while the RR unit had an accuracy of 0.57, the RL unit showed 0.13 and for this class the confusion was greater with the rock outcrop. The poorly drained organic soils, which showed an accuracy of 0.22, presented confusion mainly with the well-drained organic soil classes (OO1 and OO2), with the same confounding proportion.

Despite the confusions, which occurred mainly in classes very similar between themselves and that occupy similar positions in the landscape, the result of the prediction may be considered good. Also, when considered the spatial distribution the map discriminates well the mapping units in the landscape. The results show the importance of DSM as a potential tool to predict soil types in mountainous areas in Brazil for the purpose of environmental vulnerability analysis and land use planning.

In addition, the more detailed information, now available and with calculation of error, may contribute better to future research and implementation of the INP management plan. When analysed spatially it is possible to see that the central part of the INP, where there is a predominance of the organic soils, the climate is marked by greater precipitation, lower temperature and a relief more rugged and with higher elevation; besides the dominant geological material is quartz syenite and the geomorphology is of the escarpments type. Also, in this area, there is a predominance of vegetation cover of the Altitude Fields, characteristic of *Altomontanas* region (Soares et al., 2016). There is a great amount of rock outcropping, such as in the uppermost part of the park, where the Pico das Agulhas Negras (Figure 19) is located. Areas with poorly drained organic soils are in similar climate conditions but they are located in the lower part of the valleys among the high elevations, with alluvial sediments as parent material and fluvial plain (comprising mineral and organic sediments) geomorphology.

In the lower part of the plateau predominate deeper and better-developed soils, mainly under forest vegetation, and although they have high levels of organic matter they are not enough to identify as folic classes. These soils are placed in relatively lower slopes and elevations, developed from homogeneous gneisses and nepheline syenite geological material and with Mountain geomorphology.

3.6 CONCLUSIONS

In terms of terrain surface representation in the smallest area for defining soil sampling locations in the upper part of the INP, the best restriction rule was the 100 m buffer, In addition to having the points allocated closest to the roads and trails, it adequately represented the characteristics of the study area, and in relation to the environmental covariates, there is only a slight difference from buffers of 200 and 400 m.

One of the advantages of using the cLHS and buffer as a rule of restriction for the allocation of sampling points was the increase in work efficiency, which leads to lower cost and time, without losses of the representativeness of environmental conditions. Thus, this procedure can be indicated for areas such as INP, where access is very restrictive.

In general, the soils of the Itatiaia National Park plateau are predominantly shallow, with high levels of C, N, H + Al, and CEC; with low values of pH and bulk density; and a high capacity to store water. Dominantly, the soils are highly vulnerable to degradation, especially due to erosion, compaction and land sliding.

Soil survey and characterization supported by the quantitative techniques of soil data analysis, through the AQP and with the creation of a database in a GIS environment, are an effective way to support future researches in the park as well as INP decision making.

In the upper part of the INP, some of the soils identified were not reported there previously, specifically in the soil survey for the INP management plan. This is partly due to the low level of detail of that survey. Thus, it can be affirmed that the information produced in this study is potentially useful for several multidisciplinary researches and, in particular, to improve the INP management plan, in order to evaluate the environmental vulnerability of these soils and to possible install access restrictions for public use.

4 CHAPTER II:

MAPPING SOIL PROPERTIES IN A POORLY ACCESSIBLE AREA. CASE STUDY - ITATIAIA NATIONAL PARK

4.1 RESUMO

Os mapas de solo são importantes para avaliar as funções do solo e apoiar a tomada de decisão, particularmente para propriedades do solo, como pH, teor de carbono e capacidade de troca de cátions (CTC). A resolução espacial e a profundidade devem atender às necessidades dos usuários. A eficiência dos modelos estatísticos para criar mapas de propriedades do solo, com um nível de precisão aceitável, geralmente requer um grande número de amostras com uma distribuição apropriada na área de teste considerada. No entanto, a acessibilidade para amostragem pode ser um problema em muitas áreas remotas, como no Parque Nacional do Itatiaia (PNI). Os objetivos deste estudo foram: delinear uma estratégia de amostragem que equilibrasse a acessibilidade, os custos relativos, o tamanho da área e as covariáveis ambientais; e modelar as propriedades do solo com um número limitado de amostras. O objetivo foi produzir mapas de propriedades de solo 2D e 3D com acurácia aceitáveis e com incerteza associada. As propriedades do solo testadas foram pH, teor de carbono, CTC. A estratégia de amostragem foi projetada usando o método do Hipercubo Latino condicionado (cLHS). Diferentes métodos foram testados para produzir os mapas das propriedades do solo. Para calibração dos modelos foram usados: linear (MLR, regressão linear múltipla) e não linear (GAM, Generalized Additive Models) e Random Forest, como exemplo de um modelo de aprendizado de máquina. Os resultados mostraram diferenças no desempenho preditivo para todos os métodos estatísticos e abordagens de seleção de covariáveis. O GAM, com abordagem *scorpan*, foi o melhor método para o número limitado de amostras de solo. O modelo de RF não foi muito sensível à seleção da covariável. A maior incerteza foi associada às áreas de menor acessibilidade e, consequentemente, com baixa densidade amostral e/ou ruídos nas covariáveis. Mesmo assim, a modelagem de mapas de propriedades de solos 2D e 3D, com propagação de incerteza, contribuirá para a análise de vulnerabilidade ambiental do PNI, fornecendo informações que de outra forma não estariam disponíveis.

Palavras-chave: Função de profundidade. Modelos Aditivos Generalizados. GlobalSoilMap, Propagação de Incerteza. Seleção de Preditor.

4.2 ABSTRACT

Soil maps are important to evaluate soil functions and support decision-making particularly for soil properties such as pH, carbon content and cation exchange capacity (CEC). The spatial resolution and depth should meet the needs of users. The efficiency of statistical models to create soil properties maps, with an acceptable level of accuracy, often require a large number of samples with an appropriate distribution across the considered test area. However, accessibility for sampling can be a problem in many remote areas, such as in the Itatiaia National Park (INP). The objectives of this study were: to design a sampling strategy balancing accessibility, relative costs, area size and environmental covariates; and to model soil properties with a limited number of samples. The goal was to produce acceptable accurate 2D and 3D soil properties maps with the associated uncertainty. The soil properties tested were pH, carbon content, CEC. The sampling strategy was designed using conditioned Latin Hypercube Sample (cLHS). Different methods were tested to produce the maps of the soil properties. For calibration of the models: linear (MLR, multiple linear regression) and nonlinear (GAM, Generalised Additive Models), and Random Forest, as an example of a machine learning model, were used. The results showed differences in the predictive performance for all statistical methods and covariate selection approaches. The GAM, with *scorpan* approach, was the best method for the limited number of soil samples. The RF model was not very sensitive to the covariate selection. The greater uncertainty was associated with the areas with lowest accessibility and, consequently, with low sampling density and/or noises in covariates. Even though, the 2D and 3D soil properties maps modelling, with uncertainty propagation, will contribute to the INP analysis of environmental vulnerability by providing information otherwise not available.

Keywords: Depth function. Generalized Additive Models. GlobalSoilMap. Uncertainty Propagation. Predictor Selection.

4.3 INTRODUCTION

Soil is a vital part of the natural environment and it has crucial role in ecosystem functioning (Adhikari and Hartemink, 2016). Soil functions can be derived from soil properties and their interaction and assessment of soil functions can provide detailed spatial information particularly useful in complex mountain terrain (Jeong et al., 2017). Soil information is an essential factor for environment conservation and sustainable management, in the formulation of sustainable agricultural policies and the monitoring of impacts caused by inappropriate use of this resource (Carvalho Junior et al., 2016), especially in mountain areas.

In recent years, there was a considerable advance in digital soil mapping (DSM) due to new approaches, such as powerful predictive algorithms (Beguin., et al 2017; Sindayihebura et al., 2017); models combining machine learning and geostatistical tools (Hengl et al., 2007; Poggio et al., 2014); expert knowledge-based methods (Ashtekar et al., 2013; Poggio et al., 2016; Menezes et al., 2014; 2018) and high-resolution soil maps (Ashtekar et al., 2013; Camera et al., 2017; Forkuor et al., 2017; Meersmans et al., 2012; Mulder et al., 2016 a; Nussbaum et al., 2018). The advantage of modelling the soil properties in 3D space has been evaluated in several studies (Adhikari et al., 2013; Amirian Chakan et al., 2017; Kidd et al., 2015; Liu et al., 2013; Mulder et al., 2016b), including the assessment of associated uncertainty (Kempen et al., 2011; Malone et al., 2011; Poggio and Gimona, 2017a, 2017b, 2014). With the progress of digital soil mapping, there is a rising use of 3D modelling to provide information on soil pattern for many applications, from agricultural management to ecosystem services (Zhang et al., 2017). It is important to deliver uncertainty associated with prediction, since it can help the land management choices (Poggio and Gimona, 2017a) and the decision makers (Poggio and Gimona, 2014).

However, the limiting factor is often the reduced amount of soil data used for the model calibration (Samuel-Rosa et al., 2015; Somarathna et al., 2017), and Somarathna et al. (2017) suggested that more data is more important than a better model. However, obtaining more data can be a problem because of a large size and/or accessibility of some test areas. To facilitate DSM in poor-accessible areas, Cambule et al. (2013) proposed a methodology of sampling in a small area of greater accessibility, which is representative of the total area, and to evaluate the representativeness using, e.g., the similarity between the histogram of the covariates for the total and accessible areas. Other studies considered the costs of accessibility in soil sampling (Carvalho Júnior et al., 2014; Clifford et al., 2014; Ließ, 2015; Roudier et al., 2012; Stumpf et al., 2016) using a variation/optimization of the method known as conditioned Latin Hypercube Sampling (cLHS), proposed by Minasny and McBratney (2006). The cLHS is a robust tool for the allocation of sampling points using a set of auxiliary covariates. The idea is to be able to capture the maximum of soil variation, and its properties, by using environmental covariates as auxiliary information.

The general goal of this paper was to create 2 and 3D soil properties map, using as examples, pH, carbon content and cation exchange capacity at fine resolution (25 m) in a poorly accessible area, the Itatiaia National Park (Brazil), with spatial uncertainty. The main objectives were: to design a sampling strategy balancing accessibility, costs, area and environmental covariates; and to model soil properties, with a limited number of samples available. The maps would provide necessary information for the analysis of environmental vulnerability in the INP.

INP has limited access, due to the steep slope, dense forest cover in forested areas or by rocky outcrops in the altitude field (Barreto et al., 2013). The INP is an excellent case study, because in order to obtain a viable result at a low cost, it is necessary to use the DSM tools, ranging from optimization of the sampling site (Minasny and McBratney, 2006; Roudier et al., 2012; Stumpf et al., 2016) to the covariate selection in powerful predictive algorithms (Beguin et al., 2017; Chagas et al., 2017; Jeong et al., 2017).

4.4 MATERIAL AND METHODS

4.4.1 Data sources and environmental covariates

The environmental covariates were derived from three data sources: digital elevation model, remote sensing data (orbital image) and geology map.

Digital Elevation Model (DEM): The DEM used, with a spatial resolution of 25 cm, was generated from the contour lines with 20 m equidistance and hydrography extracted from the plani-altimetric charts, both in the 1:50,000 scale. The sheets used were SF-23-ZA-I-2 Alagoa, SF-23-ZA-I-3 Passa Quatro and SF-23-ZA-I-4 Agulhas Negras. They were obtained from the in vector format from the cartographic base of Brazilian Institute of Geography and Statistics (IBGE). The dataset was provided by the INP administration.

Satellite image: Two scenes from the RapidEye sensor (2011) were used. They have 12-bit radiometric resolution, 6.5m spatial resolution, and were orthorectified to 5m spatial resolution (RapidEye, 2012). Both images were atmospherically corrected using the 6S model (Vermote et al., 1997). The details of processing can be found in Costa et al. (2016)

Geology map: The geology map was obtained from Santos et al. (2000), and it was scanned, vectorised and georeferenced. The file was rasterized at the same spatial resolution as the DEM (25m). The environmental covariates used to model the soil properties are listed in Table 8. They were chosen to describe the main soil forming factors, according to the *scorpan* approach (McBratney et al., 2003)

Table 8. Environmental covariates, soil formation factor that represents their sources, resolution, and definition.

Formation factor	Covariate	Source	Spatial resolution	Definition	Acronym
Organism (O)	Bands (1, 2, 3, 4, 5)	RapidEye (2011)	5 m	Bands in the spectrum of 440 – 510 nm (Blue), 520 – 590 nm (Green), 630 – 685 nm (Red), 690 – 730 nm (Red Edge), 760 – 850 nm (Near IR)	Bands (1,2,3,4, 5)
	Normalized difference vegetation index	RapidEye (2011)	5 m	NDVI=(NIR–Red)/(NIR+Red);	NDVI
	Soil-adjusted vegetation index	RapidEye (2011)	5 m	SAVI=(1+0.5)(NIR–Red)/(NIR+Red+0.5)	SAVI
Relief (R)	Digital elevation model	INP managers	25 m	Digital elevation model of the area- representation of the terrain's surface made by contour lines and hydrology (scale 1:50.00, IBGE data)	DEM
	Slope	DEM	25 m	The gradient or rate of change of elevation between neighboring cells	Slope
	Aspect	DEM	25 m	Attribute representing the exposure faces, represented by values in degrees ranging from 0 to 360 °	Aspect
	Northernness	DEM	25 m	Indicates the direction of the slope relative to the northern. Northernness =abs(180°−Aspect)	Northernness
	Plan curvature	DEM	25 m	The shape of the hillside on the horizontal plane (concave, rectilinear or convex)	Plan_curv
	Profile curvature	DEM	25 m	The shape of the hillside on the vertical plane (concave, rectilinear or convex)	Prof_curv
	Convergence index	DEM	25 m	The general shape of the hillside in all directions (concave, rectilinear or convex)	Convergence
	Catchment area	DEM	25 m	It is related to the volume of flooding that reaches a certain cell	
	Topographic wetness index	DEM	25 m	Describes a tendency for a cell to accumulate water	TWI

	LS factor	DEM	25 m	Attribute equivalent to topographic factor of Revised Universal Soil Loss Equation (RUSLE)	LS_factor
	Relative slope position	DEM	25 m	Represents relative slope position based on the base channel network	RSP
	Channel network distance	DEM	25 m	Altitude above the channel network (CHNB-original elevation)	CHND
	Channel network base level	DEM	25 m	Interpolation of a channel network base level elevation	CHNB
Parent material (P)	Geology	Santos et al., 2000	25 m	Categorical map with geological information (scale 1:50.000)	Geology
Spatial position (N)	X, Y	Grid data	-----	X=longitude, Y=latitude in UTM system, zone 23S, projection Sirgas 2000	X, Y
	XY	Grid data	-----	XY= polygon of second order of X and Y	XY

4.4.2 Covariates selection approach

In order to evaluate the relationships between soil properties as pH, total carbon content (C) and cation exchange capacity (CEC), and environmental covariates, Multiple Linear Regression (MLR), Random Forest (RF) and Generalized Additive Models (GAM) were tested (Figure 21).

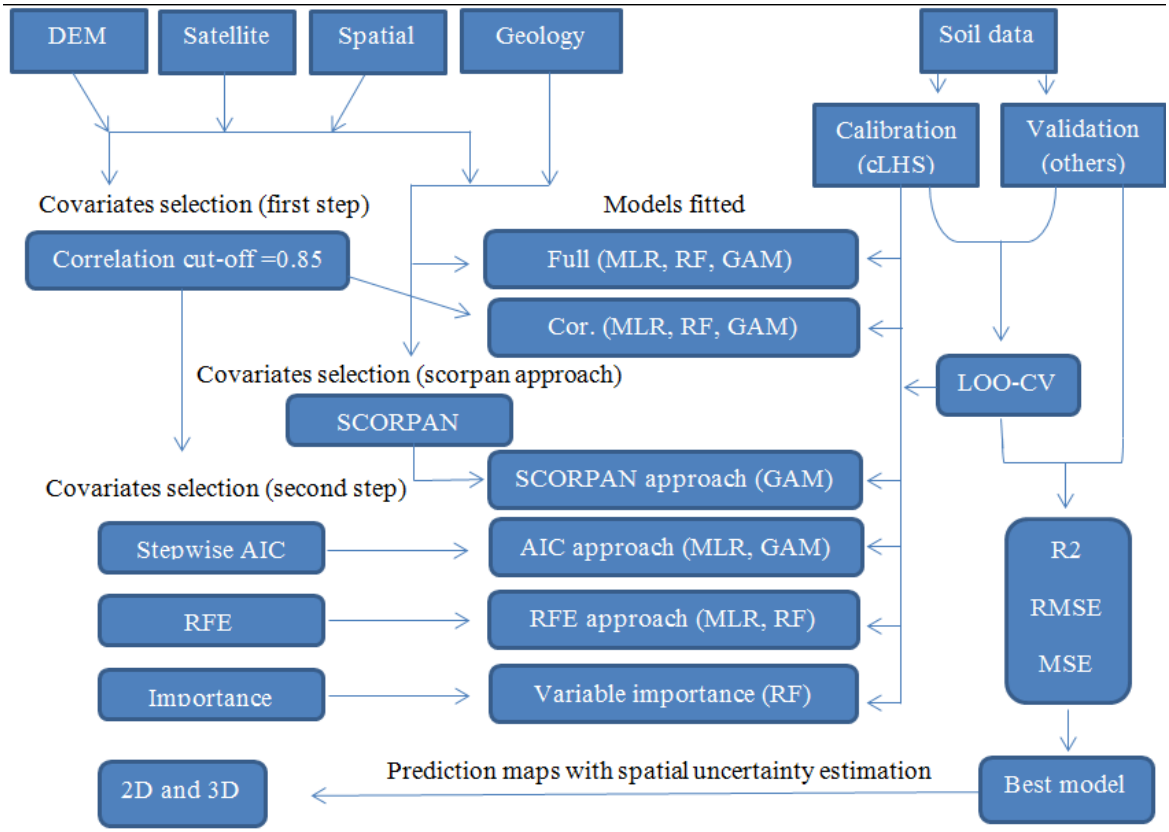


Figure 21. Covariate selection approach, model fitting, validation and prediction workflow.
Note: AIC= Akaike's information criterion; RFE= Recursive feature elimination; R^2= Coefficient of determination; RMSE=Root mean square error; MSE= Mean square error; LOO-CV= Leave-one-out cross validation

The MLR is a parametric method which assumes that the relationships between dependent variable and covariates are linear (Hastie et al., 2009). The RF techniques is a nonparametric method created by Breiman (2001) and it was based on a bootstrap aggregation (*bagging*) approach for reducing the variance of an estimated prediction function (Hastie et al., 2009). The GAM is a flexible statistical method that may be used to identify and characterize nonlinear regression effects through smoothing functions (Hastie et al., 2009; Wood, 2006).

The selection of the covariates available was carried out in order to produce simpler models with the minimum number of covariates, and still able to explain the maximum of the data variability. Different strategies were used and they are described in the sections below:

The first step evaluated the correlation between covariates. If two covariates had a correlation coefficient greater than 0.85 (the cut-off value considered for this study), only one was maintained. The covariate maintained in the model was the one that was believed to have a greater relationship with the *SCORPAN* (McBratney et al., 2003) model, i.e. greater pedological information and/or less correlation with other environmental covariates

It involved fitting models using the covariates maintained in the first step. But whereas fitted MLR and RF models were fitted with all covariates, the same was not possible for GAM,

due to limitation of degrees of freedom for many covariates and the few soil samples (Poggio et al., 2013). The purpose of the model with all covariates is to have a basis of comparison with different methods of selection commonly used**.**

a) MLR

4 models were fitted: with all covariates (MLR_full); with correlation selection inferior to 0.85 (MLR_cor); with the popular technique used in each method, stepwise AIC (Akaike's Information Criterion) (Carvalho Junior et al., 2016; Chagas et al., 2016; Meersmans et al., 2012; Samuel-Rosa et al., 2015; Vermeulen and Niekerk, 2017) in this case "forward" approach was used (MLR_step); and the technique of Recursive Feature Elimination (RFE) (MLR_RFE). This last has recently been used in soil science for variable selection in machine learning algorithm, and it is a backward selection using rank (Bachofer et al., 2015; Brungard et al., 2015; Jeong et al., 2017; Montanarella et al., 2013; Vašát et al., 2017).

b) RF

The same number of models was tested for RF that is: the full model (RF_full); correlation selection inferior to 0.85 (RF_cor), and RFE selection (RF_RFE). The variable importance, a popular technique for RF also was used (Carvalho Junior et al., 2016; Chagas et al., 2016; Rodriguez-Galiano et al., 2012; Were et al., 2015). In the RF case, the default set values have always been used, namely the number of trees equal to 500 (ntree = 500) and the number of covariates in each split equivalent to one third of the number of covariates. For example, model full mtry = 8, model with correlation selection mtry = 6 and importance of the variable mtry = 2. To build the RF with covariate selection by importance (RF_imp) the six most important covariates were always chosen, which is one-third of input covariates in RF whose importance was calculated (RF_cor).

c) GAM

The approach was different due to the degree of freedom limitation in GAM. The model´s degrees of freedom are calculated by adding up the degrees of freedom used by the parametric and non-parametric (or smooth) terms in the model, and it is not possible to fit the model if there are many covariates (terms) and a few points.

The models were fitted based on the stepwise forward approach, where covariates are added according AIC. All models began with geographic coordinates (X, Y) and geology as fixed covariates. Three models were fitted. The first using the base model, where each covariate was added in the base model individually, and then evaluated by its AIC. The model ran with all covariates and the four with lower AIC value composed the final model termed GAM_one. The second model consisted of making all possible combinations of four covariates and then run the model with all possible combinations. The combination with a smaller AIC was termed GAM_comb. This approach seeks to capture the interaction between covariates when a predictor model is fitted. In both cases (GAM_one and GAM_comb) it was included as many covariates as possible; since the base model already had 9 covariates X, Y and 7 different levels for geology, it was possible to include another 4 covariates totalling a model with a maximum of 13. The third model involved a more parsimonious model based on the *scorpan* approach (McBratney et al., 2003). In this case, in addition to the base model that already included the parent material (geology) and spatial position (X, Y), for 2D prediction and geology, and X, Y and depth (Z) for 3D prediction, different combinations of data derived from the satellite image (here representing the factor organism) and data derived from the DEM (mainly represent factor relief, topography) were tested.

In all, possible combinations were tested for each soil property using external validation, but the same procedure was repeated using cross validation. The best model in both evaluations

was selected and termed *GAM*_scorpan. A summary with all models and methods of covariate selection is presented in the Table 9.

Table 9. Summary of covariate selection method and fit for different prediction models

Models	All covariates	Cut-off only	Stepwise AIC	SCORPAN	RFE	Importance
MLR	X	X	X	------	X	------
RF	X	X	------	------	X	X
GAM	------	------	X	X	------	------

Note: The space in blank, the model does not apply

The GAM model was selected for the 3D approach due to its simplicity, and being a flexible approach that is able to deal with both linear and non-linear relationships between soil properties and the considered covariates (Poggio and Gimona, 2014). Also, the 3D smooth can provide better performance, considering non-linear relationships between covariates and soil properties (Poggio and Gimona, 2014), which are frequent when modelling natural environments.

4.4.3 Validation and uncertainty

The model's performance was evaluated in two ways. The first by external validation, where points selected by the cLHS n=74 soil profiles (Minasny and McBratney, 2006) were used to fit the models and the legacy data (retrieved from the literature), as well as data collected based on the pedological knowledge and the soil-landscape relationship (without pre-selection), were used to validate performance of the models n=16. In the training, samples were taken within a 100 m buffer in relation to roads and tracks. The validation samples include points inside and outside the buffer, defined as accessibility criterion. The second form of evaluation was leave-one-out cross-validation (LOO-CV) (Brus et al., 2011; Kempen et al., 2010; Samuel-Rosa et al., 2015). In both cases, the Mean Square Error (MSE) and Root Mean Square Error (RMSE) were computed. And a coefficient of determination was derived from a linear model between observed and predicted data (R2). For 3D soil mapping, the results of the modelling were summarized for the whole profile and at five depth layers, according to Global- SoilMap project specification (Arrouays et al., 2014), and compared with observed values from corresponding depths. Uncertainty propagation was analyzed through simulation (N=1000) from posterior distribution of the fitted GAMs to derive simultaneous confidence intervals for the derivatives of penalized splines (Ruppert et al., 2003).

4.4.4 Software used

ArcGis 10.2.2 (ESRI, 2015) was used for geology map preparation; Spring 5.2.5 (Câmara et al., 1996) and 6S (Costa et al., 2016; Vermote et al., 1997) for atmospheric correction of the satellite image.

The R software (R Core Team, 2018) was used for the covariates preparation and statistical modelling. The following packages were selected: *clhs* for conditioned Latin hypercube sampling (Roudier, 2015) *raster, rgdal, maptools* and *RSAGA* for data management, preparation and visualisation (Bivand et al., 2017; Bivand and Lewin-Koh, 2017; Brenning, 2008; Hijmans, 2016); *mgcv* for GAM (Wood, 2006), *randomForest* for RF (Liaw and Wiener, 2002); *caret* (Kuhn et al., 2017) for MLR, RF and GAM using cross validation (Kuhn et al., 2017); *aqp* for preparing soil depth function on the 3D data validation (Beaudette et al., 2013).

4.5 RESULTS AND DISCUSSION

4.5.1 Correlation analysis

It was observed a strong relationship between covariates derived from satellite image (Figure 22), most of them with a correlation greater than 0.85. They contributed in a similar way as information of vegetal coverage or land use, and their use may impair the model's fitness due to multicollinearity problems (Hengl, 2009; Kempen et al., 2009; ten Caten et al., 2011).

The covariates chnb, band1, band2, band3 and SAVI were excluded, due to a correlation greater than 0.85 with one or more covariates (Figure 22) were excluded those with less pedological information and/or more correlation with other environmental covariates and or easier to calculate manually. chnb showed a strong correlation with elevation values. One of the main reasons that lead to selection of covariates, excluding those with high correlation, is the fact that, mainly in regression models, it can jeopardize the prediction due multicollinearity among covariates (Nussbaum et al., 2018; Somarathna et al., 2017). This leads to the problem of inflating the variance of parameters, model over-fitting and even noise problems (Hengl, 2009; Kempen et al., 2009; Nussbaum et al., 2018; ten Caten et al., 2011).

It is usually necessary to decrease number of covariates in the GAM models, especially when there are a limited number of soil samples, as in this study. Since there was no high relation (no greater than the cut-off value 0.85) between covariates derived from satellite image and DEM (Figure 22), all possible combinations were tested to build the *scorpan* GAM model.

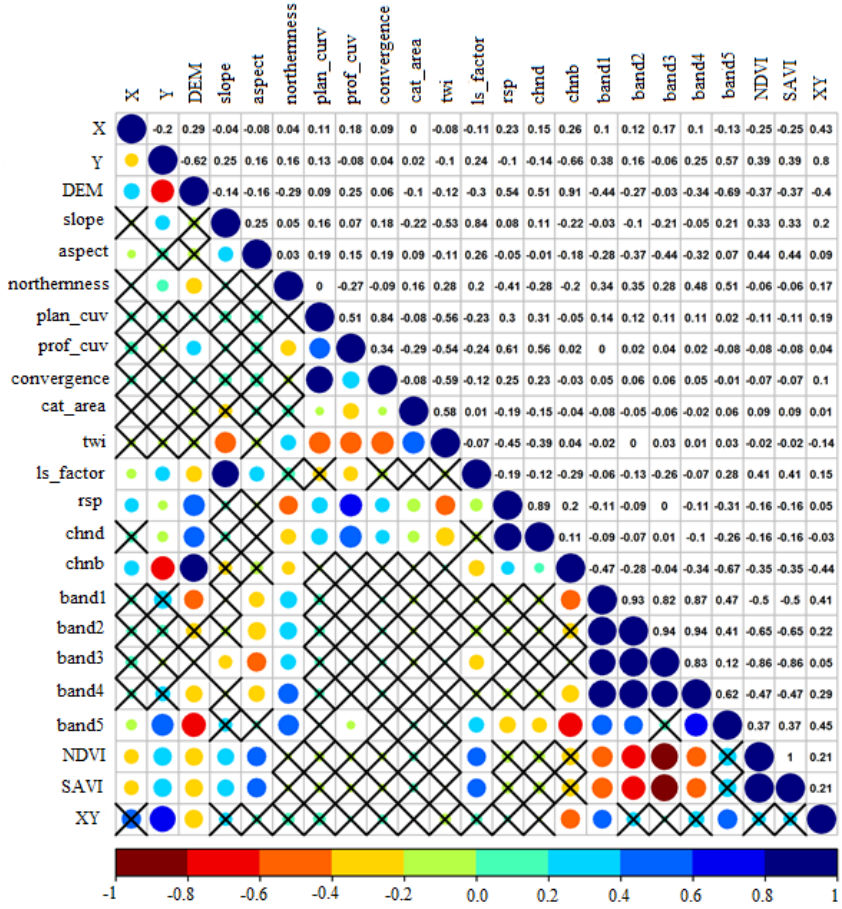


Figure 22. Matrix of correlation between environmental covariates. Correlations with "X" are not significant at 5% of confidence.

4.5.2 2D approach

4.5.2.1 Model comparison by soil attribute

For the prediction of soil pH, the RF models in top soil layer (2D approach), independent of covariate selection method, presented the worst performance compared to models MLR and GAM (Table 10) the worst RF result for external validation is related to the low representativity of the validation sample, both in the attribute space (range of variation of the soil attribute) and in the geographic space. This is because much of the external validation data is legacy data that is concentrated in a region of the park and there is a predominance of organic soils as described by Soares et al. (2016) and Silva Neto et al. (2018), which are commonly acidic with pH values that varies very little between profiles. In the specific case of soil pH, which presented the greatest difference between external and cross-validation, it is more prudent to use cross-validation to select the best mode.

For pH prediction using linear models, the best method of covariate selection was the RFE, in both cases, using external e cross-validation. The same was observed for RF models for all attributes, where the best method of covariate selection was the RFE (Tables 10, 11 and 12), despite the marginal difference between results of different covariate selection methods for RF models.

Table 10. Performance of MLR, RF and GAM models to predict soil pH

		External validation			**LOO-CV**		
Model	Cov	R^2	MSE	RMSE	R^2	MSE	RMSE
MLR_full	*	0.15	0.656	0.430	0.26	0.433	0.188
MLR_cor	**	0.21	0.623	0.388	0.30	0.406	0.165
MLR_step	1	0.22	0.477	0.228	0.24	0.423	0.179
MLR_rfe	2	**0.32**	**0.446**	**0.199**	**0.31**	**0.389**	**0.152**
RF_full	*	0.04	0.512	0.262	0.25	0.405	0.164
RF_cor	**	0.07	0.496	0.246	0.25	0.404	0.163
RF_imp	3	0.03	0.530	0.281	0.28	0.397	0.158
RF_rfe	4	**0.09**	**0.486**	**0.236**	**0.29**	**0.392**	**0.154**
GAM_one	5	0.20	0.484	0.234	0.35	0.380	0.144
GAM_comb	6	0.50	0.392	0.154	0.31	0.391	0.153
GAM_scorpan	7	**0.52**	**0.391**	**0.153**	**0.35**	**0.378**	**0.143**

Note: The best performing models are in bold. Cov = Covariates used in each model for predict soil pH
*= All covariates (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnd, chnb, band1, band2, band3, band4, band5, NDVI, SAVI, geology, XY). **= Covariates selected with correlation smaller than 0.85 with each other's (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnb, band4, band5, NDVI, geology, XY). **1**= Covariates selected in MLR by stepwise selection (X, aspect, DEM, geology, convergence, slope, band4, chnd). **2**= Covariates selected in MLR by RFE selection (plan_cuv, prof_cuv, NDVI, Y, convergence, ls_factor, X, twi, XY). **3**= Covariates selected in RF by covariate importance (X, DEM, XY, Y, aspect, NDVI). **4**= Covariates selected in RF by RFE (X, DEM, XY, Y, NDVI, geology, aspect). **5**= Covariates selected in GAM by GAM_one approach (X, Y, geology, aspect, band4, convergence, cat_area), **6**= Covariates selected in GAM by GAM_comb approach (X, Y, geology, slope, aspect, convergence, band4), **7**= Covariates selected in GAM by GAM_scorpan approach (X, Y, geology, convergence, band2). The numbers indicate the covariates that were selected in the models fitted with cross validation approach.

For the GAM models, the best method of covariate selection was the *scorpan* approach, in both cases, using external e cross validation. It was also the best model when compared with different models and approaches to select covariates. A different result was found by Jeong et al. (2017), which found that RF was better than GAM to predict soil carbon, nitrogen and phosphorus. For the total soil carbon content, the RFE method was the best for covariates

selection for RF, but for the linear model, it had the worst performance (Table 11). This corroborates Jeong et al. (2017) where the selection by RFE in the RF model improved the prediction, although with a small difference compared with methods tested especially for MLR. In the case of the linear model, the most commonly used method for covariate selection, stepwise, (Bhering et al., 2016; Chagas et al., 2016; Forkuor et al., 2017; Somarathna et al., 2017), was the best performing; both when evaluating external validation and cross-validation. The *scorpan* model remained the best approach for covariate selection in GAM models. For prediction of soil carbon content, the RF was better than the linear models, regardless of method of covariate selection. Again, the RF was barely affected by method of selection of covariates, and the results were similar.

Table 11. Performance of MLR, RF and GAM models to predict soil carbon content

		External validation			**LOO-CV**		
Model	Cov	R^2	MSE	RMSE	R^2	MSE	RMSE
MLR_full	*	0.06	7.556	57.095	0.14	6.548	42.874
MLR_cor	**	0.10	6.705	44.953	0.13	6.44	41.469
MLR_step	1	**0.17**	**5.289**	**27.97**	**0.24**	**5.253**	**27.591**
MLR_rfe	2	0.04	5.437	29.56	0.09	5.440	29.593
RF_full	*	0.24	4.804	23.082	0.38	4.486	20.128
RF_cor	**	0.25	4.653	21.651	0.36	4.557	20.769
RF_imp	3	0.23	4.776	22.811	0.40	4.390	19.276
RF_rfe	4	**0.24**	**4.683**	**21.931**	**0.40**	**4.374**	**19.131**
GAM_one	5	0.33	3.949	15.593	0.42	4.354	18.957
GAM_comb	6	0.31	4.087	16.706	0.43	4.308	18.558
GAM_scorpan	7	**0.49**	**3.851**	**14.834**	**0.45**	**4.212**	**17.742**

Note: The best performing models are in bold. Cov = Covariates used in each model for predict soil carbon content *= All covariates (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnd, chnb, band1, band2, band3, band4, band5, NDVI, SAVI, geology, XY). **= Covariates selected with correlation smaller than 0.85 with each other's (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnb, band4, band5, NDVI, geology, XY). **1**= Covariates selected in MLR by stepwise selection (DEM, Northernness, geology, X, NDVI). **2**= Covariates selected in MLR by RFE selection (plan_curv, prof_curv, NDVI, Y, X, twi). **3**= Covariates selected in RF by covariate importance (DEM, northernness, Y, chnd, geology, XY). **4**= Covariates selected in RF by RFE (DEM, Y, Northernness, geology, band4, chnd, band5, XY). **5**= Covariates selected in GAM by GAM_one approach (X, Y, geology, DEM, northernness, chnd, band5). **6**= Covariates selected in GAM by GAM_comb approach (X, Y, geology, DEM, aspect, plan_curv, cat_area). **7**= Covariates selected in GAM by GAM_scorpan approach (X, Y, geology, prof_cuv, band3). The numbers indicate the covariates that were selected in the models fitted with cross validation approach.

For pH (Table 10) and Carbon content (Table 11) the prediction using the RF_rfe showed greater performance for models that used the cross-validation method. On another hand, in the CEC prediction (Table 12) the performance was higher when using the external validation. CEC and carbon content presented similar behaviour, where RF with the RFE selection method had the best performance, and for the linear model it was the stepwise selection (Table 11 and 12).

Table 12. Performance of MLR, RF and GAM models to predict soil cation exchange capacity

	External validation			LOO-CV		
Model	R^2	MSE	RMSE	R^2	MSE	RMSE
MLR_full	0.20	14.956	223.691	0.05	12.915	166.805
MLR_cor	0.18	15.281	233.500	0.03	13.254	175.669
MLR_step	**0.19**	**14.777**	**218.369**	**0.04**	**11.179**	**124.972**
MLR_rfe	0.00	16.781	281.609	0.02	9.283	86.167
RF_full	0.30	13.283	176.432	0.25	7.823	61.205
RF_cor	0.27	13.629	185.747	0.17	8.275	68.478
RF_imp	0.29	13.344	178.058	0.26	7.832	61.335
RF_rfe	**0.36**	**13.153**	**173.007**	**0.28**	**7.657**	**58.624**
GAM_one	0.38	13.708	187.917	0.22	8.289	68.715
GAM_comb	0.32	13.168	173.406	0.17	8.581	73.626
GAM_scorpan	**0.41**	**13.604**	**185.056**	**0.27**	**7.764**	**60.285**

Note: The best performing models are in bold. Cov = Covariates used in each model for predict soil carbon content *= All covariates (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnd, chnb, band1, band2, band3, band4, band5, NDVI, SAVI, geology, XY). **= Covariates selected with correlation smaller than 0.85 with each other's (X, Y, DEM, slope, aspect, Northernness, plan_curv, prof_curv, convergence, cat_area, twi, ls_factor, rsp, chnb, band4, band5, NDVI, geology, XY). **1**= Covariates selected in MLR by stepwise selection (band5, northernness, DEM, X, chnd, geology). **2**= Covariates selected in MLR by RFE selection (plan_curv, prof_curv, NDVI, ls_factor, twi, slope, convergence). **3**= Covariates selected in RF by covariate importance (band5, northernness, DEM, Y, NDVI, XY). **4**= Covariates selected in RF by RFE (DEM, band5, northernness, NDVI, geology, Y, XY). **5**= Covariates selected in GAM by GAM_one approach (X, Y, geology, band5, northernness, DEM, NDVI). **6**= Covariates selected in GAM by GAM_comb approach (X, Y, geology, plan_curv, twi, band5, NDVI). **7**= Covariates selected in GAM by GAM_scorpan approach (X, Y, geology, chnb, band3). The numbers indicate the covariates that were selected in the models fitted with cross validation approach.

When analysing the models separately according to their respective methods of covariates selection, it is observed that, in general, the best performance results use stepwise selection for MLR (except for predicting soil pH), RF using RFE, and GAM using the *scorpan* approach. It is possible to separate the models and covariates selection approach. This appears to contradict Somarathna et al. (2017), who suggested investing in sampling rather than more robust models. However, it agrees with Beguin et al. (2017), who tested different statistical approaches and found significant differences, thus suggesting that robust methods can enhance DSM capabilities and support existing efforts for improving digital soil products, even with limited data.

For CEC prediction, all the best models in each approach had better performance when used external validation for evaluating. However, the validation samples are not completely random, despite as suggested by Brus et al. (2011), and this may overestimate the model's performance for the external validation approach. Regardless of the differences, the best model selected in the external validation was also the best model in the cross-validation.

4.5.2.2 Model summary

The RF model presented regular performance, usually better than MLR and worse than GAM. As for selection process, optimization of parameters in the RFE showed superior performance of the RF using this method. Although, in general, RF was not very sensitive to

selection of covariates. This corroborates Díaz-Uriarte and Andrés (2006) and Grimm et al. (2008), in which the algorithm is robust enough for overfitting, since each tree is trained on a bootstrap subsample of the data (Arun and Langmead, 2005; Grimm et al., 2008; Nguyen et al., 2013). In other words, the covariates are not used all at once, but via a group of them in each bootstrap sample.

Therefore, it is suggested that in the case of RFE for RF, the best performance of the model is related to optimization of selection of covariates and optimization of RF parameters. For example, the number of divisions of each tree, *mtry*. In the case of other selection methods, the *mtry* parameter was always the default, 1/3 of the covariates. In the RFE this parameter is optimized using LOO-CV validation, thus selecting an optimal value. Even so, the difference between selection methods is small, and it does not follow the pattern observed in the linear models.

The MLR with several covariates showed a tendency to have worse performance, because its effect of harmful multicollinearity in the parametric models; thereby impairing its fitting (Hengl, 2009; Kempen et al., 2009; ten Caten et al., 2011). For linear models the RFE selection method (using the RFE *lmFuncs* function) did not present good results; it was the model with the worst result, except for soil pH. Furthermore, it almost always selects the same covariates, regardless of the soil attributes tested. This suggests that selection algorithm using the function for linear models in the *caret* package should be used carefully. In general, the best way to select covariates for MLR is the stepwise selection with AIC criteria, a common method to select models in linear regression (Carvalho Junior et al., 2016; Chagas et al., 2016; Meersmans et al., 2012b; Samuel-Rosa et al., 2015; Vermeulen and Niekerk, 2017).

The GAM_scorpan was the most appropriate model for prediction of all soil's attributes. It presented the best performance, in both ways, when evaluated using external validation and with cross-validation (Tables 10, 11, 12). However, this was the only model where, in all soil attributes, performance in the cross-validation was lower than in the external validation. This result may be due to the fact that external validation samples (mostly legacy data) are not random and do not overlap the geographic and attribute space of the data variation.

For the three soil attributes and selection methods, the linear models showed inferior performance to GAM. This is probably because relationships between soil attributes and covariates are not linear, and models such as the MLR fail to capture the nonlinear relationships efficiently (Guo et al., 2015; Jeong et al., 2017; Lagacherie et al., 2013; Poggio et al., 2013; Poggio and Gimona, 2014). In contrast, GAM models, where it is possible to model nonlinear relationships (Chartin et al., 2017; de Brogniez et al., 2015; Jeong et al., 2017; Poggio et al., 2013; Poggio and Gimona, 2017a), had the best performance when combined a nonlinear modelling approach and the concept of soil formation factors for covariate selection (expert knowledge).

Pedological knowledge was used by Nussbaum et al. (2018) to exclude covariates with low spatial variation, and aggregate levels of categorical variables with low sample density per level. The knowledge of soil forming factors as well as of study area is a powerful tool, and when associated with computational tools it may improve predictions of soils properties and classes.

4.5.2.3 Soil formation factors and pedological elicitation

When the importance of the variable is considered in the RF models, the RF_cor, for just to see the more important covariates (Figure 23), elevation (DEM) is a factor deemed highly important for the three evaluated attributes. It is always listed in the top three when evaluated in the MSE (mean squared error).

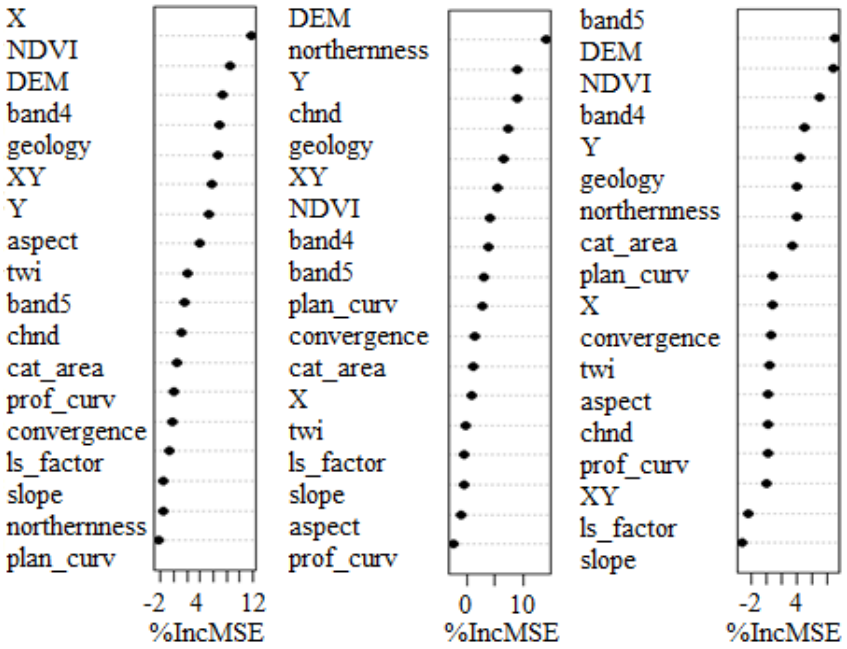


Figure 23. Importance of the environmental covariates derived from the RF for pH (left), soil carbon content (middle) and CEC (right). %IncMSE— % increase in mean squared error

Similarly, the geological material always appears as a relevant covariate in the RF model, and for all soil attributes elevation and geology are included in the RF_imp model; that is to say, they are among the six most significant covariates. For pH, the spatial component was a relevant factor in which both X and Y coordinates were selected and also the covariate XY. For soil carbon content the Y coordinate and XY covariates were selected among the six most important ones. For CEC only, the Y was selected to represent spatial information among the six most important covariates. At least one attribute derived from satellite image was selected, via pH and CEC prediction, usually NDVI and band4. For soil carbon content, attributes from satellite image were placed as seventh and eighth (NDVI and band4, respectively) (Figure 23). The correlation analyses show a direct relationship between elevation and the soil attributes, with values of 0.35, 0.59 and 0.45 for pH, C and CEC, respectively; and a negative correlation with the reflectance (for example the Red Edge band) with values of -0.16, -0.48, and -0.59 for pH, C and CEC, respectively (Figure 17).

All soil factors are related, for example, the relief has spatial variation, and the highest part is in the centre; in turn, the elevation influences the weather, that is cold and wet in the PNI, and that leads to a distinct distribution of plant species. This environment favours accumulation and preservation of soil organic matter, due to low temperatures, leading to formation of the organic soils in the high altitudes (Benites et al., 2007; Soares et al., 2016).

This agreed with the results found in the GAM_scorpan models, the models selected as the best since they combine the most important covariates, related to parent material, relief, organism and position in the geographic space. For example, the higher carbon and CEC contents, attributes strongly related to each other, were predicted with highest values in the areas of INP with altitudinal fields coverage (predominant species are Poaceae and Cyperaceae), which are concentrated in the plateau central region, where the dominant geology is composed by quartz-syenites and related sediments (Santos et al., 2000; Soares et al., 2016).

4.5.2.4 Spatial prediction and uncertainty propagation

When the predicted values (in the grid) and the observed values were compared, it was observed a tendency of the MLR to extrapolate results, especially lower values being more negative; as in the carbon and CEC (Table 13 and Figures S2, S3 and S4). The RF model focuses on values closer to the mean and with less amplitude, which is a characteristic of this model, that have not capacity for extrapolation in the attribute space (Table 13 and Figures S2, S3 and S4). Similar results were observed by Chagas et al. (2016) and Carvalho Junior et al. (2016), when comparing RF and MLR.

The GAM models placed between RF and MLR regarding ability to interpolate and extrapolate. For carbon content, the extreme values (minimum and maximum) were similar to prediction made by MLR, with maximum values close to measured values, but the minimum showed negative values (Table 13).

Table 13. Descriptive statistics of the observed values (original data) and predicted values (grid) for soil attributes using the best covariate selection approach for MLR, RF and GAM models

pH	Min	Mean	Max
Original data	3.24	4.4	5.49
MLR_rfe	0.69	4.22	7.19
RF_rfe	3.63	4.32	4.99
GAM_scorpan	2.33	4.21	5.61
Total carbon			
Original data	3.13	11.3	27.8
MLR_step	-4.31	9.79	22.78
RF_rfe	4.29	10.23	18.95
GAM	-4.14	8.98	23.54
CEC			
Original data	9.39	19.39	69.01
MLR_step	-4.61	16.47	31.57
RF_rfe	11.37	17.82	46.57
GAM_scorpan	1.38	16.67	36.63

Despite the best performance of the GAM model, evaluated by the R2, RMSE and MSE metrics, there was an extrapolation of values predicted in the grid. This reinforces the importance of evaluating the spatial propagation of uncertainty in DSMs (Merrill et al., 2016; Poggio and Gimona, 2014; Stumpf et al., 2016; Truong and Heuvelink, 2013; Vaysse and Lagacherie, 2017). The uncertainty in the predictions of soil attributes for the superficial layer was mainly associated with extrapolation of values for regions not sampled, and the INP boundaries with greater limitation of access. Besides some geology classes occurred in small areas, consequently they had fewer soil samples (Figure 24). A similar pattern was observed by Cambule et al. (2014), predicting carbon stocks in the Limpopo National Park, where they observed high uncertainty values and it was suggested that it was due to short-range spatial structure combined with the sparse sampling.

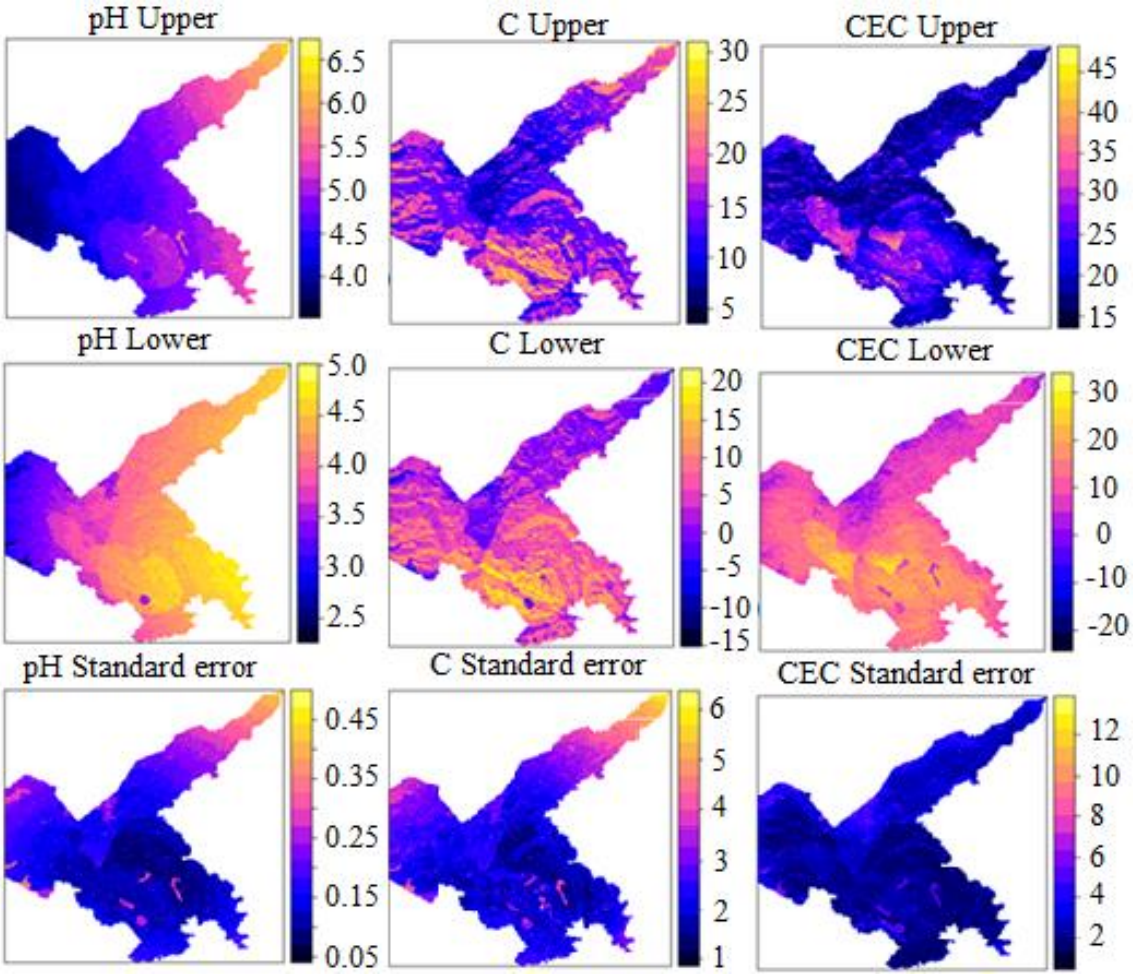


Figure 24. Standard error lower and upper values derived from a Bayesian posterior distribution of each GAM_scorpan model fitted.

4.5.3 3D approach

Three-dimensional quantitative modelling is relatively new in soil science especially in Brazil. Soil prediction in 3D space it is nothing more than prediction in three-dimensional space, in other words, prediction in 2D space (Latitude [Y] and Longitude [X]) plus soil depth (Z) or commonly described as a prediction for the whole profile (Poggio and Gimona, 2017b). And is possible by using manly continuous depth functions in advanced algorithms (Malone et al., 2009) or geostatistical interpolation using 3D variogram (Poggio and Gimona, 2017a).

To run 3D modelling using continuous depth functions, is needed to add depth of horizons as covariate in the model and for that is used a small function to assign depth values as the center depth of each horizon. As it is known where the horizons start and end, using the mentioned function is known at what depths the values of the properties change and consequently, it is possible to predict the target soil property in any depth using in the grid the central point of the horizon to be estimated, see Hengl and MacMillan (2019) for more information. Normally the predictions are done according to Global- SoilMap project specification, 0-5, 5-15, 5-30, 30-60, 60-100 and 100-200 cm (Arrouays et al., 2014).

4.5.3.1 Continuous depth function using GAM scorpan

The Figure 25 shows the vertical distribution of size fractions for each order of soil, according to the slice in layers of 1 cm and harmonized in five layers of predefined depths (0-

5, 5-15, 5-30, 30-60, 60-100 cm). The dark blue line represents the mean value of the soil attribute and light blue spot represents the 25 and 75 percentiles of the attribute at a given depth.

Based on results for topsoil layer prediction, the GAM *scorpan* approach was chosen to predict the soil attributes for the whole profile. In this case, besides the base model of 2D GAM with covariates X, Y and geology, the soil depth (Z) was added as a covariate in the base model to create a smoother 3D (Poggio and Gimona, 2014). As with 2D modelling, the base model was used to test different combinations of attributes derivate from DEM and satellite image. Since most of the soils in the INP are shallow, it was considered for prediction up to 100 cm depth. A greater depth than that represents less than 23% of the total data (Figure 25).

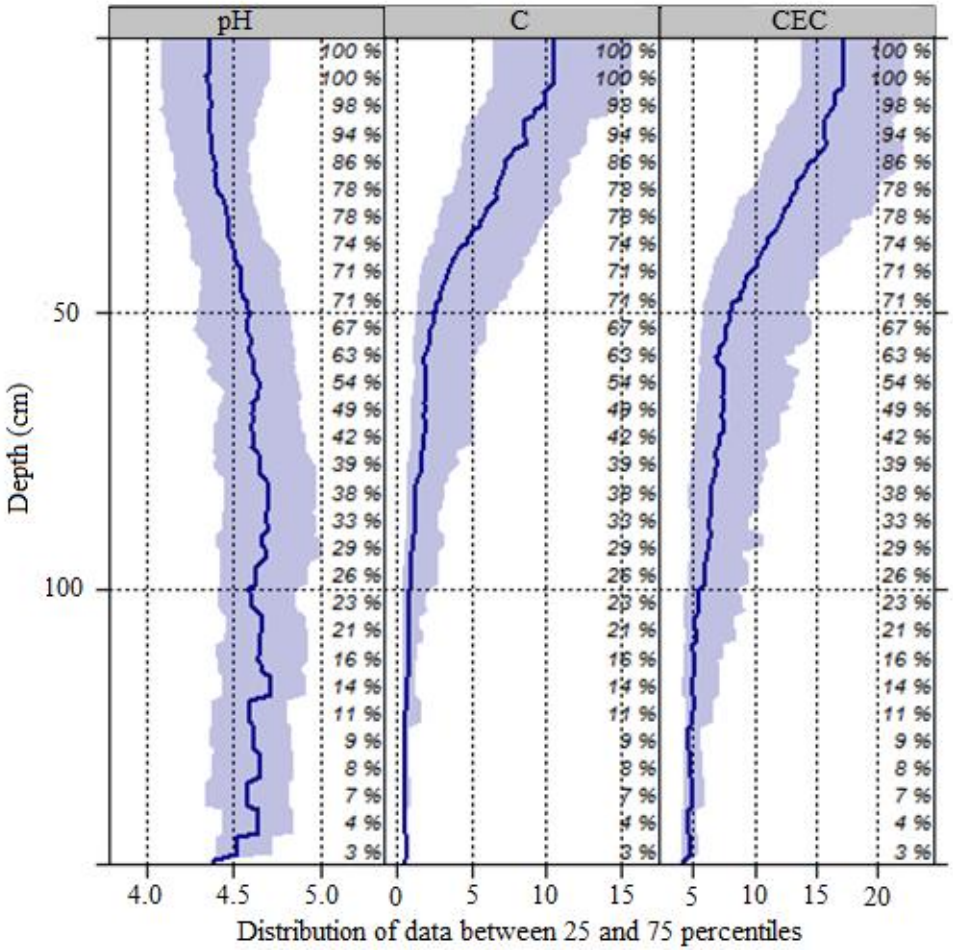


Figure 25. Distribution of pH, Carbon content (%), and CEC (cmolc.dm^{-3}) for the data collection. The percentage values represent the relative number of profiles that contributed to the estimates in each layer.

4.5.3.2 Model evaluation

The descriptive statistic for the whole soil profile was evaluated, and the predictions values for soil pH are very close (Tables 14) to the observed (Table 16), especially when using the cross-validation approach (Table 15). Values of determination coefficient for carbon content and CEC are higher among observed and predicted values than for pH, especially in cross-validation (Table 15). The magnitude of errors, RMSE and MSE, shows a tendency to extrapolate low carbon contents.

Table 14. Descriptive statistics of predicted values for the whole profile using external validation dataset

Attribute	R^2	RMSE	MSE	Min	Mean	Max
pH	0.27	0.384	0.148	3.80	4.49	5.00
C	0.26	5.729	32.820	-2.98	10.53	22.58
CEC	0.42	10.749	115.540	6.61	17.67	31.52

Table 15. Descriptive statistics of predicted values for the whole profile using cross-validation

Attribute	R^2	RMSE	MSE	Min	Mean	Max
pH	0.45	0.294	0.087	3.42	4.51	5.14
C	0.60	3.633	13.202	-3.81	6.43	20.96
CEC	0.59	5.947	35.362	-0.03	13.65	46.95

Although there is a positive relationship between predicted and observed carbon values, they decrease in depth (Figure 25) and there is a tendency for values to become very low for depths greater than 30 cm (Tables 17 and 18). This is especially true for soils that begin with low levels of soil carbon, such as mineral soils, which have almost 0% of soil carbon at the greatest depth (Figure S5). Particularly in the deeper layers, the low number of points to represent these layers is a factor that affects the prediction (Tables 17 and 18). The same results were observed by (Kempen et al., 2011; Mulder et al., 2016b; Poggio and Gimona, 2017a) where the performance was much better for the top layer than for subsurface layers.

Table 16. Descriptive statistics of complete and validation dataset

Attribute	All data			Data validation		
	Min	Mean	Max	Min	Mean	Max
pH	3.24	4.51	5.72	3.72	4.69	5.46
C	0.24	6.42	29.48	0.43	7.95	17.46
CEC	3.00	13.68	69.01	4.35	19.04	69.01

The form of the model evaluation, external data or cross validation, leads to different suggestions of better result by depth. In external validation, the best performance was in the 30-60 cm layer for all attributes, despite sub estimation of carbon content (Table 17). For cross-validation, in which better results were obtained, the best performance was for 5-15 cm, for pH and C, and 60-100 cm for CEC (Table 18).

Table 17. Descriptive statistics of predicted values for each depth using external validation dataset.

Attribute	R^2	RMSE	MSE	Min	Mean	Max	Depth (cm)	n*	n**
pH	0.31	0.45	0.203	3.78	4.39	5.08	0-5	90	16
	0.33	0.439	0.192	3.81	4.41	5.00	5-15	90	16
	0.29	0.385	0.148	3.87	4.45	4.88	15-30	85	16
	0.28	0.334	0.112	4.00	4.50	4.88	30-60	73	14
	0.42	0.302	0.091	4.14	4.57	5.09	60-100	51	10
C	0.15	5.437	29.564	2.64	14.31	22.90	0-5	90	16
	0.14	5.216	27.203	2.02	13.38	21.93	5-15	90	16
	0.11	5.708	32.582	1.02	11.85	20.34	15-30	85	16
	0.49	3.306	10.929	-0.41	8.38	12.95	30-60	73	14
	0.32	4.430	19.624	-1.80	6.97	13.86	60-100	51	10
CEC	0.29	12.978	168.417	14.47	22.61	33.27	0-5	90	16
	0.31	13.171	173.471	13.62	21.59	32.08	5-15	90	16
	0.41	12.707	161.461	12.11	19.84	30.08	15-30	85	16
	0.61	10.420	108.579	9.01	16.60	26.62	30-60	73	14
	0.52	7.214	52.045	7.52	14.78	23.44	60-100	51	10

Note: n* number of observations in all data for each depth, and n** number of observations in validation dataset for each depth.

Table 18. Descriptive statistics of predicted values for each depth using LOO-CV dataset.

Attribute	R^2	RMSE	MSE	Min	Mean	Max	Depth (cm)	n*	n**
pH	0.43	0.354	0.126	3.78	4.39	5.08	0-5	90	16
	0.46	0.332	0.110	3.81	4.41	5.00	5-15	90	16
	0.41	0.299	0.089	3.87	4.45	4.88	15-30	85	16
	0.32	0.272	0.074	4.00	4.50	4.88	30-60	73	14
	0.38	0.282	0.079	4.14	4.57	5.09	60-100	51	10
C	0.32	9.416	88.661	2.64	14.31	22.90	0-5	90	16
	0.35	9.504	90.328	2.02	13.38	21.93	5-15	90	16
	0.30	9.805	96.145	1.02	11.85	20.34	15-30	85	16
	0.27	9.907	98.139	-0.41	8.38	12.95	30-60	73	14
	0.28	9.027	81.489	-1.80	6.97	13.86	60-100	51	10
CEC	0.40	7.152	51.150	14.47	22.61	33.27	0-5	90	16
	0.41	7.154	51.173	13.62	21.59	32.08	5-15	90	16
	0.52	6.805	46.313	12.11	19.84	30.08	15-30	85	16
	0.58	6.292	39.594	9.01	16.60	26.62	30-60	73	14
	0.65	4.360	19.012	7.52	14.78	23.44	60-100	51	10

Note: n* number of observations in all data for each depth, and n** number of observations in validation dataset for each depth.

4.5.4 3.3.3 Spatial prediction and uncertainty propagation

When comparing the spatial prediction of the models using external and cross-validation (CV), the first had worse performance (Table 19). This is related, among other factors, to limitations in access, consequently, smaller number of points to calibrate the models.

Table 19. Descriptive statistics for grid values prediction using external validation model and LOO-CV model

Attribute	Depth (cm)	External validation			LOO-CV		
		Min	Mean	Max	Min	Mean	Max
pH	0-5	3.78	4.39	5.08	3.78	4.39	5.08
	5-15	3.81	4.41	5.00	3.81	4.41	5.00
	15-30	3.87	4.45	4.88	3.87	4.45	4.88
	30-60	4	4.5	4.88	4.00	4.5	4.88
	60-100	4.14	4.57	5.09	4.14	4.57	5.09
C	0-5	-54.62	8.83	34.89	-6.29	11.16	24.38
	5-15	-55.29	7.9	33.43	-6.89	10.04	22.97
	15-30	-56.41	6.41	31.14	-8.31	8.26	20.81
	30-60	-58.18	4.24	30.01	-10.26	5.67	19.18
	60-100	-58.34	2.98	30.47	-10.59	4.01	19.19
CEC	0-5	14.47	22.61	33.27	14.47	22.61	33.27
	5-15	13.62	21.59	32.08	13.62	21.59	32.08
	15-30	12.11	19.84	30.08	12.11	19.84	30.08
	30-60	9.01	16.6	26.62	9.01	16.60	26.62
	60-100	7.52	14.78	23.44	7.52	14.78	23.44

Thus, the cross-validation seems to be more appropriate when there is a limited number of samples to fit the 3D function (Amirian Chakan et al., 2017; Taghizadeh-Mehrjardi, 2016; Taghizadeh-Mehrjardi et al., 2016). Even if more data are available, this is a common approach used in recent 3D soil properties modelling (Mulder et al., 2016a, 2016b; Veronesi et al., 2014).

The evaluation of spatial uncertainty propagation showed a greater uncertainty for the higher values given by the standard error (Figure 26), due to noise in the very steep areas (above 100%) and/or shadow detected in the satellite images. Also, areas that had greater uncertainty occurred when the predicted values are associated with higher access limitation, with little or no soil samples (see Figure 2 and 26).

Most soil attributes were produced with acceptable modelling diagnostics and uncertainty ranges, delivering realistic soil–landscape spatial patterns, extrapolated in limited access areas, especially for soil carbon content in deeper layers (Table 19). The same results were observed by Kidd et al. (2015), suggesting that maps should be produced with continuous improvements, from the input of newly collected data. Prediction uncertainty can help to choose supplemental sampling to improve the DSM (Li et al., 2016).

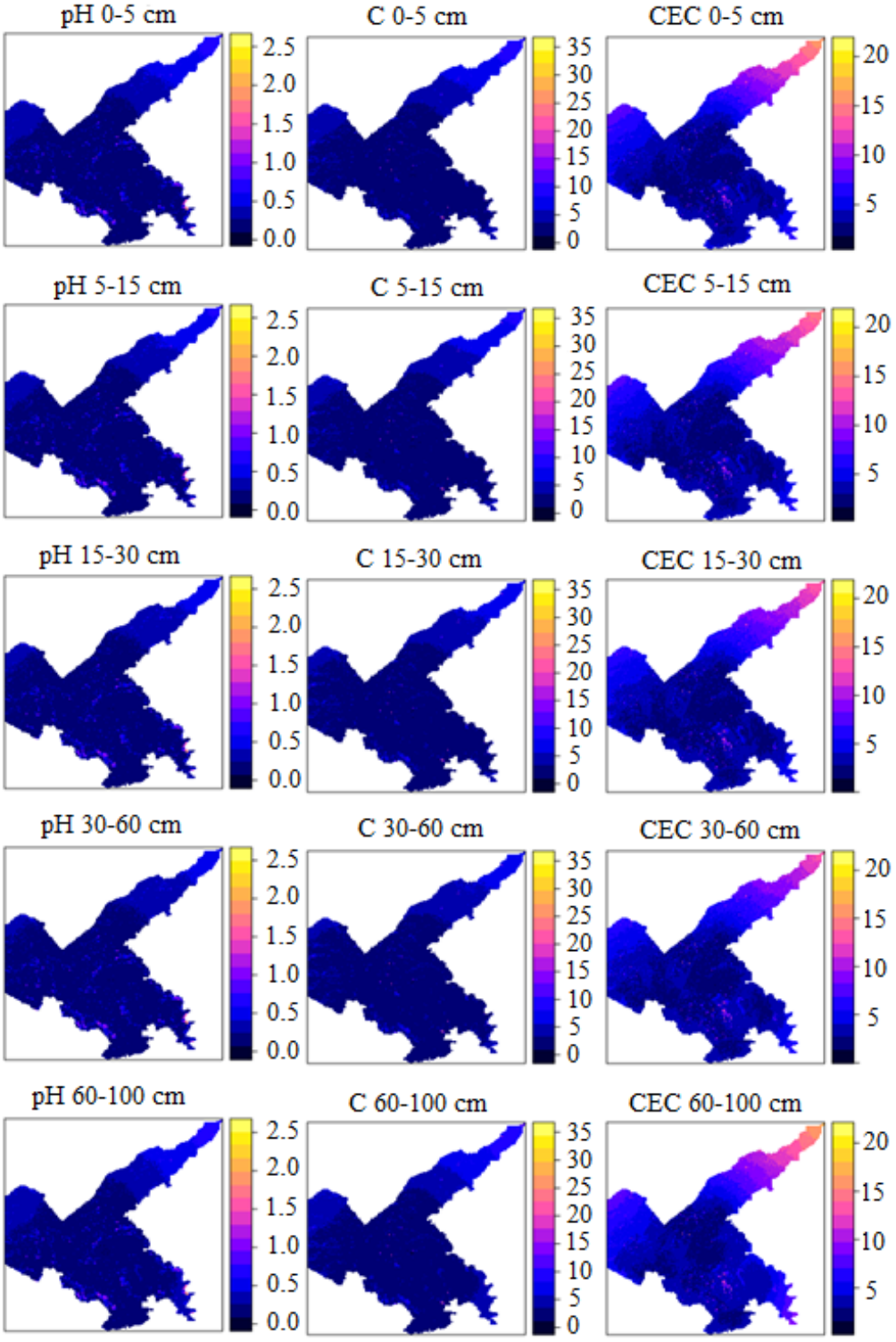


Figure 26. Standard error propagation derived from a Bayesian posterior distribution of each 3D GAM model fitted for pH (left) Carbon content (%, middle), and CEC (cmolc.dm^{-3}, right).

4.6 CONCLUSIONS

In general, the GAM model had superior performance than RF and MLR. The approach based on soil forming factors showed to be a simple and viable method for covariates selection in the GAM model, especially considering limitations regarding degrees of freedom due to limited number of soil samples.

The selection by variable importance did not present a significant improvement in relation to the model with all the covariates, and it is a subjective method as to number of covariates to be maintained in the final model.

RF models had not been sensitive to covariate selection methods. RFE proved to be a viable alternative only to RF, especially since the model has optimized parameters when RFE is used.

Because soil attributes do not have linear relationships with environmental covariates, the models that captured these relationships tend to be better.

The elevation, parent material and covariates from the RapidEye sensor were the factors that most influenced the soil properties of the Itatiaia National Park plateau.

The greater uncertainty was associated with the low accessibility areas, which had low sampling density and/or noises in the covariates. The 2 and 3D soil properties modelling with uncertainty propagation can be useful to attend the demand of INP for ecosystem management.

The high resolution of soil attributes and uncertainties produced for INP in the 3D space are an important step in developing a comprehensive soil database, allowing to deliver quantitative soil-information on a scale adequate to the INP demands.

5 CHAPTER III:

SPATIAL BAYESIAN BELIEF NETWORK: A PARTICIPATORY APPROACH FOR MAPPING ENVIRONMENTAL VULNERABILITY AT THE ITATIAIA NATIONAL PARK

5.1 RESUMO

Uma Rede de Crenças Bayesiana (BBN) foi implementada para avaliar a vulnerabilidade ambiental no Parque Nacional do Itatiaia (PNI). Informações sobre solos, uso/cobertura da terra, clima, relevo e material parental foram usadas para criar a BBN, e os nós de entrada foram aqueles presumidos como tendo influência direta na análise de vulnerabilidade ambiental. A revisão da literatura e uma abordagem participativa serviram de base para construir a estrutura da rede e definir as relações de dependência entre os nós através das tabelas de probabilidade condicional (CPTs, inglês). As áreas mais frágeis foram identificadas como as que apresentam solos com os seguintes atributos: altos teores de carbono orgânico, baixa densidade, baixo pH, alta capacidade de troca catiônica, pouco drenado, perfis menos desenvolvidos (rasos), declives acentuados, cobertura vegetal com gramíneas herbáceas (campos de altitude), localizados próximos às trilhas / fazendas e com alta suscetibilidade ao fogo. Também recebeu uma alta probabilidade de forte e muito forte vulnerabilidade ambiental, solos formados a partir de material parental do tipo coluvial. Apesar da complexidade da área de estudo, a BBN conseguiu produzir um resultado significativo para a distribuição espacial da vulnerabilidade ambiental. Além da abordagem da BBN ser menos subjetiva do que a convencionalmente utilizada em estudos de vulnerabilidade no Brasil, foi possível obter a propagação da incerteza associada à predição. Os resultados ajudarão os tomadores de decisão a identificar áreas prioritárias de intervenção, reduzir a degradação do solo nas áreas altamente vulneráveis e ajudar os gerentes do PNI a alcançar um equilíbrio entre as necessidades de conservação da natureza e as oportunidades de recreação. Além disso, o modelo da BBN pode ser atualizado à medida que novos conhecimentos ou dados são produzidos; e pode ser usado para apoiar o processo do plano de manejo adaptativo, bem como para contribuir com outras pesquisas, especialmente aquelas relacionadas aos serviços ecossistêmicos em áreas montanhosas.

Palavras-chave: Degradação do solo. Raciocínio probabilístico. Sistemas baseados em especialistas. Incerteza. Plano de manejo

5.2 ABSTRACT

A Bayesian Belief Network (BBN) was implemented to assess environmental vulnerability in the Itatiaia National Park (INP). Information of soils, land use/cover, climate, relief and parent material were used to create the BBN, and the input nodes were those presumed as having direct influence in the environmental vulnerability analysis. The literature review and a participatory approach were the basis to construct the structure of the network and to define the relations of dependence between the nodes through the conditional probability tables (CPTs). The most fragile areas were identified as having soils showing the following attributes: high levels of organic carbon, low bulk density, low pH, high cation exchange capacity, poorly drained, profiles less developed (shallow), accentuated slopes, vegetation cover with herbaceous graminoid plants (high altitude fields), located close to the trails/farms and with high fire susceptibility. Also received a high probability for strong and very strong environmental vulnerability, soils formed from colluvial parent material. Despite the complexity of the study area, BBN was able to produce a significative result of the environmental vulnerability spatial distribution. In addition to the BBN approach being less subjective than that conventionally used in vulnerability studies in Brazil, it was possible to obtain the propagation of the uncertainty associated with the prediction. The results will help decision-makers to identify priority areas for intervention, to reduce soil degradation in the highly vulnerable areas and help the INP managers in achieving a balance between nature conservation needs and recreational opportunity. Moreover, the BBN model may be updated as new knowledge or data is produced; and it can be used to support the process of the adaptive management plan, as well as to contribute to other researches, especially those related with ecosystem services in mountainous areas.

Keywords: Soil degradation. Probabilistic reasoning. Expert-based systems. Uncertainty. Management plan

5.3 INTRODUCTION

The “Serra da Mantiqueira” region, where the Itatiaia National Park (INP) is located, is considered to be highly vulnerable to erosion processes, mass movements and landslides, due to their geo-environmental characteristics, such as high slope, shallow soils, high precipitation in a short period of time (Barreto et al., 2013; Delgado et al., 2018). In addition, these factors are aggravated by anthropic pressures (land usage).

One of the sources of pressure in the INP is the tourist visitation. According to Magro et al. (2004) and Barros et al. (2007) in the last years and mainly in the last decades the INP received a significant increase of visitation, especially for recreational activities, the practice of sports and ecotourism; which depend on access by trails that interconnect the various regions of the Park. The lack of information on land use planning, to support the park's management plan, associated with intensive usage may lead to the environmental degradation, especially in areas with high natural risks related to landslide. The pressure by people visitation contributes to environmental imbalances, caused by the successive trampling and compacting of the soil, thus reducing the soil porosity, which associated with the degree of slope can trigger other impacts, for instance, intensifying soil loss by erosion processes (Olive et al., 2009; Tomczyk et al., 2013).

The concept of environmental vulnerability has several interpretations, for example for Nguyen et al. (2016) it is defined and governed by four factors: hydro-meteorology signatures, land resource, social economics (human activities), and topography condition. De Lange et al. (2010) considered that, in general, the environmental vulnerability is a function of exposure to a stressor, effect (also termed sensitivity or potential impact) and recovery potential (also termed resilience or adaptive capacity). Besides that, environmental vulnerability is one of the shortcomings commonly associated with tourism pressure in developing regions (Geneletti and Dawa, 2009),

In Brazil, there are basically two approaches to vulnerability/fragility assessment in environmental projects, and they are based on the principle that nature presents an intrinsic functionality between its physical and biotic components which consider the relief analysis focused on the geomorphologic application. In one of the methodologies, the authors define the fragility of the environment as a function of relief, soil types/attributes, vegetation cover/land use and climate (rainfall) (Ross, 1994, 2012); and the second defines basic territorial units that consider all those factors plus parent material (Crepani et al., 2001). In both models the elaboration of the final map is based on map algebra and weights are given to each factor considered (Crepani et al., 2001; Ross, 1994; Spörl and Ross, 2004). However, these methodologies have limitations because the weights are arbitrarily and subjectively attributed (Spörl and Ross, 2004). Thus, the result is strongly dependent on the person who elaborates the map and there is no estimate of the uncertainty, basically one expert tells the "truth". However, these authors were chosen in this study of the INP, a conservation unit on the Atlantic Forest with a mountainous landscape, due to the parameters they take in consideration in the models.

There are several works based on mentioned studies, such as Manfre et al. (2013), Rovani et al. (2016), Valle et al. (2016), and Choudhary et al. (2017). They follow the same principle, with small adaptations in accordance with the area of study, the experience of the specialist, and available data. To minimize the subjectivity, some studies implement the methodology of algorithms, such as artificial neural network (Spörl et al., 2011) and fuzzy logic (Cereda Junior and Röhm, 2013), but there is no description of application of these methodologies in models such as the Bayesian Belief Network (BBN), based on probability and specialist knowledge through participatory approach (Chen and Pollino, 2012; Landuyt et al., 2013; Gonzalez-Redin et al. 2016). BBN modelling may facilitate the decision-making processes by managers, the characterization in a spatial context using GIS layers makes the

information more accessible (Gonzalez-Redin et al. 2016) and it is easily updated with new data entry (Bashari et al., 2016).

The assumption of this study was that by using the BBN approach it is possible to combine the knowledge of several experts from different areas, implemented in a probabilistic model and based on studies involving vulnerability, to reduce the subjectivity of the vulnerability analysis process, as well as to produce results with the uncertainty associated with the prediction, which is essential in environmental analysis.

The aims of this study were:

i) to assess the soil vulnerability in the INP by integrating information of physical environment with knowledge of experts, in order to reconcile the demand for public use with ecosystems conservation;

ii) to reduce subjectivity of the commonly used analysis of environmental vulnerability, by incorporating expert knowledge and results from literature in a quantitative/probabilistic approach, the BBN;

iii) to compare Ross (1994) and Crepani et al. (2001) methodologies to assess the environmental vulnerability of INP areas, using quantitative/probabilistic approach.

5.4 MATERIAL AND METHODS

5.4.1 Study area characterization

The study area comprises the upper part (plateau) of the Itatiaia National Park, with an approximate area of 164 km^2 (Figure 27), high elevation and relief characterized as very steep (Barreto et al., 2013). The Itatiaia National Park is located in the southeastern region of Brazil in the *Serra da Mantiqueira*, between the States of Minas Gerais, Rio de Janeiro and São Paulo.

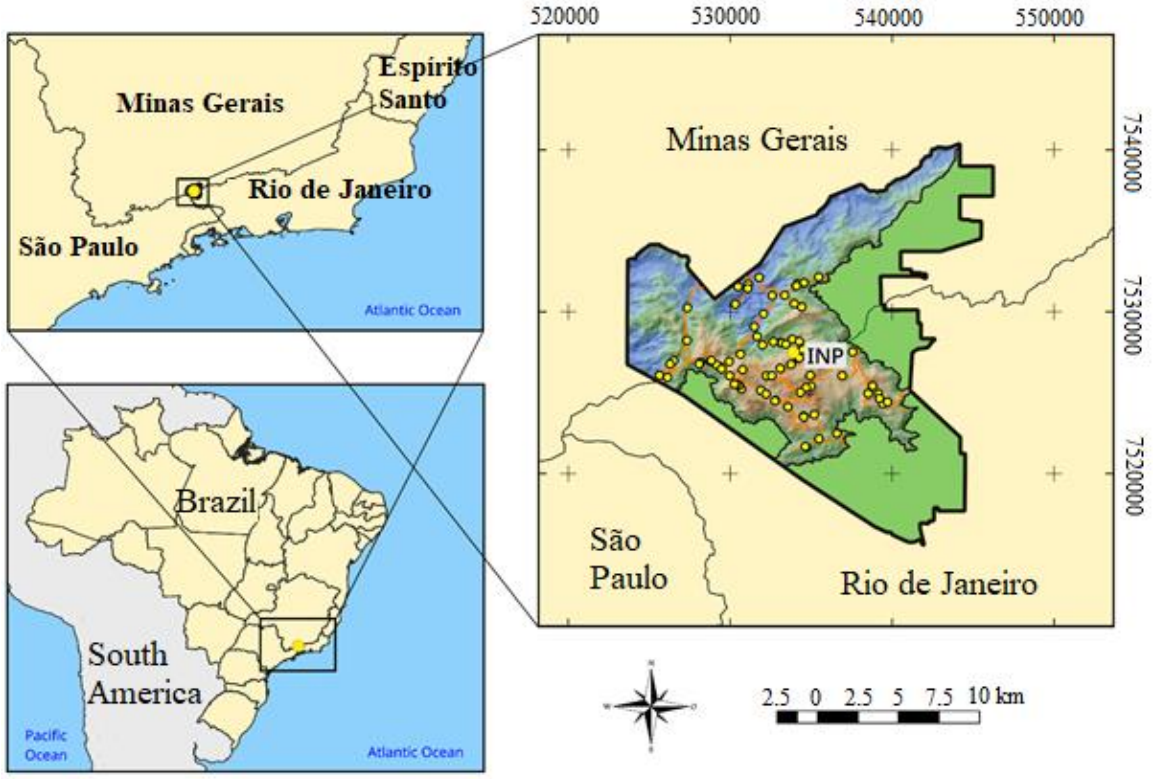


Figure 27. Location of Itatiaia National Park (INP) in the south-eastern region of Brazil. In detail, the total area of INP, and the polygon with relief limits the upper part. Yellow points mark soil sampling

The climate in the region is of mesothermic type (Cwb) according to the Koppen classification (Alvares et al., 2013), and the INP climatic domain has two mesothermal types. The upper part of the landscape shows a mild and rainy season during summer, and the yearly average temperature is 11.5°C with 8.4°C average during the winter (Barreto et al., 2013). The temperature may reach values below zero sporadically, for example in the winter of 2010 it was recorded a negative 8º C (Tomzhinski et al., 2012).

The geology at the upper part of INP is dominated by quartz syenites and magmatic breccia, some alkaline granites, and in the total area homogeneous gneisses and nepheline syenite. Colluvial and alluvium sediments vary from sand to clayey and have blocks and boulders of alkaline rocks. In the upper part of INP, sediments in the valleys and bottom of slopes are rich in organic matter (Santos et al., 2000). Relief varies from mountainous to steep (Barreto et al., 2013), with slopes ranging from flat, on plains and valleys, to very steep in the rocky outcrops.

The predominant soils on the INP are: Leptosols, Regosols, Folic Umbrisols, Cambic Folic Histosols, Cambic Umbrisols, Umbrisols+Ferrasols, Cambisols, Folic Histosols and Histosols. In the upper part of the park the classes of Leptosols, Histosols and Folic Histosols dominate. These soils are mainly a result from the relief and high elevation, which create environmental conditions for accumulation of organic matter along the valleys and slopes among the rock outcrops (see chapter 1).

5.4.2 BBM implementation, network construction and participatory process

A Bayesian Belief Network (BBN) was developed to assess environmental vulnerability in the upper part of the INP (Figure 28). Information of, relief, parent material, climate soils and land use/cover were used to create the BBN, and the input nodes were those presumed as having direct influence in the fragility factors (environmental vulnerability).

The main factors involved with the INP environmental vulnerability where identified based on literature review. The whole structure of the network was set up based on commonly used and well-established procedures to evaluate environmental vulnerability/fragility (Crepani et al., 2001; Ross, 1994; Ross, 2012), with the addition of a participatory approach (questionnaire, see attachment). Although these studies use mainly map algebra as a mapping approach (Adami et al., 2012; Manfré et al., 2013; Valle et al., 2016; Choudhary et al., 2017; Calderano Filho et al., 2018), which is believed to be a simplistic process of the environmental relationships, they provide a detailed description of the factors that involve vulnerability. This description was then used to create an online questionnaire, where researchers (total of 26), from various fields of knowledge, answered questions elaborated to capture the opinion about factors involved in the INP environmental vulnerability. Due to the large dispersion of information, a second selection was made to create and populate the conditional probability tables (CPTs). It was considered mainly the literature review and the expert opinion of researchers working directly with the environmental vulnerability project at the INP.

It is noteworthy that for BBN implementation the results were not only considered quantitatively and that qualitative factors were also taken into account. In addition to the qualitative factors, the experience of the professional working in the INP and their expertise in the area on which the questions referred were taken into account when eliciting the specialist's knowledge (Hemming et al., 2017). Details about results of the participatory process, expert specialization, institution and years of work or research in the INP are presented in the figures from S14 to S19 and table S1.

It is important not to use only quantitative data, because the same soil attribute may have very different interpretations depending on the environment (study area) and the experts. For example, the lowest vulnerability was considered by many as occurring in the sites with higher levels of soil organic matter, because these experts consider that the soil properties are more stable with the increase of the organic matter content of the soil, which favors aggregation, helps to retain cations and soil moisture, among other beneficial functions. This applies to mineral soils and agricultural conservationist systems, but in the case of INP, the highest levels of organic matter are found in the organic soils, identified by many experts as very vulnerable, because they are precisely that - fragile and on susceptible areas. These are the areas with the highest soil degradation by compaction and erosion (See Figure S13) especially close to the trails.

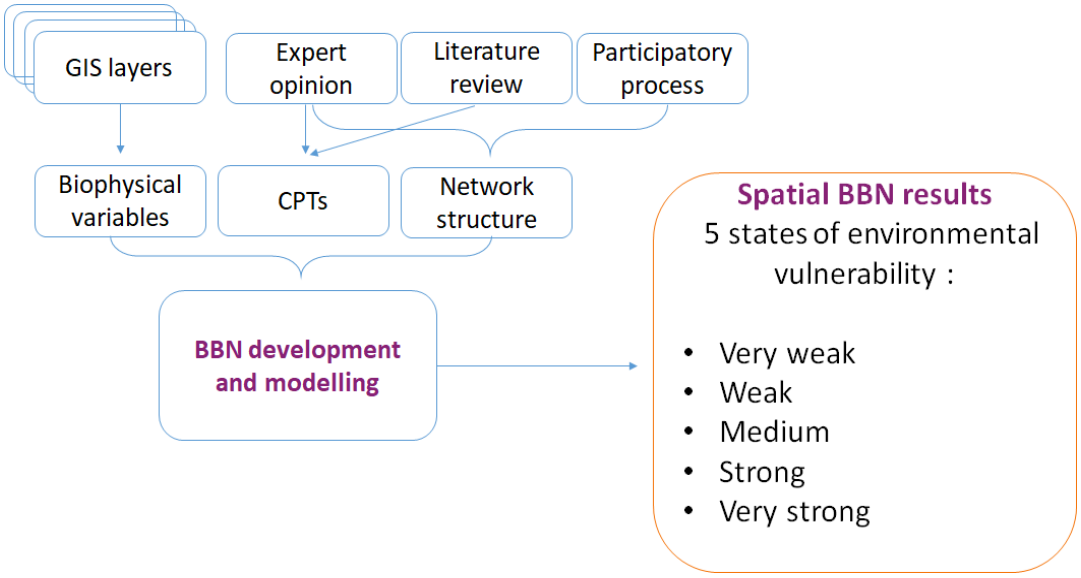


Figure 28. Flowchart of the BBN development with the input nodes (in our case all spatial) and biophysical variables. The network was based on expert opinion, literature review and participatory process. The CPTs population was based on expert opinion and literature review. The spatial BBN modelling that performed the inference of probability at the pixel level to obtain the final spatial outcomes is identified as environmental vulnerability probability (from very weak to very strong).

Two basic networks were created, both having in common the following fragility factors: soil attributes (description of the main soil attributes that influence environmental vulnerability), relief (slope classes), land coverage or usage, climate (mainly the rainfall, volume and distribution) (Crepani et al., 2001; Ross 1994, 2012) (Figure 29).

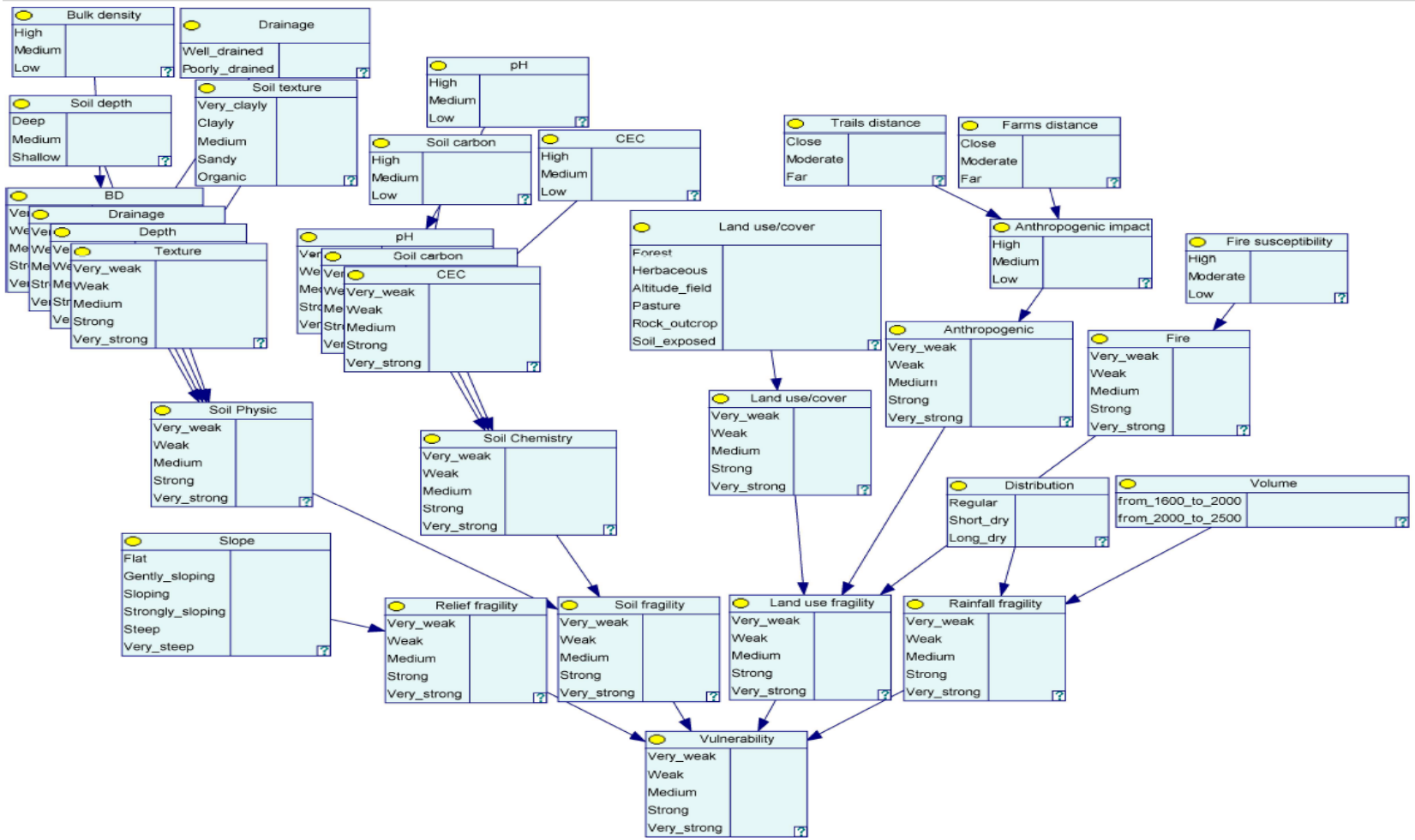


Figure 29. BBN structure with the input nodes (spatial) on the top layers; and the intermediate nodes (not spatial) some intermediate nodes are used specially to translate the information into the five states of vulnerability (from very weak to very strong). The base of the network are the fragility factors. The bottom node is the BBN outcome (five states of environmental vulnerability).

In addition, the network based on Crepani et al. (2001) uses the concept of territorial units and the geology is considered in the environmental vulnerability analysis (Figure 30).

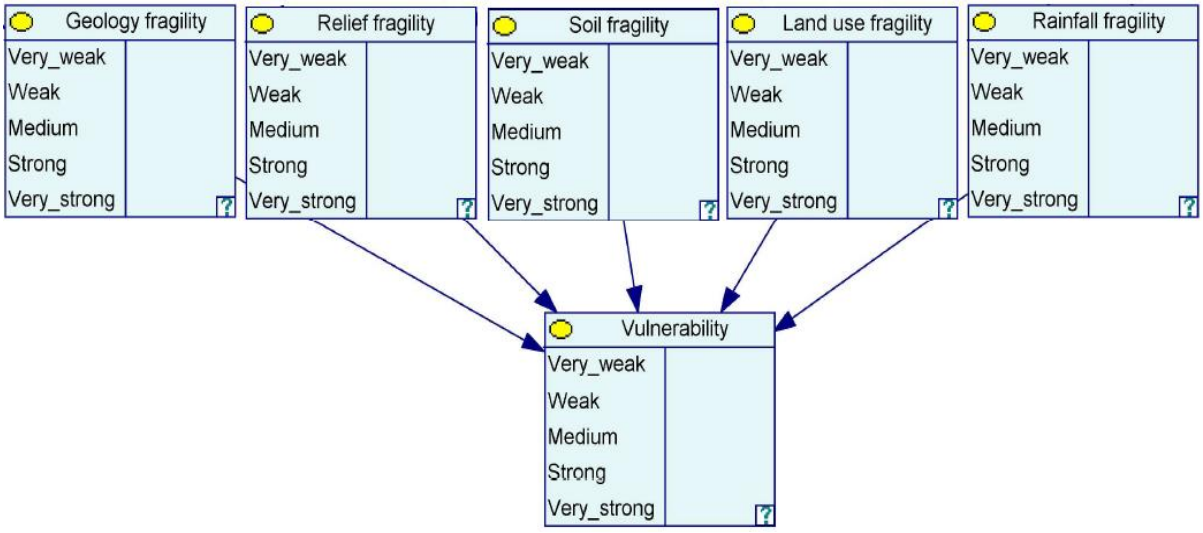


Figure 30. The BBN structure using basic territorial units according to Crepani et al. (2001). In this case there are five fragility factors, with the inclusion of geology**.**

The probability of nodes that are directly linked with the environmental fragility was termed here as fragility factors, to differentiate them from the other nodes because they are the base of the network. The fragility factors, commonly used in the analysis of environmental vulnerability (Crepani et al., 2001; Ross, 1994; Ross, 2012 Adami et al., 2012, Manfré et al., 2013, Valle et al., 2016; Choudhary et al., 2017, Calderano Filho et al., 2018) receive the probabilities of the input and intermediate nodes and transfer to the output (environmental vulnerability). The CPT for the input and intermediate nodes were based mainly on the participatory process and expert opinion. The CPT for the base nodes, fragility factors, was adapted from Valle et al. (2016) and filled based on expert opinion.

Based in Valle et al. (2016) each factor is part of the sum that influences vulnerability, and the environmental vulnerability varies from 1 (very weak) to 5 (very strong). The best possible condition observed are for the sum 4 (Table 20, without geology according to Ross, 1994), and 5 (Table 21, with geology according to Crepani et al., 2001) in this case all the fragility factors have a very weak environmental vulnerability. The worst possible condition is for all the factors with very strong environmental vulnerability, layers sum 20 and 25, with and without geology respectively (Table 20 and 21).

Table 20. Sum of scores, completion of the CPT to define the probability and description of the environmental vulnerability (Ross, 1994).

Layers' sum	Environmental Vulnerability					
	Very weak	Weak	Medium	Strong	Very strong	Highest probability
4	0.99	0.01	-	-	-	Very weak
5	0.30	0.70	-	-	-	Weak
6	0.15	0.80	0.05	-	-	Weak
7	0.10	0.80	0.10	-	-	Weak
8	-	0.70	0.30	-	-	Weak
9	-	0.30	0.70	-	-	Medium
10	-	0.15	0.80	0.05	-	Medium
11	-	0.10	0.80	0.10	-	Medium

12	-	-	0.70	0.30	-	Medium
13	-	-	0.30	0.70	-	Strong
14	-	-	0.15	0.80	0.05	Strong
15	-	-	0.10	0.80	0.10	Strong
16	-	-	-	0.70	0.30	Strong
17	-	-	-	0.30	0.70	Very Strong
18	-	-	-	0.15	0.85	Very Strong
19	-	-	-	0.10	0.90	Very Strong
20	-	-	-	0.01	0.99	Very Strong

Table 21. Sum of scores, completion of the CPT to define the probability and description of the environmental vulnerability (Crepani et al., 2001).

Layers' sum	Environmental Vulnerability					
	Very weak	Weak	Medium	Strong	Very strong	Highest probability
5	0.99	0.01	-	-	-	Weak
6	0.40	0.60	-	-	-	Weak
7	0.25	0.75	-	-	-	Weak
8	0.10	0.80	0.10	-	-	Weak
9	0.05	0.80	0.15	-	-	Weak
10	-	0.70	0.30	-	-	Weak
11	-	0.40	0.60	-	-	Medium
12	-	0.25	0.75	-	-	Medium
13	-	0.10	0.80	0.10	-	Medium
14	-	0.05	0.80	0.15	-	Medium
15	-	-	0.70	0.30	-	Medium
16	-	-	0.40	0.60	-	Strong
17	-	-	0.25	0.75	-	Strong
18	-	-	0.10	0.80	0.10	Strong
19	-	-	0.05	0.80	0.15	Strong
20	-	-	-	0.70	0.30	Strong
21	-	-	-	0.40	0.60	Very Strong
22	-	-	-	0.25	0.75	Very Strong
23	-	-	-	0.15	0.85	Very Strong
24	-	-	-	0.10	0.90	Very Strong
25	-	-	-	0.01	0.99	Very Strong

5.4.3 Biophysical variables (GIS layers)

To infer the basic fragility factors that influence environmental vulnerability, several biophysical variables were prepared and defined as input node in the BBN (Table 22).

In order to generate the soil properties maps, the INP database that comprises a total of 90 profiles and 359 horizons was used. To generate the maps several predictions models were calibrated and tested. The two best models selected from chapter one, Generalized Additive Model (GAM) with selection of covariates by the SCORPAN approach (GAM_scorpan) and Rando Forest (RF) with covariate selection by recursive feature elimination (RFE), were used to predict continuous and categorical soil attributes, respectively. Also, for land use/cover prediction the RF_rfe was used to predict the current land use. The fragility factors are based on Ross (1994; 2012) and Crepani et al. (2001). The other layers are based in the literature, and they were defined in this study for the environmental vulnerability assessment proposal.

Table 22. Description of the biophysical variables used on the Bayesian network as input nodes to assess the vulnerability at the INP.

Fragility factor	Variable/node	Description and literature sources	Classes / States
Relief (R)	Slope (%)	Calculated from the digital elevation model (DEM) of INP, and classified according to Santos et al. (2015)	Flat (0-3), gently sloping (3-8), sloping (8-20), strongly sloping (20-45), steep (45-75), very steep (> 75)
Soil (S)	Bulk density (g.dm-3)	Continuous soil property (physical) predicted using GAM_scorpan model	High (>0.9), medium (0.5-0.9), low (< 0.5)
	Depth (cm)	Continuous soil property (physical) predicted using GAM_scorpan model	High (>100), medium (50-100), low (< 50)
	Drainage (dimensionless)	Categorical soil property (physical) predicted using RF_rfe model	Well-drained, poorly drained
	Texture (dimensionless)	Categorical soil property (physical) predicted using RF_rfe model	Clay, loam, organic
	pH (dimensionless)	Continuous soil property (chemical) predicted using GAM_scorpan model	High (> 5), medium (4-5), low (< 4)
	Carbon content (%)	Continuous soil property (chemical) predicted using GAM_scorpan model	High (> 8), medium (3-8), low (< 3)
	CEC ($cmolc.dm^{-3}$)	Continuous soil property (chemical) predicted using GAM_scorpan model	High (> 20), medium (10-20), low (< 10)
Land cover/use (LU)	Land cover	Current land use predicted using RF_rfe. The six classes were based on literature, soil points observation and Google Earth®	Forest, herbaceous, altitude fields, pasture, rock outcrop, and exposed soil
	Trails (m)	Distance from the trails	Near (0 to 100), medium (100 to 200), far (> 200)
	Farms (m)	Distance from the farms	Near (0 to 150), medium (150 to 300), far (> 300)
	Fire susceptibility (dimensionless)	Derived from land cover, DEM terrain attributes DEM, fire history, socio-economic factors (Tomzhinski et al., 2012)	High, medium, low
Climate (C)	Rainfall (mm)	Obtained from the WorldClim Vers2. The INP upper part was divided into two classes (Fick et al., 2017)	1600 to 2000 2000 to 2500
Parent material (P)	Geology	Map with six classes in the INP upper part, from Santos et al. (2000)	Alluvial sediments, colluvium sediments, nepheline syenite, quartz syenite, alkaline granite, magmatic breccia, homogeneous gneisses

Note: For the purpose of modeling soil textural classes, the soil material defined as organic was included as a texture class, as in Poggio et al. (2017a)

5.4.4 BBN parameterization

Data from literature review, expert opinion, and participatory approach were used to populate the CPTs. The first step was to translate the biophysical variables states into the five states of fragility, in order to make the links between the networks easier, especially in the nodes that receive a lot of parent nodes. The bottom of the network (the outcome node) is based in a combination of the states of environmental vulnerability for each factor (Valle et al., 2016) and in the CPT there is a probability for each combination (Tables 20 and 21). This approach takes into account the uncertainty into the environmental vulnerability analysis (not considered in previous studies), and at the same time, it makes easier to populate the CPT.

Based on the literature review, expert opinion, participatory process and expert elicitation, the INP environmental vulnerability was defined, in a wide-ranging sense, as a function of the fragility factors: relief, parent material, climate, soil, and land use/cover. Intermediate nodes were created to translate the information into five fragility classes, very weak (VW), weak (W), medium (M), strong (S) and very strong (VS), which were used to combine factors in the network and populate CPTs.

In this way, the areas with very strong environmental vulnerability in the INP were those with soils having low bulk density, high soil carbon content, high cation exchange capacity, low pH, with texture identified as organic, altitude fields as land coverage, and specially the soils identified in the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018) as *Organossolos* (Folic Histosols, according to WRB, 2015). There are two classes of these soils in the INP: *Organossolos Háplicos* (Histosols with a histic horizon), with poor drainage; and *Organossolos Fólicos* (Histosols with a folic horizon) that have a good drainage and are often shallow. Other areas with strong environmental vulnerability were those with very steep slopes, close to the trails/farms, and with highest rainfall volume (Figure 31). On the other hand, areas with soils having medium values of bulk density, soil organic carbon, and cation exchange capacity; high pH; loamy texture; forest land cover; mineral soils well developed, with deep profiles and good drainage; normally on flat slopes; far away from trails and farms; and with medium volume of rainfall had the most stable environmental condition (Figure 32).

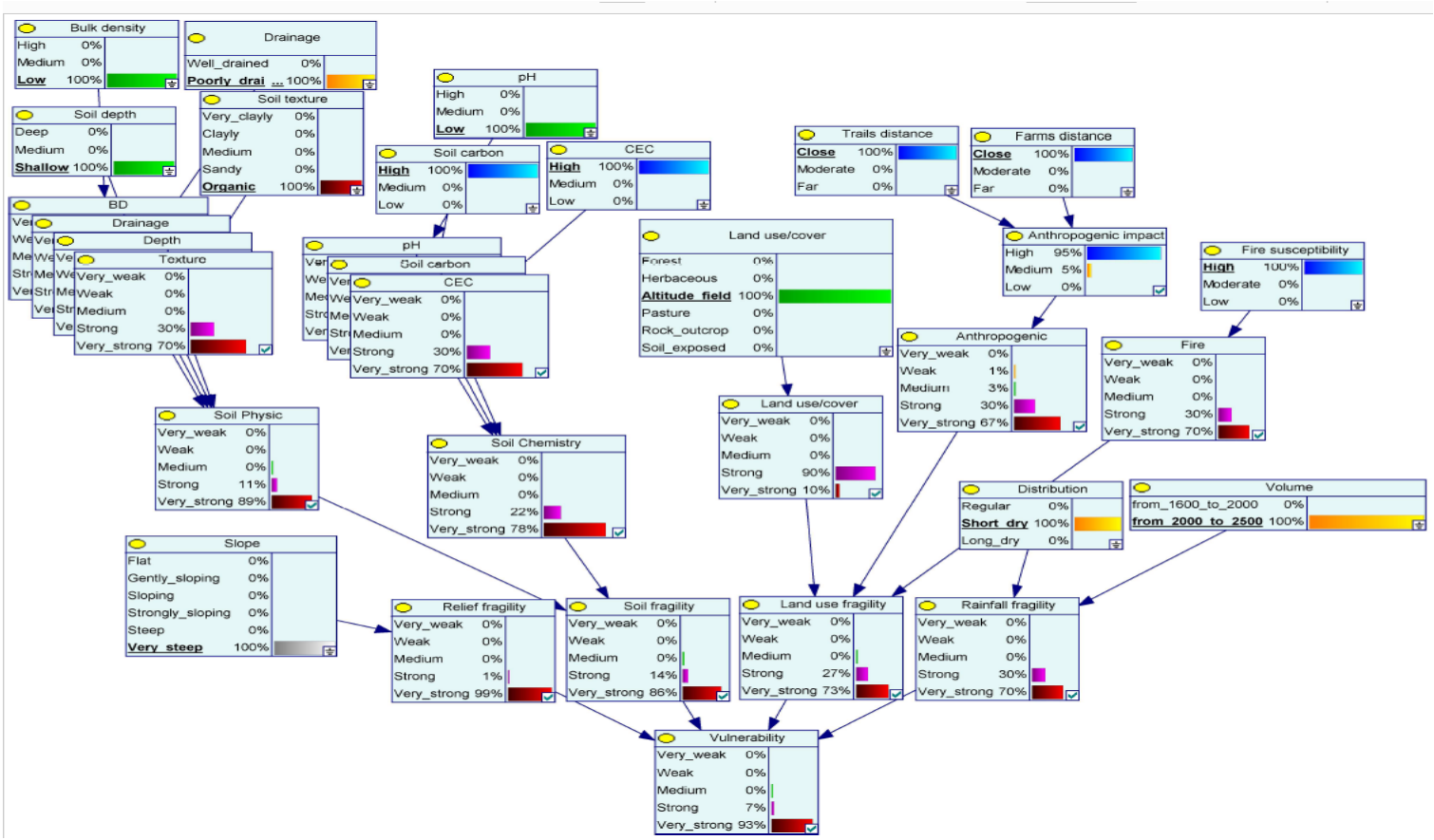


Figure 31. Illustrated case of a feasible combination and the relative probability of the most unstable environmental conditions

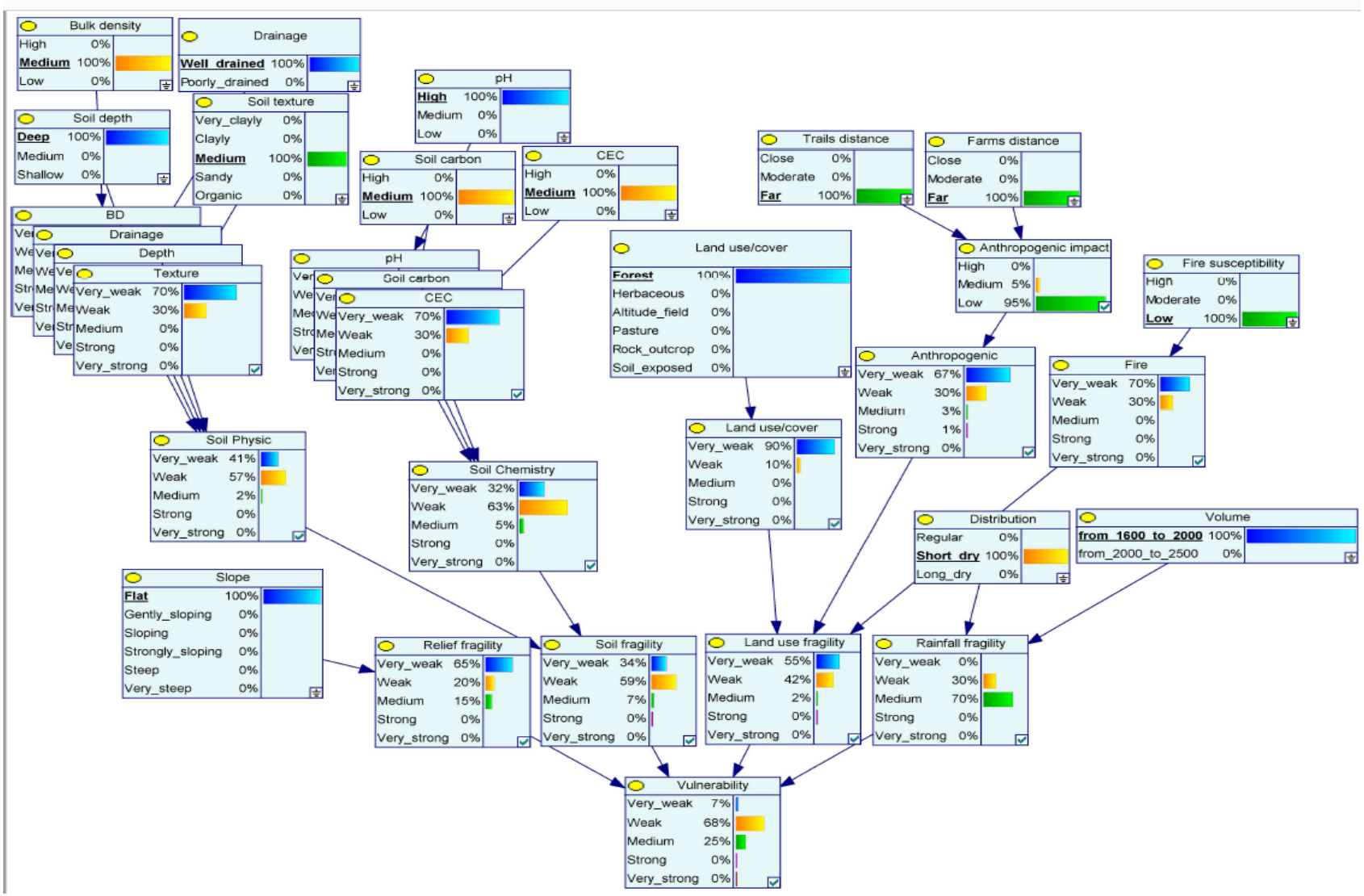


Figure 32. Illustrated case of a feasible combination and the relative probability with the most stable environmental condition

5.4.5 Spatial prediction and uncertainty propagation

Spatial explicit outputs were obtained through the CPT and evidence propagation of the probability by using the BBN tools in GeNIe® and R software. The final spatial output of the BBN modelling was the environmental vulnerability probability, divided in five states: very weak, weak, medium, strong and very strong. Based on the BBN implementation, five sets of maps were obtained (one for each state), showing the probability of environmental vulnerability (very weak, weak, medium, strong and very strong), also a map showing the environmental vulnerability (from very weak to very strong) with highest probability per pixel (each pixel returned the most probable state of environmental vulnerability). The uncertainty of the spatial prediction of environmental vulnerability, a random discrete (categorical) variable C in a spatial location s was quantified using the Shannon entropy (Shannon, 1948).

$$H(s) = \sum_{i=1}^{n_i} \pi(c_i, s) log_k \pi(c_i, s) \qquad \text{Eq. 06}$$

Where (ci, s) is the estimated probability that the random variable C, at location s, takes the value ci among the k possible values (Agresti, 2002). The use of the logarithm with base k scales the value of H(s) between 0 and 1, where 0 means no uncertainty – one of the k categories has the probability of occurrence equal to 1, and the value 1 means maximum uncertainty, all categories have an equal probability of occurrence (Kempen et al., 2009). All maps were computed with a spatial resolution of 25 m.

5.4.6 Software used

The R software (R Core Team, 2018) with the packages: raster, rgdal, maptools and RSAGA for data management, preparation and visualisation (Bivand et al., 2017; Bivand and Lewin-Koh, 2017; Brenning, 2008; Hijmans, 2016). For the BBN inference it was used the James Hutton Institute tool, BayesGIS (DEMO version), a shiny app that is supported by the R packages bnlearn (Scutari, 2010) and gRain (Højsgaard, 2012). The GeNIe® 2.2 BBN Modeler software (BayesFusion LLC 2018) was used to develop the BBN structure and to fill the CPTs.

5.5 RESULTS AND DISCUSSION

5.5.1 Literature review, participatory process e expert elicitation

The relief was represented by slope classes (Table 23), since they influence strongly the water infiltration and flow velocity, and the slope is one of the main factors in the risk analysis of landslides, soil erosion and silting of watercourses (Calderano Filho et al., 2018; Valle et al., 2016).

Table 23. Environmental vulnerability and probabilities (CPT) by variables slope and geology

Slope classes (%)	CPT	Highest probability	Geology	CPT	Highest probability
Flat (0-3)	0.99 VW; 0.01 W	Very weak	Alluvial sediments	0.3 S; 0.7 VS	Very strong
Gently sloping (3-8)	0.7 W; 0.3 M	Weak	Colluvium sediments	0.45 S; 0.55 VS	Very strong
Sloping (8-20)	0.7 M; 0.3 S	Medium	Nepheline syenite	0.45 VW; 0.55 W	Weak
Strongly sloping (20-45)	0.7 S; 0.3 VS	Strong	Quartz syenite	0.7 VW; 0.3 W	Very weak
Steep (45-75)	0.3 S; 0.7 VS	Very Strong	Alkaline granite	0.55 W; 0.45 M	Weak
Very steep (> 75)	0.01 S; 0.99 VS	Very Strong	Magmatic breccia	0.45 W; 0.55 M	Medium
			Homogeneous gneisses	0.55 VW; 0.45 W	Very weak

The somewhat flat valleys are assigned to very weak fragility, while steep and very steep areas are described as having high fragility. Due to the predominantly mountainous relief, the INP shows a high degree of environmental vulnerability, according to the slope information.

When analyzed the geology variable, the sediments show the greatest fragility and gneisses and quartz syenite are the most stable are (Table 23). The classes of fragility for parent material are related to the geological evolution of the INP and the degree of cohesion of the rocks that compose the formation where INP is inserted (Crepani et al., 2001).

Climate was evaluated based on the volume and distribution of rainfalls. Although the CPT has more classes, for the INP the spatial information grouped the influence of rainfall water surface runoff in two classes of environmental vulnerability based on highest probabilities, defined as: Medium (1600 to 2000 mm of precipitation, with CPT 0.3 W; 0.7 M); and Very strong (2000 to 2500 mm of precipitation, with CPT 0.3 W; 0.7 VS), with a short dry period distribution according to Ross (2012). The location of the INP favors occurrence of orographic rains, intensified in the summer. This explains the greater vulnerability (Ross, 2012), attributed especially to the higher part of the INP, where the rainfall is heaviest.

The soil fragility was represented by soil properties reported by literature and they were: bulk density, soil depth, drainage, texture, pH, carbon content, and CEC. These characteristics were selected based on their relationship to land capability and predisposition to erosion, mass movements, and superficial landslides (Calderano Filho et al., 2018). For populating the CPTs the participatory process and expert knowledge were used, where very weak fragility is

associated with soils having: medium values of bulk density, organic carbon and CEC; deep profiles; well-drained; loamy texture; and with high pH. These properties are mainly found in the mineral soils of the study area. In the upper part of INP, the *Organossolos* (Histosols) have a common occurrence, and this class is described in the literature and confirmed in the participatory process as soils with high vulnerability. Thus, the soils with high fragility (very strong vulnerability) in the INP are those with low bulk density, shallow profiles, poor drainage, organic material, low pH, high organic carbon content and high CEC (Table 24), attributes found in the organic soils.

Table 24. Environmental vulnerability and probabilities (CPT) variable soil

Soil properties	States	CPT	Highest probability
Bulk density	High	0.3 M; 0.7 S	Strong
	Medium	0.7 VW; 0.3 W	Very weak
	Low	0.7 VS; 0.3 S	Very strong
Depth	High	0.7 VW; 0.3 W	Very weak
	Medium	0.1 W; 0.8 M; 0.1 S	Medium
	Low	0.3 S; 0.7 VS	Very strong
Drainage	Well-drained	0.7 VW; 0.3 W	Very weak
	Poorly drained	0.7 VS; 0.3 S	Very strong
Texture	Clay	0.3 M; 0.7 S	Strong
	Loam	0.7 VW; 0.3 W	Very weak
	Organic	0.3 S; 0.7 VS	Very strong
pH	High	0.3 VW; 0.7 W	Weak
	Medium	0.1 W; 0.8 M; 0.1 S	Medium
	Low	0.7 S; 0.3 VS	Strong
Carbon content	High	0.3 S; 0.7 VS	Very strong
	Medium	0.7 VW; 0.3 W	Very weak
	Low	0.3 M; 0.7 S	Strong
CEC	High	0.3 S; 0.7 VS	Very strong
	Medium	0.7 VW; 0.3 W	Very weak
	Low	0.3 M; 0.7 S	Strong

For the land use/cover six classes were identified, where very weak and weak fragility were associated with forest and herbaceous cover (Table 25), and strong and very strong fragility associated with altitude fields, pasture, rock outcrop and exposed soils. The exposed soil had the highest probability of having a very strong environmental vulnerability.

Table 25. Environmental vulnerability and probabilities (CPT) variable land use/cover

Land use/cover	CPT	Highest probability
Forest	0.9 VW; 0.1 W	Very weak
Herbaceous	0.9 W; 0.1 M	Weak
Altitude fields	0.9 S; 0. 1 VS	Strong
Pasture	0.3 S; 0.7 VS	Very strong
Rock outcrop	0.1 S; 0.9 VS	Very strong
Exposed soil	0.01 S; 0.99 VS	Very strong

5.5.2 Model results, spatial interpolation and uncertainty propagation

The model results for environmental vulnerability in the upper part of INP indicated very strong vulnerability in about 15% of the area, and strong in about 75%, for the model without geology (Ross, 1994); and when geology was taken into account (Crepani et al., 2001) the areas with strong vulnerability were more than 69 %, and very strong of 0.3 % (Table 26). When the geology map is used as model input there is a considerable increase in the medium vulnerability and a decrease of areas with very strong class.

The predominance of strong vulnerability (Table 26) is mainly related to the slope in the mountainous relief. For the most sloping areas, both approaches (Ross, 1994; and Crepani et al., 2001) coincided with the higher class of environmental vulnerability. In addition, the soil types *Organossolos* (Histosols) are predominantly fragile. The vulnerability class very strong is found when a greater rainfall volume, presence of rock outcrops, and altitude fields as soil cover are combined (Table 26).

Table 26. Areas of the upper part of INP with their environmental vulnerability classes according to Ross (1994) and Crepani et al. (2001).

Vulnerability Class	**Ross (1994)**		**Crepani et al. (2001)**	
	Area (ha)	Area %	Area (ha)	Area %
Weak	2.75	0.02	10.19	0.06
Medium	1592.81	9.71	5011.88	30.55
Strong	12254.88	74.71	11333.12	69.09
Very strong	2553.25	15.56	48.50	0.30
Total	**16403.69**	**100%**	**16403.69**	**100%**

The spatial BBN, developed using GeNIe and R software, provided a useful framework to represent relationships between spatial variables and environmental vulnerability. Even if there is a degree of uncertainty (Gonzalez-Redin et al., 2016) in the process, BBN is a better approach because it can deal with uncertainty (Landuyt et al., 2015; Marcot, 2012). The maps obtained through the spatial BBN (Figure 33) represent the highest probability of environmental vulnerability for each class. The results from the BBN models, according to the methods without the geology (Ross, 1994), and with geology (Crepani et al., 2001), led to different maps for INP (Figure 33a and 33b), and consequently different uncertainties (Figure 33c and 33d).

The map of areas in the upper part of INP elaborated according to Crepani et al. (2001) approach shows the very strong environmental vulnerability along the valleys (Figure 33b); where soils with high organic matter and poor drainage predominate. For the map according to Ross (1994), the areas with very strong vulnerability are mainly related to organic soils or soils with a high content of organic matter but a shallow profile (*Neossolos Litólicos*, Leptosols) and near rock outcrops. They also include areas of altitude fields vegetation and with higher slope and elevations, where the volume of rainfall is greater.

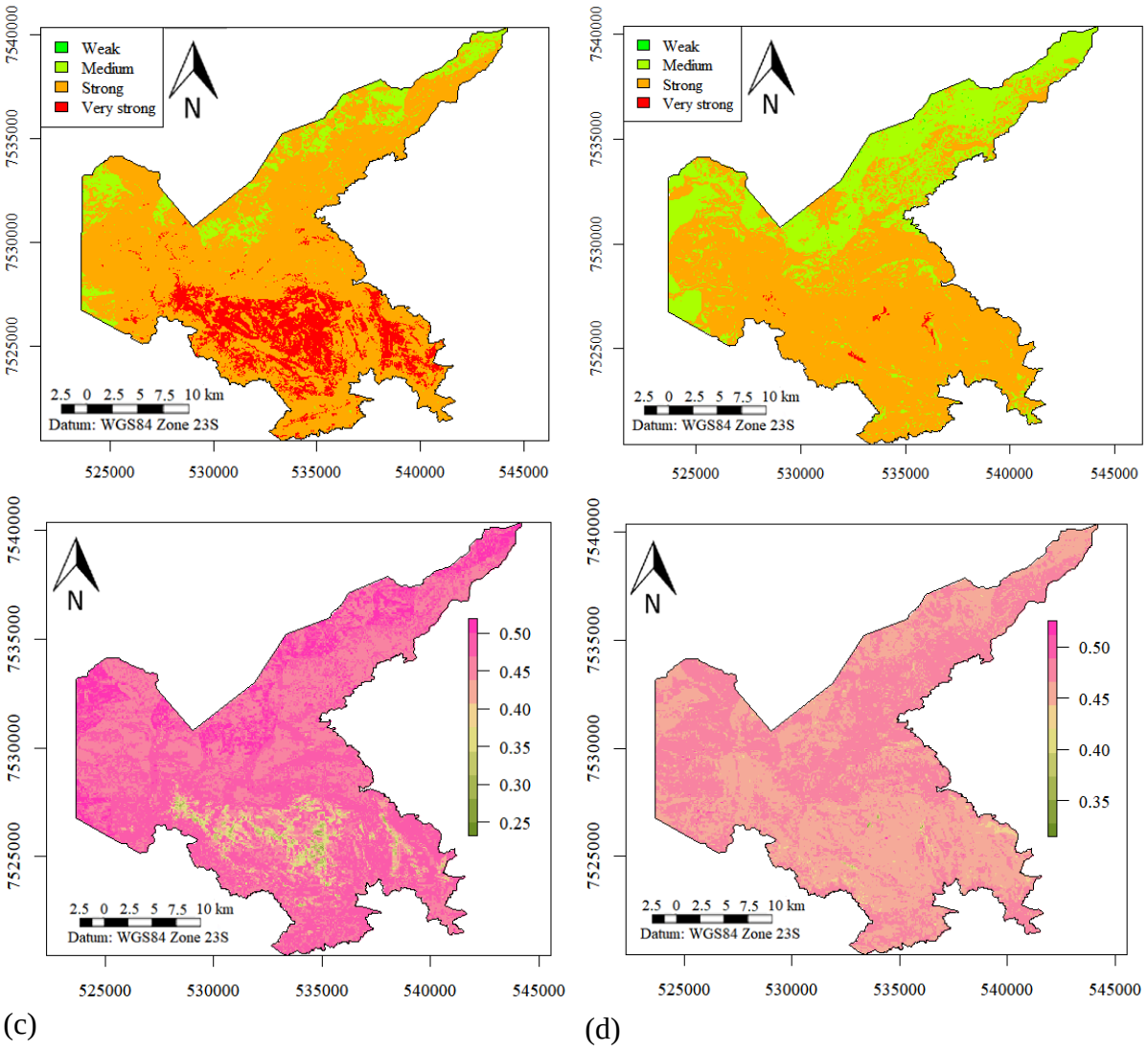


Figure 33. Environmental vulnerability maps of the upper part of INP (a and b), and Shannon entropy (uncertainty) of environmental vulnerability predictions (c and d) (left - Ross, 1994; right - Crepani et al., 2001)

The main difference between the two approaches was in the proportion not in the number of classes, as it was in Manfré et al. (2013). According to this author, the Crepani et al. (2001) methodology generalized the environmental vulnerability for the geomorphic component, which may decrease the map accuracy, as it is shown in the uncertainty map (Figure 33c and 33d). The results show the importance of establishing a methodology for validating the results using local samples, and to complement the method applied in this study where it was used uncertainty measurements. However, different from soil classes, which are well stablished by the classification system adopted, vulnerability classes have a conceptual subjectivity, which will affect the validation process.

It was not possible to validate the vulnerability maps (by calculating estimates of error as accuracy and Kappa) as it was done in the digital soil mapping, because the degrees of vulnerability are not well defined as the soil classes. However, the methodology here proposed brings an advance by decreasing the subjectivity of the process, and it provides an element not previously available in this type of mapping, which is the uncertainty associated with the predictions. It can be observed, for example, that the regions that were described as having the

greatest vulnerability (central part of the park) were also the regions with the lowest uncertainty associated with it (or more certainty of classification), especially in the methodology proposed by Ross (1994), which is recommended for the INP. Not coincidentally these are the areas that already show greater soil degradation by compaction and erosion (Figure S13).

In situ observations showed that human influence is higher on the lowest areas of INP, and mainly right outside the park boundaries, where land cultivation and access are possible. The conflicts over the territory, where a portion of the INP land is still used by farmers (mainly with pasture), and the socio-economic aspects also increase the vulnerability. Although the landscape and soils are not as fragile as in the upper central part of the park. Overall, the main goal should be protecting the most vulnerable environments at the upper part of INP, due to the resilience of endemic species and importance of Organossolos (Histosols) for water resources.

5.6 CONCLUSIONS

The analysis of the environmental vulnerability results, using the BBN classes, did not show areas with the very weak class, and most of the INP upper part areas had a strong vulnerability. The most fragile areas were identified with soils with high levels of carbon, profiles less developed (shallow), accentuated slopes, and a vegetation cover with herbaceous graminoid plants (high altitude fields).

The Ross (1994) model presented larger areas with less uncertainty when compared with Crepani et al. (2001), and it showed a better agreement with the results of the evaluation of INP environmental vulnerability based on the expert's knowledge. BBN models add value to environmental vulnerability mapping by integrating uncertainties and expert knowledge in the analysis.

The results of this study will allow decision-makers to identify priority areas for intervention, to reduce soil degradation in the highly vulnerable areas and to subsidize the management plan, as well as defining possible sites that should have limited access in the park.

6 GENERAL CONCLUSIONS

This work investigated the use of environmental covariates from sources such as digital elevation models, satellite imagery, geology, geomorphology and climatic data to map soil attributes and types, and to better understand the soil-landscape relationship in the upper part of Itatiaia National Park (INP). Overall, the plateau has low accessibility, and the usage of techniques to optimize the sampling points selection in the areas with best accessibility was a viable alternative, in economic terms, and efficient considering the results produced.

In general, the soils of the plateau part of the INP are predominantly shallow, with high levels of organic matter and low natural fertility. The organic soils have a high capacity to store carbon and water, on the other hand, they are very fragile. Even the mineral soils often have superficial horizons with high levels of organic matter and, in general, show lesser pedogenetic development. Many soil classes, especially those in the central part of the plateau, were not previously reported in the generalized soil map that is part of the INP management plan. With the update of the information, several classes have been included and this data will be available to future research projects and for environmental planning and preservation.

The machine learning algorithm tested and the methods of covariates selection showed that even with few data good results could be obtained. Also, for the prediction of soil attributes (2D and 3D) it is indicated to use the Generalized Additive Model with covariates selection based on the soil forming factors equation *scorpan* and Random Forest; while the selection by recursive feature elimination method can be used for class prediction. Yet the predicted maps and associated uncertainty are important steps in the development of a comprehensive database, providing quantitative soil information in a scale more appropriate to the park's demands.

Although the participatory process for evaluating environmental vulnerability requires reliability of judgment that depends on the experience of the experts consulted, taking advantage of this knowledge together with the literature review on the subject can be an efficient alternative to current vulnerability assessment methods. Models such as the Bayesian Belief Network (BBN) can integrate interdisciplinary knowledge, not only to ensure a useful model according to the needs of final users but to increase acceptance of vulnerability maps, thus adding value to current research. Also, BBN should be used whenever possible, as it allows the capture of complex relationships between vulnerability criteria in a clear way.

The most fragile areas were those with the soils with greater capacity to store or lose carbon and water depending on the usage given. Finally, the methods tested and the results obtained will help decision-makers to identify priority areas for intervention to reduce soil degradation in the highly vulnerable areas. Thus, it will be possible to reconcile public use and conservation. The data provides a scientific background to INP projects requesting environmental services payment, thus increasing the efforts for Atlantic Forest Biome conservation and, consequently, the maintenance of ecosystem services in the park.

7 BIBLIOGRAPHICAL REFERENCES

AALDERS, I.; HOUGH, R. L.; TOWERS, W. Risk of erosion in peat soils – an investigation using Bayesian belief networks. **Soil Use and Management**, v. 27, n. December, p. 538–549, 2011.

ADAMI, S. F.; COELHO, R. M.; CHIBA, M. K.; MORAES, J. F. L. DE. Environmental fragility and susceptibility mapping using geographic information systems: applications on Ribeirão do Pinhal watershed (Limeira, State of São Paulo). **Acta Scientiarum**, v. 34, n. 4, p. 433–440, 2012.

ADHIKARI, K.; HARTEMINK, A. E. Linking soils to ecosystem services - A global review. **Geoderma**, v. 262, p. 101–111, 2016.

ADHIKARI, K.; KHEIR, R. B.; GREVE, M. B.; MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B.; GREVE, M. H. High-resolution 3-D mapping of Soil texture in Denmark. **Soil Science Society of America Journal**, v. 77, p. 860–876, 2013.

AGRESTI, A. Categorical data analysis, **New york: John Wiley & Sons. Gainesville.** 2002.

AGUILERA, P. A.; FERNÁNDEZ, A.; FERNÁNDEZ, R.; RUMÍ, R.; SALMERÓN, A. Bayesian networks in environmental modelling. **Environmental Modelling and Software**, v. 26, n. 12, p. 1376–1388, 2011.

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; GONÇALVES, J. L. DE M.; SPAROVEK, G. Koppen' s climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.

AMIRIAN CHAKAN, A.; TAGHIZADEH-MEHRJARDI, R.; KERRY, R.; KUMAR, S.; KHORDEHBIN, S.; YUSEFI KHANGHAH, S. Spatial 3D distribution of soil organic carbon under different land use types. **Environmental Monitoring and Assessment**, v. 131, n. 3, p. 1–16, 2017.

ANTUNES, M. A. H.; DEBIASI, P.; SIQUEIRA, J. C. DOS S. Avaliação espectral e geométrica das imagens Rapideye e seu potencial para o mapeamento e monitoramento agrícola e ambiental. **Revista Brasileira de Cartografia**, v. 66, n. 1, p. 105–113, 2014.

ARAÚJO FILHO, J. C.; JACOMINE. Utilidade dos mapeamentos de solos e possíveis relações custo/benefício das iniciativas realizadas no país. Boletim iformativo da SBCS, v. 39, n. 1, p. 15–19, 2014.

ARROUAYS, D., GRUNDY, M., HARTEMINK, A.E., HEMPEL, J.W., HEUVELINK, G.B., HONG, S.Y., LAGACHERIE, P., LELYK, G., MCBRATNEY, A.B., MCKENZIE, N.J., D.L MENDONCA-SANTOS, M., MINASNY, B., MONTANARELLA, L., ODEH, I., SANCHEZ, P., THOMPSON, J., ZHANG, G.L. **Chapter three - GlobalSoilMap**: toward a fine-resolution global grid of soil properties. Vol. 125 of advances in agronomy. Academic Press 93–134. 2014.

ARUN, K.; LANGMEAD, C. Structure based chemical shift prediction using Random Forests non-linear regression Proceedings of the Fourth Asia-Pacific Bioinformatics Conference. Anais...Taipei, Taiwan: 2005

ASHTEKAR, J. M.; OWENS, P. R. Remembering knowledge: An expert knowledge-based approach to digital soil mapping. **Soil Horizons**, v. 54, n. 5, p. 0, 2013.

AXIMOFF, I. A.; ALVES, R. G.; RODRIGUES, R. DE C. Campos de altitude do Itatiaia: Aspectos ambientai, biológico e ecológicos. **Boletim do Parque Nacional do Itatiaia N° 18**, p. 74, 2014.

AXIMOFF, I. A.; RODRIGUES, R. D. C. Histórico dos incêndios florestais no Parque Nacional do Itatiaia. **Ciencia Florestal**, v. 21, n. 1, p. 83–92, 2011.

BACHOFER, F.; QUÉNÉHERVÉ, G.; HOCHSCHILD, V.; MAERKER, M. Multisensoral topsoil mapping in the semiarid lake Manyara region, Northern Tanzania. **Remote Sensing**, v. 7, p. 9563–9586, 2015.

BARBERENA, F. F. V. A.; BAUMGRATZ, J. F. A.; CHIAVEGATTO, B. Melastomataceae No Parque Nacional Do Itatiaia, Sudeste Do Brasil: Tribos Bertolonieae E Merianieae. **Rodriguésia**, v. 59, n. 2, p. 381–392, 2008.

BARBOSA, H. S. L.; TEXEIRA, P. H. S.; CAMPOS, A. C.; FIGUEIREDO, M. DO A.; ROCHA, L. C.; NEGREIROS, A. B. Aspectos da degradação ambiental de uma trilha recreacional na Serra do Lenheiro, São João del-Rei, MG. **Territorium Terram**, v. 5, n. 1, p. 32–40, 2015.

BARRETO, C. G.; CAMPOS, J. B.; ROBERTO, D. M.; ROBERTO, D. M.; SCHWARZSTEIN, N. T.; ALVES, G. S. G.; COELHO, W. Plano de Manejo: Parque Nacional do Itatiaia.**Encarte 3. Relatório Técnico Instituto Chico Mendes**, 2013.

BARROS, A.; GONNET, J.; PICKERING, C. Impacts of informal trails on vegetation and soils in the highest protected area in the Southern Hemisphere. **Journal of Environmental Management**, v. 127, p. 50–60, 2013.

BARROS, M. I. A.; MAGRO, T. C. Visitors' experience and lack of knowledge of minimum impact techniques in the highlands of Brazil's Itatiaia National Park. **USDA Forest Service Proceedings RMRS**, p. 374–379, 2007.

BASHARI, H.; NAGHIPOUR, A. A.; KHAJEDDIN, S. J.; SANGOONY, H.; TAHMASEBI, P. Risk of fire occurrence in arid and semi-arid ecosystems of Iran: an investigation using Bayesian belief networks. **Environmental Monitoring and Assessment**, v. 188, n. 9, 2016.

BAYESFUSION, L. GeNIe Modeler Software. **© Copyright 2018 BayesFusion, LLC**, 2018.

BEAUDETTE, D. E.; ROUDIER, P.; O'GEEN, A. T. Algorithms for quantitative pedology: A toolkit for soil scientists. **Computers and Geosciences**, v. 52, n. March, p. 258–268, 2013.

BEGUIN, J.; FUGLSTAD, G.; MANSUY, N.; PARÉ, D. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. **Geoderma**, v. 306, n. November, p. 195–205, 2017.

BENITES, V. M.; SCHAEFER, C. E. G. R.; SIMAS, F. N. B.; SANTOS, H. G. Soils associated with rock outcrops in the Brazilian mountain ranges Mantiqueira and Espinhaço. **Revista Brasileira de Botânica**, v. 30, n. 4, p. 569–577, 2007.

BHERING, S. B.; CHAGAS, C. DA S.; CARVALHO JÚNIOR, W.; PEREIRA, N. R.; CALDERANO FILHO, B.; PINHEIRO, H. S. K. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. **Pesquisa Agropecuaria Brasileira**, v. 51, n. 9, p. 1359–1370, 2016.

BIVAND, R.; KEITT, T.; ROWLINGSON, B. **rgdal: Bindings for the geospatial data abstraction library**, 2017. Disponível em: <https://cran.r-project.org/package=rgdal>

BIVAND, R.; LEWIN-KOH, N. **maptools: Tools for reading and handling spatial objects**, 2017. Disponível em: <https://cran.r-project.org/package=maptools>

BRADE, A. C. A flora do Parque Nacional do Itatiaia. **Boletim do Parque Nacional do Itatiaia Nº 5**, 1956.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRENNING, A.; BLASCHKE, T.; MONTANARELLA, L. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. *In*: **SAGA -- Seconds Out (= Hamburger Beitraege zur Physischen Geographie und Landschaftsoekologie, vol. 19)**. [s.l.] J. Boehner, T. Blaschke, L. Montanarella, 2008. p. 23–32.

BROGNIEZ, D. DE; BALLABIO, C.; STEVENS, A.; JONES, R. J. A.; MONTANARELLA, L.; WESEMAEL, B. VAN. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. **European Journal of Soil Science**, v. 66, n. 1, p. 121–134, 2015.

BRUNGARD, C. W.; BOETTINGER, J. L.; DUNIWAY, M. C.; WILLS, S. A.; EDWARDS, T. C. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239, p. 68–83, 2015.

BRUS, D. J. Balanced sampling: A versatile sampling approach for statistical soil surveys. **Geoderma**, v. 253–254, p. 111–121, 2015.

BRUS, D. J.; KEMPEN, B.; HEUVELINK, G. B. M. Sampling for validation of digital soil maps. **European Journal of Soil Science**, v. 62, n. 3, p. 394–407, 2011.

CALDERANO FILHO, B.; POLIVANOV, H.; CARVALHO JUNIRO, W.; CHAGAS, C. DA S.; CALDERANO, S. B. Avaliação da vulnerabilidade ambiental de regiões tropicais montanhosas com suporte de SIG. **Revista de Geografia**, v. 35, n. 3, p. 269–288, 2018.

CÂMARA, G.; SOUZA, R.; FREITAS, U.; GARRIDO, J. Spring: integrating remote sensing and gis by object- oriented data modelling. **Computers & graphics**, v. 2, n. 3, p. 395–403, 1996.

CAMBULE, A. H.; ROSSITER, D. G.; STOORVOGEL, J. J. A methodology for digital soil mapping in poorly-accessible areas. **Geoderma**, v. 192, n. 0, p. 341–353, 2013.

CAMBULE, A. H.; ROSSITER, D. G.; STOORVOGEL, J. J.; SMALING, E. M. A. Soil organic carbon stocks in the limpopo national park, mozambique: Amount, spatial distribution and uncertainty. **Geoderma**, v. 213, p. 46–56, 2014.

CAMERA, C.; ZOMENI, Z.; NOLLER, J. S.; ZISSIMOS, A. M.; CHRISTOFOROU, I. C.; BRUGGEMAN, A. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. **Geoderma**, v. 285, p. 35–49, 2017.

CARVALHO JUNIOR, W.; CALDERANO FILHO, B.; CHAGAS, C. DA S.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1428–1437, 2016.

CARVALHO JÚNIOR, W.; CHAGAS, C. D. S.; MUSELLI, A.; PINHEIRO, H. S. K.; PEREIRA, N. R.; BHERING, S. B. Método do hipercubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos. **Revista Brasileira de Ciência do Solo**, v. 38, n. 2, p. 386–396, 2014.

CARVALHO JÚNIOR, W.; LUMBRERAS, J. F.; LEMOS, A. L.; SANTOS, R. D.; FILHO, B. C.; WITTERN, K. P. **Mapa Mapa de Solos do Estado do Rio de Janeiro, Escala 1:500.000**, 2000.

CAVAZZI, S.; CORSTANJE, R.; MAYR, T.; HANNAM, J.; FEALY, R. Are fine resolution digital elevation models always the best choice in digital soil mapping? **Geoderma**, v. 195–196, p. 111–121, 2013.

CELIO, E.; KOELLNER, T.; GRÊT-REGAMEY, A. Modeling land use decisions with Bayesian networks: Spatially explicit analysis of driving forces on land use change. **Environmental Modelling and Software**, v. 52, p. 222–233, 2014.

CEREDA JUNIOR, A.; RÖHM, S. A. Analysis of environmental fragility using multi-criteria analysis (MCE) for integrated landscape assessment. **Journal of Urban and Environmental Engineering**, v. 8, n. 1, p. 28–37, 2014.

CHAGAS, C. DA S.; CARVALHO JUNIOR, W.; BHERING, S. B.; CALDERANO FILHO, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. **Catena**, v. 139, p. 232–240, 2016.

CHAGAS, C. DA S.; PINHEIRO, H. S. K.; CARVALHO JUNIOR, W. DE; ANJOS, L. H. C. DOS; PEREIRA, N. R.; BHERING, S. B. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. **Geoderma Regional**, v. 9, p. 47–55, 2017.

CHARTIN, C.; STEVENS, A.; GOIDTS, E.; KRÜGER, I.; CARNOL, M.; WESEMAEL, B. VAN. Mapping soil organic Carbon stocks and estimating uncertainties at the regional scale following a legacy sampling strategy (Southern Belgium, Wallonia). **Geoderma Regional**, v. 9, p. 73–86, 2017.

CHEN, S. H.; POLLINO, C. A. Good practice in Bayesian network modelling. **Environmental Modelling and Software**, v. 37, p. 134–145, 2012.

CHOUDHARY, K.; BOORI, M. S.; KUPRIYANOV, A. Spatial modelling for natural and environmental vulnerability through remote sensing and GIS in Astrakhan, Russia. **The Egyptian Journal of Remote Sensing and Space Sciences**, v. 31, n. May, p. 1–9, 2017.

CLIFFORD, D.; PAYNE, J. E.; PRINGLE, M. J.; SEARLE, R.; BUTLER, N. Pragmatic soil survey design using flexible Latin hypercube sampling. **Computers and Geosciences**, v. 67, p. 62–68, 2014.

COSTA, E. M.; ANTUNES, M. A. H.; DEBIASI, P.; ANJOS, L. H. C. DOS. Processamento de imagens RapidEye no mapeamento de uso do solo em ambiente de Mar de Morros. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1417–1427, 2016.

CREPANI, E.; MEDEIROS, J. S. DE; HERNANDEZ FILHO, P.; FLORENZANO, T. G.; DUARTE, V.; BARBOSA, C. C. F. Sensoriamento remoto e geoprocessamento aplicados ao zoneamento ecológico-econômico e ao ordenamento territorial. **INPE(INPE-8454-RPQ/722)**, p. 103, 2001.

D'ANTONIO, A.; MONZ, C.; NEWMAN, P.; LAWSON, S.; TAFF, D. Enhancing the utility of visitor impact assessment in parks and protected areas: A combined socialeecological approach. **Journal of Environmental Management**, v. 124, p. 72–81, 2013.

DE LANGE, H. J.; SALA, S.; VIGHI, M.; FABER, J. H. Ecological vulnerability in risk assessment – A review and perspectives. **Science of the Total Environment**, v. 408, n. 18, p. 3871-3879. 2010.

DELGADO, R. C.; PEREIRA, M. G.; TEODORO, P. E.; SANTOS, G. L. DOS; CARVALHO, D. C. DE; MAGISTRALI, I. C.; VILANOVA, R. S. Seasonality of gross primary production in the Atlantic Forest of Brazil. **Global Ecology and Conservation**, v. 14, p. e00392, 2018.

DÍAZ-URIARTE, R.; ANDRÉS, S. A. Gene selection and classification of microarray data using random forest. **BMC bioinformatics**, v. 7, p. 3, 2006.

DONAGEMMA, K. G.; CAMPOS, D. V. B.; CALDERANO, S. B.; TEIXEIRA, W. G.; VIANA, J. H. M. **Manual de métodos de Análise de solo. Embrapa Solos**. Rio de Janeiro. 2011. 225p.

ESRI. **ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. Software: ArcGIS e ArcINFO**Nova York, EUA, 2015. Disponível em: <http://www.esri.com/software/arcgis/arcgis-for-desktop/free-trial>

FERNANDES FILHO, E. I.; SCHAEFER, C. E. G. R.; ABRAHÃO, W. A. P. **Mapa de solos do estado de Minas Gerais: Legenda expandida**, 2010.

FICK, S. E.; HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, v. 37, n. 12, p. 4302–4315, 2017.

FIGUEIREDO, M. DO A.; BRITO, Í. DE A.; TAKEUCHI, R. C.; ALMEIDA-ANDRADE, M.; ROCHA, C. T. V. Compactação do solo como indicador pedogeomorfológico para erosão em trilhas de unidades de conservação: Estudo de caso no parque nacional da Serra do Cipó, MG. **Revista de Geografia. Recife**, v. 8, n. 3, p. 236–247, 2010.

FORKUOR, G.; HOUNKPATIN, O. K. L.; WELP, G.; THIEL, M. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: A comparison of machine learning and multiple linear regression models. **Plos One**, v. 12, n. 1, p. e0170478, 2017.

FREIRE, E. DOS S.; LEMOS, L. DE O. Uso Público no Parque Nacional do Itatiaia. *In*: **VII Congresso Brasileiro de Geógrafos**. p. 1–12, 2014

GENELETTI, DAVIDE & DAWA, DORJE. Environmental impact assessment of mountain tourism in developing regions: A study in Ladakh, Indian Himalaya. **Environmental Impact Assessment Review**. V. 29, n. 4, p. 229-242. 2009.

GONZALEZ-REDIN, J.; LUQUE, S.; POGGIO, L.; SMITH, R.; GIMONA, A. Spatial Bayesian Belief networks as a planning decision tool for mapping ecosystem services trade-offs on forested landscapes. **Environmental Research**, v. 144, p. 15–26, 2016.

GRIMM, R.; BEHRENS, T.; MÄRKER, M.; ELSENBEER, H. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. **Geoderma**, v. 146, n. 1–2, p. 102–113, 2008.

GUO, P. T.; LI, M. F.; LUO, W.; TANG, Q. F.; LIU, Z. W.; LIN, Z. M. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. **Geoderma**, v. 237–238, p. 49–59, 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: Data mining, inference and prediction**. 2º ed. Stanford, California: Springer Series in Statistics, 2009.

HEMMING, V.; BURGMAN, M. A.; HANEA, A. M.; MCBRIDE, M. F.; WINTLE, B. C. A practical guide to structured expert elicitation using the IDEA protocol. **Methods in Ecology and Evolution**, v. 9, n. 1, p. 169–180, 2017.

HENGL, T. **A Practical guide to geostatistical mapping**. 1º ed. Amsterdam: p. 293, 2009

HENGL, T. E.; HEUVELINK, G. B. M. New challenges for predictive soil mapping**. Anais. Global Workshop on Digital Soil Mapping. Anais**. 2004.

Hengl, T., MacMillan, R.A., Predictive Soil Mapping with R. **OpenGeoHub foundation**, Wageningen, the Netherlands, www.soilmapper.org, ISBN: 978-0-359-30635-0. p. 370, 2019.

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; SHEPHERD, K. D.; SILA, A.; MACMILLAN, R. A.; JESUS, J. M. DE; TAMENE, L.; TONDOH, J. E. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. **PLOS ONE**, v. 10, n. 6, p. 1–26, 2015.

HENGL, T.; HEUVELINK, G. B. M.; ROSSITER, D. G. About regression-kriging: From equations to case studies. **Geoderma**, v. 33, p. 1301–1351, 2007.

HIJMANS, R. J. **raster: geographic data analysis and modeling**, 2016. Disponível em: <https://cran.r-project.org/package=raster>

HØJSGAARD, S. Graphical Independence networks with the gRain package for R. **Journal of Statistical Software**, v. 46, n. 10, p. 1–26, 2012.

HUANG, J.; MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B.; TRIANTAFILIS, J. Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. **Science of the Total Environment**, v. 609, n. July, p. 621–632, 2017.

HUETE, A. R. A. Soil-Adjusted Vegetation Index (SAVI). **Remote Sensing of Environment**, v. 25, p. 295–308, 1988.

IBGE. **Manual Técnico de Pedologia**. 3º Edição ed. Rio de Janeiro: p. 425, 2015

IUSS WORKING GROUP WRB. **World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps. Update 2015**. Rome: 2015. 1-192 p.

IWAMOTO, P.K.; RODRIGUES, M.G. Uma proposta de delimitação da zona de amortecimento do Parque Nacional do Itatiaia, Rio de Janeiro, Brasil. **Revista Nordestina de Ecoturismo**, v. 4, n. 2, p. 5–14, 2011.

JEONG, G.; OEVERDIECK, H.; PARK, S. J.; HUWE, B.; LIESS, M. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. **Catena**, v. 154, p. 73–84, 2017.

JEUNE, W.; FRANCELINO, M. R.; SOUZA, E. DE; FERNANDES FILHO, E. I.; ROCHA, G. C. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. **Revista Brasileira de Ciencia do Solo**, v. 42, p. 1–20, 2018.

KEMPEN, B.; BRUS, D. J.; HEUVELINK, G. B. M.; STOORVOGEL, J.J. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. **Geoderma**, v. 151, p. 311–326, 2009.

KEMPEN, B.; BRUS, D. J.; STOORVOGEL, J. J. Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. **Geoderma**, v. 162, n. 1–2, p. 107–123, 2011.

KEMPEN, B.; HEUVELINK, G. B. M.; BRUS, D. J.; STOORVOGEL, J. J. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. **European Journal of Soil Science**, v. 61, n. 3, p. 333–347, 2010.

KIDD, D.; MALONE, B.; MCBRATNEY, A. B.; MINASNY, B.; WEBB, M. Operational sampling challenges to digital soil mapping in Tasmania, Australia. **Geoderma Regional**, v. 4, p. 1–10, 2015.

KIDD, D.; WEBB, M.; MALONE, B.; MINASNY, B.; MCBRATNEY, A. Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. **Soil Research**, v. 53, n. 8, p. 932–955, 2015.

KUHN, M.; WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; BENESTY, M.; LESCARBEAU, R.; ZIEM, A.; SCRUCCA, L.; TANG, Y.; CANDAN, C.; HUNT, T. **caret:** classification and regression training, 2017. Disponível em: <https://cran.r-project.org/package=caret>

LAGACHERIE, P.; SNEEP, A. R.; GOMEZ, C.; BACHA, S.; COULOUMA, G.; HAMROUNI, M. H.; MEKKI, I. Combining Vis-NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). **Geoderma**, v. 209–210, p. 168–176, 2013.

LANDUYT, D.; BIEST, K. VAN DER; BROEKX, S.; STAES, J.; MEIRE, P.; GOETHALS, P. L. M. A GIS plug-in for Bayesian belief networks: Towards a transparent software framework to assess and visualise uncertainties in ecosystem service mapping. **Environmental Modelling & Software**, v. 71, n. June, p. 30–38, 2015.

LANDUYT, D.; BROEKX, S.; ROB, D.; ENGELEN, G.; AERTSENS, J.; GOETHALS, P. L. M. A review of Bayesian belief networks in ecosystem service modelling. **Environmental Modelling and Software**, v. 46, p. 1–11, 2013.

LI, Y.; ZHU, A.; SHI, Z.; LIU, J.; DU, F. Supplemental sampling for digital soil mapping based on prediction uncertainty from both the feature domain and the spatial domain. **Geoderma**, v. 284, p. 73–84, 2016.

LIESS, M. Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. **Spatial Statistics**, v. 13, p. 106–122, 2015.

LIMA, W. G.; GUEDES-BRUNI, R. R. Myrceugenia (Myrtaceae) ocorrentes no Parque Nacional do Itatiaia, Rio de Janeiro. **Rodriguésia**, v. 55, n. 85, p. 71–94, 2004.

LIU, F.; ZHANG, G.-L.; SUN, Y.-J.; ZHAO, Y.-G.; LI, D.-C. Mapping the three-dimensional distribution of soil organic matter across a subtropical hilly landscape. **Soil Science Society of America Journal**, v. 77, n. 4, p. 1241, 2013.

MAGRO, T. CRISTINA; BARROS, M. I. A. **Understanding use and users at Itatiaia National Park**. 1º ed. Ralf Buckley. (Org.), 2004.

MALONE, B. P.; MCBRATNEY, A. B.; MINASNY, B. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. **Geoderma**, v. 160, n. 3–4, p. 614–626, 2011.

MANFRÉ, L. A.; SILVA, A. M. DA; URBAN, R. C.; RODGERS, J. Environmental fragility evaluation and guidelines for environmental zoning: A study case on Ibiuna (the Southeastern Brazilian region). **Environmental Earth Sciences**, v. 69, n. 3, p. 947–957, 2013.

MARCOT, B. G. Metrics for evaluating performance and uncertainty of Bayesian network models. **Ecological Modelling**, v. 230, p. 50–62, 2012.

MARRA, G.; WOOD, S. N. Practical variable selection for generalized additive models. **Computational Statistics and Data Analysis**, v. 55, n. 7, p. 2372–2387, 2011.

MCBRATNEY, A. B.; MENDONÇA-SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, p. 3–52, 2003.

MCBRATNEY, A. B.; MINASNY, B.; CATTLE, S. R.; VERVOORT, R. W. From pedotransfer functions to soil inference systems. **Geoderma**, v. 109, p. 41–73, 2002.

MEDEIROS, J. C.; COOPER, M.; ROSA, J. D.; GRIMALDI, M.; COQUET, Y. Assessment of pedotransfer functions for estimating soil water retention curves for the Amazon region. **Revista Brasileira de Ciência do Solo**, v. 38, n. 4, p. 730–743, 2014.

MEERSMANS, J.; MARTIN, M. P.; RIDDER, F. DE; LACARCE, E.; WETTERLIND, J.; BAETS, S. DE; BAS, C. LE; LOUIS, B. P.; ORTON, T. G.; BISPO, A.; ARROUAYS, D. A novel soil organic C model using climate, soil type and management data at the national scale in France. **Agronomy for Sustainable Development**, v. 32, n. 4, p. 873–888, 2012.

MEIER, M.; SOUZA, E.; FRANCELINO, M. R.; FERNANDES FILHO, E. I.; SCHAEFER, C. E. G. Digital soil mapping using machine learning algorithms in a tropical mountainous area. **Revista Brasileira de Ciência do Solo**, v. 42, p. 1–23, 2018.

MENEZES, M. D. DE; SILVA, S. H. G.; MELLO, C. R. DE; OWENS, P. R.; CURI, N. Knowledge-based digital soil mapping for predicting soil properties in two representative watersheds. **Scientia Agricola**, v. 75, n. 2, p. 144–153, 2018.

MENEZES, M. D. DE; SILVA, S. H. G.; MELLO, C. R. DE; OWENS, P. R.; CURI, N. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. **Scientia Agricola**, v. 71, n. 4, p. 316–323, 2014.

MERRILL, H. R.; GRUNWALD, S.; BLIZNYUK, N. Semiparametric regression models for spatial prediction and uncertainty quantification of soil attributes. **Stochastic Environmental Research and Risk Assessment**, p. 1–14, 2016.

MEYER, S. R.; JOHNSON, M. L.; LILIEHOLM, R. J.; CRONAN, C. S. Development of a stakeholder-driven spatial modeling framework for strategic landscape planning using Bayesian networks across two urban-rural gradients in Maine, USA. **Ecological Modelling**, v. 291, p. 42–57, 2014.

MEZABARBA, V.; VIANA FILHO, M. D. M.; BORGES, R. A. X.; MANSANO, V. D. F. Ericaceae do Parque Nacional do Itatiaia, RJ, Brasil. **Hoehnea**, v. 40, n. 1, p. 115–130, 2013.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers and Geosciences**, v. 32, n. 9, p. 1378–1388, 2006.

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

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. *In*: OLIVEIRA, S. R. (Ed.). **Sistemas inteligentes: fundamentos e aplicações**. 1◦ ed. Barueri-SP: Sistemas Inteligentes Fundamentos e Aplicações, 2003. p. 89–114.

MONTANARELLA, L.; WESEMAEL, B. VAN; STEVENS, A.; NOCITA, M.; TO, G. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. **PLOS ONE**, v. 8, n. 6, p. 1–13, 2013.

MORIM, MARLI PIRES BARROSO, G. M. Leguminosae arbustivas e arbóreas da Floresta Atlântica do Parque Nacional do Itatiaia, sudeste do Brasil; subfamilias Caesalpinioideae e Mimosoideae. **Rodriguesia**, v. 58, p. 423–468, 2007.

MULDER, V. L.; LACOSTE, M.; ARROUAYS, D. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two-meter depth. **Science of the Total Environment**, v. 573, p. 1352–1369, 2016.

MULDER, V. L.; LACOSTE, M.; RICHER-DE-FORGES, A. C.; MARTIN, M. P.; ARROUAYS, D. National versus global modelling the 3D distribution of soil organic carbon in mainland France. **Geoderma**, v. 263, p. 16–34, 2016.

MUTANGA, O.; ADAM, E.; CHO, M. A. High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm. **International Journal of Applied Earth Observation and Geoinformation**, v. 18, n. 1, p. 399–406, 2012.

NGUYEN, A. K.; LIOU, YUEI-AN; LI, MING-HSU; TRAN, T. A. Zoning eco-environmental vulnerability for environmental management and protection. **Ecological Indicators**, v. 69, p. 100-117. 2016.

NGUYEN, C.; WANG, Y.; NGUYEN, H. N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. **Journal of Biomedical Science and Engineering**, v. 06, n. 05, p. 551–560, 2013.

NUSSBAUM, M.; SPIESS, K.; BALTENSWEILER, A.; GROB, U.; KELLER, A.; GREINER, L.; SCHAEPMAN, M. E.; PAPRITZ, A.; DYNAMICS, P.; LABORATORIES, R. S. Evaluation of digital soil mapping approaches with large sets of environmental covariates. **SOIL Discussions**, v. 4, p. 1–22, 2018.

OLIVE, N. D.; MARION, J. L. The influence of use-related, environmental, and managerial factors on soil loss from recreational trails. **Journal of Environmental Management**, v. 90, p. 1483–1493, 2009.

OLIVEIRA, J. G. R.; TAVARES FILHO, J.; BARBOSA, G. M. D. C. Qualidade física do solo das trilhas do parque estadual do Cerrado – PR. **Semina: Ciências Agrárias,** v. 34, n. 4, p. 1715–1722, 2013.

OLIVEIRA, S. N.; CARVALHO JÚNIOR, O. A.; MARTINS, É. DE S.; SILVA, T. M. DA; GOMES, R. A. T.; GUIMARÃES, R. F. Identificação de unidades de paisagem e sua implicação para o ecoturismo no parque nacional da Serra dos Órgãos, Rio de Janeiro. **Revista Brasileira de Geomorfologia**, v. 8, n. 1, p. 87–107, 2007.

PINHEIRO, H. S. K.; CARVALHO JUNIOR, W.; CHAGAS, C. DA S.; ANJOS, L. H. C.; OWENS, P. R. Prediction of topsoil texture through regression trees and multiple linear regressions. **Revista Brasileira de Ciência do Solo**, v. 42, p. 1–21, 2018.

PINHEIRO, H. S. K.; CHAGAS, C. DA S.; CARVALHO JUNIOR, W.; ANJOS, L. H. C. Ferramentas de pedometria para caracterização da composição granulométrica de perfis de solos hidromórficos. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1326–1338, 2016.

PINHEIRO, H. S. K.; HELENA, L.; ANJOS, C.; XAVIER, P. A. M.; CESAR, S. Quantitative pedology to evaluate a soil profile collection from the Brazilian semi-arid region. **South African Journal of Geomatics**, v. 1862, p. 1–11, 2018.

POGGIO, L.; GIMONA, A. 3D mapping of soil texture in Scotland. **Geoderma Regional**, v. 9, p. 5–16, 2017a.

POGGIO, L.; GIMONA, A. Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. **Science of the Total Environment**, v. 579, p. 1094-1110, 2017b.

POGGIO, L.; GIMONA, A. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland. **Geoderma**, v. 232–234, p. 284–299, 2014.

POGGIO, L.; GIMONA, A.; BREWER, M. J. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. **Geoderma**, v. 209–210, p. 1–14, 2013.

POGGIO, L.; GIMONA, A.; BROWN, I.; CASTELLAZZI, M. Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. **Geoderma**, v. 160, n. 2, p. 175–188, 2010.

POGGIO, L.; GIMONA, A.; SPEZIA, L.; BREWER, M. J. Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. **Geoderma**, v. 277, p. 69–82, 2016.

POGGIO, L.; SIMONETTI, E.; GIMONA, A. Enhancing the WorldClim data set for national and regional applications. **Science of the Total Environment**, v. 625, p. 1628–1643, 2018.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**Vienna, Austria, 2018. Disponível em: <https://www.r-project.org/>

RAD, M. R. P.; TOOMANIAN, N.; KHORMALI, F.; BRUNGARD, C. W.; KOMAKI, C. B.; BOGAERT, P. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. **Geoderma**, v. 232–234, p. 97–106, 2014.

RAPIDEYE. **RapidEye Mosaic™ Product SpecificationsRapidEye Delivering the World**, 2012.

RICHTER, M.; SOUZA, E. M. F. DE R. DE. Avaliação de impactos ecológicos e sociais do uso público no Parque Nacional do Itatiaia - Trilha Alto dos Brejos. **Boletim de Geografia**, v. 31, n. 1, p. 91–100, 2013.

RODRIGUES, K. R. Geoambientes e solos em ambientes altimontanos nos parques nacionais de Itatiaia e Caparaó-MG. **Tese (Doutorado em Ciência do Solo) Universidade Federal de Viçosa**. p. 124, 2011.

RODRIGUEZ-GALIANO, V. F.; CHICA-OLMO, M.; ABARCA-HERNANDEZ, F.; ATKINSON, P. M.; JEGANATHAN, C. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. **Remote Sensing of Environment**, v. 121, p. 93–107, 2012.

ROSS, J. L. S. Análise empírica da fragilidade dos ambientes naturais e antropizados. **Revista do Departamento de Geografia**, v. 8, p. 63–74, 1994.

ROSS, J. L. S. Landforms and environmental planning: Potentialities and fragilities. **Revista do Departamento de Geografia**, p. 38–51, 2012.

ROUDIER, P. **Package 'clhs'R Package**, 2015.

ROUDIER, P.; HEWITT, A. E.; BEAUDETTE, D. E. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. **Digital Soil Assessments and Beyond**, n. September 2015, p. 227–231, 2012.

ROVANI, F. F. M.; VIERA, M. Vulnerabilidade Natural do Solo de Silveira Martins-RS. **Floresta e Ambiente**, v. 23, n. 2, p. 151–160, 2016.

RUPPERT, D.; P. WAND, M.; CARROLL, R. **Semiparametric Regression**. [s.l: s.n.]. v. 101

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MIGUEL, P. Building predictive models of soil particle-size distribution. **Revista Brasileira de Ciência do Solo**, v. 37, n. 2, p. 422–430, 2013.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M.; ANJOS, L. H. C. Do more detailed environmental covariates deliver more accurate soil maps? **Geoderma**, v. 243–244, p. 214–227, 2015.

SANTOS, H. G.; JACOMINE, P. K. T.; ANJOS, L. H. C. DOS; OLIVEIRA, V. Á.; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A.; CUNHA, T. J. F.; OLIVEIRA, J. B. **Brazilian soil classification system**. 5◦ ed. Brasília: Embrapa, 2018.

SANTOS, R. D.; LEMOS, R. C.; SANTOS, H. G.; KER, J. C.; ANJOS, L. H. C.; SHIMIZU, S. H. Manual de descrição e coleta de solo no campo. **7°**ed**. Revisada e ampliada. SBCS. Viçosa**. p. 100, 2015

SANTOS, R. F. DOS; PIRES NETO, A. G.; CSORDAS, S. M. O Parque Nacional do Itatiaia. **Fundação Brasileira para o Desenvolvimento Sustentável**, p. 09-19, 2000.

SCUTARI, M. Learning Bayesian Networks with the bnlearn R Package. **Journal of Statistical Software**, v. 35, n. 3, p. 1–22, 2010.

SENA, Í. S. DE; TEIXEIRA, H. W.; FIGUEIREDO, M. DO A.; ROCHA, L. C. Degradação dos solos ao longo de uma trilha de destino a atrativos do monumento geoturístico Serra de São José, Tiradentes, Minas Gerais, Brasil. **Geonomos**, v. 22, n. 2, p. 70–76, 2014.

SHANNON, C. E, A mathematical theory of communication. **Bell System Technical Journal**. v.27, n.3, p.79–423, 1948.

SILVA NETO, E. C.; SANTOS, J. J. S.; PEREIRA, M. G.; MARANHÃO, D. D. C.; BARROS, F. DA C.; ANJOS, L. H. C. Paleoenvironmental characterization of a high-mountain environment in the Atlantic forest in Southeastern Brazil. **Revista Brasileira de Ciência do Solo**, v. 42, p. 1–17, 2018.

SILVA NETO, S. J.; PEIXOTO, A. L. Rubiaceae do Parque Nacional de Itatiaia, Rio de Janeiro, Brasil. **Boletim do Parque Nacional do Itatiaia N◦ 14**, 2012.

SILVA, S. H. G.; TEIXEIRA, A. F. DOS S.; MENEZES, M. D. DE; GUILHERME, L. R. G.; MOREIRA, F. M. DE S.; CURI, N. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence analyzer (pXRF). **Ciência e Agrotecnologia**, v. in press, n. 6, p. 648–664, 2017.

SINDAYIHEBURA, A.; OTTOY, S.; DONDEYNE, S.; MEIRVENNE, M. VAN; ORSHOVEN, J. VAN. Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus. **Catena**, v. 156, n. April, p. 161–175, 2017.

SOARES, P. F. C.; ANJOS, L. H. C.; PEREIRA, M. G.; PESSENDA, L. C. R. Histosols in an Upper Montane Environment in the Itatiaia Plateau. **Revista Brasileira de Ciência do Solo**, v. 40, p. 1–13, 2016.

SOMARATHNA, P. D. S. N.; MINASNY, B.; MALONE, B. P. More data or a better model? figuring out what matters most for the spatial prediction of soil carbon. **Soil Science Society of America Journal**, v. 0, n. 0, p. 0, 2017.

SOUSA, G. M. DE; FERNANDES, M. DO C.; COSTA, G. A. O. P. DA. Classificação da susceptibilidade à ocorrência de incêndio através de mineração de dados e geobia. **Revista Brasileira de Cartografia**, v. 3, n. 67, p. 555–567, 2015.

SOUZA, E. DE; BATJES, N. H.; PONTES, L. M. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. **Scientia Agricola**, v. 73, n. 6, p. 525–534, 2016.

SPÖRL, C. **Metodologia para elaboração de modelos de fragilidade ambiental utilizando redes neurais**. p.185, 2007

SPÖRL, C.; CASTRO, E.; LUCHIARI, A. Aplicação de redes neurais artificiais na construção de modelos de fragilidade ambiental. **Revista do Departamento de Geografia - USP**, v. 21, n. 0, p. 113–135, 2011.

SPÖRL, C.; ROSS, J. L. S. Análise comparativa da fragilidade ambiental com aplicação de três modelos. **GEOUSP - Espaço e Tempo**, v. 15, p. 39–49, 2004.

STUMPF, F.; SCHMIDT, K.; BEHRENS, T.; SCHÖNBRODT-STITT, S.; BUZZO, G.; DUMPERTH, C.; WADOUX, A.; XIANG, W.; SCHOLTEN, T. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. **Journal of Plant Nutrition and Soil Science**, v. 000, p. 1–11, 2016.

TAGHIZADEH-MEHRJARDI, R. Digital mapping of cation exchange capacity using genetic programming and soil depth functions in Baneh region, Iran. **Archives of Agronomy and Soil Science**, v. 62, n. 1, p. 109–126, 2016.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; KERRY, R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. **Geoderma**, v. 266, p. 98–110, 2016.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; MINASNY, B.; TRIANTAFILIS, J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. **Geoderma**, v. 253–254, p. 67–77, 2015.

TEN CATEN, A. DALMOLIN, R. S. D.; PEDRON, F. A.; MENDONÇA-SANTOS, M. DE L. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v. 35, n. 4, p. 53–62, 2011.

TOMCZYK, A. M. A GIS assessment and modelling of environmental sensitivity of recreational trails: The case of Gorce National Park, Poland. **Applied Geography journal**, v. 31, p. 339–351, 2011.

TOMCZYK, A. M.; EWERTOWSKI, M. Planning of recreational trails in protected areas: Application of regression tree analysis and geographic information systems. **Applied Geography**, v. 40, p. 129–139, 2013a.

TOMCZYK, A. M.; EWERTOWSKI, M. Quantifying short-term surface changes on recreational trails: The use of topographic surveys and ‘digital elevation models of differences’ (DODs). **Geomorphology**, v. 183, p. 58–72, 2013b.

TOMZHINSKI, G. W.; COURA, P. H. F.; FERNANDES, M. DO C. Avaliação da detecção de focos de calor por Sensoriamento Remoto para o Parque Nacional do Itatiaia. **Biodiversidade Brasileira**, v. 1, n. 2, p. 201–211, 2011.

TOMZHINSKI, G. W.; RIBEIRO, K. T.; FERNANDES, M. DO C. Análise geoecológica dos incêndios florestais do Parque Nacional do Itatiaia**. Boletim do Parque Nacional do Itatiaia N° 15**, 2012.

TRUONG, P. N.; HEUVELINK, G. B. M. Uncertainty quantification of soil property maps with statistical expert elicitation. **Geoderma**, v. 202–203, p. 142–152, 2013.

VALLE, I. C.; FRANCELINO, M. R.; PINHEIRO, H. S. K. Mapeamento da fragilidade ambiental na bacia do rio Aldeia Velha, RJ. **Floresta e Ambiente**, v. 23, n. 2, p. 295–308, 2016.

VAŠÁT, R.; KODEŠOVÁ, R.; BORŮVKA, L.; JAKŠÍK, O.; KLEMENT, A.; BRODSKÝ, L. Combining reflectance spectroscopy and the digital elevation model for soil oxidizable carbon estimation. **Geoderma**, v. 303, n. May, p. 133–142, 2017.

VAYSSE, K.; LAGACHERIE, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. **Geoderma**, v. 291, p. 55–64, 2017.

VERMEULEN, D.; NIEKERK, A. VAN. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. **Geoderma**, v. 299, p. 1–12, 2017.

VERMOTE, E. F.; HERMAN, M.; MORCRETTE, J. Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: An Overview. **IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING**, v. 35, n. 3, p. 675–686, 1997.

VERONESI, F.; CORSTANJE, R.; MAYR, T. Landscape scale estimation of soil carbon stock using 3D modelling. **Science of the Total Environment**, v. 487, n. 1, p. 578–586, 2014.

WERE, K.; BUI, D. T.; DICK, Ø. B.; SINGH, B. R. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, v. 52, p. 394–403, 2015.

WOOD, S. Generalized Additive Models: An Introduction with R. **CRC Texts in Statistical Science**, 2006.

XIONG, X.; GRUNWALD, S.; MYERS, D. B.; KIM, J.; HARRIS, W. G.; BLIZNYUK, N. Assessing uncertainty in soil organic carbon modeling across a highly heterogeneous landscape. **Geoderma**, v. 251–252, p. 105–116, 2015.

YANG, R. M.; ZHANG, G. L.; LIU, F.; LU, Y. Y.; YANG, F.; YANG, F.; YANG, M.; ZHAO, Y. G.; LI, D. C. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. **Ecological Indicators**, v. 60, p. 870–878, 2016.

ZHANG, G. LIN; LIU, F.; SONG, X. DONG. Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture, v. 16, n. 12, p. 2871–2885, 2017.

8 SUPPLEMENTARY MATERIAL

Chapter II

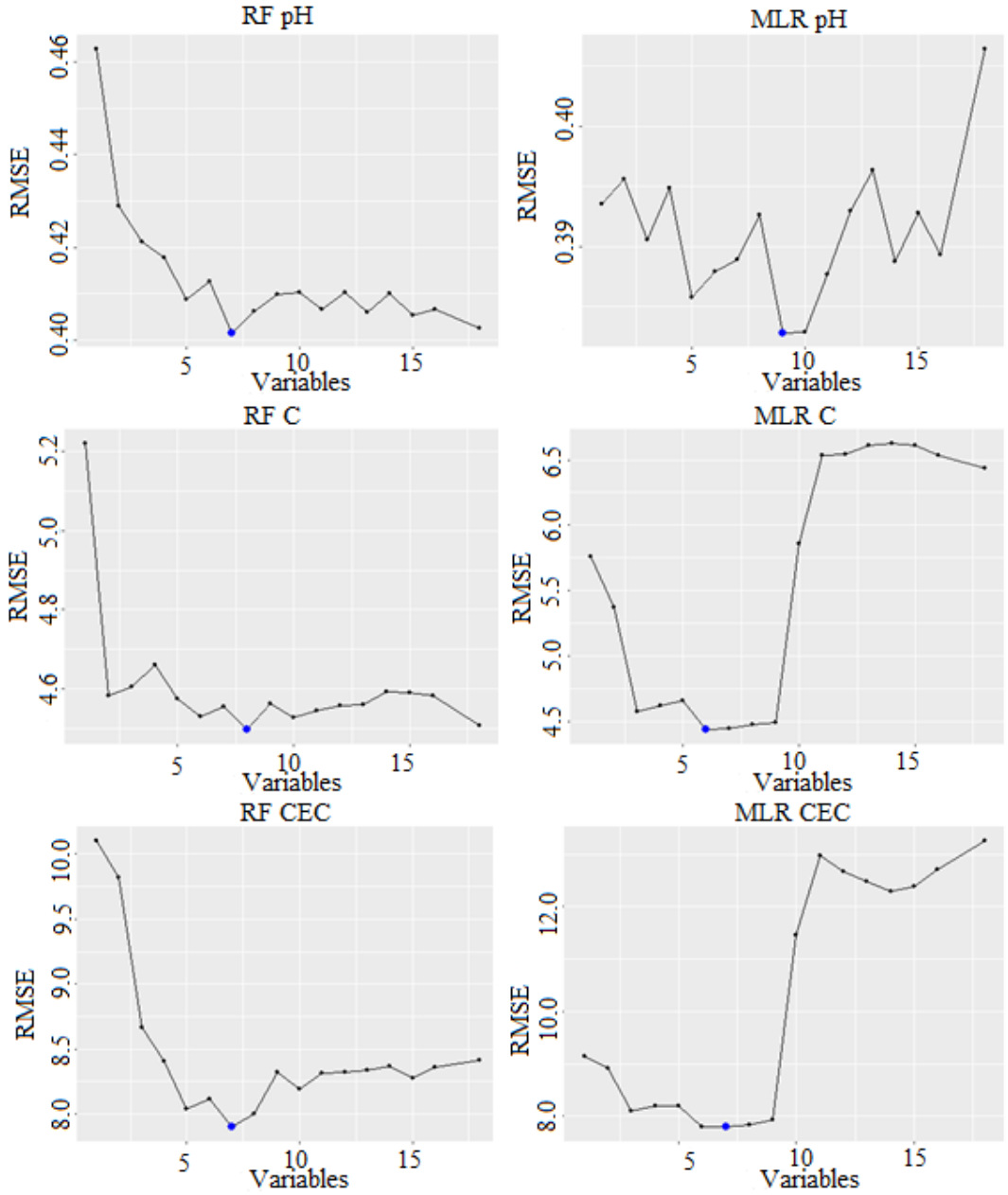


Figure S1. Optimum number of covariates selected by the RFE for each soil attributes in RF an MLR models.

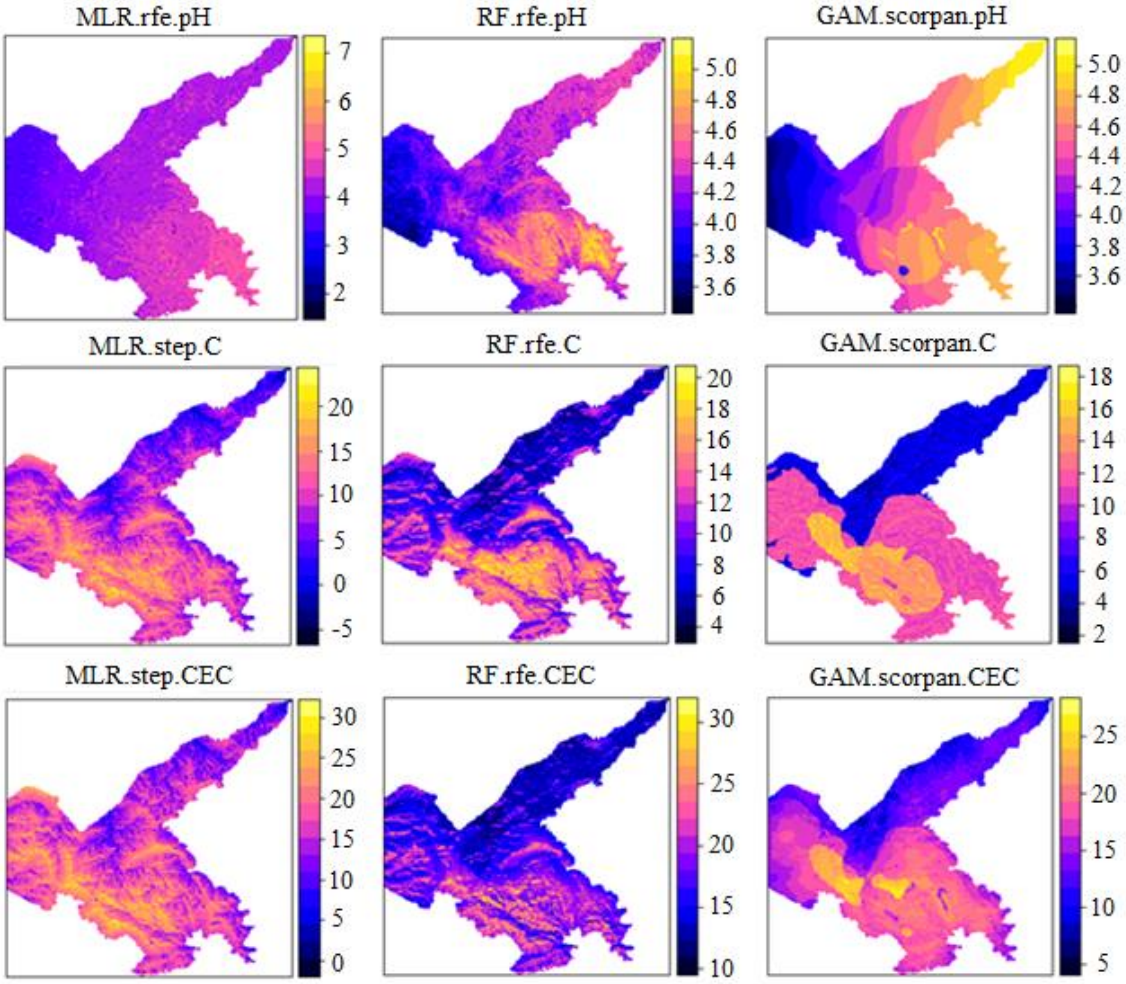


Figure S2. Spatial prediction of soil attributes using the best model evaluated by external validation for each method tested. For pH, Carbon content (%) and CEC (cmolc.dm^{-3}).

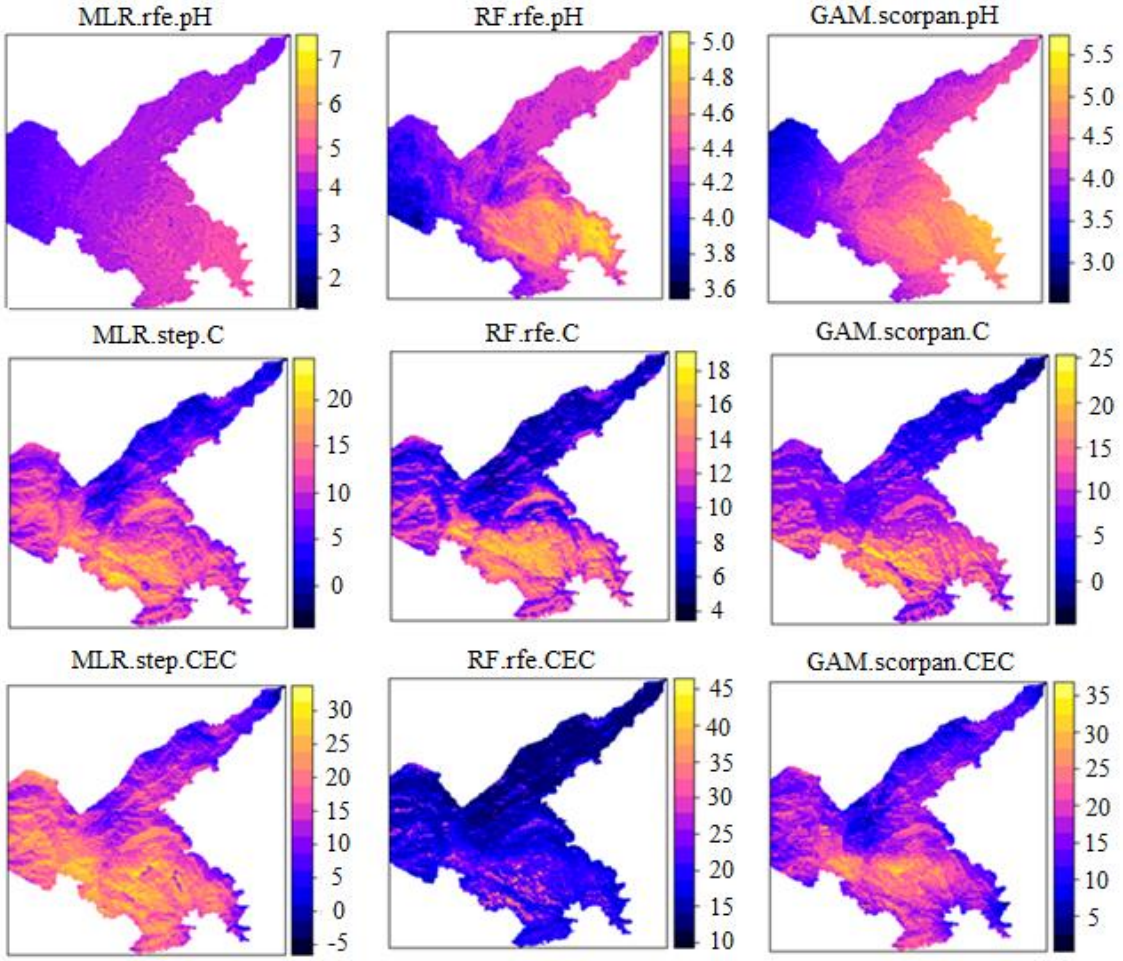


Figure S3. Spatial prediction of soil attributes using the best model evaluated by LOO-CV for each method tested. For pH, Carbon content (%) and CEC (cmolc.dm^{-3}).

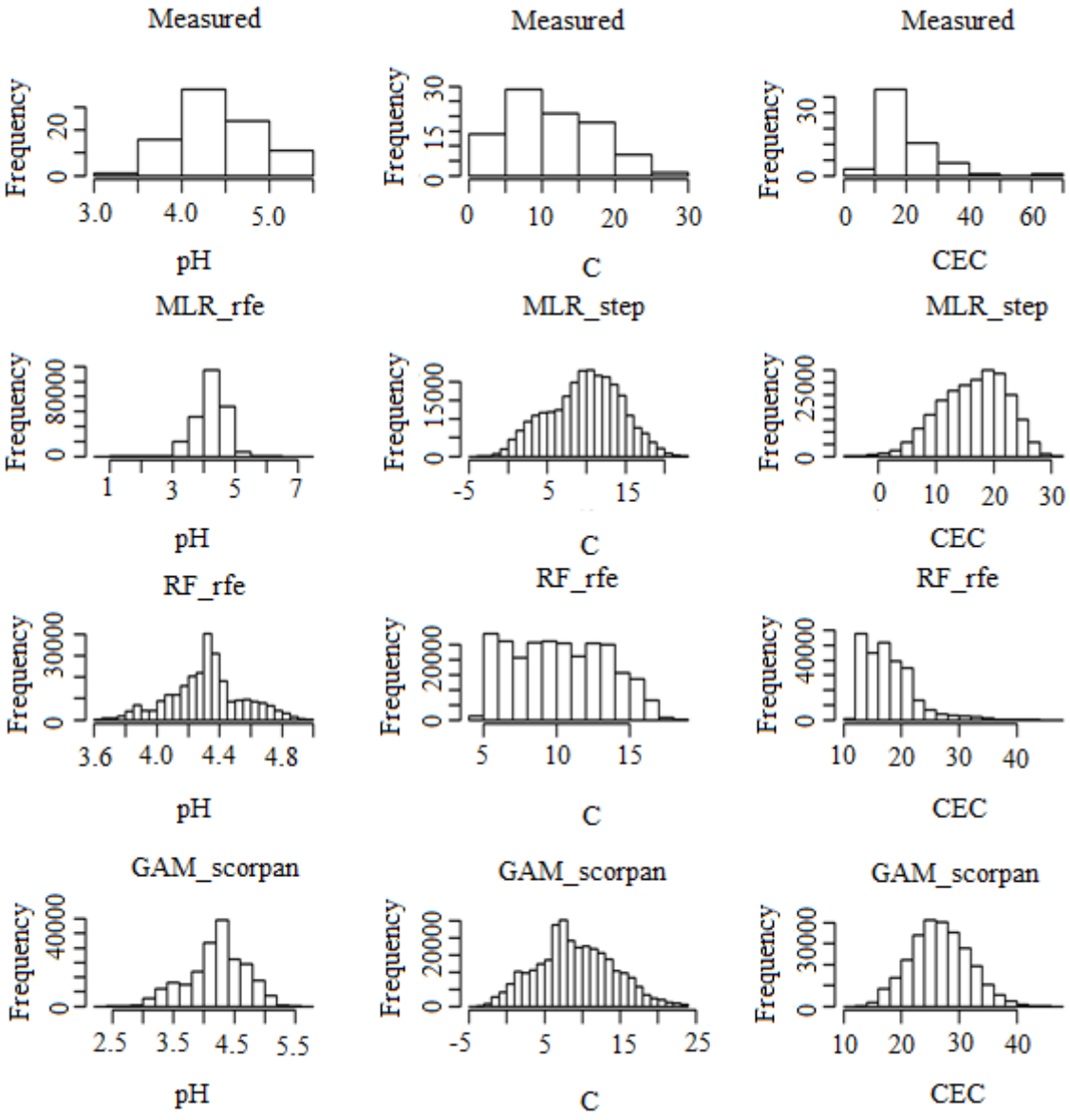


Figure S4. Histogram of measured data and predicted values in the grid by better models

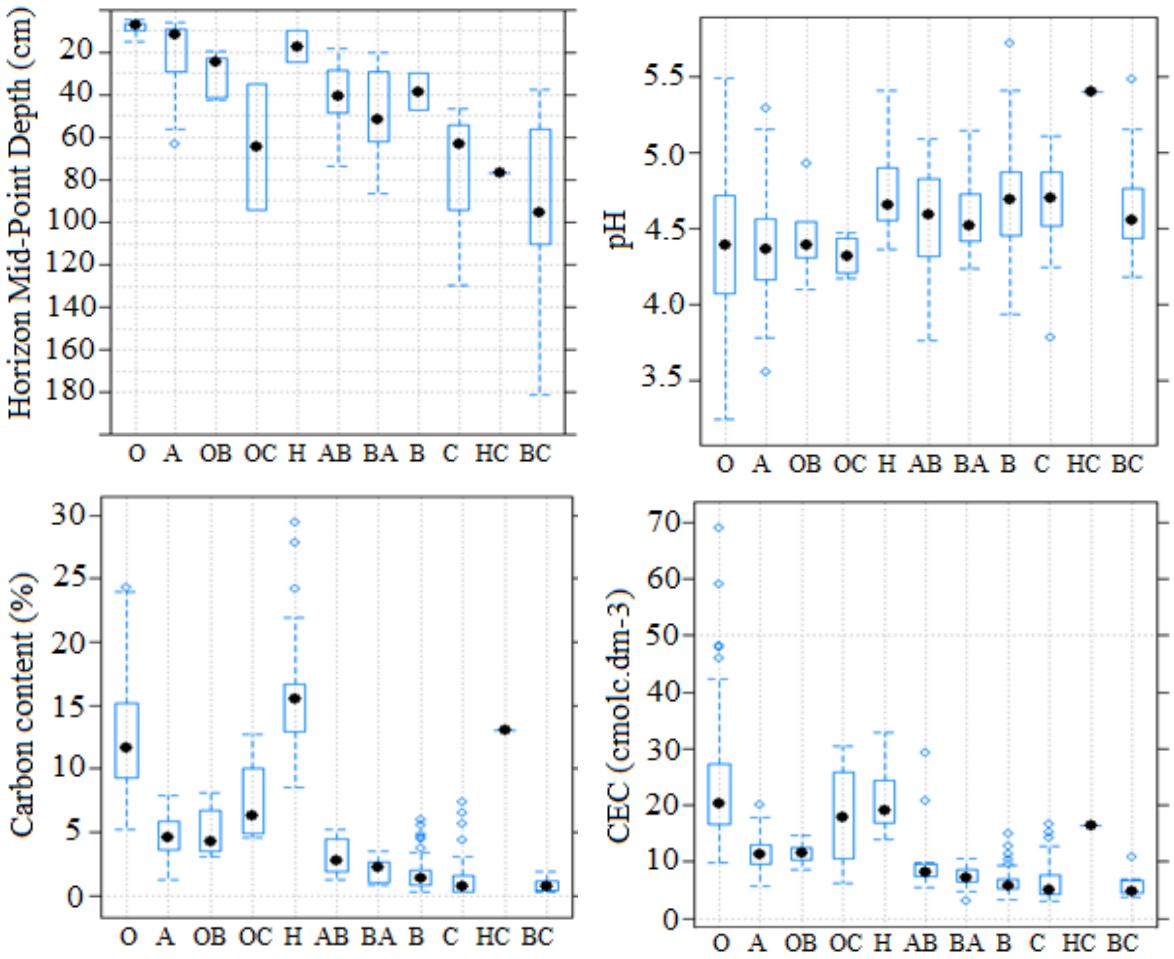


Figure S5. Box-plot of measured values for each type of horizon. Top left horizon mid-point (cm), top right pH, bottom left soil carbon content (%), and bottom right CEC

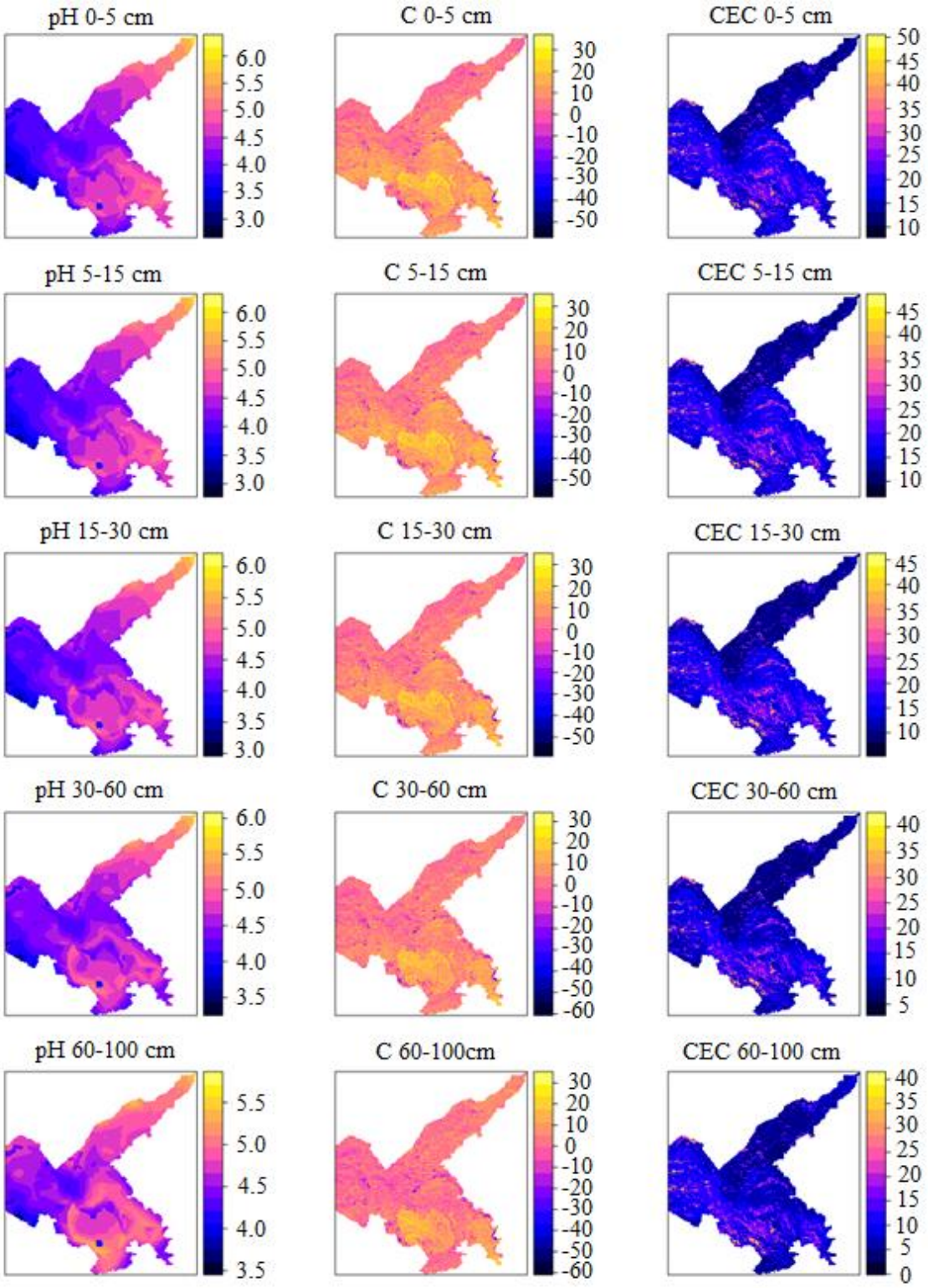


Figure S6. Maps of the soil attributes at five depths (pH left; Carbon content, %, centre; CEC, $cmol_c.dm^{-3}$, right). Predicted with models evaluated with external validation. The five depths are 0–5, 5–15, 15–30, 30–60, 60–100cm.

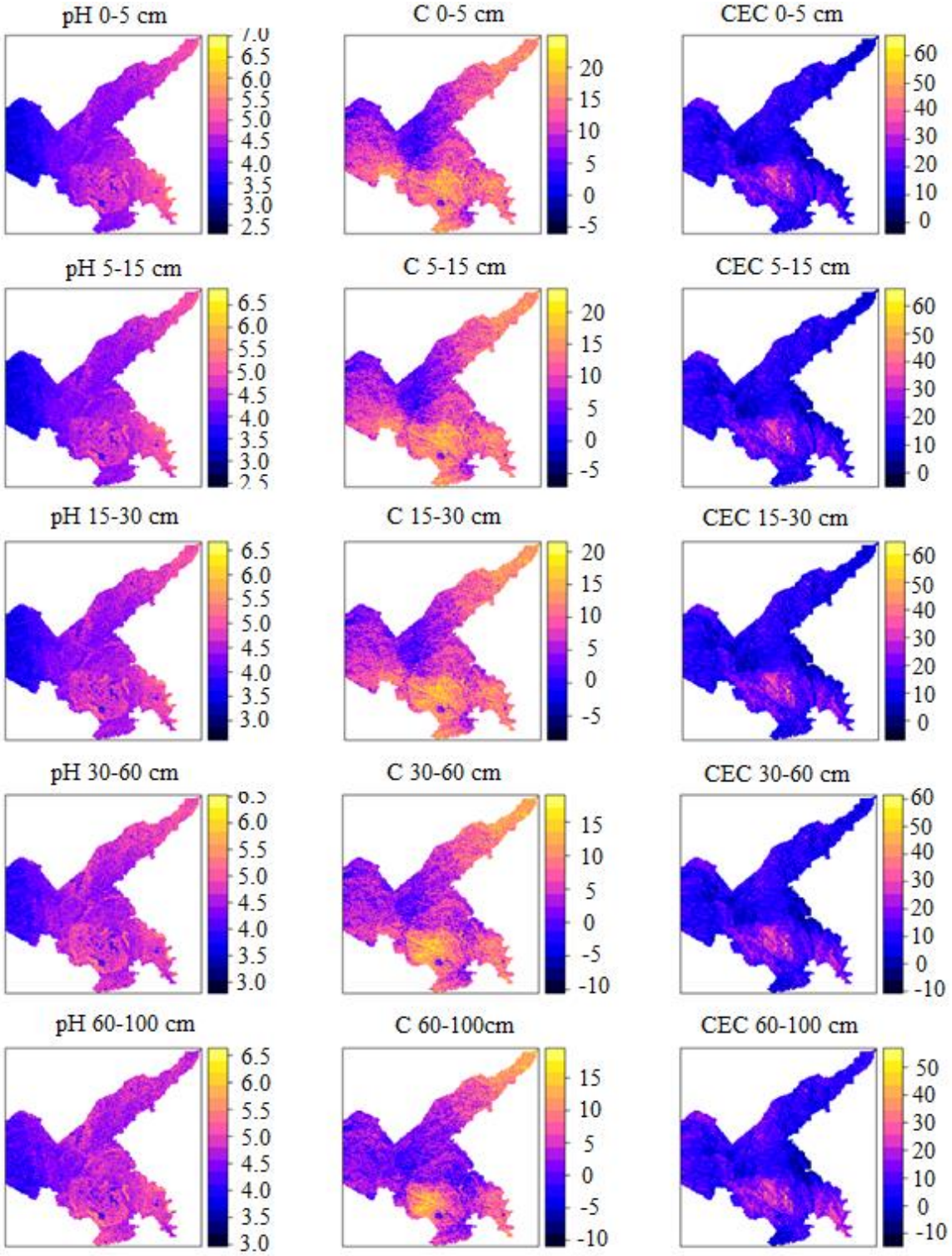


Figure S7. Maps of the soil attributes at five depths (pH left, Carbon content (%) centre, CEC ($cmol_c.dm^{-3}$) right). Predicted with models evaluated with cross validation. The five depths are 0–5, 5–15, 15–30, 30–60, 60–100cm.

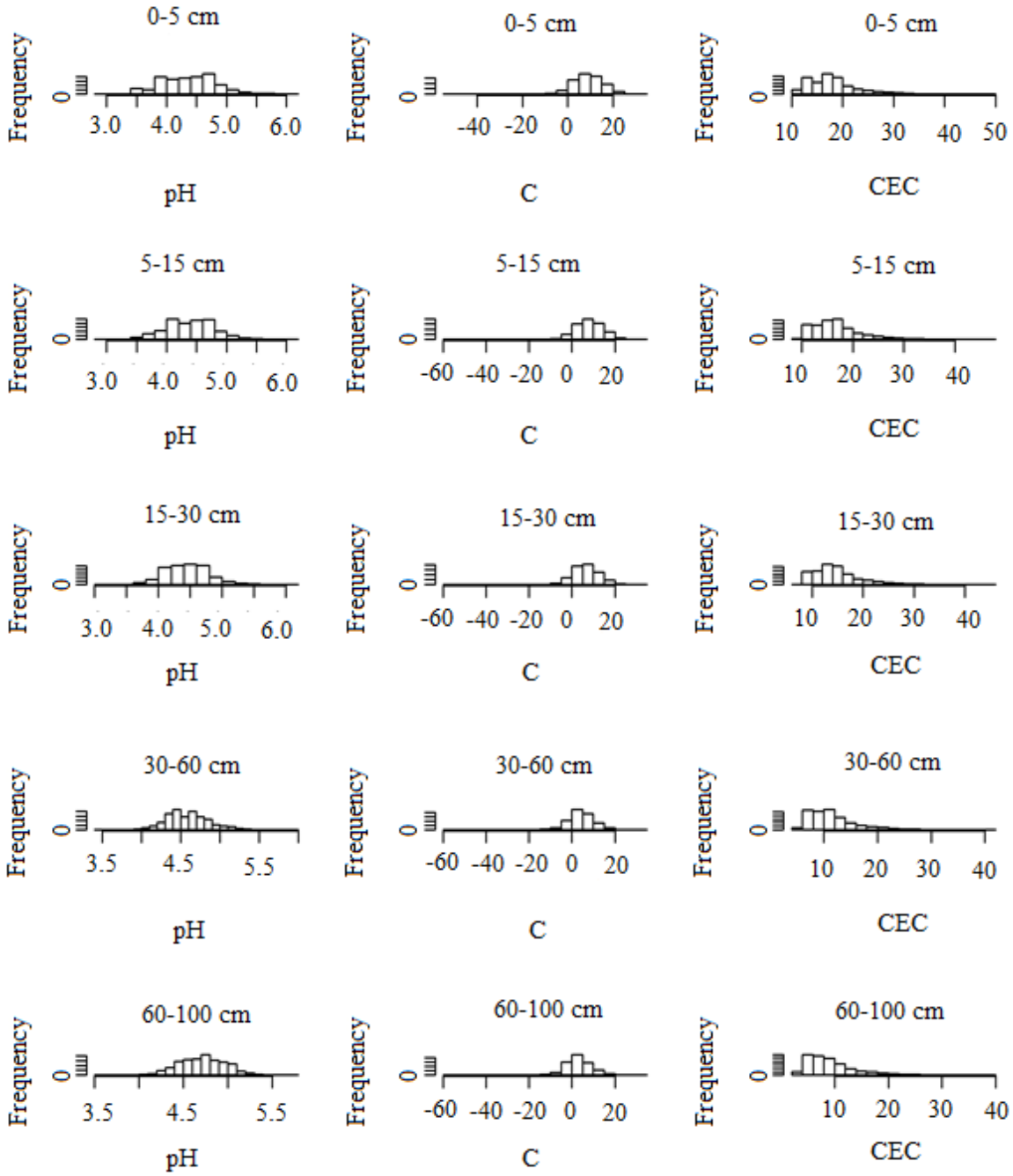


Figure S8. Histogram of predicted values in the grid by the model selected in external validation approach. (pH left, Carbon content (%) centre, CEC ($cmol_c.dm^{-3}$) right)

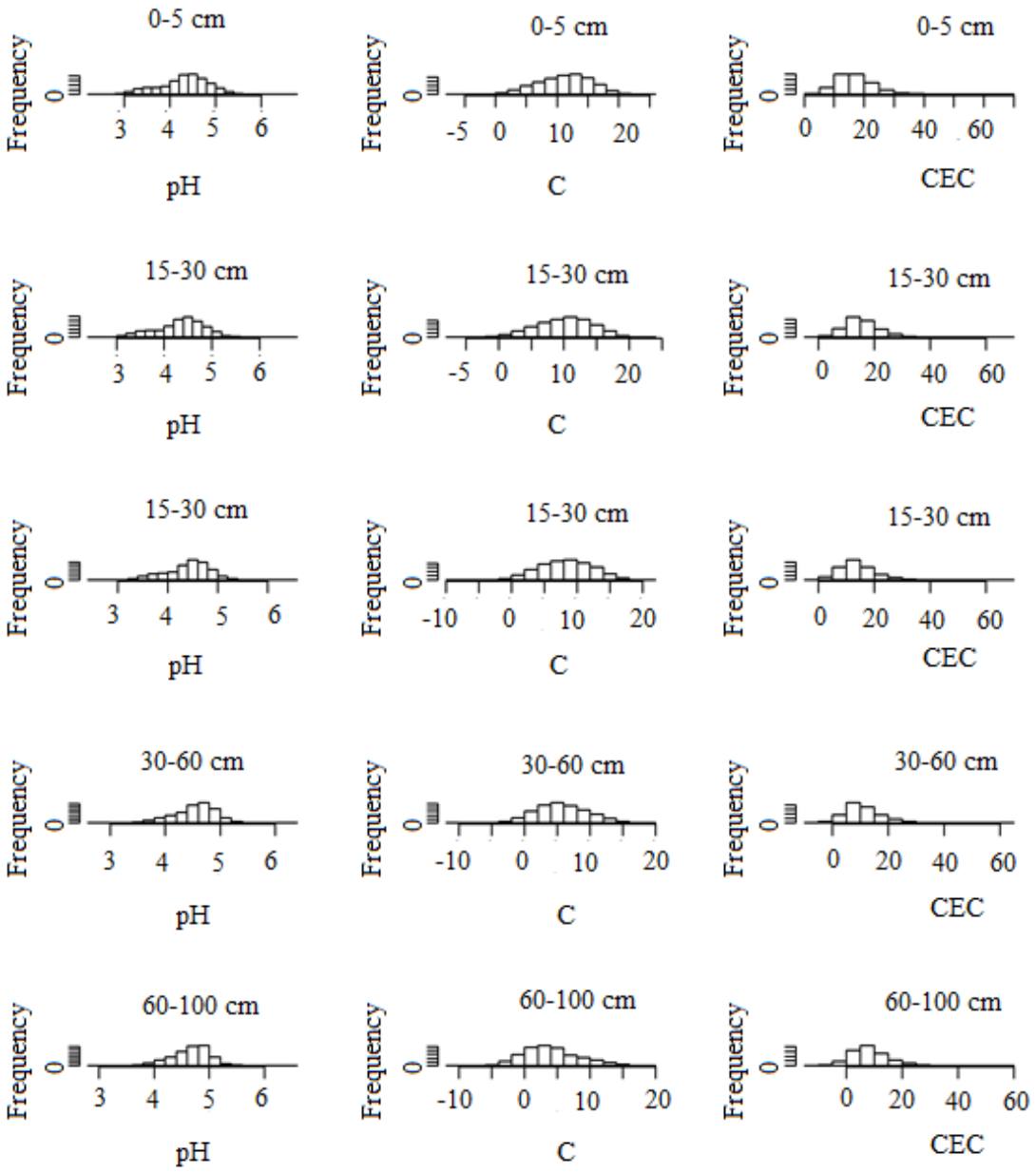


Figure S9. Histogram of predicted values in the grid by the model selected in LOO-CV approach. (pH left, Carbon content (%) centre, CEC ($cmol_c.dm^{-3}$) right)

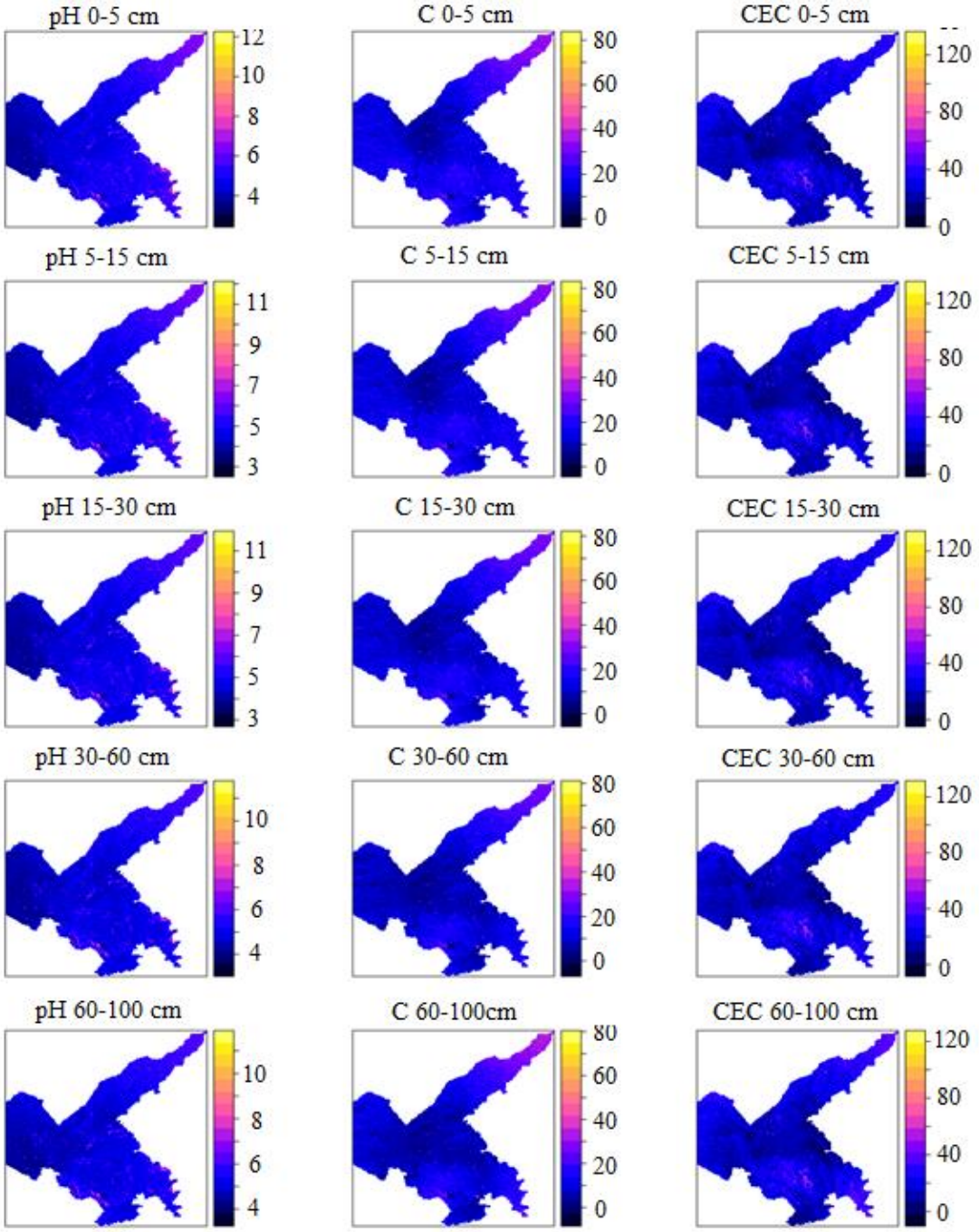


Figure S10. Upper limit prediction derived from a Bayesian posteriori distribution of each 3D GAM model fitted. pH left, Carbon content (%) centre, CEC ($cmol_c.dm^{-3}$) right

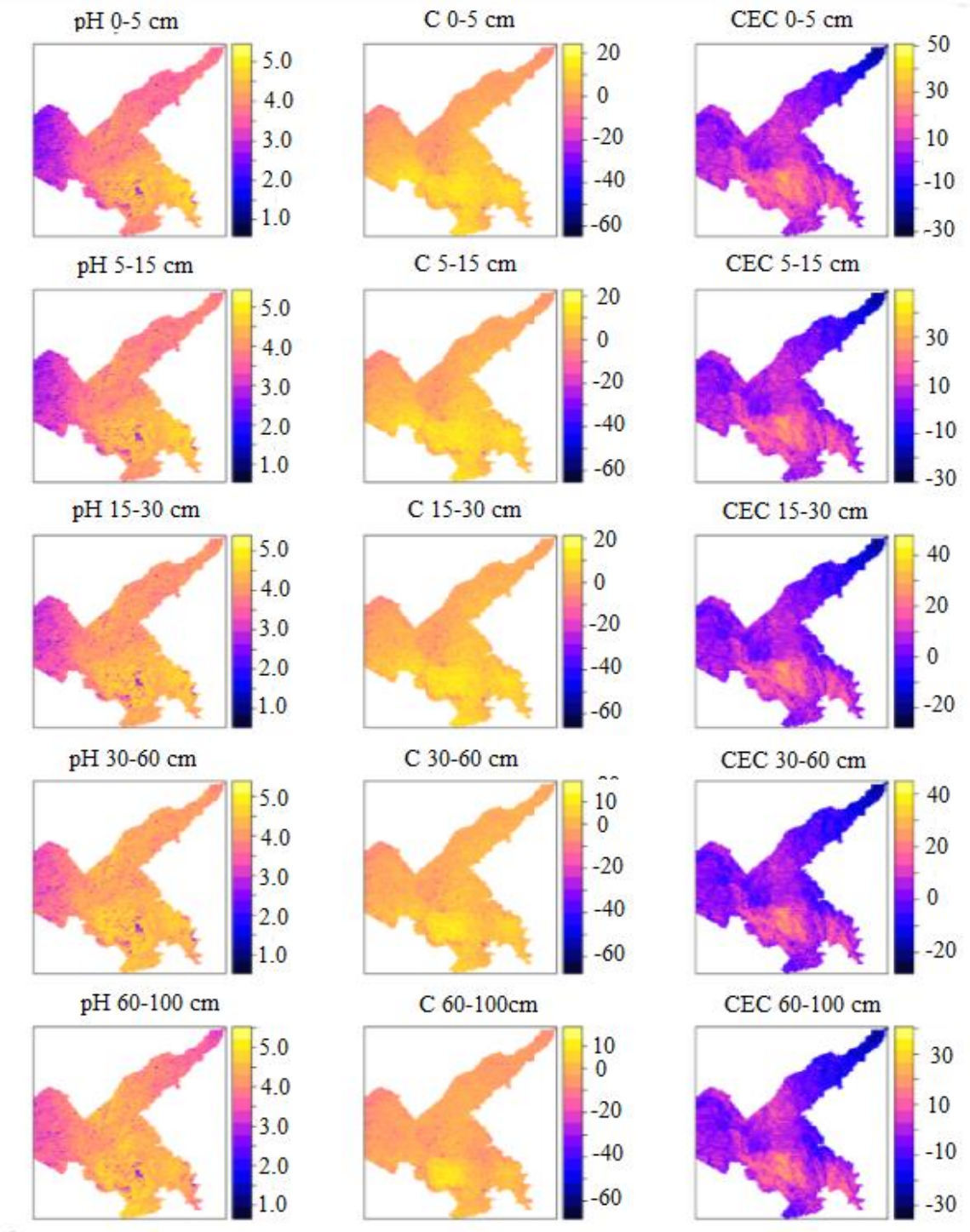


Figure S11. Lower limit prediction derived from a Bayesian posteriori distribution of each 3D GAM model fitted. pH left, Carbon content (%) centre, CEC ($cmol_c.dm^{-3}$) right.

Figure 12S are intermediate nodes that represent the fragility factors.

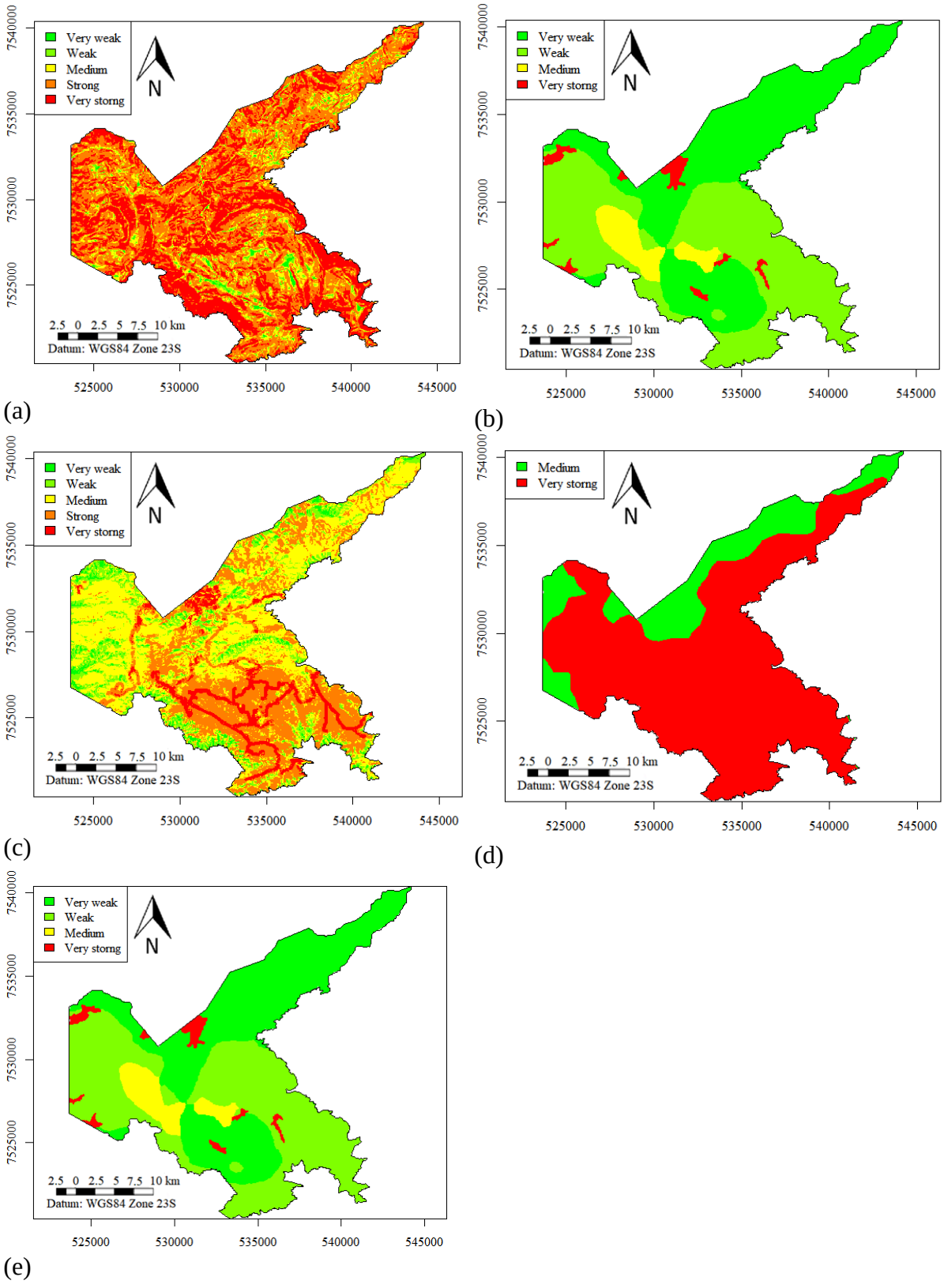


Figure S12. The environmental vulnerability by factors (fragility factors). (a) = relief; (b) = soil; (c) = Land use/cover; (d) = climate; (e) = parent material

Figure S13. Points where soil degradation is evident. They are mainly present in the upper part of the INP where it was classified as having very strong environmental vulnerability.

Dataset about the questionnaire

Some of the results of the questionnaire are showed from figure S14 to S19 and table S1. The form is available at: https://docs.google.com/forms/d/1iN4JybcqtoMrjtstJVf_uYhmQ_8xzmtvnMXYs8G5qj4/edit (Portuguese language).

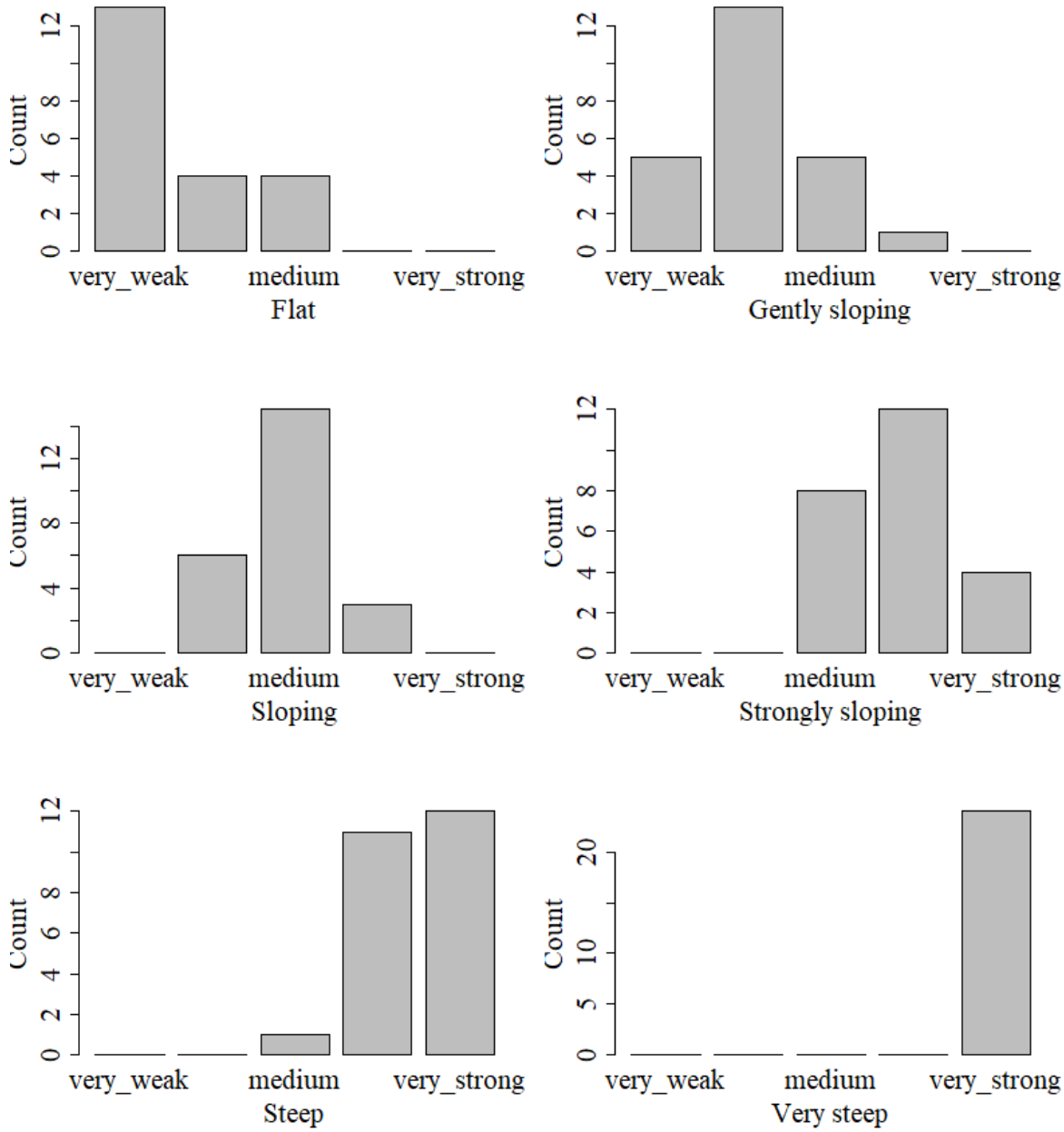


Figure S14. Classes of environmental vulnerability per slope classes as defined by the questionnaire answers.

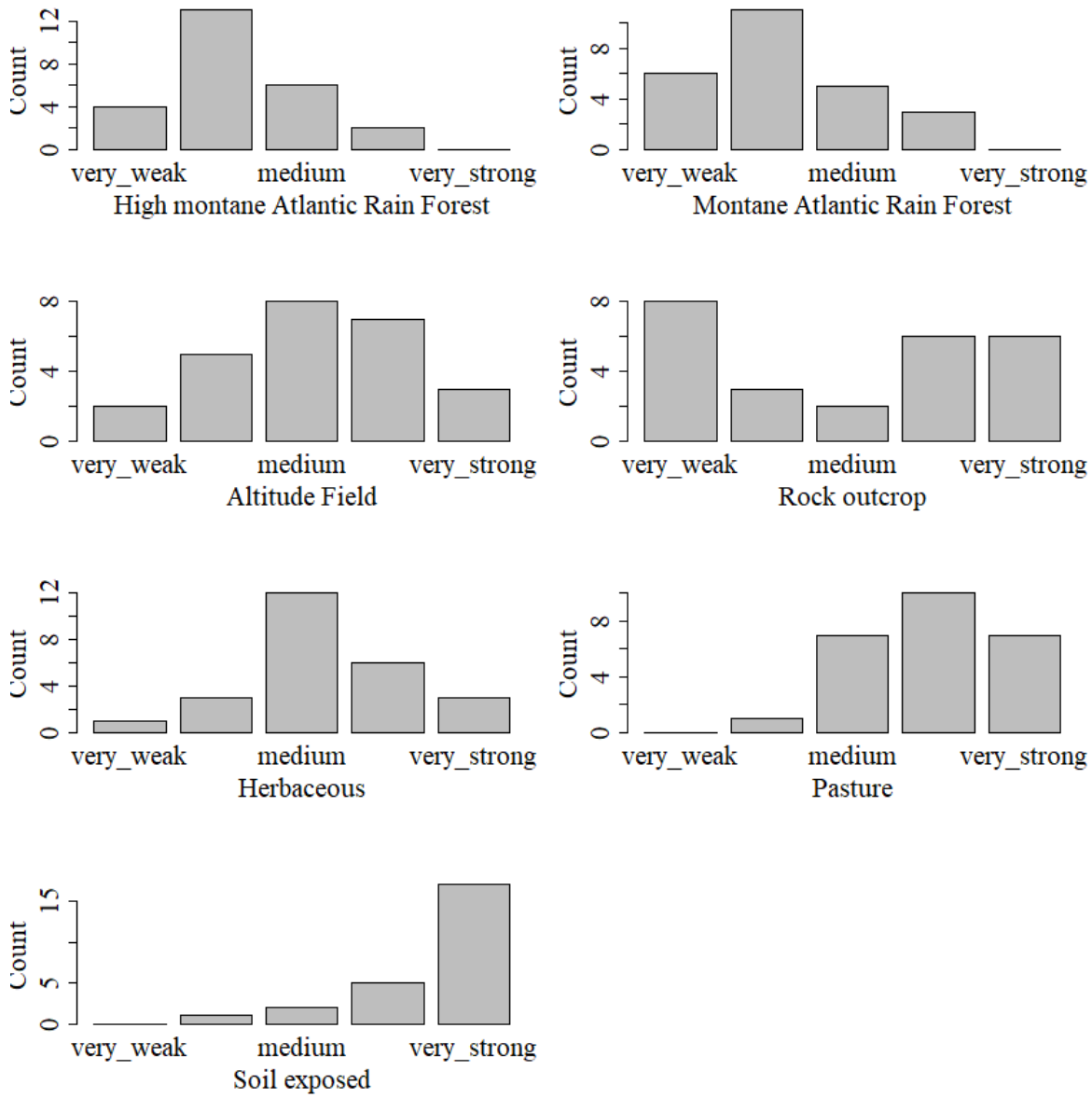


Figure S15. Classes of environmental vulnerability per land use classes as defined by the questionnaire answers.

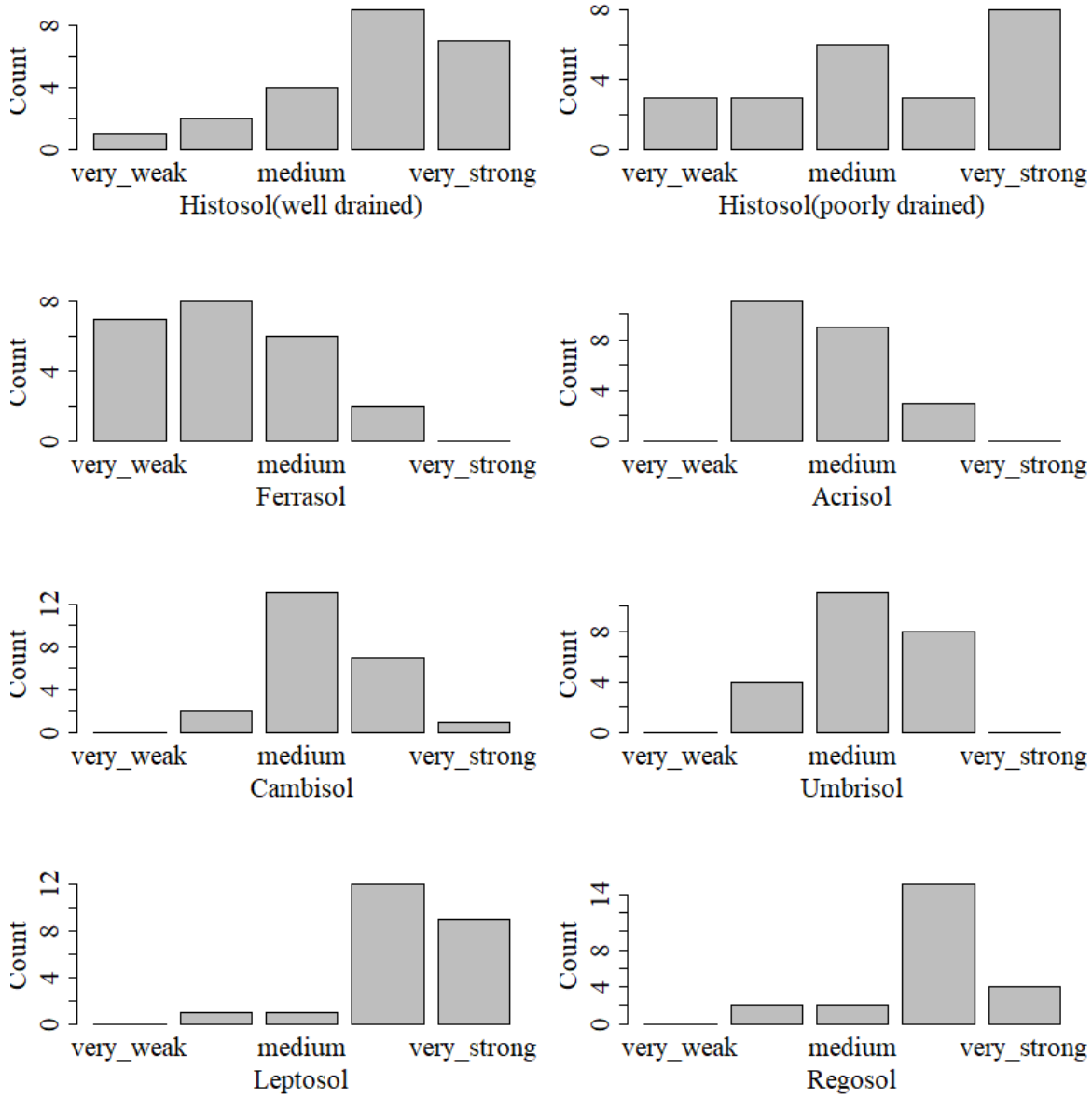


Figure S16. Classes of environmental vulnerability per soil classes as defined by the questionnaire answers.

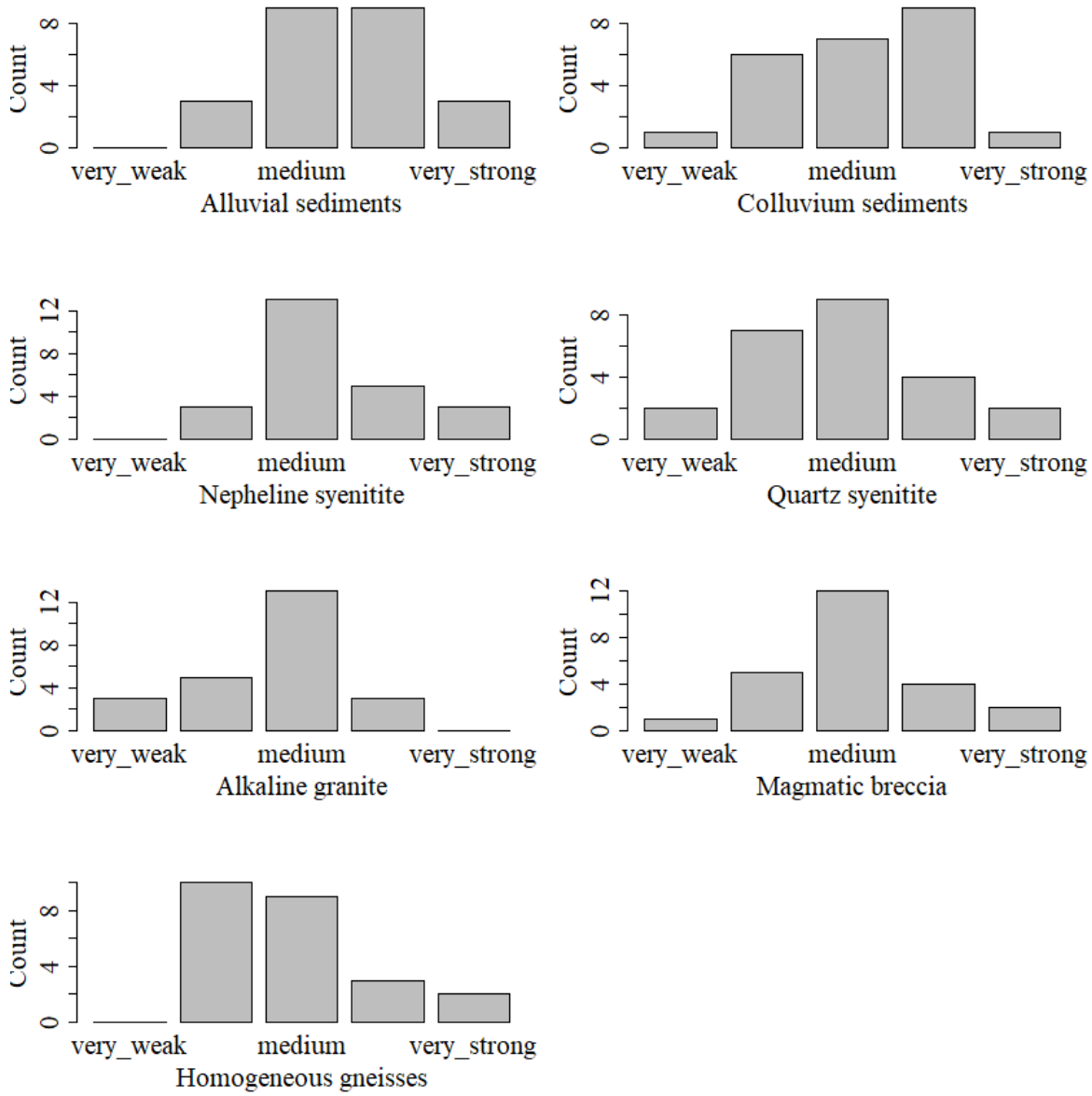


Figure S17. Classes of environmental vulnerability per geology classes as defined by the questionnaire answers.

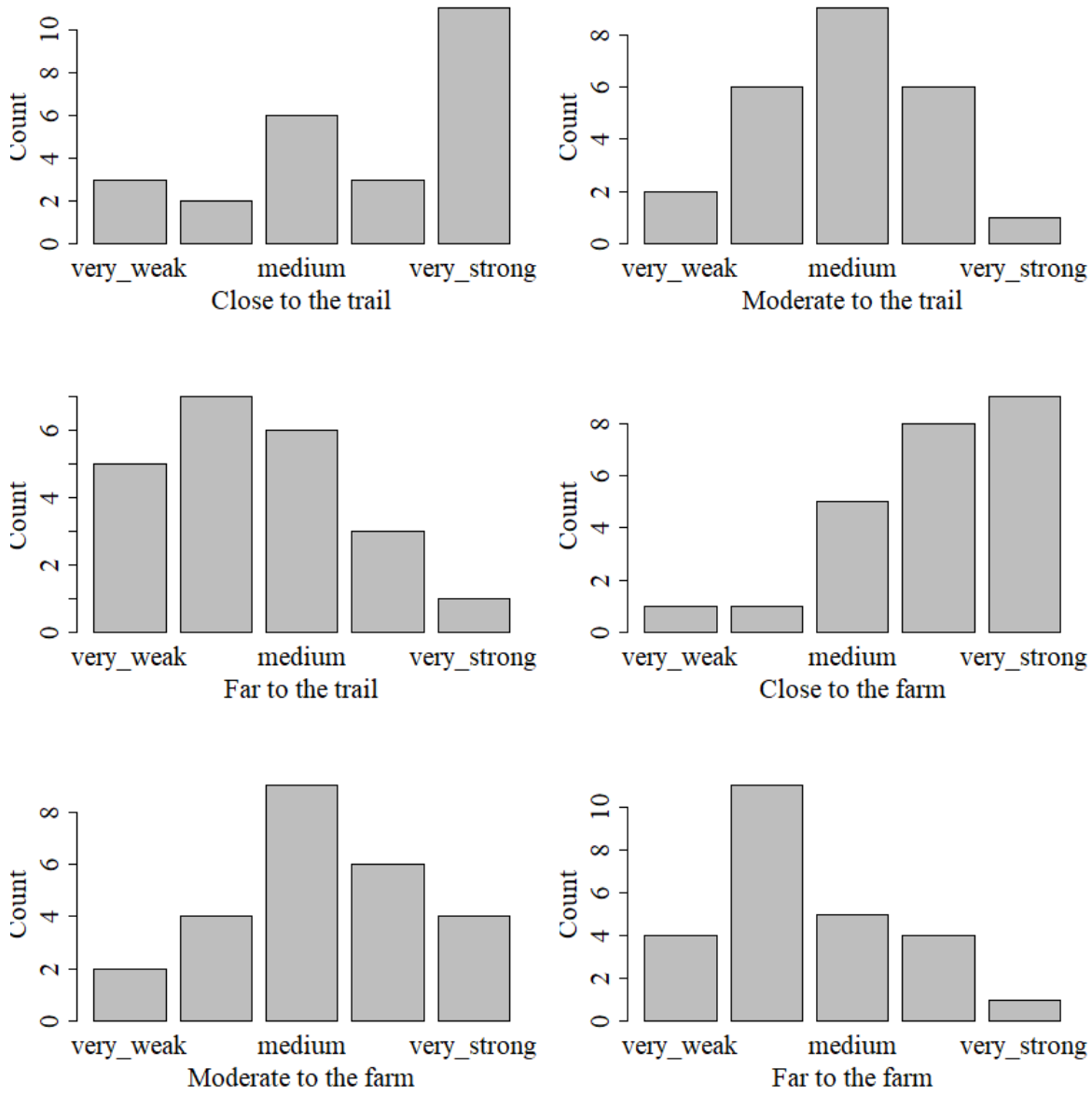


Figure S18. Classes of environmental vulnerability per anthropogenic impact as defined by the questionnaire answers.

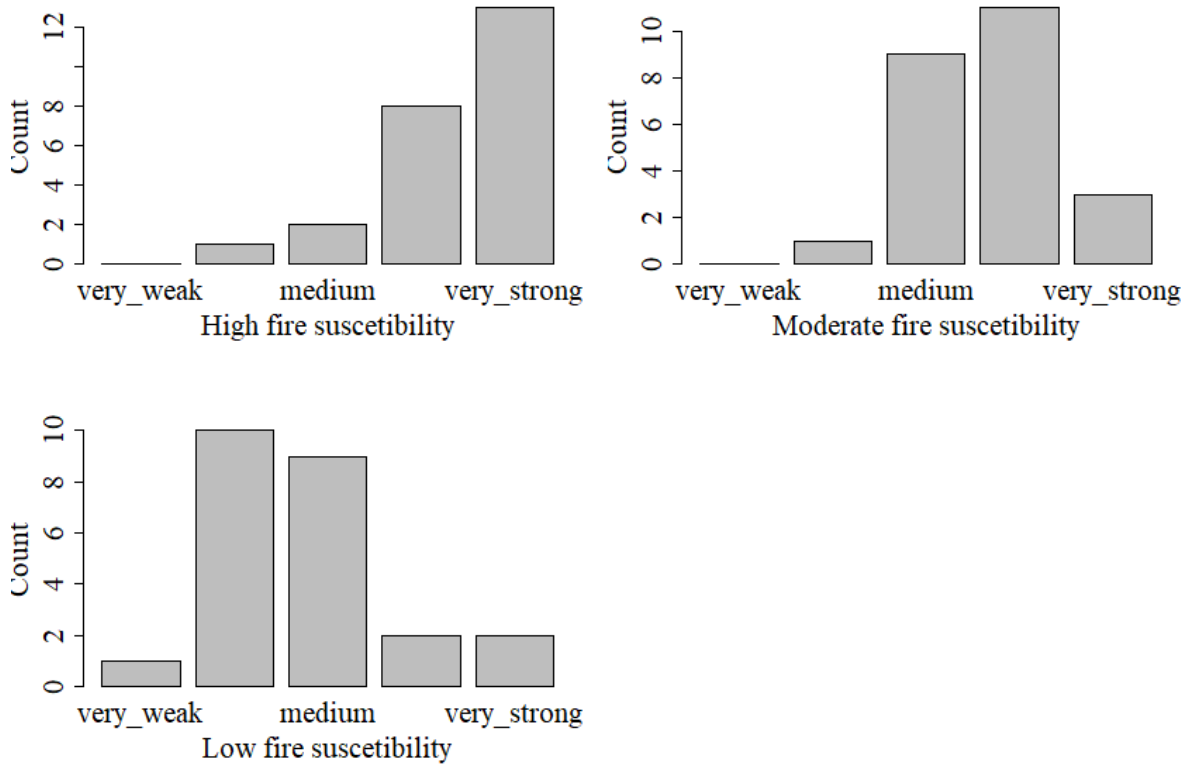


Figure S19. Classes of environmental susceptibility per fire as defined by the questionnaire answers.

Table S1. Participants' degree and specialization, institutions and years of work/research at the Itatiaia National Park

ID	Degree / specialization	Institution	Years
1	Agronomy / soil science	UFRRJ	6
2	Environmental remote sensing and applied climatology	UFRRJ	3
3	Environmental analyst	INP-ICMBIO	18
4	Forestry engineering / soil science	UFRRJ	3
5	Botany	AEDB	12
6	Agronomy / soil science	UFRRJ	8
7	Agronomy / soil science	UFRRJ	0
8	Soil and plant nutrition	UFG	0
9	Forestry engineering	UFRRJ	3
10	Agronomy / soil science	UFRRJ	2.5
11	Tourism and environment	CEFET-MG	10
12	Agronomy / soil science	UFRRJ	2
13	Education	Not identified	4
14	Agronomy / soil science	UFRRJ	2
15	Zootechny / soil science	UFRRJ	1.5
16	Environmental engineering	ESALQ-USP	6
17	Botany	JBRJ	15
18	Physical geography	UFF	0
19	Agronomy	Autonomous	0
20	Forestry engineering / soil science	IF-Sudeste-MG	18
21	Geography / geotechnology	UFRRJ	7
22	Pedometrics	UFTPR	0
23	Health sciences / epidemiology	UFRRJ	6
24	Agronomy / soil science	UFRRJ	0
25	Agronomy / soil science	UFPI	4
26	Environmental analyst	INP-ICMBIO	7

Note: **Years**= Means years of work or research in the INP. **0** for only visitor and/or sporadic field work or field lessons or planning to do some working in the park. **Institution acronym**: UFRRJ = Federal Rural University of Rio de Janeiro; INP= Itatiaia National Park; ICMBIO= Chico Mendes Institute for Biodiversity Conservation; AEDB= Don Bosco Educational Association; UFG= Federal University of Goiás; CEFET-MG= Federal Centre for Technological Education of Minas Gerais; ESALQ-USP School of Agriculture "Luiz de Queiroz"; University of Sao Paulo; JBRJ= Botanical Garden of Rio de Janeiro; UFF= Federal Fluminense University; IF Sudeste MG= Federal Southeast Institute of Minas Gerais; UTFPR= Federal Technological University of Paraná